Theory of Ensemble Forecasting
- with Applications in Transport Modeling

Hao Wu

September 23, 2021

A thesis submitted to fulfill requirements for the degree of
Doctor of Philosophy

Advisor: David M. Levinson

Faculty of Engineering
School of Civil Engineering
The University of Sydney

Statement of Originality

This is to certify that the content of this thesis is my own work. This thesis has not been submitted for any degree, and all assistance received in preparing this thesis and sources have been acknowledged.

- Hao Wu

Contents

List of Figures

List of Tables

Acknowledgement

Doing a PhD is often described as a dark and lonely journey, that only the most motivated and determined souls would undertake. But my journey was illuminated by individuals I met along the way, who helped and supported me with the kindness of their hearts. I would like to use this rare opportunity to express my thanks and gratitude to those who helped me along the way.

My journey would not have been possible without the support and guidance from my advisor, Prof. David Levinson. I'm forever grateful to his teachings, both in life and in research, which greatly enhanced my skills, and my understanding of the world. Prof. Levinson takes great care in mentoring, as he has always supported my research with great patience and encouragement. His teachings shaped my views in many ways, and I came to realize what "enlightenment" really means. I appreciate this academic mentor-apprentice relationship, where no other mode of learning can possibly provide a substitute. I also thank Prof. Levinson for all the good books he sent me to read.

I would like to extend my gratitude to members of the TransportLab at the University of Sydney. I was fortunate to have the opportunity to collaborate with many brilliant researchers on transport projects, which are especially enriching experiences - special thanks go to Dr. Emily Moylan and Dr. Somwrita Sarkar. And I thank Dr. Mohsen Ramezani for his interest and support in my research.

I also would like to thank my friends for their friendship and company along the way - Bahman Lahoorpoor, Amir Valadkhani, we had a great time enjoying local food together - and Hema Rayaprolu, Ang Ji, and Mengying Cui. I feel indebted to their company, and for the great working environment they created.

In memory of my grandfather, Shumo Wu, who passed away at the time of my writing this dissertation. My grandfather worked on the Chinese aerospace program as an electrical engineer in his early years, and he once taught me that "If you cannot fully understand something from a single book, go find another book on the same topic", which was my earliest indoctrination to "ensemble forecasting", and one of the inspirations for this research.

Abstract

Ensemble forecasting is a modeling approach that internalizes uncertainties, combining models with different assumptions or pattern recognition methods, data from different sources, and different methods of combining models. Compared to the prevalent single-model procedure, ensemble model predictions are more useful as decision support tools.

The use of ensemble forecasting has significantly improved forecast accuracy in weather forecasting, and is increasingly adopted in other fields. We find a lack of awareness, or application of ensemble models in transport, so the benefits of ensemble forecasting are not being realized.

In this research we establish a systematic framework for ensemble forecasting, and propose the 'ensemble of ensembles' to combine uncertainties in different ensemble methods. Ensemble models are applied to transport-related cases to examine the performance of different ensemble methods, and to compare ensemble models with single-model forecasts.

We find ensemble models can improve forecast accuracy by a notable degree beyond the best single model. Simple and weighted average ensemble models have mixed results. Meta-learner ensemble models provide significant improvement upon base models, but require sufficient training data to calibrate. We find the linear meta-learner to be robust and have good performance even with small training data. Ensemble of ensembles method combining different ways of combining models improves performance upon ensemble models, and generally has the best performance.

We conclude that ensemble models, if properly applied, are able to improve model performance. We posit that transport modeling can benefit enormously from the wider adoption, and awareness of ensemble forecasting methods. We hope that this research opens the door to methodically adopting ensemble models into transport modeling, that future transport research can build upon.

Chapter 1

Introduction

1.1 The Role of Forecasting in Transport

The ability to *predict* future or unknown events has long been a dream. At different times in history, forecasting follows different standards; forecasting has been conducted by oracles in ancient Greece, and by performing the Sun Dance by the indigenous people of North America. Currently, statistical models are the main instruments with which forecasts are carried out. The word 'model' in contemporary academic language often refers to Ronald Fisher's statistical methods, emphasizing objective models that are untouched by human judgement. Models are instruments of extrapolation, and the extraction of patterns; for models to achieve these functions, simplified statistical relationships are used to represent real-world mechanisms, connecting the outcome of an event (dependent variable) with factors affecting that event (explanatory variables). More recent machine learning algorithms make no attempt in understanding the data generation process, but instead focus on exploring patterns within the data (Breiman et al., 2001).

Models can either be used for *forecasting*, where the focus is on predicting future or unknown cases; or used for *analysis*, where the emphasis is on analyzing contributing factors of the dependent variable, examining the relationship between variables to better understand the data generation process. Correlation between variables, and identifying sub- or super-linear scaling, are among the analysis role of models. In this research we focus primarily on the forecasting role of models.

The concept underlying ensemble forecasting has existed since time immemorial, such that most cultures have expressions for the 'wisdom of the crowd' in their language. The same subject may be described by different books, each containing a portion of the whole knowledge; my grandfather once told me, that "If you cannot fully understand something from a single book, go find another book on the same topic" (Wu, 2009). In the real world, rarely would a single entity contain all the information; peer entities often need to complement each other. Most transport modeling focuses on finding the 'best' model, while ensemble forecasting emphasizes combining multiple models and data sources in order to combine information, and to consider different possibilities.

Models also have their own limitations. Most models start with imposing assumptions on the data generation process, or methods of pattern recognition. For instance, the logit discrete choice model (Ben-Akiva and Watanatada, 1981, McFadden et al., 1973) imposes

linear relationship for the 'utility' function, which supposedly measures the usefulness of each choice option to an individual. But the actual relationship does not have to be linear, or exponential, or even any mathematical expression at all. The linear (or any other) assumption is preferred, because this single-thread world view aligns well with how humans perceive the world. Humans subconsciously prefer, and substitute simpler solutions for complex ones (Kahneman, 2011), even when modeling complex events that require equally complex models. We may think of Greek oracles or North American Sun Dance as absurd ways of predicting the future, but people at that time held different views.

In transport, models are useful in understanding the inner-workings of transport phenomena, and in making quantitative predictions for future scenarios. Modeling provides transport planners with the means of analyzing and evaluating transport systems throughout different stages of a transport project, and communicating information to decision makers (Meyer and Miller, 2001). Transport modeling guides policies, infrastructure investment and management in order to solve problems. For instance, predicting the amount of social, economic and environmental benefit provides the necessary justification for potential transport projects; estimating future land use development and travel demand has been an important starting point in transport planning, and a major focus of transport modeling; other objectives of modeling include estimating the amount of traffic and levels of service on specific transport infrastructure, measuring system performance, and impacts of different transport improvement measures (Meyer and Miller, 2001).

Transport infrastructure, such as roads, ports, bridges, railways and airports, generally take years of planning, environmental evaluation, design, and construction to materialize, so there is a lengthy delay between the emergence of actual travel demand, and supply of transport infrastructure. Transport modeling is often used to overcome that delay. Travel demand predictions inform the scale and timing of transport investments, so that, in theory, the right transport infrastructure is built at the right time to meet travel demand, and with capacity corresponding to the level of demand. Forecast of travel demand has implications for the routing and scheduling of fixed route transport services, such as transit and airlines. Inaccurate traffic forecasts carry significant costs: under-building transport infrastructure results in significant social costs, and lost sales, while over-building based on inaccurate forecasts causes waste with excess capacities.

The social, economical and environmental impact of transport investment can be felt for prolonged periods of time, and remedial actions for any fault in transport projects can be costly to implement. Therefore in transport modeling, it is important to obtain models that are both accurate and reliable. Inaccurate, or misleading transport model predictions can result in large financial and social losses. As the transport equivalent of the Hippocratic oath, 'first do no harm', model predictions should be accurate, informative of future events, and avoid leading to stray decisions as much as possible. An ideal transport model would make the best use of all available information, so that every known possibility is reflected in the model output. Transport models shall be *honest* with its reduced accuracy in long-range predictions, and for predictions with a large amount of uncertainties, by presenting in some form, the amount of uncertainties in model predictions.

1.2 Track Record of Transport Models

Most transport predictions use the single-model doctrine, where a single model uses one set of data input, and produces a single number as the model prediction. Models are mere statistical imitations of real-world mechanisms, so different models are wrong by different degrees. As Box (1976) famously put it, 'all models are wrong, but some are useful'. The ultimate goal for modeling is to make predictions useful. The accuracy and reliability of these models have not historically been satisfying. Hoque et al. (2021) reviewed the performance of 1291 traffic forecasts in the US and Europe between 1960 and 2017, and found a 17% mean absolute difference between forecasts and traffic counts, and 90% of opening-year traffic volumes were in the range of -38% to 37% of traffic forecasts.

Although more recent traffic forecasts appear to be more accurate, it is still unclear whether this is due to better data and forecast methods, or the 2009 economic recession bringing down travel demand (Hoque et al., 2021); and a significant amount of forecasts still resulted in large errors. Transport modeling methods have not seen significant changes for some time, while the forecast accuracy in many other areas, notably in weather forecasting, has been steadily improving through the use of ensemble models.

In weather forecasting, it is typical for modelers to dissect and validate the model with observed data later on, and examine why the model did or didn't work, and also test other alternatives to analyze how the model can be improved; the result is a steady increase in the accuracy of weather forecasts (Blum, 2019). In transport modeling, modelers compare model predictions against observations, but don't often go back to examine and dissect past models. Transport model predictions can extend decades into the future, while weather forecasts extends mere days or weeks into the future; therefore weather models can be iterated at a faster rate, and by the same modelers with a fresh memory of past models.

Examples of inaccurate transport forecasts are plenty, and some of these inaccurate forecasts caused significant harm. In the text box we show several examples where inaccurate transport models proved detrimental to their intended purposes. These examples are meant to provide a contextual understanding for the state of practice in transport modeling. While there have been many other cases of erroneous transport forecasts (Boyce and Williams, 2015), we don't aim to enumerate all of them. Decisions based on erroneous forecasts turn out costly in many cases, and the cost and lost opportunities are borne by all members of the society, who are sometimes better-off without these model predictions in the first place.

The failure of transport models can be attributed to many reasons, among which are faulty model assumptions, inaccurate measurement data, and exogenous factors affecting the dependent variable. Transport modeling generally uses one set of data, and focuses on finding the one 'true' model that generally produces the highest overall accuracy for specific applications, and discarding the other models. This practice has resulted from both expediency, and the belief that the assumptions from the 'best' model better capture the mechanism than any other model, and that the model simply performs better in all aspects. Although a single 'best' model has merits when *analysis* is the principal objective of modeling, discarding information contained in other data set or models is not the best approach when *forecast* becomes the main goal of modeling, especially when different models differ in their assumptions about the relationship between variables (Bates and Granger, 1969).

In addition to selecting one set of model assumptions, transport models often use data

Cases of Inaccurate Transport Models

Case 1 Vehicle Miles Travelled Automobile travel demand forecasts provide guidance on the funding of road construction and maintenance. In the US Department of Transportation's 2002 Report to Congress, *The Status of the Nation's Highways, Bridges and Transit, Conditions and Performance* (FHWA, 2002), the vehicle miles traveled on highways in 2019 was forecasted to be 4.1 billion miles, which overshot the actual observed value (Federal Reserve Bank, 2020) by 27%. The model consistently over-estimates vehicle miles from the beginning, and the magnitude of error increased over time. In Figure 1.1 the 2002 US Department of Transportation prediction is plotted against the observed moving 12-month vehicle miles traveled. The divergence of the forecasted values away from the model predictions over time shows accumulated forecast error over time.

Case 2 Transit Travel Demand Transit demand forecasts scale the amount, and the timing of transit infrastructure investment and construction. In Sydney, Australia, the conventional strategic transport planning model (four-stage travel demand model) is used to forecast travel demand, which predicts transit patronage to increase by 48% between 2011 and 2031 (Transport for NSW, 2012). However, between 2011 and 2018, the observed transit patronage had already increased by 46% (Transport for NSW, 2020), reaching the predicted level of transit patronage over a decade earlier than the model predicted.

Case 3 Vehicular Traffic on Toll Road The short-term traffic predictions are not spared from large forecast errors. In 2012 the purchase decision of a toll road in Brisbane, Australia (Airport Link Tunnel) was partially based on the traffic forecast made by the Arup (a consulting company), which suggests profitability from the volume of traffic that would use the road. But only 25% of the traffic predicted by the model ever materialized (New Civil Engineer, 2018)(meaning a 400% forecast error), and the purchase decision led to a significant financial loss. Arup was subsequently sued for its role in mismanaging the modeling.

inputs without incorporating measurement uncertainties within the data. The measurement uncertainties in economic, population, and demographic data have been widely acknowledged, but solutions to these uncertainties are not incorporated systematically into transport modeling. For example, the various sources of uncertainties in travel demand models are acknowledged by Rasouli and Timmermans (2012). De Jong et al. (2007) reviewed literature on the order-of-magnitude of how much the uncertainties affect transport model output, and found a combined effect of 4% to 16% variations in model output from the data input and model uncertainties in traffic forecasts, and 5% to 15% on predicting passenger kilometers; these numbers are from simulations that assume models completely capture real events in a deterministic process, and should be interpreted as the floor in the magnitude of errors from uncertainties in data measurement.

Transport models are useful to the extent they function as decision support tools. Historically, most transport models don't have a good track record for making accurate predictions, especially in long range forecasts. Transport modeling techniques such as the conventional four-stage travel demand model originally developed in the 1950s (Meyer and Miller, 2001) have been left largely unchanged for over half a century. This is not to say there have been no advances. The introduction of equilibrium traffic assignment and discrete choice modeling,

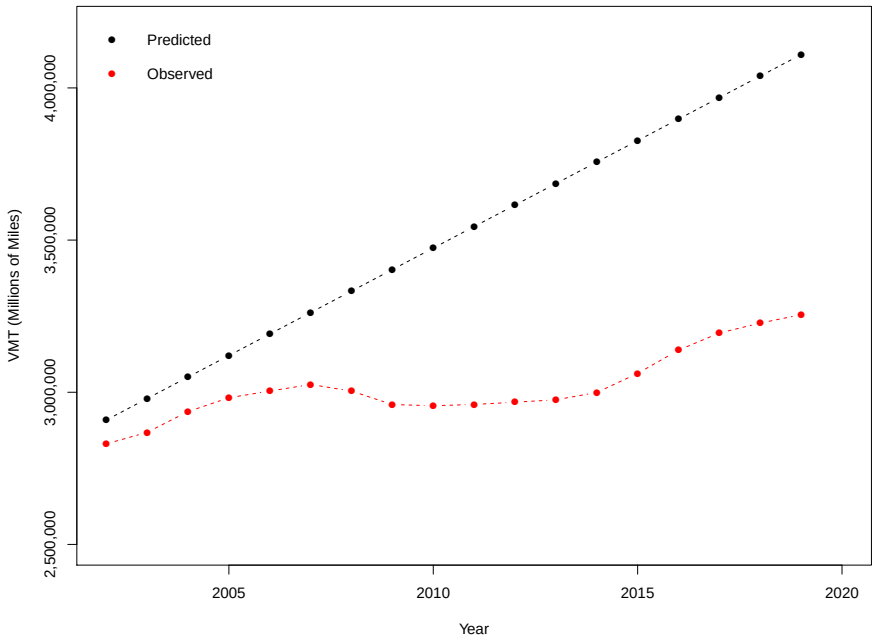


Figure 1.1: Vehicle miles traveled: observed moving 12-month (Federal Reserve Bank, 2020) versus prediction by the US Department of Transportation (FHWA, 2002) in 2002

as well as a move to activity-based models have occurred over this period. In addition the geographical representation of networks and places has become more detailed. However those changes have relied on the same kinds of data and analysis, and have not improved forecast accuracy.

While weather forecasters produce predictions that are validated against observations on a daily basis, and are under constant pressure to improve accuracy, it is common for transport practitioners to reach retirement before predictions made during their careers can be validated. So a lack of accountability, and motivation to improve transport models might also be responsible for the trail of inaccurate forecasts in transport. In weather forecasts, observed weather data provide daily opportunities to validate and improve models; transport models on the other hand, take many years before the results of their predictions can be validated. The frequent failures of transport models as decision support tools raise the question of how should transport forecasts be conducted in the future.

The concept of 'modeling' is often taken for granted, and correctness of conventional modeling methods are often enshrined and not questioned, despite the relatively short period of time (relative to the length of history) those models have existed. The widespread use of controlled experiments and statistical models are still recent history, so the suboptimal performance of some transport models, as we show in the above examples, are largely expected. If the real-world events were to behave exactly as the models, we would suspect that we live in a simulation. In modeling, it is fairly common for early attempts to fail, or to struggle with the right methodology (Blum, 2019); we should be open to new ideas, and new methods in order to improve our forecasts.

1.3 Theory-driven and Data-driven Models

There are 'two cultures' (Breiman et al., 2001) in modeling, namely theory-driven models, and data-driven models (terminology from Van Cranenburgh et al. (2021)). Theory-driven models have predetermined assumptions on the data generation process (e.g. linear, logit, etc.), and the model calibration aims to find parameters that suits the data. An illustration of theory driven models, would be the assumptions made at the beginning of many academic research papers, for example, "Assuming linear relationship between factor A and response B".

Most transport models are based on theoretical understandings of how different factors interact, and impose assumptions on the data-generation process; therefore these models are theory-driven models. Because assumptions are based on modelers' limited observation, knowledge and understanding, not all hidden relationship can be included in the model, and there is also no guarantee for these assumptions to be correct. It is safe to say that all assumptions are wrong, albeit by different degrees.

The entry of data-driven machine learning models provides an alternative to theory-driven models, in that machine learning models don't impose assumptions on the data generation process in the same way that theory-driven models do, and instead focus on the data to extract and reproduce patterns in the data. The rationale for machine learning models is that, if a model were capable of predicting out-of-sample data, then it should have identified the right patterns in the training data (Breiman et al., 2001, Van Cranenburgh et al., 2021).

Although data-driven models don't directly impose assumptions on the data generation process, other restrictions are still present in how the model recognizes patterns in data. Philosophically any method with a fixed shape and form will possess some restriction or drawbacks, so no single model will be perfect; for instance, the filament enables an incandescent light bulb to shine, but it will also cast a shadow of itself onto the ground. In terms of model performance, data-driven models tend to be more accurate than theory-driven models (Breiman et al., 2001), although data-driven models are less interpretable. Data-driven machine learning models are not commonly used in transport modeling, and the rate of adoption in transport has been slow (Van Cranenburgh et al., 2021).

Each of these two cultures has some merits, and under the single-model doctrine, modelers have to pick sides: either select one assumption to fit the model, or choose one pattern recognition method. And the vast majority of transport modelers sided with making assumptions. It would be ideal to combine models from both cultures, which ensemble models are capable of.

1.4 Misuse of Newtonian Physical Models

Newtonian physical models are a class of theory-driven models. In this section we discuss the misuse of Newtonian physical models in transport modeling.

Newtonian physical models describe real-world phenomenon with a single mathematical equation, as if these phenomenon were Platonic physical entities. Although the exact origin of the theory-driven single-models remains unclear, it is possible that the idea came through originally from the field of Newtonian physics, where the applied methods share many sim-

ilarities with transport modeling. Linear models were initially used to describe planetary orbits (Legendre, 1805). Newtonian physical phenomenon are deterministic, and can be described by single mathematical models with a high degree of precision. Single mathematical models in Newtonian physics have an 'unreasonable effectiveness' (Wigner, 1990) in describing physical phenomenon, and these models indeed seem to represent the 'true' underlying mechanisms. It appears that physical entities are truly underpinned by single models; gravity, motion, and energy, etc. are all well described by single equations, with a high degree of precision, to the extent that it would almost be silly to introduce other less accurate equations to describe them. The Newtonian physical models are intended for problems with a single deterministic process, and worked fine until the introduction of quantum physics (Schrödinger, 1935), which describes sub-atomic particles as 'cloud of probability', and stating the exactly location of particles are not knowable (Heisenberg, 1985). The question arises as to whether some of the problems are non-computational in nature (Penrose and Mermin, 1990), and a 'true' single model may never exist for these problems.

Most transport models are theory-driven Newtonian physical models. This is noted by Garrison and Levinson (2014). Transport model specifications are generally dictated by theories (e.g. utilities in discrete choice models), that are concerned about finding the true underlying data generation process; different model specifications are produced by different theories. Newtonian physical models are used, naively, to represent what is theorized to be the true mechanism, and models with lower performance are labelled simply as 'incorrect'. The rationale for applying physical models in transport is the assumption that a 'true' model really exists, and can be described using physical models.

Newtonian physical models work only in idealized circumstances, such as isolated reference systems, no surface or air friction, etc., and where the system mechanism is well understood, which is not possible with transport modeling problems involving complexities and uncertainties. So following the same school of thought as in Newtonian physics, and applying single models in transport modeling might be a misuse of physical models. Transport-related cases such as flow, travel demand, and mode choice are clearly not Platonic physical entities, and past attempts by transport models to describe these complex real-world phenomenon (as we described in the previous section 1.2) have resulted in some spectacular failures. This high rate of failures is unlikely to be caused by incompetence alone, therefore the conclusion must be that the tools used by transport modelers are either flawed, or have applicability issues.

Expediency also played a role in the prevalence of the single-model doctrine. Choosing a single model, and one set of data input is notably more convenient than calibrating, and combining an ensemble of models. In an era without electronic computers in which physical models grew, applying ensemble models can be prohibitively expensive. Developments in computation power is making such constraints less relevant, but using ensemble models still requires significantly more manual labor and computation time than single models.

The mystery of why Newtonian physical models ended up dominating modeling may never be solved. In transport modeling, this probably began in the 1950s with the emergence of the four-stage transport demand model (Boyce and Williams, 2015), and continued to this day. This research is not about archaeology in transport modeling. We often have to live with vestiges of previous eras, for example, the width of U.S. railway gauges is derived from the width of horse drawn wagons, which is not optimum for rail operations; some internet traffic is still carried by old telephone lines, which are not as fast or reliable as optical fiber.

The development of modeling does not have much physical vestige that stunts its progress, however, the single-modeling doctrine, and Newtonian physical models have become hard-wired into the modeling practice.

1.5 Concept of Ensemble Forecasting

Ensemble forecasting intends to extract more information out of available data, and to incorporate uncertainties in modeling. The resulting ensemble models have higher accuracy, better reliability, and with model outputs that are more useful as decision support tools. The defining characteristic of ensemble models is the combination of outputs from different models, and data from different sources. Philosophically this combination of data and models constitutes an aggregation of information, since different models can extract different pieces of information embedded within the data (Winkler, 1989); data from different sources also contain non-overlapping pieces of information, that can be combined by ensemble models.

Uncertainties are unknown parts in the data, and the data generation processes, that cannot be ascertained with all available information. Various sources of uncertainties in the modeling process, including in the measurement of explanatory variables, in different assumptions of the data generation process, and in different methods of combining models, all contribute to inaccuracies in model predictions. The widely used single model procedure is generally not compatible with incorporating uncertainties. Expediencies related with single models, such as arbitrary choice of one set of measurement data, or one model over other possible type of models would subjugate model performance to random chances.

Sensitivity and scenario analysis of explanatory variables can test, to some extent, the effect of measurement variations on the model output, but is limited to testing one set of explanatory variables at a time, and the model outcome is still typically presented as a single number, relying on a single model formulation. Therefore the sensitivity and scenario analyses do not consider the wider range of modeling uncertainties, and are still not a reliable method of modeling.

Ensemble forecasting is not a specific method in modeling, but a doctrine that acknowledges, and accounts for uncertainties in forecasting. Ensemble models consider all possibilities stemming from uncertainties, and include these uncertainties both in the model calibration, and in presenting model outputs. The implementation of ensemble forecasting often involves multiple parallel models of different model types, using different datasets; outputs from these parallel models are combined into the ensemble model output. The term 'ensemble' means multiple things being considered together, and is applied here to refer to the combining of information, and uncertainties, so the ensemble forecast represents the 'best available' inference by the model, within the confines of all available information. Ensemble model outputs can indicate the possible range of prediction variation from the available data and models.

Ensemble model outputs can include a range of possible outcomes from parallel base models, instead of a single number. Real-world events have many possibilities, to which models only provide an 'estimate' for what is likely to happen. In this light, different models rely on different assumptions, that provide different perspectives for a prediction. The performance of models are measured in probabilities of being correct, so even the lowest performing model still has a small chance of being correct. The job of ensemble model is to incorporate these

uncertainties into an ensemble forecast.

Ensemble forecasting is not yet the mainstream practice in modeling. The foundational work in ensemble models was in the 1960s (Bates and Granger, 1969, Reid, 1968), and the the first operational ensemble model was introduced in weather forecasting in 1985 (Palmer, 2019), specifically to address the impact of initial measurement error on the accumulated error over time (i.e. chaos). The use of ensemble methods has expanded ever since. The use of ensemble forecasting is relatively rare in transport, but a number of other fields have already adopted some form of ensemble model as a state of practice. The field of meteorology uses ensemble forecasting as standard practice to improve accuracy in weather forecasts. Weather forecasters acknowledge the persistence of errors in the measurement, and in their models, so the purpose of an ensemble forecast is to dilute the accumulated error over time in long-range forecasts. The output of the weather forecast is presented as a probability of precipitation, instead of a deterministic result; the forecasted trajectories of hurricane/typhoons from different agencies/nations are often superimposed to show the path that the most models or scenarios expect (the statistical mode), but also all possible paths. Similarly ensemble methods are used by some political forecasters, where the measurement data is noisy, and future events have multiple possible paths (Silver, 2012). Business (Ashton and Ashton, 1985) and economic forecasts (Armstrong, 2001, McNees, 1990) have been using ensemble modeling for more reliable predictions. Ensemble modeling in these fields has been progressively transforming into the state-of-the-practice. There have been attempts from the sociological (Silver, 2012), medical (Winkler and Poses, 1993), biological (Van der Laan et al., 2007), and agricultural models (Kronvang et al., 2009) in adopting ensemble models.

Transport modeling shares in the same uncertainties in data and in models as weather, economic, and political forecasting, and yet ensemble forecasting remains rare in transport. Through the use of ensemble models, different theories, different assumptions on data generation processes, and data from different sources with slight (or significant) variations can be combined to present multiple possibilities in an ensemble forecast. Ensemble forecasting provides an opportunity for improving transport models.

1.6 Sources of Forecast Uncertainties

Uncertainties are unknown pieces of information not covered by forecast models, and it is the job of ensemble models to incorporate and represent uncertainties. Models are best with the 'known knowns', which are strictly deterministic, and are better with *risk* (the 'known unknowns' to follow the Rumsfeld framework (Rumsfeld, 2011)) than *uncertainty* (the 'unknown unknowns'), where risk can be quantified and measured but uncertainty cannot. Transport problems include significant uncertainties from various sources, which is the root cause of forecast errors. Newtonian physical models, on the other hand, are deterministic, and do not allow for such uncertainties.

Different types of uncertainties affect forecast models in different ways, and some of the uncertainties can be better accounted for than others. This section discusses and categorizes major sources of forecast uncertainties, and clarifies the role of ensemble forecasting in accounting for uncertainties.

Uncertainties in Data and Measurement. Measurement data for real-world contin-

uous variables will always include errors, so no measurement or observation can ever describe the true state of explanatory variables. Small differences in initial measurements have been shown to accumulate, and produce drastically different outcomes; this is known as 'chaos' (Thietart and Forgues, 1995), which is responsible for much of the failure in early weather forecasts (Blum, 2019, Silver, 2012). Traffic forecasts with longer time span are less accurate (Hoque et al., 2021). In a sense all measurement data contain errors, because any real number is a 'mathematical idealization' rather than 'actual physically objective quantity' (Penrose and Mermin, 1990). Data from different sources may use different sampling and measurement techniques, so may contain errors of different magnitude, in different directions. Using data from a single measurement process, and without testing the effects of data perturbations is not a prudent practice in modeling.

Ensemble forecasting accounts for uncertainties in measurements by testing data from different sources, measurements, or with perturbations. There are two pathways by which this is implemented, namely by averaging repeated measurements, or by repeatedly running models with different measurements. But instead of averaging repeated measurements (and there will still be measurement error, albeit smaller in magnitude), ensemble weather models opted for combining models trained on perturbed measurement data. This is because in non-linear systems, even small measurement errors will produce a large difference over time, so producing model output based on measurement perturbations becomes more helpful, than averaging data sources.

In the transport context, measurement errors, and data on the same subject but from different sources, are a common problem faced by transport modelers. Repeated measurements, measurements made at different points in time, and measurements by different researchers are among examples of possible data sources for a single explanatory variable. Averaging multiple data sources into a single number is more common than testing the effect of different data sources on the model output (as is done in weather forecasting).

Uncertainties in Models. This research adopts the view that among various models, there does not exist a 'true' model (Winkler, 1989), that can exactly replicate the data generation process, or identify patterns within data. To take this view further, the non-existence of a true model is not solely due to the complexities that are intrinsic in real-world phenomena, but also their non-computability. Both conventional (statistical), and machine learning models rely on computational procedures. Leaving aside the debate on determinism versus free will (Margenau, 1967), the nature of many real-world problems are non-computational (Penrose and Mermin, 1990, Turing, 1936), so applying computational procedures, no matter how complicated these procedures are, will not be able to exactly replicate real-world events or patterns. The 'true' computational mechanism may never be found by a model, because the underlying computational mechanism may not even exist. This is especially the case for transport problems where human behavior and consciousness are involved.

There is uncertainty in the choice of one model over other models. McCullagh and Nelder (1989) wrote: "Data will often point with almost equal emphasis on several possible models, and it is important that the statistician recognize and accept this." For an unknown or future event, each model provides a different description, and different descriptions can have near identical error rates (based on model performance metrics), although these descriptions come from models with very different formulations. This has been termed the 'Rashomon

Effect' by Breiman et al. (2001), which describes that different models (witnesses) may arrive at the same conclusion, although the paths (story) taken by these models (witnesses) may differ. Further to this point, small perturbations in the training data, which will very likely be caused by measurement errors, may cause the 'best model' to shift from one model to another (Breiman et al., 1996, 2001). This instability in the choice of best model is corroborated by other sources, which stated that because of the noise (or 'accidental features') in a particular dataset, model formulations selected based on highest R^2 are often wrong (Bacon, 1977, Kennedy, 2003, Mayer, 1975). This means that it won't be known until after the fact which model would have had the highest performance in future cases, and the best model for alternative history (unbuilt scenario) cases are unknowable even in retrospect.

The choice of one model, or one set of assumptions over others raises a serious question about the validity of model outputs, especially in Markovian type problems where history is irrelevant: among models with similar performance measures, if the arbitrary choice of a single model would substantially impact the prediction, or significantly affect subsequent models using its prediction as an input (such as in the conventional four-stage transport planning model), then the model will have a multitude of different possible predictions, and any single one of these predictions should be treated not deterministically, but as some probability.

Uncertainties in models means that different assumptions (models) on the data generation process or different methods of extracting patterns from the data cannot be judged simply as right or wrong, but each as an integral part of the multiple paths constituting a prospective event outcome. Selecting one of the paths essentially ignores other possibilities. Therefore what can be known from all available information is not a deterministic number, as in Newtonian physical models, but a probability distribution for all likely outcomes.

The role of ensemble forecasting is to acknowledge multiple paths in transport problems, that the future trajectory of an event is a probability cloud, instead of a single path. And ensemble forecasting accounts for model uncertainties by combining different models.

1.7 Limitations of Ensemble Forecast, the 'Unknown Unknowns'

Ensemble forecasting is not a panacea for all modeling problems. By combining models and data, ensemble models are only able to consider uncertainties that are already known to exist in model assumptions and in measurement data. Ensemble forecasting obviates the need to make arbitrary determinations as to which model or data source (unknown possibility) is more likely.

The ensemble forecast does not automatically account for factors outside of model specifications, and is not immune to external shocks to the system. This is exemplified by Donald Rumsfeld's 'unknown unknowns' (Rumsfeld, 2011), as shown in Table 1.1. The Table 1.1 categorizes model uncertainties into four categories:

- the 'known-known': The part of model that the modelers know they have knowledge of, that there is no uncertainty.

		Nature of the Uncertainty	
		Known	Unknown
Modeler's Awareness	Known	Known-Known	Known-Unknown
	Unknown	Unknown-Known	Unknown-Unknown

Table 1.1: Known-unknown matrix for modeling, based on Donald Rumsfeld's memoir (Rumsfeld, 2011)

- the 'known-unknown': The aspects of the model that is known to modelers to have uncertainties; this is what the ensemble forecast specializes in.
- the 'unknown-known': This is common knowledge and tacit understanding, that modelers do unknowingly.
- the 'unknown-unknown': The aspects of model that the modelers are not aware that they have no knowledge of.

The 'unknown unknowns' are situations where the modelers are not aware that uncertainties exist, and therefore cannot prepare for the impact of these factors in model formulations beforehand. The 'unknown unknowns' often alter the course of future events. For example, the emergence of new transport technology or modes can throw off the travel demand forecast; congestion, economic downturns, and pandemic outbreaks can change how, and whether people still commute or go shopping or out to eat. Behavioral changes caused by pandemics (and working from home) may permanently change land use and travel patterns, so the models calibrated on previous data may need to be re-estimated. Such 'unknown unknowns' occur with low frequency, and so are more likely to effect a long range forecast, which has more time to accumulate such events, than a short run forecast.

1.8 Research Scope and Objectives

This research introduces ensemble forecasting to transport modeling. We establish a theoretical framework for methods of ensemble forecasting, bringing together methods of ensemble forecasting from different fields, to gain a bird eyes' view for the various methods, purposes, as well as capabilities and limitations of ensemble forecasting. This framework should facilitate in the understanding and adoption of ensemble models by transport professionals.

We recognize there are ensemble models at different levels, that range from ensemble of models of the same type, to higher level meta-learners, and all the way to ensemble of ensemble models. In this research, we primarily treat basic types of ensemble models as base models. This research is not a technical study to devise a specific type of ensemble model, or aimed at performance optimization of models. Instead, the focus is on theorizing, and testing whether the idea of ensemble forecasting can be useful for transport applications.

Ensemble forecasting and single-model prediction have many similarities, and may be perceived as two closely related methods, but they represent two entirely different views of the world, and modeling approaches. The single-model doctrine has a naive view of the world, treating events as deterministic physical processes, while ensemble forecasting accounts for

inherent uncertainties. This research attempts to raise awareness to differences between this two modeling approaches.

In this research we first examine the literature on ensemble models, looking at methods of ensemble forecasts in different fields, and the commonalities of these methods; the state of ensemble forecast in transport modeling is also examined. We modify, combine, and improve upon existing methods of ensemble forecasting, and test these ensemble methods on a number of transport problems, including trip generation, attraction, flow of for-hire vehicles (FHV), real estate prices. A number of base models are included, including linear, Classification Tree (CT), Random Forest (RF), Gradient Boosting Machine (GBM) and Neural Network (NN). Performance of ensemble models are examined in terms of accuracy, reliability, and the usefulness of model output as decision support tools. Conclusions of this research are based on reviewing the literature, and testing ensemble model performance.

Many more questions and potential research directions are raised than answered in this work, and we are not able to solve all of them due to time and other constraints. Considering the short history since the advent of ensemble forecasting, we are likely still in early stages of its development; future research will almost certainly find more effective ensemble methods than what are presented in this research. This research nonetheless builds way-points through accumulated knowledge and empirical evidence, that enables future research to grow on the shoulders of the past. It is hoped that by showing the case for ensemble forecasting, this research will expand available options in transport modeling, and make itself useful to future developments in modeling.

Chapter 2

Literature Review and Synthesis

This review examines the literature for applications of ensemble models. The scope of this review covers ensemble models both within and outside of transport research, covering both methods of ensemble forecasting, and author-stated objectives of applying ensemble methods, to ascertain what types of ensemble models are in existence, and what these models are used for. The term 'ensemble' refers broadly to the doctrine of combining data and models, sometimes involving human intervention in model outputs; so the scope of this review is not limited to one specific method, but a broader range of methods that incorporate the idea of combining information and uncertainties. Ensemble models consist of multiple different models, which are termed 'base models'. Ensemble models differ both in the composition of base models, the amount of data incorporated, and in methods of combining base models.

There are different levels of ensemble models. Degenerate forms of ensemble models that include only one model formulation are treated as single models, in the context of this research. Algorithms such as bootstrap aggregating (e.g. random forest), and boosting methods (e.g. boosting machines) are themselves single models, but consist of an ensemble of a simpler type of models; these models are covered in this review, but herein are not regarded as ensemble models by themselves. Special attention is given to the discipline of weather forecasting, where ensemble models first originated, and the extensive use of ensemble models improved weather forecast accuracy over the past decades.

It is noted during this review that modeling practices are highly siloed across different disciplines, and the terms used for ensemble models vary significantly. The standard practice of combining data and models in one discipline may not be used, or even properly acknowledged in another field. This review aims to bring to light niche practices used in different fields.

This chapter begins with a review of various ensemble methods, tallying the specific areas and applications where the ensemble methods have been applied. Some of these ensemble methods are more suited for transport problems than others. This review is then followed by a synthesis of various ensemble methods. We synthesize different levels, and methods of ensemble forecasting across different fields into a unifying framework for ensemble models, with the addition of our understanding of ensemble models. This review and synthesis itself serves as an ensemble of knowledge from different disciplines.

There is a broad literature on ensemble models, and wide spread application of ensemble models in some areas. For instance, use of ensemble models in weather forecasting has

significantly improved forecast accuracy over time. However, the idea of ensemble forecasting is still largely foreign to the field of transport. In fact, transport modeling suffers from such an inadequate appreciation of ensemble models, and deep dependency on the single-model approach, that transport models are not delivering their due benefits. It is hoped that the adoption of ensemble forecasting might be helpful in that respect.

This review contributes to the literature with its wider scope than technical reviews of ensemble methods. Even in weather forecasting where ensemble methods originated, its ensemble methods are not comprehensive. This review covers both ensemble models that make a single simultaneous prediction, and iterative models that use model outputs as new inputs, where forecast uncertainties resulting from initial condition and accumulated error (i.e. the Chaos theory) tend to accumulate. These two types of ensemble models are generally discussed separately in different contexts, because historically, the Chaos theory and error accumulation are more specific to weather forecasting; the combination of different model formulations and assumptions are not a major interest in weather forecasting. However, these two types of ensemble models are both relevant to transport related cases, therefore the broad scope of this review is necessary for transport application of ensemble models. We also review ensemble methods used in different disciplines, which includes the use of expert opinions, judgemental adjustments to model predictions, and meta-analysis, etc. covering both methodical, and what would be considered empirical ensemble methods.

The purpose of this review is a comprehensive introduction of ensemble forecasting to the field of transport, pointing out the inadequacies of the single-model practice, and showing common ensemble methods in their basic forms, that transport practitioners can readily adopt.

2.1 Ensemble Forecasting in Transport Modeling

This review uses a systematic review method for ensemble forecasting in transport modelling. Unfortunately the use of language in the literature in describing ensemble models, and different definitions for what constitutes ensemble models are very messy. In some cases ensemble methods are included in papers with titles that appear unrelated to ensemble models; in other cases, papers titled "ensemble models" use degenerate ensemble methods, such as random forest, or gradient boosting machines, that are regarded as single algorithms and not ensemble models in this review. Therefore in this review, we first identified a number of papers with actual transport applications of ensemble models, and use them as seeds, then went through related literature in these papers using a snowball method (Wohlin, 2014) to expand our literature. There is indeed very limited application of actual ensemble applications in transport.

There have been a limited number of transport research projects using the ensemble forecast. Ensemble models in transport applications present model outputs as single number, not as probabilities (as in some other disciplines). The list of transport applications of ensemble forecasting is shown in Table 2.1; numerical performance of classification models are presented with brackets. This section reviews the use of (or the lack thereof) ensemble forecast in transport modeling. We also added ensemble models' performance from our experiments in this research, where sufficient training data is available, and ensemble forecasting can be

properly applied.

Machine learning algorithms are sometimes used in transport forecasts. It has been shown that single machine learning algorithms that incorporate the idea of ensemble forecasting can outperform statistical models (Zhang and Haghani, 2015), and can outperform boosting algorithms (Zhang et al., 2020). However, applications of single machine learning algorithms, such as the random forest and the gradient boosting machine, are not discussed in this section, since they are considered base models in this research. The performance of machine learning algorithms has often been compared with statistical methods, and the average accuracy has been the major consideration (Karlaftis and Vlahogianni, 2011), different algorithms are rarely combined in transport applications.

Applications of ensemble forecasting in transport mostly focus on accuracy improvement, and there is a clear lack of understanding in the mechanism of ensemble models, or other potential benefits from the use of ensemble models.

Flow, traffic state

The traffic state of freeway segments can be predicted by different models. Li et al. (2014) argues that each traffic flow model has imperfections, and different methods of combining traffic flow models using weights are tested, but no further attempt is made to combine different methods of combining forecasts. The combined forecasts are more accurate than single-model forecasts. Tan et al. (2009) generates forecasts from three different models in short-term traffic flow forecasting, and uses these forecasts as inputs for a meta-learner; this method improves model accuracy in most cases, and provides similar accuracy when the single model forecast was better. Tselentis et al. (2014) uses a meta-learner to combine six different models in short-term traffic flow predictions, and noted that the combined forecast to be more accurate in most instances, and the risk with a combined forecast is lower than selecting a single model. In another study of short-term traffic forecasting, Pavlyuk (2020) uses the majority vote from an ensemble of three different models to predict the relationship between flow characteristics of road segments with time delays on arterial roads in Minneapolis, and the ensemble forecast slightly improves model accuracy.

Different configurations of graph neural network (GNN) come from different fields and are based on different theories (Qi and Kwok, 2020). In predicting traffic state, Qi and Kwok (2020) use simple average to combine intermediate model outputs from different GNN configurations, which improves model accuracy.

A peculiar adaptive ensemble method is used by Stathopoulos et al. (2008) in predicting future traffic flow, in which the performance of two candidate models are compared with real-time data, and the model with better performance is selected to predict traffic flow in the next time interval. There are different preferences in individual oute choice, for example, between any two points there can be different paths with the shortest travel time, distance, and routes that minimize emission, or maximize traffic safety (Cui and Levinson, 2021); basing a model on any single criterion might be problematic.

Study	Ensemble Method	Criterion	Forecast Accuracy
Short-term traffic flow			
Stathopoulos et al. (2008)	Select model for t+1	RMSE	-
Tan et al. (2009)	Meta-learner	MAPE	Mixed, generally improves MAPE
Tselentis et al. (2014)	Meta-learner	RMSE	Mixed, generally improves RMSE
Pavlyuk (2020)	Majority vote	MAE	+1.43%
Traffic state, flow rate and density			
Li et al. (2014)	Weighted sum	RMSE	(density+11.1%; flow rate+16.5%)
Qi and Kwok (2020)	Simple average	MSE	+2.07%
Mode Choice			
Cheng et al. (2019)	Majority vote	% cases	(+18.3%)
Rasouli and Timmermans (2014)	Majority vote	-	-
Ride-splitting choice in on-demand taxi			
Chen et al. (2017)	Weighted sum	ROC curves	-
Demand for on-demand taxi			
Liu et al. (2019)	Meta-learner	MAE	+ 2.7%
Short-term transit demand			
Wei and Chen (2012)	Weighted sum	MAPE	-
Ma et al. (2014)	Weighted sum	MAE	+35%
Safety			
Ji and Levinson (2020)	Meta-learner	% cases	(+4.57%)
Stop Detection for Travel Surveys			
Servizi et al. (2020)	Stacking	% cases	(+3.0% to +6.5%)
Willingness to pay			
Layton and Lee (2006)	Weighted sum	-	-
Airport passenger traffic			
Xiao et al. (2015)	Weighted sum	MAE	+29.4%

Table 2.1: Transport applications of ensemble forecast procedures (performance of classification models displayed within parentheses)

Mode choice

The mode choice of an individual is generally predicted by one model in transport modeling; Rasouli and Timmermans (2014) proposes using multiple decision trees on each individual, and let the decision trees vote on the mode choice, to account for uncertainties in the individual mode choice. This idea is followed through by Cheng et al. (2019) in predicting the mode choice in a travel demand model, where an ensemble of multinomial logit models are trained, each with a random sample of training data, and a random selection of explanatory variables (random multinomial logit model). When applied to household survey data and using the 'majority vote' rule to combine the model forecasts, the ensemble model becomes more accurate than both the mixed logit, and the base multinomial logit models.

Ride-splitting choice in on-demand taxi

Chen et al. (2017) combines multiple classifier models using weights into an ensemble, which is then used to predict the ride-splitting choice in on-demand taxi services. The ensemble model has improved accuracy in the training dataset.

Demand for on-demand taxi

Travel demand is conventionally estimated through a linear model, which is a component in the four-stage transport planning model. The 'on-demand taxi' (e.g. Uber, Lyft) is a relatively recent occurrence, and so are the methods of estimating its travel demand. The demand for 'on-demand taxi' is estimated by different base models in Liu et al. (2019), including a linear model. A meta-learner is then trained to combine these base models using weights, which improved the forecast accuracy.

Short-term transit demand

Some of the ensemble methods in transport modeling originated from the mode decomposition of signal processing, which views a single data generation process as having multiple sources, each requiring a different model. In predicting the short term travel demand for metro, Wei and Chen (2012) decompose the demand data into multiple components for an analogy model, and a meta-level model is used to combine these historical data for new predictions. In research predicting the short-term travel demand for buses (Ma et al., 2014), the weekly, daily, and hourly demand model forecasts are combined using an interactive meta-model with real-time data input; the meta-model estimates the probabilities of the most effective model at different times, and combines model forecasts using weights.

Safety

Ensemble models have been used in the transport safety literature. In Ji and Levinson (2020), stacking models are used for vehicle crash injury severity prediction, which improved accuracy of injury classification.

Stop Detection for Travel Surveys

In stop detection for smartphone-based travel surveys, Servizi et al. (2020) use a stacking method where the output of one machine learning (ML) model becomes the input of another model, and raised model accuracy by between 3% to 6.5%.

Willingness to pay

In estimating the willingness to pay using stated preference (SP) survey data, Layton and Lee (2006) use the weighted sum from a range of models instead of selecting only one model; the weights are based on relative statistical fit. It is argued that this method provides forecasts with better robustness (Layton and Lee, 2006).

Airport passenger traffic

This method of dissecting data to be predicted by different models is extended by Xiao et al. (2015) in forecasting the number of airline passengers at an airport. The passenger traffic is decomposed as the sum of three components: trend (representing the natural growth of the airline industry over time), seasonal oscillations, and irregularities. Each of the three components is predicted with a different type of model. The ensemble model is trained on past data, and applied to future dates for the Hong Kong airport; forecast accuracy (MAE) from the ensemble model is higher than other single-model alternatives.

2.2 Types of Ensemble Models

Ensemble models are based on rules specifying how data and models should be combined. These rules vary significantly among different ensemble models, that differ in the ease of implementation, computation cost, and model performance. Some of the rules are used more often than others, such as assigning weights to different models based on model performance; other rules, such as the 'stacked generalization', are less commonly used in practice.

Combining rules are often developed by different disciplines in their own silos, to deal with disciplinary-specific problems. Some of the ensemble methods require significant scientific knowledge, and additional effort to tailor for each dataset, which limits the wider adoption of these methods. In fields such as meteorology, ensemble modeling is used almost by default in every forecast, mostly out of the necessity to deal with environmental uncertainties, and accumulated errors that would otherwise render the model forecasts useless; other fields use ensemble modeling less often. Ensemble modeling makes only rare appearance in transport problems. The benefits and constraints of different ensemble methods, and their use in different transport problems are not well understood. This section reviews different ensemble models from the literature.

2.2.1 Non-parametric, Voting, Equal-weight Averages

Non-parametric rules combining model outputs are the simplest form of ensemble forecasting. Under non-parametric rules, model outputs are combined regardless of the performance of

each model. Typical examples of non-parametric ensemble rules include equal-weight simple arithmetic average (Wichard and Ogorzalek, 2004, Winkler, 1989, Zhou et al., 2010), median Zhou et al. (2010), or mode of different model predictions. Equation 2.1 shows the paradigm of non-parametric combining rules.

$$x_c = \frac{\sum_i^I x_i}{I} \tag{2.1}$$

Where,
x_i : Forecast from model i
x_c : Combined forecast from aggregating I different models

Non-parametric rules are among the most frequently used ensemble methods, and have a wide range of applications. For instance, ensemble models using equal-weight of different models have been used in sales forecasting, which tend to be more accurate than single forecasts (Ashton and Ashton, 1985), and prevents large prediction errors when trusting a single model. In agriculture, the average of different model predictions on soil nutrient load is robust, and enables analysis and comparison between different empirical models (Kronvang et al., 2009).

In classification models, the simple averages combining rule translates directly to 'majority vote'. In majority vote, the type of model output shared by the majority of base models will become the ensemble model output. The majority voting scheme is non-parametric, and does not need a separate validation dataset (Zhou et al., 2010). When the prediction from different models disagree on a specific case, the ensemble model output takes the value of the majority. The majority voting scheme is found to improve model performance in classifying traffic flows' spatiotemporal structure (Pavlyuk, 2020), and in travel demand forecasts (Cheng et al., 2019).

The literature finds good robustness in non-parametric ensemble rules. In Monte Carlo analyses, it has been found that the model performance-based weights for combining models can be unstable, and a simple arithmetic average may produce better accuracy (Elliott, 2011, Kang, 1986). The arithmetic average has been shown to be robust, have good performance in empirical studies (Winkler, 1989), and can be practical to implement. The use of simple averages is supported by much analytical and empirical evidence, and has the benefit of being easily understood and implemented (Graefe et al., 2014).

2.2.2 Weighted Average

The weighted average of several model predictions is widely used for combining model outputs, and is the earliest method used in the foundational work on ensemble models (Bates and Granger, 1969, Reid, 1968). Output from each base model is assigned a weight, which depends on some performance criteria of the base models; and the weights from all base models add up to one. The rationale for weighted average ensemble rules is to give better models higher weight in the ensemble model output. One distinct feature of the weighted average method is that one weight applies to all forecasts from a single model, so there is no differentiation of weights from case to case.

In practice, it is thought that the unweighted simple averages is more robust than weighted averaged based on model performance, and there have been cases favoring simple averages over weighted averages combining rule. For instance, the optimal weight for minimizing the variance of error (average square error of the forecast) proposed by Bates and Granger (1969) reportedly does not work well in practice, and is often outperformed by simple average (Winkler, 1989).

Weight for each model in the combined ensemble model output can be based on a variety of criteria, and can be easily customized for specific applications. Here we discuss three commonly used weighting schemes: weights based on the variance of error, contribution of each model in reducing variance of error, and percentage of correct classifications.

There is no general consensus as to which weighting scheme will produce the best outcome. One commonality among different weighting schemes is that models with good performance have a higher weight, and no model is dropped completely from the ensemble.

Variance of error

The variance of error provides a performance measure for base models, and a criteria for assigning weights (Bates and Granger, 1969, Wann and Lin, 2004). When base models present a large variance of error, suggesting low accuracy, then weights for these models are scaled back. This weighting scheme applies to a minimum of two base models (Wann and Lin, 2004), and can be extrapolated to multiple base models, as shown in Equation 2.2; the weights are derived in Equation 2.3 (Meier, 1953). Models with lower variance of error obtain higher weights in the combined forecast. Inverse of the forecast RMSE in the training data has also been used as weight (Gneiting et al., 2005).

$$x_c = \sum_{i=1}^{I} W_i \cdot x_i \tag{2.2}$$

$$W_i = \frac{\frac{1}{\sigma_i^2}}{\sum_{i=1}^{I} \frac{1}{\sigma_i^2}} \tag{2.3}$$

Where,
W_i : Weight of x_i in x_c
σ_i, σ_j : Variances of error in model i and j's forecast

Low correlation in the variance of error between models ensures that the variance of error in the combined model to be lower than constituent models, and the combining rule in Equation 2.3 assumes zero correlation between errors of different models. This type of combining rule becomes less ideal when there exists significant correlation between errors of constituent models (Bates and Granger, 1969). This means in combining model forecasts using a weighted average method, it is desirable to include only models with independent information. The attempt by Yu et al. (2005) to exclude models from the ensemble that provide redundant information using Principal Component Analysis (PCA) effectively fulfills that purpose.

Weights based on the variance of error have been used in aggregating different wireless location system outputs to achieve better location accuracy (Wann and Lin, 2004). In weather

forecasting, base models are combined using weights from the variance of error (Fay and Ringwood, 2010).

Contribution of models in reducing variance of error

In Delen et al. (2017), instead of using the calibrated coefficients, the sensitivity of adding each member model i, in reducing the variance of error in the dependent variable, S_i, is calculated, and this sensitivity becomes the weight for each member model. The formulation is shown in Equation 2.4.

$$S_i = \frac{\sigma(E_{t-i})}{\sigma(E_t)} \tag{2.4}$$

S_i: Sensitivity; contribution of model i in reducing the variance of the dependent variable;
E_t, E_{t-i}: Ensemble of t models; Ensemble of t models less the model i.

Percentage of correct classification

For classification models, the weights are applied directly to the votes by different models, which are combined to produce the ensemble model classification. The weight on each of the model vote can be based on the percentage of correct predictions (Zhou et al., 2010), which will generally require an additional validation dataset.

2.2.3 Super Learner, Stacked Generalization, Stacking

The stacking method goes by various names in the literature, including super/meta-learner, stacked generalization, stacking etc., although these methods follow essentially the same methodology. The stacking method (Wolpert, 1992) introduces the idea of a higher level meta-learner that combines predictions from multiple base models. The meta-learner is trained on how to combine forecasts from different base models; the training data is often divided into different parts for base models, and for meta-learners (Sewell, 2008), in order to reduce over-fitting. The meta-learner generally takes one of the two forms: meta-regressor, or meta-classifier.

A meta-regressor combines model predictions from different models into an ensemble model forecast. This can be viewed as an advanced form of cross-validation of model performance (Wolpert, 1992). Linear combination of different models is a common way of combining base models in the ensemble (Breiman, 1996), producing weights for forecasts from different models (Van der Laan et al., 2007). The stacking method differs from the weighted average methods, in that weights for different base models don't have to add up to one.

The stacked generalization can be used even for a single model prediction (Wolpert, 1992), in which case the meta-learner corrects errors from this single base model, and becomes similar to a boosting algorithm.

A meta-classifier attempts to identify which of the base models is likely to be the most accurate in each case, using the same set of explanatory variables as the base models. The meta-classifier selects one prediction from one of the base models as the ensemble model

output. The rationale for the meta-classifier is that different models may have varying performance under different circumstances. For example, in mapping applications, some roads will be missing from the maps due to various reasons, and algorithms are used to identify these missing links; it is found that no single algorithm is best or worst overall, and that the forecast can be improved by using the random forest as a classifier to select the most suitable model for the situation (Ghasemian et al., 2020).

The meta-learner reportedly has lower computation cost, compared to training each individual model on the full dataset (Van der Laan et al., 2007). However, there is concern that in practice, the stacking method may over-fit the data (Van der Laan et al., 2007). Freund et al. (1997) suggests that results from stacking will be no worse than the best single model prediction.

In social science, it is a common practice that different models are combined by assigning weights (from a meta-regressor) to different predictors, and is termed 'multiple regression' (Kahneman, 2011). In practice, complex super learners often have little or no improvement to model accuracy (Dawes, 1979), the equal-weight combining rule often works well in practice, because assigning equal weights is not affected by sampling errors (Kahneman, 2011).

Compared to simple average and weighted average combining rules, the stacking method provides a more targeted approach for improving model accuracy, by training meta-learners. However, a common concern among the users of stacking methods is that the meta-learner may over-fit. Therefore a sufficiently large training data may be a prerequisite for the use of stacking ensemble models.

2.2.4 Segmented Models

Segmented regression is a basic form of selecting different models for different conditions. The segmented regression method partitions the regression line into multiple connected segments, each with different incline to fit data points within different intervals of the explanatory variable. Without the meta-classifier, the segmentation is generally based on a single variable.

In Haider (2019), regression line segments with different slopes and intercepts are used to fit the changing tends in the transport mode share in relation to increasing population density, to account for the plateauing transit mode split, once the population density becomes large enough.

2.2.5 Bootstrapping and Bagging

Bootstrapping and bagging refer to the practice of repeatedly taking random small samples, as well as subsets of explanatory variables, from the larger training data, with replacement (bootstrap samples), then fitting models to each sample, and combining these models (bagging).

The idea behind the bootstrapping and bagging methods is that, according to Perrone and Cooper (1992), there is the probability that some other models may perform better than a single model on some unseen data, so the average performance of these models has better reliability than relying on a single model. The well known Random Forest algorithm is an ensemble model based on combining multiple classification trees.

The philosophy behind bootstrapping and bagging can be extended to any model. Practices such as repeatedly taking different training samples from the larger dataset to train the models (Wichard and Ogorzalek, 2004, Zhou et al., 2010), or randomize model parameters (e.g. the number of nearest neighbors in the k-nearest-neighbor, the structure, and type of activation functions in the neural network, the kernel function in the support vector machine) to train different models generally improves model accuracy. Bootstrapping and bagging applied to neural networks alleviates the local minima problem, and improves the classification accuracy by combining predictions from multiple networks (Perrone and Cooper, 1992). When applied to multinomial logit models in forecasting travel demand (Cheng et al., 2019), the ensemble has higher accuracy, and lower computation cost when compared to conventional multinomial logit models.

2.2.6 Boosting

Boosting refers to a class of ensemble modeling that turns multiple weak learning algorithms into a model with high accuracy. A weak model (either regressor or classifier) is first trained, which is barely better than random guessing; successive models are trained to predict the remaining error from the previous model, in an iterative manner (Schapire, 1990).

There have been many variants of boosting techniques; these variants share the same objective, but have slight variations in implementation. Enumerating all the boosting algorithms is not within the scope of this review, as the numbers are high, and more algorithms are still undergoing development. Zhang and Ma (2012) provide an account for different boosting algorithms. Among the myriad of boosting algorithms, gradient boosting is a basic type of boosting algorithm; adaptive boosting (Adaboost) is a common boosting algorithm, which repeatedly uses the same training data, and adaptively increases the weights on difficult cases and decreases weights for already well predicted cases (Sewell, 2008).

2.2.7 Probabilistic, Non-deterministic Ensemble Models

Probability, Maximum Likelihood

There are different ways to present ensemble model outputs, and one of these methods is to show model output as a distribution of the dependent variable, instead of a single value. It is common practice in weather forecasting to obtain multiple path predictions, and convert these numbers into a distribution. To obtain the parameters for the distribution from multiple model predictions, one method is to sort the model predicted values (e.g. temperature) into bins, then assuming same probability of the true value falling into each bin, and conduct a maximum likelihood estimate (Anderson, 1996).

The asserted benefit of such probabilistic models is better interpretability, and the ease of extracting information from the forecast (Anderson, 1996). For example, Hamill et al. (2004) uses the ensemble mean (average value of forecasts) of temperature anomaly as the explanatory variable in a logistic regression, to predict the probability of the anomaly exceeding a certain value.

Model outputs from multiple probabilistic base models can be combined using weighted averages. For example, the base model output can be expressed as a normal distribution,

instead of a discrete value, and parameters of that distribution (mean, variance) expressed using weighted sum from members of the ensemble, shown in Equation 2.5; the weights are then estimated through a maximum probability estimation of the training data, given the parameters in the normal distribution (Gneiting et al., 2005).

$$\mathcal{N}\left((a + \sum_{i=1}^{I} W_i \cdot x_i), \sigma^2\right) \tag{2.5}$$

Bayesian Model Averaging

Bayesian model averaging has many variants, and is intended to average distributions from base models of the ensemble, rather than averaging discrete values. It can be considered as a more complex version of weighted average, but with probabilities. In Bayesian model averaging it is assumed that all the data is generated by the same underlying mechanism, and different models reflect the uncertainty as to which one represents the true model (Sewell, 2008).

The forecast probability density function (PDF), p_y, from Bayesian model averaging is shown in Equation 2.6, per Raftery et al. (2005), where the $p_{(y|M_k)}$ is the forecast PDF using only the model M_k, and the $p_{(M_k|y^T)}$ is the probability of model M_k being correct in the training data. Weights of (the PDF of) each model depends on the probability of each model being correct (Raftery et al., 2005) or closest to the observed data. The parameters are estimated using maximum likelihood estimation. Bayesian mode averaging is reported to have good performance with weather forecast (Gneiting et al., 2004); another research finds simple averaging to have higher accuracy than Bayesian averaging (Graefe et al., 2015).

$$p_y = \sum_{k=1}^{K} p_{(y|M_k)} p_{(M_k|y^T)} \tag{2.6}$$

$$\sum_{k=1}^{K} p_{(M_k|y^T)} = 1$$

y : Data to forecast
y^T : Training data

2.2.8 Analog Ensemble

The Analog Ensemble method draws historical data with a high degree of similarity, and uses these historical values as an ensemble prediction.

The Analog Ensemble is used in short term weather forecasting (NCAR, 2020), which draws on past data that are similar to the observed situation, and takes the observed past values into the ensemble; these past observations constitute the ensemble prediction for the current forecast (Delle Monache et al., 2013). The Analog Ensemble technique has the benefit of low computation cost (Delle Monache et al., 2013), and the forecast is often expressed as a probability distribution (Shahriari et al., 2020).

2.2.9 Judgemental Adjustments

Judgemental adjustment combines model predictions with human judgement. Human judgements are based on experience, and can sometimes be a useful addition to improving model predictions.

In weather forecasting, combining output from computer clusters with human judgement is able to improve the forecast accuracy by 25% for precipitation, and 10% for temperature (NOAA, 2012a,b). Human judgement has some value in pattern recognition in weather forecast (Silver, 2012). In many other disciplines, making judgemental adjustments also improves model accuracy (McNees, 1990, Silver, 2012). However, in psychology, and economic predictions, judgemental adjustments don't always improve model predictions, and in some cases can even reduce forecast accuracy (McNees, 1990). In predicting economic statistics (four quarters) such as the interest rates and consumer prices, human judgements improves model forecast in 75% of the cases; but in predicting the level of inventories and the amount of imports, incorporating human judgements reduces forecast accuracy in 66% of the cases (McNees, 1990).

Human judgement by itself is pretty inaccurate in some disciplines. For example, human judgement performs worse than statistical models in clinical applications (Meehl, 1954), and mostly useless for candidate selection in military recruiting (Kahneman, 2011). Whether human judgements can improve upon model prediction has not been tested in these cases.

Empirical evidence on the effectiveness of judgemental adjustments to model outputs has been very mixed. It appears that the value of human judgement depends both on the discipline, and on individual level skill and experience. In many cases, however, the mixed effectiveness in judgemental adjustment is likely a result of regression to the mean.

2.2.10 Ensemble from Perturbation of Initial Condition

Perturbation of initial condition is a practice that adds small disturbances to the measurement data (i.e. explanatory variables), and subsequently combining an ensemble of models trained on the datasets with different perturbations (Silver, 2012, Xu et al., 2014). This practice recognizes uncertainties in the measurement data, and is intended to display a range of possible outcomes, that resulted from measurement uncertainties.

The perturbation method is developed and used initially in weather forecasting, where the behavior of the weather system, and models describing the weather system are non-linear (Gustafsson, 2002), and slight errors in the initial measurement can accumulate over time to render the forecast useless. Models trained on many sets of slightly perturbed initial conditions are able to produce many possible paths (ECMWF, 2013), and it is hoped that the true trajectory of events will follow the mode of the ensemble forecast (Leutbecher and Palmer, 2008).

Accounting for perturbation in measurement data improves prediction accuracy in weather forecasting. In predicting the trajectory of Hurricane Katrina in 2005, an ensemble model consisting of one model using the best guess of the initial state, and 50 parallel models with slightly perturbed initial condition measurements were used (ECMWF, 2013); the true trajectory of the hurricane came close to the mode of the ensemble predictions (Leutbecher and Palmer, 2008). The added perturbations represent the amount of measurement uncertainties,

the effect of which is explored by multiple path predictions from ensemble models. In practice, the perturbation can be taken from a distribution, and the standard deviation of that distribution can either be based on expert judgement (Ollinaho et al., 2017), or obtained from other models measuring the extent of measurement uncertainty (Christensen et al., 2015).

In the transport context, perturbing the census data does not appear to have a large or systematic effect on the performance of transport models (Zador and Levinson, 2013). It is possible that the extent of perturbation needed to protect respondents' privacy is too small to have any major effect on the model. Transport models also don't experience the same rapid accumulation of error over time that weather models do, so the effect of measurement uncertainties may take a longer time span to pose significant problems.

2.2.11 Federated Learning

Federated learning (Konečnỳ et al., 2015) is a breed of ensemble models that aim at protecting the data privacy while maintaining a level of model performance. Instead of collecting data and training the models at one centralized location, federated learning distributes algorithms to individual devices, or to a third party server where the data are stored; after the algorithms were calibrated, only the calibrated parameters are collected and sent back. Improving model performance is not the major motivation of federated learning. Gupta and Raskar (2018) report similar performance in federated learning compared to conventional models.

2.2.12 Balanced Scorecard

The balanced scorecard (Kaplan and Norton, 1998) is an ensemble performance metric for complex systems, where the systems often have multiple objectives, and a single number is not sufficient for measuring performance. The balanced scorecard can graphically present different performance metrics, and relation between these metrics, to facilitate decision making processes. The balanced scorecard is intended initially for making decisions in business management applications (Kaplan, 2009), but the idea of the ensemble performance metric applies to areas outside business management.

In the context of modeling, average forecast accuracy and forecast reliability are two performance metrics for models, and converting these two performance metrics into a single number is difficult, and sometimes misleading (e.g. the root mean squared error, RMSE is an attempt to penalize large errors, to account for forecast reliability). The balanced scorecard idea can be incorporated into model performance metrics to show different aspects of model performance.

2.2.13 Expert Systems (Delphi Method)

The Delphi Method is a process where a group of experts communicate and exchange views in a 'structured' way (as opposed to casual exchange of ideas) while following a formal protocol, to allow each individual make their own independent judgement without much influence from a dominant individual. The individual judgements are then synthesized into a group judgement (Linstone et al., 1975).

The Delphi Method is able to inquire and analyze a broad range of possibilities (Linstone et al., 1975). The use of Delphi Methods provide 'checks and balances' from a panel of experts, which reduces the possibility of erroneous prediction made by a single expert.

The term 'expert systems' has other meanings depending on the context. In computer science 'expert systems' refer to automated systems, with a set of rules built on a pool of knowledge and past experience, that can suggest solutions (Lucas and Van Der Gaag, 1991). This is a scaled down version, and substitute for a group of live experts, and uses past knowledge and experience.

2.2.14 Meta-analysis

Meta-analysis (or meta-combination) is a broad category of ensemble methods that combine forecasts on the same subject from different studies, organizations, or individuals. Meta-combinations can be carried out with simple rules, using algorithms, or with ad hoc methods.

Empirical evidence form the literature shows that forecasts using meta-combinations are generally more robust and accurate than relying on a single prediction, since individual errors tend to cancel each other. Average or aggregate predictions from a group of individuals are generally more accurate than forecasts from a typical individual (Silver, 2012). Forecast algorithms developed by people of different demographics are found to exhibit correlations in predictions errors, so meta-combination by cross-demographic averaging may improve accuracy (Cowgill et al., 2020).

The 'wisdom of the crowd' (Surowiecki, 2005, Tetlock and Gardner, 2016) is a peculiar type of meta-combination, that combines subjective judgements, with the hope that biases among individual judgements will cancel out. An example is given by the market economy, where the collective judgement by a large group of people can often efficiently set the proper price for stocks and commodities, so that the market is efficient (Fama, 1960) most of the time. Although bubbles do exist in market economies, nations that implemented market economies generally fare better than the socialist planned economies (Soviet Union, North Korea, and China between 1950 and 1978).

Commercial weather forecasting companies combine forecasts from different sources. For example, The Weather Company combines 162 different model inputs (which are themselves ensemble models) in producing its own real-time weather prediction (Blum, 2019).

Meta-combination has a wide range of applications in economics and business analyses. In a survey by Dalrymple (1987), about 40% of US companies frequently combine sales forecasts from different sources, and another 28% occasionally combine forecasts. Combining two forecasts each based on a different surveys for the home buyers' intentions reduces MAPE by 1% from the one more accurate single forecast (Armstrong, 2001, Okun, 1960); combining any two of three forecasts from the U.S. Department of Commerce's Bureau of Economic Analysis (BEA), McGraw-Hill, and Merrill Lynch Economics reduces error by 11.8%, and 20% if all three forecasts are combined (Armstrong, 2001, Landefeld and Seskin, 1986). Equal-weight combination of earnings per share forecasts produced by different financial analysts from multiple financial agencies reduces forecast error (Lobo, 1992).

The components of meta-combination can themselves be combined forecasts (Armstrong, 2001). For example, when judgements on the probability of survival from multiple groups of physicians are combined, the resulting prediction accuracy on the survival of 231 patients

admitted to intensive care units improved substantially (Winkler and Poses, 1993).

2.3 Choice of Base Models

It is noted that a large number of models in the ensemble does not equate to better ensemble model performance (Zhou et al., 2010). Particularly bad models in an ensemble may become detrimental to the ensemble model performance. Sometimes it becomes necessary to selectively include models in the ensemble.

Methods to select models for the ensemble include picking models with good performance, such as small mean square errors in the validation data (Wichard and Ogorzalek, 2004). For a group of support vector machine models, Zhou et al. (2010) select models with small covariance (within threshold) with other models, and include them in the ensemble; the small covariance criteria signals a model needs to be largely in agreement with predictions of other models.

2.4 Review Summary

This review finds a range of different ensemble methods used in many disciplines outside transport modeling. Most applications of ensemble forecasting use ad hoc methods, and methods of ensemble forecasting are referred to using different names in different disciplines; ensemble methods of different levels, some degenerate, are all covered under the broad term "ensemble", which may cause confusion, and hinder communication between disciplines. This review also attempts to clear up some of the confusion, and what is meant by "ensemble" in the literature.

Accuracy improvement is often cited as the major benefit of ensemble forecasting. The full apparatus of ensemble methods are often not explored, and the benefit of ensemble models in potentially improving forecast accuracy and robustness have not been fully realized.

Applications of ensemble forecasting in transport modeling remain very limited. The small number of applications in transport modeling suggest it remains a niche practice in the field, and the use of the one best model only (what we refer to as 'the single model doctrine') continues to be the standard practice that guides transport modeling. Average accuracy metrics are often used as model performance, and the overall usefulness of the model as a decision support tool, including the forecast robustness, and the distribution of possible model outputs, are often overlooked. Awareness of the ensemble forecast as an available option appears to be lacking in the transport community.

This review examines the literature for real-world performance of ensemble models both in and outside of the transport field, and finds ensemble models to generally improve forecast accuracy beyond base models. In situations where the ensemble forecast doesn't improve forecast accuracy, it either slightly under-performs, or maintains the same level of accuracy as base models. There has not been any case reported in the transport literature where ensemble methods significantly worsens forecast accuracy.

The lack of cases where ensemble models reportedly worsen forecast accuracy can either be a sign for the superiority of ensemble models, or due to publication bias (Easterbrook

et al., 1991) making null results less likely to get published, therefore inflating the performance of ensemble models. Because of the publication bias, it is impossible to deduce the true effectiveness of ensemble forecasting from reviewing the literature. In evaluating the performance of ensemble models in this research, we present results from each experiment with ensemble forecasting, including both positive, and any null result where the ensemble models either aggravate, or have no effect on model performance.

2.5 Synthesis of Ensemble Methods

Within transport modeling, there are currently no systematic methods, or rules of thumb, for identifying suitable ensemble forecasting solutions for different scenarios. Ensemble forecasting generally does not work 'out-of-the-box', because each specific transport problem has its unique data availability, different sources of uncertainties, and requirements for forecast accuracy and reliability; different ensemble methods can also produce different types of ensemble model output, from a single number to a range of possible outcomes. In order to methodically apply ensemble forecasting methods, and to achieve its full potential for improving forecast accuracy and reliability, theory and methods on the application of ensemble forecasting are needed.

In this section we synthesize methods of ensemble forecasts, categorizing ensemble methods into two broad classes: combining data and combining models, which act as two pillars of ensemble forecast. These two classes are not mutually exclusive, and can be used in conjunction, albeit with increased complexity and computation cost. In Figure 2.1 we list different combinations of ensemble methods. In addition to combining data, and combining models, we propose further combining alternative methods of combining data and models, which can be iterated until gains are exhausted – which forms an ensemble of ensemble models. The ensemble of ensembles recognizes that a single rule for combining models may not be optimal, and is in line with the theory of ensemble forecasting, that all uncertainties should be accounted for. This section formalizes the theoretical framework of ensemble forecasting for transport problems; this framework is shown in Figure 2.2.

Ensemble forecasting is a general concept in modeling, and depending on which pieces are combined, there are different extent and levels of ensemble models. At the one end of the ensemble spectrum, there are ensemble models closely related to single-model predictions. For example, coefficients from the mixed logit model are based on distributions, and random draws from these distributions can generate a range of possible outcomes, which produces a result better at dealing with risk than uncertainty; sensitivity tests on explanatory variables can similarly generate different model outcomes based on different scenarios, but cannot of itself tell which of those scenario outcomes is more likely. At the other end of this ensemble spectrum, models can combine both data and different model formulations, add perturbations to data, or form an ensemble of ensemble models.

Here we formulate the traditional single-model approach as in Equation 2.7. The model output (Y_i) is a particular instance produced by a single model specification (f_s)) that uses a single set of measurement data (X_d). This single-model, single-data modeling approach is the basis for ensemble models in later sections. Model specification involves both selection of model variables and the statistical technique for combining those variables (linear regression,

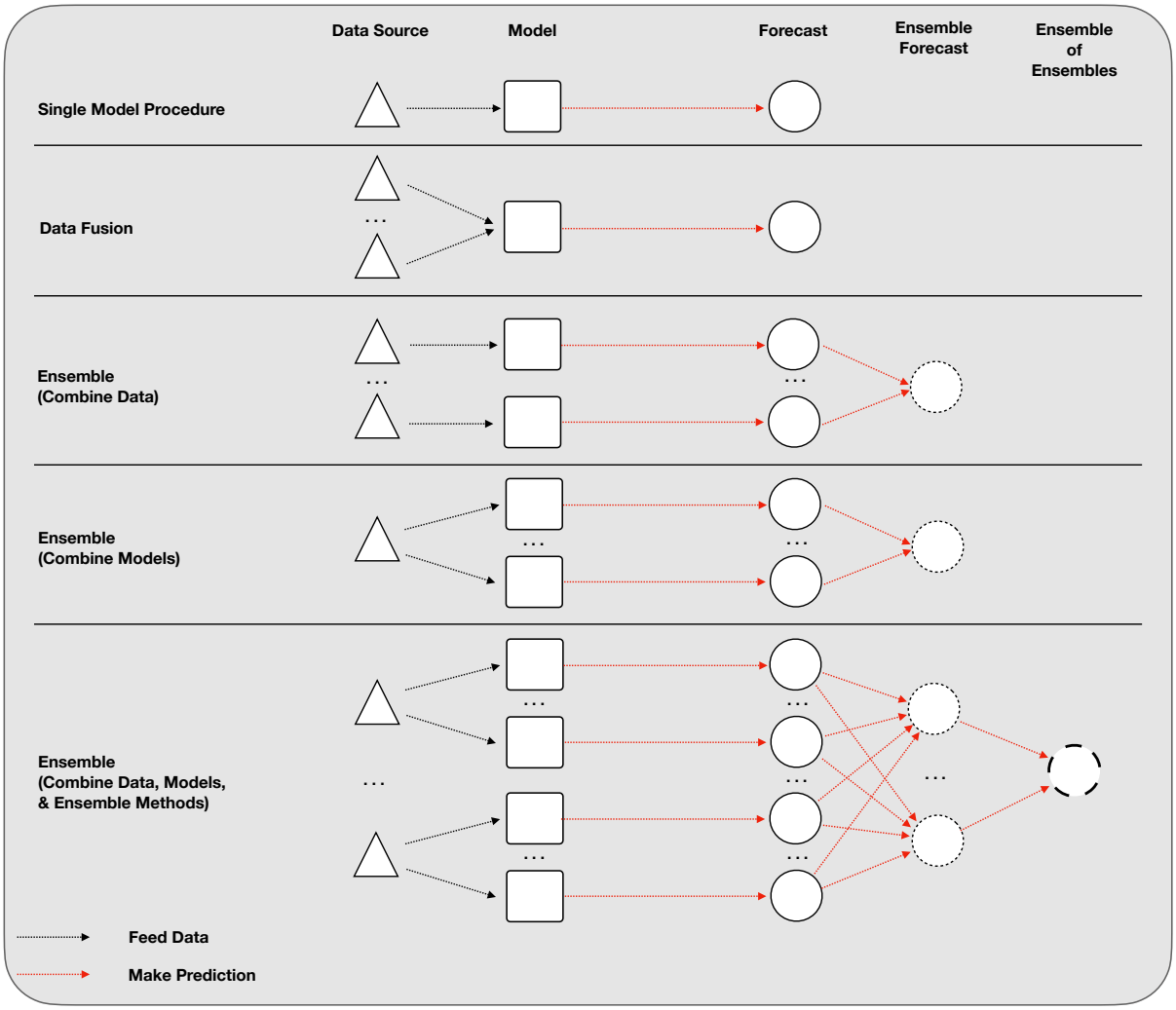


Figure 2.1: Methods of Combining Data and Models

log-linear, logit, ML, etc.)

$$Y_i = f_s(X_d) \tag{2.7}$$

2.5.1 Combine Models

Combining models is a general purpose ensemble method that synthesizes information from different types of models, and considers multiple model assumptions, and methods of pattern recognition. A major motivation in combining models is to pool information from different models, to filter out noise, which improves model accuracy. The practice of combining models also makes up for deficiencies in relying on a single assumptions on the data generation process; the ensemble model output comes from a group of different base models, providing checks and balances in case one or more base models went wrong.

Methods for combining models focus on combining outputs from models with different formulations. Each base model would function independently in producing its own predictions,

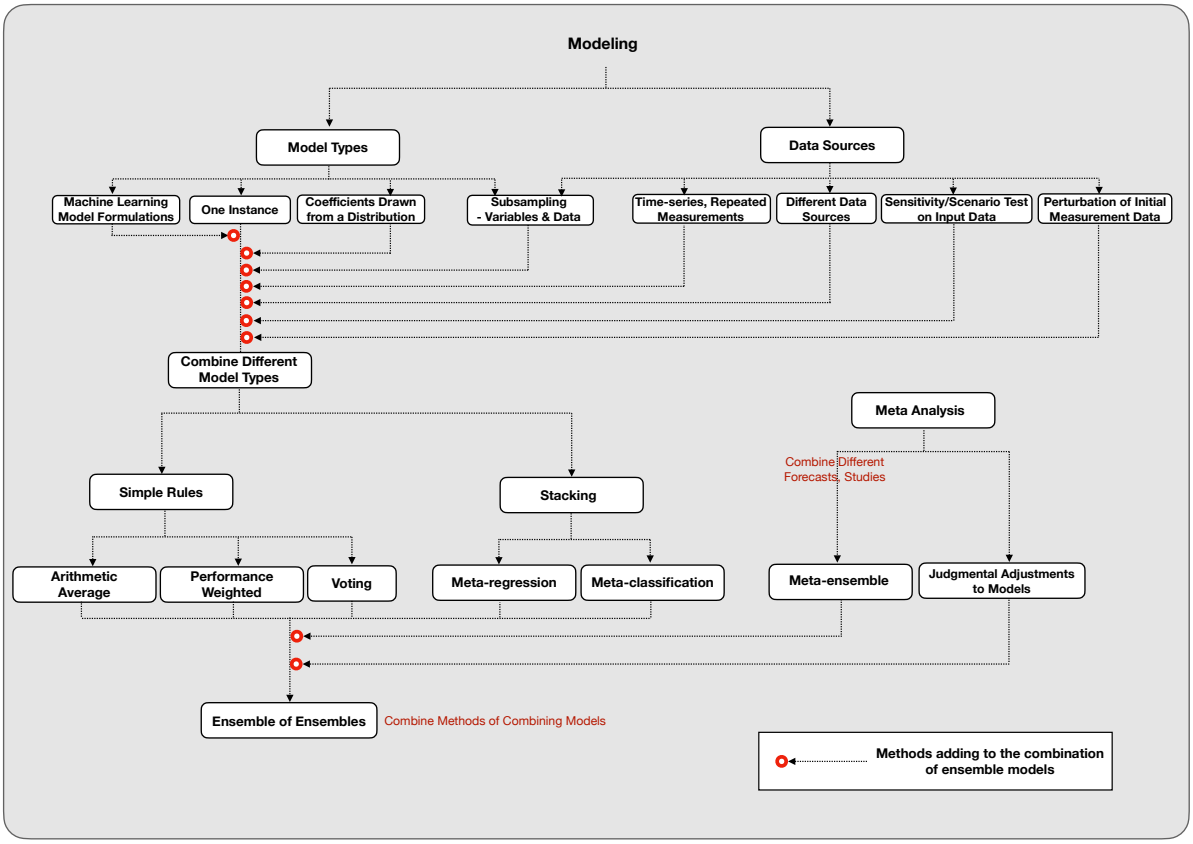


Figure 2.2: Framework for ensemble forecasting

and these predictions are combined following a set of predetermined rules.

Equation 2.8 shows the formulation for combining models. Here the model output (Y_I) is based on outputs from different model instances (Y_i, $i \in I$) that each result from a particular selection of model specification and data, that are combined through ensemble methods (e).This section discusses rules for combining models.

$$Y_I = e[\{Y_i\}] \forall i \in I \tag{2.8}$$

Simple Rules

Application of simple rules constitutes a set of aggregated methods for combining models. Different base models are calibrated independently using the same set of training data, and then applied to a distinct testing dataset. Each base model produces an independent forecast, which are then combined using simple rules.

Simple rules are based on weights, which are assigned to each base model in the ensemble, and the final ensemble forecast is a weighted sum of outputs from different base models. The weights discriminate only between models, so all outputs from a single model are assigned the same weight. Weights assigned to different models sum to one.

We categorize simple rules ensemble methods into three categories:

- **Weighted average**. With the weighted average rule, a rule for generating weights is required. For instance, the weights can be based on model performance in the training data, so models with better performance in the training data are assigned more weight. Alternatively, each model can have an equal weight, which is then equivalent to a **Simple average**.

- **Weighted stacking**. 'Weighted stacking' differs from the 'weighted average' method, in that the training data is divided into two parts in 'weighted stacking', with one part used for training the base models, and the other part used for acquiring the weight.

  The benefit of weighted stacking is in more accurate weights for actual model performance. Some algorithms would have better performance in training data, than in testing data, so using weights from the training data may be deceiving, and not actually reflect model performance. The weighted stacking trades off the size of training data for more accurate weights, so the resulting models might be less accurate as a result.

- **Voting**. The voting scheme is intended to resolve disagreement among classification models.

Stacking

The stacking method (Wolpert, 1992) calibrates a higher level model (meta-learner) to combine forecasts made by different base models. Stacking models consist of two layers of models:

- base models, that are calibrated using training data to produce predictions just as normal models;

- a meta-learner, which learns how to combine base model forecasts by using a portion of the training data. To avoid over-fitting, training data used by base models and the meta-learner must not overlap.

Each time, the training data is divided into two parts, with one part used to train the base models, and the other part used to calibrate the meta-learner. To compensate for the reduced size of training data for base models, multiple base models and meta-learners are trained by repeatedly dividing the training data (e.g. k-fold cross-validation (Chand et al., 2016)). Figure 2.3 shows the process of dividing the training data, and calibrating both base and meta-learner models, which will be repeated multiple times.

The stacking method can be further categorized into meta-regression, where the meta-learner is a regression, and the meta-classifier, where the meta-learner is a classification (discrete or qualitative outcomes) model.

Meta-regression The meta-regression method uses the outputs from the base models as the new set of explanatory variables to predict the event outcome, in which case the coefficients from the meta-learner can be viewed as weights, that can be different in each case. Alternatively, the original set of explanatory variables can be added to the meta-learner, to provide information on the context of each event; the coefficients from the meta-learner can no longer be viewed as weights.

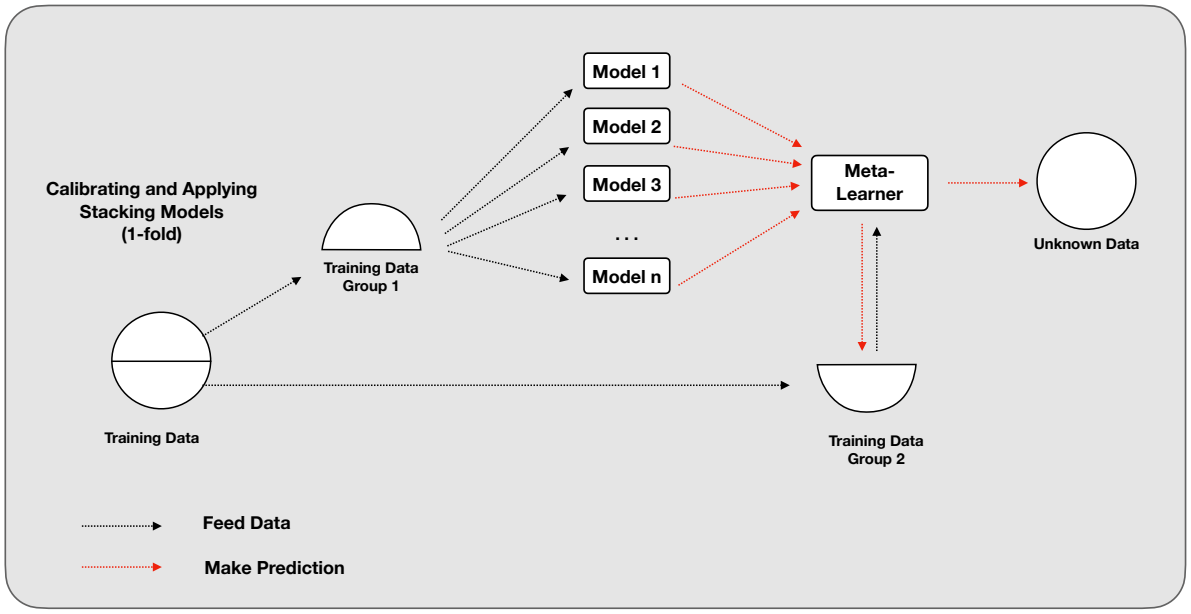


Figure 2.3: Ensemble methods, meta-learners

The meta-regression method has the risk of mistaking noise for signal, especially with a small sample, in which case applying equal weights to different model outputs provides a more robust solution. The meta-regression generally uses only the different model outputs from the base models, although it is possible to add the variables used in the base models.

Meta-classifier When one model performs better than other models under some specific circumstances, it becomes rational to select models based on the circumstances. The meta-classifier uses the same explanatory variables as in the base models; these variables are used by the classifier in determining the specific conditions faced by the base models that would favor adopting certain models. The basic idea of a meta-classifier is similar to that of the segmented linear regression (Haider, 2019) where different models are fitted to different segments of the data. In cases the weights in the meta-regression are discrete '0/1', the meta-regression becomes similar to the meta-classifier.

The performance of the meta-classifiers depends on whether there is easily recognizable patterns with the performance of models and scenarios, and if such patterns are consistent. Examples are abound for where selecting models may work better than single models, including the lumpiness in human behavior to outside stimulus, and seasonal growth of plant and animal populations.

Ensemble of of Machine Learning (ML) Models with Different Formulations

Formulation of machine learning models affects performance. There are complex algorithms for finding optimal formulations for specific problems and datasets (Ren and Zhao, 2002, Sanger, 1989), as well as rules of thumb (Wanas et al., 1998) for selecting the optimal model formulation. In practice, model formulations are generally obtained through a grid search, by testing different formulations and corresponding model performance on a testing dataset.

In linear regression, model formulations selected based on the highest R^2 are often wrong (Bacon, 1977, Kennedy, 2003, Mayer, 1975); in the same light, machine learning formulations based on model performance in limited testing data may also be wrong. Since the 'best' machine learning model formulation may not exist, or cannot be obtained, we note the possibility of combining machine learning models of the same type, but with different formulations. Examples include neural networks with different number of neurons and number of layers, with different activation functions, or with different learning rates, and gradient boosting machine with differing interaction depth, and random forest with different number of trees.

One example of ensemble ML models with different formulations can be found in the Random Forest model, which is an ensemble of Classification Trees with different subsets of explanatory variables and training data. In our experiments, Random Forest has consistently outperformed the Classification Tree. The performance of other ensemble ML models with different formulations is also worth testing for other practical reasons. For instance, the number of possible model formulations might be too large, or too computationally intensive to enumerate; the amount of available data might also be insufficient to test for different model formulations. An ensemble of different ML model formulations provides an alternative for such circumstances.

2.5.2 Combine Data using Ensemble Models

Transport models sometimes encounter data from different sources, from multiple measurements, or time-series data; in some cases significant variations exist between these different data sources. For instance, population and employment data collected by the Census Bureau and local governments may be in disagreement. Alternatively, perturbations added to the census data (for privacy protection) may have limited effect on some transport models (Zador and Levinson, 2013), but its impact on other longer term forecasts remains unknown. Therefore it often cannot be known which of the data sources are more accurate than others, or if the measurement data include errors. Combining data from different sources reduces the risk of relying on a single data source, and potentially incorporates more information into the model.

Even in completely deterministic processes, such as when the data generation process can be completely captured by the model, small errors in the measurement of the initial condition can accumulate over time in the modeling process, and cause drastically different model outcomes. This is known as chaos (Lorenz, 1995), or more widely known as the butterfly effect (Lorenz, 2000). And because it is impossible to measure continuous real-world events with discrete numbers, it often becomes necessary to create multiple datasets, by adding random perturbations to the initial measurement data, to test the effect of each perturbation. The model outcome is expressed as a range of possibilities.

There are at least two avenues for combining different data: (a) averaging measurement data, and, (b) combining models that use different measurement data (these models may have the same functional form and specification, or may differ). Ensemble forecasting enbables the second avenue for combining data.

The advantage of combining data with ensemble models is in showing the range of possibilities (including the best to worst scenarios), which is not possible with averaging the

measurement data, which may improve the estimate of the central tendency at the cost of knowing about the tails of the distribution. In non-linear systems perturbations of identical amounts but in different directions have different degrees of impact on the outcome of an event. The reasoning is that different models based on each measurement dataset have unique paths leading to an event outcome, and averaging the measurements prematurely does not necessarily average the event outcome.

To combine data using ensemble models, data from different sources can be fed one-at-a-time into parallel models, simultaneously calibrating multiple models; these parallel models are generally of the same type, but can have different formulations, in which case it becomes an ensemble of both data sources and different models. Figure 2.1 shows the difference between combining data using ensemble models, and combining data by merging data. In ensemble models, each of the data source calibrates the model once, each producing one forecast.

Uncertainties in the measurement data is one of the reasons for the multitude of possible model outcomes. When there is need for the ensemble model output to be a single number, averaging these model outputs is one way to reflect the combined effect of the different data sources on the model. When the data sources are equally valid, the model outputs resulting from these measurements should have equal weights.

The method to combine data in ensemble forecasting is the same as combining models Equation 2.8. The instances are just variations in input data rather than variations in model specification. In more complex applications, there can be different types of models, resulting in the variations of both data and models in the model forecast instances that are combined with the ensemble method.

Different Data Sources

Data sources can include data from different institutes, from repeated measurements, or time-series data, etc. Parallel models each using a different data source can produce a distribution of model predictions. Since the accuracy of each data source is unknown, this distribution considers different hypothetical outcomes, assuming each of the data measurement were true.

Ensemble models with perturbed initial measurement data is perhaps the most frequently used type of multiple path prediction, which is also the standard practice in weather forecasting. Because small errors in initial measurement tend to expand rapidly in weather models, the method is intended to plot all likely results from measurement uncertainties, and to learn which outcome has a higher probability. Similarly, traffic simulations typically are run multiple times with different random seeds to get better estimates of traffic states than relying on a single model run. This is seldom done with strategic planning models, which are assumed to be deterministic rather than stochastic, but testing shows results tend to converge (Zhao and Kockelman, 2002).

Sensitivity, Scenario Testing

Sensitivity and scenario testing are degenerate methods for combining data through ensemble methods. Although useful in some cases, these tests don't consider the possibility of other potential model assumptions, and rely on a single model formulation, and models are

calibrated using a single dataset.

The number of possibilities explored is also limited in sensitivity and scenario testing. The widely used single model procedure is limited in accounting for uncertainties in its own model assumptions, or accumulated error over time. Outputs from sensitivity and scenario testing reflect the limited effect of a single input variation in the model output, assuming the input explanatory variables will vary by exactly the amount tested. This however does not account for the compounding effect of uncertainties, or the wide range of possible inputs.

Perturbation of Initial Measurement Data

An enhanced version of sensitivity and scenario testing is provided though ensemble models with perturbations of initial measurement data. The perturbation method is used by default in weather forecasting, where a range of variations with different degrees are added to the initial measurement data. In iterative processes, each of these perturbed data produces even more perturbed data points with each iteration, causing the number of possible data combinations to grow exponentially.

Ensemble models using this method displays a wide range of possible model outcomes, and the computation time also grows exponentially with longer forecast time frame. An engineering solution to reduce computation time is to reduce the number of additional perturbed data points (Blum, 2019), which inevitably reduces the resolution of the model.

This process of data perturbation also incorporates the amount of uncertainties into the modeling process. For long-term forecasts, or in volatile models, the range of possible model outcomes will also become wider. In contrast to deterministic model outputs in most sensitivity and scenario analysis, model outputs from the perturbation method are presented as probabilities.

2.5.3 Multiple Path Prediction

Methods for combining data and combining models can interweave to produce multiple path predictions. Models cannot reflect reality with absolute certainty, so uncertainty is an intrinsic part of modeling, which means that models cannot provide definitive answers in predicting future events. Real-world events have a range of possibilities. Predicting something with a single number assumes a level of accuracy that does not actually exist. As a decision support tool, model predictions giving only a single number are generally inadequate, because they do not inform other likely outcomes, such as the best to worst scenarios, and how likely are these scenarios.

Model outputs as a single number are expedient for data processing, which can sometimes be useful. However, there are instances where we would want models to produce a range of possible outcomes, instead of a single number. For example, when predicting an event that has significant importance, it would be useful to know both the predicted value, and the likely range of variations as a result of uncertainties. The narrower the spread of model predictions the more reliable the forecast.

A distribution of possible model outcomes can be produced by a number of methods, each accounting for different aspects of uncertainties, and having different interpretations for the model output. Interpreting ensemble model output as a distribution of possibilities

will require some understanding of how this distribution is obtained, and which sources of uncertainties have been considered. Using multiple path prediction out-of-context poses a potential danger.

In this section we synthesize a number of methods for producing multiple path prediction. Some of the ensemble methods mentioned earlier in this chapter are also capable of producing multiple path predictions, although having multiple paths may not be the main objective of these methods. We note that the list of methods for producing multiple path prediction is not exhaustive, and different methods can also be combined.

Subsampling

Subsampling, or resampling (Politis et al., 1999), is a practice that takes subsets from sampled data, or from the list of explanatory variables, or from both data and explanatory variables. In cases where models are unable to handle vast amount of data, or to reduce unrealistic computation time to an acceptable level, subsampled data is used to calibrate models. In other cases subsampling is used where a probability distribution is preferable to a single number as model output. Subsamples can be taken repeatedly, to calibrate an ensemble of models, which produces a range of possibilities in the model output.

Subsampling potentially filters out noise, and 'accidental' features in the training data. By repeatedly taking subsamples, random noise tends to cancel out, and actual features tend to be reinforced, which is reminiscent of the law of large numbers. Among other methods, bootstrapping (Politis et al., 1999) is an ensemble method that is able to produce multiple path predictions. For example, in the random forecast (RF) model, subsamples are taken repeatedly from both the training data, and from the set of explanatory variables; the result is that different classification trees are calibrated using different subsets of the training data, forming an ensemble of classification trees.

We extend the use of subsampling from single algorithms, to ensemble forecasting. Ensemble models can be calibrated repeatedly using subsamples from a larger dataset, obtaining multiple parallel ensemble models. Predictions from these parallel ensemble models form a distribution from an ensemble of ensemble models, showing multiple possible paths for the event's outcome.

Repeated subsampling is computationally expensive, but depending on computer architecture, it is possible to parallelize model training processes by each subsample, so that the total computation time would be lower than training models with the entire dataset without subsampling.

Multiple Model Assumptions, Pattern Recognition Methods

An ensemble of different parallel models can produce a distribution of different predictions. Although calibrated on the same training data, different data-driven machine learning models explore data patterns in different ways, and different theory-driven models will take different paths in reaching the same conclusions (dependent variables in the training data). Therefore with future or unknown data, each model will produce a different prediction, forming a range of possible model outcomes.

For example, in the early days of the COVID-19 pandemic in early 2020, when data were

scarce, and the growth pattern of infection was not yet clear, different scaling models were used, mostly by non-professionals, to estimate the trend of the pandemic. Models that were most often applied include exponential, linear, and power functions. Although these models were extrapolations (Lammers et al., 2020, Petropoulos and Makridakis, 2020) at best, they were jointly useful in showing that, regardless of which model was fitted, the rate of new cases was increasing at a faster rate than the 2003 SARS virus, which raised concern.

2.5.4 Meta-combination, Judgemental Adjustment

The meta-combination of different studies combines different beliefs, judgements, and mistakes of different individuals and organizations, that generally use different sets of data, and models that differ in the underlying assumptions about the data generation process. Combining independent judgements from a diverse group often result in more accurate predictions, and are more reliable from time to time, than picking a single individual (Surowiecki, 2005). In addition to pooling errors, the diversity (Zhou et al., 2010) in models being combined is also important to the ensemble model performance. Groups of diverse problem solvers can outperform groups of high-ability problem solvers (Hong and Page, 2004).The meta-combination of models is in line with the approach of ensemble forecasting, that the uncertainties in both models and in human judgement (from different studies) should be combined.

Results from different studies are often combined using simple-averages, since it is robust, and ideal for situations where the reliability of each study is unknown. Human judgements play a significant role in the meta-combination of studies, since whether or not to include the forecast from a particular study into the ensemble remains mostly a subjective decision.

Judgemental adjustment to model outputs is a type of meta-combination model, and it sometimes improves model accuracy (Silver, 2012). However, the effectiveness of judgemental adjustments to model outputs depends heavily on the individuals making judgements, and it often require years of experience to be able to identify instances where the model output is obviously wrong.

Human judgements also have value for variables that are difficult to quantify. For example, the hedonic model for house prices can take into account variables describing its location and functional attributes, such as distance to the city center, and number of bathroom; however, the aesthetics of the residential property cannot be easily quantified without human judgement (though one imagines future AI applications trained to do that). The price for most commodities consists essentially of human judgement from the ensemble of all potential buyers.

In a sense, predictions from any type of model will go through judgemental adjustment, in that unreasonable model predictions will likely get dismissed by users of the model.

2.5.5 Ensemble of Ensembles

The *Ensemble of Ensembles* concept is our addition to the ensemble forecasting approach. In ensemble models, data and base model forecast instances are combined in different ways following predefined rules. For example, once a number of base models have been calibrated, and each produces one forecast instance, combining rules such as simple average, linear combination, or random forest models etc. can be used to combine these individual forecasts

into an ensemble model output. However, models that are used to combine data and base models are themselves single algorithms, and therefore have limitations of their own. The idea of ensemble of ensembles is to account for, and alleviate uncertainties within ensemble methods themselves.

There are different ways to implement ensemble of ensembles, ranging from simplistic averages, to more complex iterative methods. Regardless of the specific method used in ensemble of ensembles, initial outputs from ensemble models, which are combinations of base model instances (e.g. linear, random forest meta-learners) are further combined. The goal is to combine different ensemble methods to reduce dependence on any single ensemble method.

The framework for ensemble of ensembles is shown in Figure 2.4. Ensemble of ensembles itself has many levels. A simple type of ensemble of ensembles combines different ensemble models with a single algorithm (iteration 1). More complex forms of ensemble of ensembles can be applied iteratively, using each of the ensemble model outputs from the previous iteration as an input in the next cycle. We expect there to be diminishing returns with each iteration, and so this process has a natural culmination when overall predictive performance ceases to improve. In our limited experiments with up to 200 iterations in the For-Hire-Vehicle (FHV) flow dataset, we found the accuracy improvement to be mostly exhausted in the first 2 iterations, and subsequent iterations have marginal to no improvement. Future research can test the effect of more iterations on other datasets. With more ensemble methods, and more iterations, the dependence on any single model formulation is further reduced.

Iterations in combining models might cause models to pick up noise specific to the training data, which potentially produces over-fitting in the testing data later on. However, we posit that with sufficient training and testing data, this problem is self-limiting. A large number of iterations also require more training data. This is because in previous ensemble methods, k-fold cross validation can re-use the same training data multiple times, because folds are independent; however, iterations in ensemble of ensembles are not independent, and later interactions depends on earlier ones, so the training data cannot be reused.

The most basic type of ensemble of ensembles is the simple average of different ensemble model forecasts (with iteration 1 in Figure 2.4). This ensemble of ensembles method has low risk of over-fitting, is relatively simple to implement, and obviates the need to arbitrarily select one of the combining rules over another. We apply this type of ensemble of ensembles in testing the performance of ensemble of ensembles.

Ensemble of ensembles obviates simultaneously the need to select a single base model, and the choice of a single ensemble methods. There is no limit on the type of ensemble models included in the ensemble of ensembles, nor is there any restriction on the algorithms involved in combining ensemble models. Models with different shapes and forms, including expert systems and judgemental adjustments can all become inputs to ensemble of ensembles; these inputs are combined and validated automatically against a sample dataset. In a sense the model dependency on a single data source, or a single assumption is reduced.

Equation 2.9 formulates the ensemble of ensembles. The combining rule is (E). The building blocks of "ensemble of ensembles" $(Y_{\mathbf{I},\mathbf{n}})$ are themselves ensemble models (Y_I). Multiple iterations (an ensemble of ensembles of ensembles ..., where ensemble of ensembles can become elements of other ensembles) is denoted with subscript (n) to indicated the iteration number.

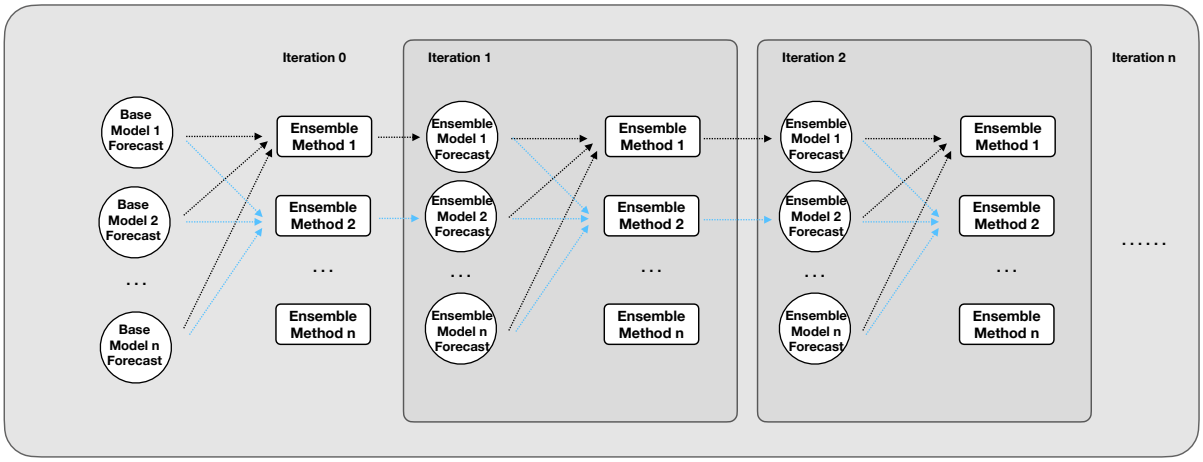


Figure 2.4: Framework for ensemble of ensembles. Repeat until converged.

$$Y_{\mathbf{I,n+1}} = E_n[\{Y_{\mathbf{I,n}}\}] \forall I \in \mathbf{I} \tag{2.9}$$

Note that when:

- $n = 0$, the Equation 2.9 becomes Equation 2.8 (and $E_0 = e$), where $Y_{I,1}$ is just an ensemble of base models.
- $n = 1$, the $Y_{I,2}$ is an ensemble of ensembles
- $n \geq 2$, the $Y_{I,n+1}$ is a multi-layered ensemble of ensembles

2.5.6 Synthesis Summary

As shown in this synthesis, ensemble models can take on different forms and with different levels of complexity, ranging from a simple combination of different models, using different data sources, to complex ensemble models combining different methods for combining models (ensemble of ensembles). Future research may establish more methods for ensemble forecasting, but the adoption of ensemble forecasting does not absolve modelers from the burden of choosing appropriate models. The choice of ensemble method will likely depend on the modeling purpose, data availability, and the level of expertise of the modelers.

Combining different models doesn't require too much data or expertise from the modelers, and can be implemented as a standard practice in modeling. Combining data from different sources can become a better alternative to arbitrarily choosing a single data source when there are multiple data describing the same event; data sources with lesser quality may still be useful for modeling, since these data contain different sets of information. Model predictions using perturbation of initial measurement data will likely be useful for long range transport planning models.

Chapter 3

Model Evaluation Criteria

There are different measures for the performance of a model in producing forecasts, and some of these performance measures are based on the modeling context. Model performance measures generally include the average accuracy of forecasts, the relative frequency of large and small errors, and computation cost. The robustness of a model in predicting future out-of-sample cases is sometimes considered. Models with good performance do not equate being the best model in all aspects, or to possess all desirable properties. Generally models that are good at recognizing patterns are also prone to mistaking noise for signal; naive models are less accurate overall, but are also less likely to produce large errors. A model would be adopted, if some of its properties meet the requirements for a specific application.

Two broad categories of model evaluation criteria are discussed:

- Accuracy
- Reliability

The *accuracy* metric measures the average magnitude of error in model forecasts, which is a measure of a model's ability in picking out signals over noise. Models with good accuracy metrics are desirable for forecasts with large volumes, so on average, these models will correctly predict most of the cases. The accuracy metric alone may become insufficient for applications with large penalties for large forecast errors, or for predicting 'mission-critical' events.

The accuracy of a model, when applied to future datasets, under different scenarios, or predicting events with external shocks to the system, etc., can also be a measure of model robustness.

The *reliability* metric measures the consistency of forecast accuracy, reflecting the extent to which model forecasts can be trusted. Models with good average accuracy, but with frequent large errors indicate low reliability. For 'mission-critical' events, the distribution of forecast errors, and how often large errors occur becomes more important than the average accuracy of the model.

There is no guarantee that models will simultaneously be accurate and reliable, so there are trade-offs. Some of the model performance metrics include only the accuracy measure, such as the mean absolute error; but often accuracy and reliability can be mixed into a single measure, such as the mean square error, which penalizes large errors.

Models picking up noise for signal often display high statistical goodness of fit in the training dataset, even when the model formulations were wrong (Bacon, 1977, Kennedy, 2003, Mayer, 1975). Hence it is important to evaluate the model performance using a separate testing dataset, that is mutually exclusive from the training dataset. Model performance in the training dataset is meaningless and misleading in reflecting the ability of the model in making predictions (although still useful when the purpose of the model is analysis). This chapter discusses different measures of the model performance.

3.1 Proportion of Variation Explained, R^2

R-Squared (R^2) is a frequently used goodness-of-fit metric for linear regression models. The R^2 measures the proportion of variation in the data that is explained by the model, when compared to a naive prediction using the average value. Maximizing R^2 is equivalent to minimizing the sum of squared errors (Kennedy, 2003). The formulation of R^2 is shown in Equation 3.1. An adjusted R^2 is often used to account for the number of explanatory variables.

$$R^2 = 1 - \frac{\sum_{t=1}^{k}(F_t - A_t)^2}{\sum_{t=1}^{k}(\bar{A} - A_t)^2} \tag{3.1}$$

k: Number of forecast/observations
F_t: Model predicted value at the t^{th} instance
$\bar{A}$: Average of observed value (Naive prediction)
A_t: Actual value at the t^{th} instance

The R^2 itself does not measure the magnitude that predictions vary from actual values, so it does not measure directly the accuracy of model forecasts. The R^2 uses the naive prediction as a benchmark, measuring the extent to which linear models are better than using the sample average as prediction. Since the R^2 in a linear regression is averaged over individual predictions, it does not show the relative frequency of large and small errors.

In non-linear regressions, the R^2 measure is invalid. Hence the R^2 measure can only be used in measuring the performance of linear regressions, and in comparing the performance of different linear models.

3.2 Average Accuracy

One common measure of model performance is the average accuracy. The average accuracy reflects the average difference between the model's prediction, and the true value of an unknown event. The average accuracy can be measured in the same nominal units as the variables, such as the Mean Absolute Error (MAE); or by the squared difference between the prediction and the true value, such as the Root Mean Squared Error (RMSE). Both the MAE and RMSE can be applied to compare the performance of different models.

The Mean Absolute Error (MAE) measures the average nominal distance between the prediction and the true value. The formulation of MAE is shown in Equation 3.2.

$$MAE = \frac{1}{k}\sum_{t=1}^{k}|F_t - A_t| \tag{3.2}$$

A variation of the MAE, the Mean Absolute Percentage Error (MAPE, Equation 3.3) measures the magnitude of error in terms of the size of the error relative to the true value. The MAPE shows the average size of the forecast error as a percentage of its true value.

$$MAPE = \frac{1}{k}\sum_{t=1}^{k}\frac{|F_t - A_t|}{A_t} \tag{3.3}$$

The Root Mean Squared Error (RMSE, Equation 3.4) measures the average difference between prediction and actual value, by the root of squared differences. The RMSE penalizes large errors by first taking the squares of the differences, before taking the square root. Taking the squares of errors may appear arbitrary, as there are many other candidate numbers to use as index (e.g. quadruple); the RMSE nonetheless provides a convenient way of partially accounting for large errors.

$$RMSE = \sqrt{\frac{1}{k}\sum_{t=1}^{k}(F_t - A_t)^2} \tag{3.4}$$

3.3 Reliability - Distribution of Absolute Error Sizes

The reliability of model forecast can be measured using the standard deviation of the absolute forecast errors (shown in Equation 3.5). The standard deviation shows the spread of forecast errors,

$$SD = \sqrt{\frac{\sum[|F_T - A_t| - \bar{E}]^2}{k}} \tag{3.5}$$

where $\bar{E}$: Average size of errors; MAE

The indicative figure in Figure 3.1 shows an example of the distribution of the standard deviation of absolute errors by different models. Model 1 has the highest accuracy and reliability; although Model 2 is more accurate than the Model 3 on average, forecast errors of Model 2 has a wider spread, and the longer tail of Model 2 shows significantly more large errors than Model 3.

3.4 Computation Cost

The computation cost concerns the amount of computation required of a model; the amount of computation is reflected in the time required to calibrate the model, and to produce predictions.

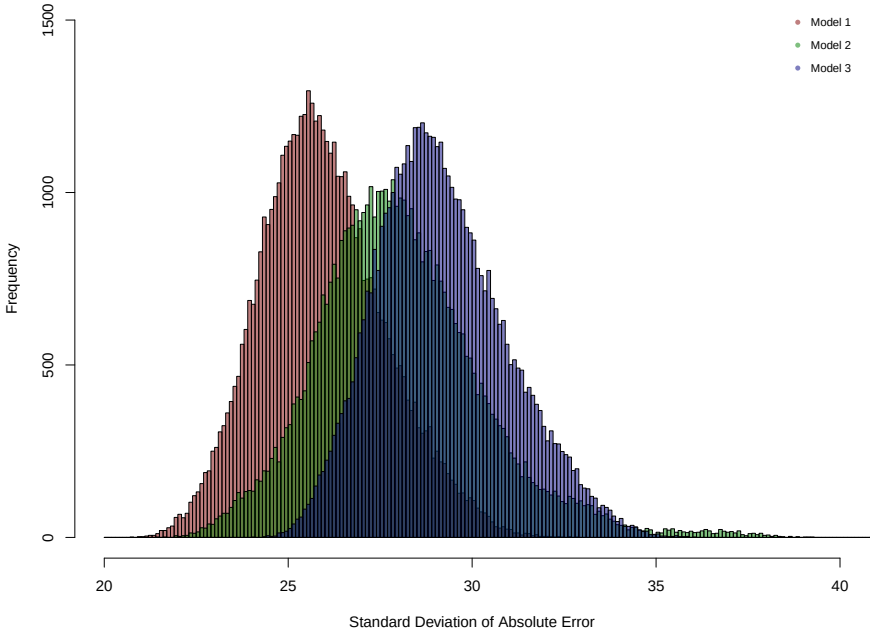


Figure 3.1: Distribution of standard deviation of absolute errors

The computation complexity (CC) definition, per Makridakis et al. (2018), compares computation cost between models. The CC is a ratio of computation time between the model and its naive prediction, as shown in Equation 3.6. For models with similar performance in all other aspects, lower computation cost is preferable.

$$CC = \frac{ComputationTimeModel}{ComputationTimeNaive} \tag{3.6}$$

In ensemble models, the computation time depends highly on the level of parallelization. Different base models can be calibrated either in a queue, or simultaneously in parallel computations; so the actual computation time will differ significantly depending on the amount of computation hardware available. Theoretically the computation time of ensemble models will not notably exceed base models.

3.5 Choice of Model Performance Measures

The objective of this research is to compare the performance of models from different aspects, so the model performance metrics have to be applicable to different model formulations. Some of the performance metrics, such as R^2, applies only to specific types of models, so will not be compatible with comparing different model formulations.

Different measures of model performance have different goals, and examine different aspects of the model's ability in predicting future or unknown events. In this research, we use Mean Absolute Error (MAE), and Mean Square Error (MSE) to measure a model's general

ability to produce accurate forecasts, including both small and large errors. The relative frequency of large and small errors are measured with standard deviation of absolute error, to examine the reliability of model forecasts.

Chapter 4

Predicting Housing Prices with Ensemble Models

4.1 Introduction

In this chapter Sydney housing price data is used as a case study to examine the performance of ensemble models, and to compare ensemble models against base models. The topic of housing prices is closely related to transport, because residential properties derive part of their value from the convenience of transport, so transport related variables have good predictive ability for housing prices. In order to prevent publication bias (Easterbrook et al., 1991), this chapter will report results from all applications of ensemble models, regardless of whether ensemble models improve model performance.

Hedonic models for real estate (Chau and Chin, 2003, Hoehn et al., 1987) relate prices to preferences and characteristics of residential properties. The transaction price of any specific residential property may depend on a particular buyer or seller, and is affected by the market sentiment (Shiller, 2005), but the overall market price of specific types of properties have underlying patterns that determine their value. In this chapter we apply ensemble forecasting to hedonic models, and compare single model predictions with ensemble model outputs in terms of their accuracy, reliability, and usefulness as decision support tools (Wu and Levinson, 2021).

In one of the foundational works of hedonic models, Rosen (1974) posits that goods are sold as, and their prices depends on a package of attributes. In the case of residential properties, this package of attributes includes both its functional characteristics as living spaces, and access, or the convenience in reaching urban opportunities, which depends on the transport network and distribution of places. The combined price of residential property and the cost of transport are subject to constraints of the buyers' budget, creating a trade-off between property price, and transport cost; so similar properties are cheaper at more remote locations, because the cost of transport would be higher (Alonso et al., 1964). This trade-off is backed by empirical data (Nelson, 1977). The positive effect of access to employment opportunities on the land value has long been corroborated, where distance to CBD (Brigham, 1965), or to highway (Mohring, 1961) is used as a proxy for access to jobs or other important destinations. The value for the convenience of transport is reflected in the sales price of residential properties, and accounting for this location-bestowed value is essential in estimating

the value of properties.

Public transport is a significant mode within Sydney, Australia, accounting for 21% of commute trips in the 2016 census (Australian Bureau of Statistics, 2017), so transit access to jobs will be an important consideration in property prices. Bus rapid transit (Mulley and Tsai, 2017) and light rail (Mulley et al., 2018) have been identified to have positive effect on Sydney property prices. Both automobile and transit access to jobs contribute to higher property price in Sydney, and transit has a stronger effect than automobile (Rayaprolu and Levinson, 2019). In addition to functional characteristics and convenience of transport, residential property prices are affected by building aesthetics, availability of house amenities, social-demographics of the neighborhood, and adjacency to undesirable entities (e.g. landfill, high voltage lines), as summarized by Sirmans et al. (2005) in a review of different hedonic models.

Since the determinants of residential property prices are clear, and most are quantifiable, both models and human judgements are capable of estimating residential property prices. However, there are major differences between model predictions and human judgements. The actual convenience of transport can be systematically quantified, but may be perceived differently, or incorrectly or inconsistently by humans. In terms of a building's functional properties, models may not capture factors outside of model specifications (e.g. number of rooms), while humans perceive surroundings with multiple senses that can provide a more detailed picture (Hawkins, 2021). But human judgements are affected by sentiments, and how the house is presented (therefore human judgements can change in a re-test), and obviously vary across people. In most cases where the forecast problem can be quantified, statistical models outperform individual human judgements that are based on past experience (Meehl, 1954, Silver, 2012).

Sufficiently reliable model predictions, provide objective estimates for the value of a residential property based on its location and functional characteristics. The actual sales price of properties often deviate away from its modeled value for various reasons. Cases where the price of a specific property deviates from model prediction provide opportunities to further investigate the cause. Instances in which the listed price significantly exceeds model prediction may signify positive factors outside of the model specification affecting the price, such as aesthetics, in the catchment of highly ranked public schools, etc. Factors responsible for the price premium may not be valued by certain groups of buyers, so it would be wise for them to avoid certain areas. Cases where the listed price being notably below the model prediction may signal hidden undesirable factors, or under-valued properties. In both cases hedonic model provides a useful tool to estimating the value of residential properties.

However, most applications of hedonic models rely on the assumptions of a single model, or compare outcomes from individual models. Ensemble forecasting is a different modeling approach that systematically accounts for uncertainties in modeling, and aims to improve forecast accuracy by combining data and different model outputs. Ensemble models are also capable of presenting model predictions as a range of possible outcomes instead of a singular deterministic number, in order to reflect inherent modeling uncertainties, which makes ensemble models more useful as decision support tools than the single-model approach. Ensemble models have been applied in other fields, most notably in weather forecasting where it significantly improved forecast accuracy (Blum, 2019).

We model the sales price of houses in Sydney, Australia, using ensemble models (Wu

and Levinson, 2021), and variables describing the functional characteristics, and location of the house. We compare ensemble model outputs with single model predictions in terms of accuracy, reliability, and usefulness as decision support tools. The objective of this chapter is to test different models' ability in extracting information, and to examine the effectiveness of ensemble models in predicting residential property prices.

4.2 Data

Property Attributes and Transaction Data

The 'house' is defined as single detached residential units in this study; terraces, townhouses, and attached housing units are excluded from the 'house' category. The 'house' is the only residential property type included in this study, which aims to test the success of various model and ensemble specifications, rather than make predictions of every type of structure. By selecting one type of structure, we can minimize variation in the comparisons. Future studies can look at other structure types, including attached houses, apartment units, and various kinds of commercial property.

Property attributes data describe the characteristics of a house as living space. The distribution of each attribute of the houses is examined, and outliers (e.g. extremely cheap, or large and expensive houses) are excluded.

The Sydney property transaction data (Australian Property Monitors, 2019) is obtained from the Australian Urban Research Infrastructure Network (AURIN), which records the transaction date, price, location and basic attributes of residential properties. This data records property transactions, which the models are trained on.

We narrow down the date of transaction to between January 2017 to May 2019, to obtain a sizable dataset, and to limit the effect on housing price from economic fluctuations, and thereby also limiting the analysis to avoid the COVID-19 period. The transaction data includes house attributes, such as number of rooms, and size of the land parcel. The property transaction data is combined with the Geoscape data (Geoscape, 2020) to obtain additional building attributes, including the height of the eaves and roof truss, the size of building footprint.

Local culture and environment and socio-economic and demographic makeup affect the property price. Although the causality between culture and the value of property is not clear, we use the composition of the nation of origin (percentages in various categories at the SA2 geography from 2016 Australian Census data) as an explanatory variable in the house price.

The surrounding environment where a residential property is located can either add or detract from its value. Empirical data suggests that airport flight noise has a negative impact on residential property values (Espey and Lopez, 2000). To measure the effect from flight noise, we obtained Sydney flight noise contour data (Sydney Airport, 2018) to categorize Sydney Mesh Blocks into 4 categories based on the daily frequency of flight noise exceeding 70dB: no effect from flight noise (less than 5 events per day), low impact (5-20 events), medium impact (20-70 events), and high impact areas (>70 events) with immediate proximity to the Sydney airport. Proximity to desirable amenities, such as rivers and coastlines add to residential property values (Powe et al., 1995). Both the view, and the use of seaside amenity is a significant factor for residential properties in Sydney. We use the Euclidean distance from

Category	Variables
House	Size of non-building area
	Size of the building footprint
	Number of bedrooms
	Number of bathrooms
	Number of onsite parking spaces
	Binary: pools
	Binary: air conditioning
	Binary: fireplace
	Height of outer-wall (excluding roof)
	Height of roof truss
	Height of adjacent trees
Neighbourhood	Percentage of people with foreign origin
Access	Number of jobs within 10 minutes by driving
	Number of jobs within 10 minutes by walking
	Number of jobs within 45 minutes by transit
	Number of hospitals within 20 minutes by walking
	Driving time from Sydney Town Hall
	Euclidean distance to the nearest coastline
Disamenities	Median speed of the adjacent road as token for road noise levels
	Flight noise > 70dB per day: no effect (<5 events/day)
	Flight noise > 70dB per day: low (5-20 events/day)
	Flight noise > 70dB per day: medium (21-70 events/day)
	Flight noise > 70dB per day: high (>70 events/day))

Table 4.1: Explanatory variables for house transaction price

residential properties' Mesh Block centroids to the nearest coastline to measure proximity to the coastline, which includes harbors, and sections of the Parramatta River.

Because this research uses historical data, the model output for house prices should be interpreted as the predicted house price for the period from January 2017 to May 2019, and would need to be adjusted to predict present value.

Data Measuring the Convenience of Transport, Accessibility Data

The convenience of transport from each location is measured by accessibility. We use the *cumulative accessibility* to jobs, and to urban amenities (hospitals used as an example) to measure the ease of reaching desired destinations, and the convenience of transport from a particular location. The location of hospitals are obtained from the Department of Health (Department of Health, 2019) lists; job locations are based on the 2016 census data (Australian Bureau of Statistics, 2016). The access to jobs is calculated as the cumulative number of jobs reachable, under a predefined travel time threshold. In Equation 4.1 and Equation 4.2 we show the equation for calculating accessibility. Residential properties are assigned accessibility values based on the Mesh Block in which the property is located.

$$A_{i,m} = \sum_{j=1}^{J} O_j f(C_{ij,m}) \tag{4.1}$$

$$f(C_{ij,m}) = \begin{cases} 1 & \text{if } C_{ij,m} < t \\ 0 & \text{if } C_{ij,m} \geq t \end{cases} \tag{4.2}$$

$A_{i,m}$: Access measure for zone i, by mode m
O_j: Number of jobs at zone j
$f(C_{ij,m})$: Travel time between zone i and j for mode m
t: Travel time threshold

Accessibility is calculated for walking, transit and automobile separately for all 58,819 Mesh Blocks in the Greater Sydney area. Jobs and population data comes from the 2016 census. Automobile accessibility is based on a Mesh Block level travel time matrix. Automobile travel times between locations are calculated using the QNEAT3 network analysis tool in the QGIS software, and based on the street network from OpenStreetMap. The link speed data employed here was purchased from a third party (Compass, 2019) and contains the median speed for automobile traffic that occurred between November 11 to November 25, 2019. Though accessibility was computed only for one period of time, we assume it is stable for the entire period (2017-2019).

For the walking and transit mode, accessibility is measured using the isochrone coverage area of each Mesh Block, under different travel time thresholds. Transit travel time covers all stages of a transit trip (with walking access and egress), including on-board travel time, waiting time, and walking phases during the station access, egress, and necessary transfers. OpenStreetMap provides the pedestrian network for both walking and transit. The isochrone coverage area originating from the centroid of each Mesh Block is drawn by Open Trip Planner (OTP) (Open Source, 2020). These coverage areas are then overlaid with a point layer of all destination Mesh Block centroids; each destination Mesh Block centroid covered within the isochrone is marked as 'reachable' from the origin Mesh Block. The list of reachable Mesh Blocks for each origin Mesh Block under each travel time threshold is combined with the land use data to produce accessibility values. Transit accessibility is based on transit schedule on July 8, 2020, and 08:00 departure time; we calculated in addition to the 08:00 departure time, transit accessibility to jobs using departure times between 07:45 and 08:15, in 5-minute increments, to form a time-series measure of transit accessibility. Accessibility at different travel time thresholds, and by different modes of transport are often correlated.

An example of accessibility measurement is shown in Figure 4.1, which plots the Mesh Block level 45-minute transit accessibility to jobs. Job locations provide more than employment opportunities, in that jobs also represent urban opportunities, such as banks, restaurants, schools and hospitals, etc. The accessibility variables are able to describe the location of residential properties in terms of the convenience of transport. Similar accessibility measures are applied in three technical reports (Levinson et al., 2020, Wu and Levinson, 2019, 2020).

4.3 Methods

We calibrate base and ensemble models to predict Sydney house transaction prices. Each base model consists of a single model, one dataset, and produces a single number as model output; both linear and machine learning models are included as base models. Ensemble models combine outputs from base models. Models used in this research only consider explanatory variables that are related to the location and functional characteristics of the

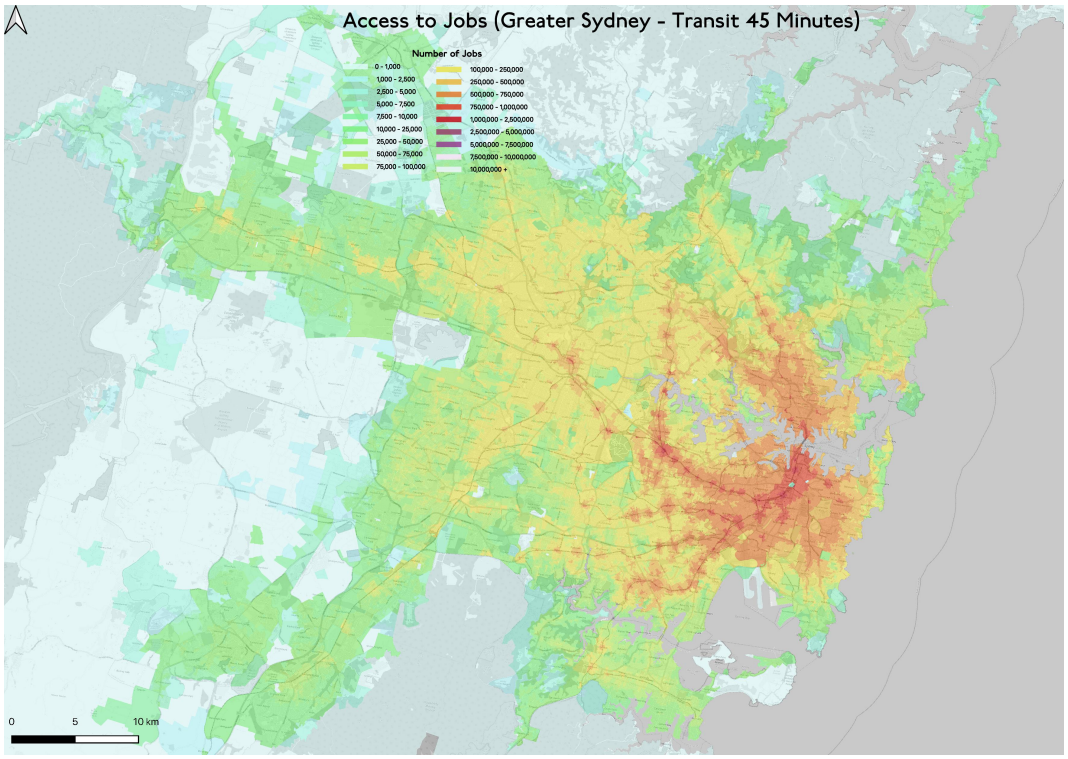


Figure 4.1: Transit access to jobs in Sydney, Australia. 45-minute threshold, Wednesday 8 am trip departure time.

property. Factors that are difficult to quantify, for example, building aesthetics, good quality public schools, are also excluded.

The ideal for hedonic models is to be accurate, reliable, and useful in property valuation for applications such as assessment and purchase decisions. Generally the purchase decision is made for a single property (assuming individual buyers), so it is critical that large forecast errors be avoided, even at the cost of some reduced forecast accuracy. Model prediction presented as a probability distribution of different prices may be more useful for decision making than providing a single number as model output.

4.3.1 Training and Testing Models

Models are calibrated using the training data. During this process, models find coefficients that fit the data, or extract patterns from the data, which are later used for predicting cases in the testing data. Testing data provides validation for the performance of models. Machine learning models can 'memorize' certain features as labels. In transport applications, spatial characteristics have the potential to be used as labels. For instance, the social demographic statistics can be used as a label to group houses from the same geographical areas, which gives machine learning models an unfair advantage over the linear model. It would be desirable to mix houses from different geographical areas in order to maximize accuracy. Since this research aims to evaluate model performance, a clear separation between training and testing data based on geographies is required.

To test different models for their actual capability in extracting information, we split the training and testing data based on different geographical areas, so machine learning models cannot use areal statistics to substitute for actual pattern recognition. We run repeated experiments to test the performance of models. The Greater Sydney has 312 Statistical Area Level 2 (SA2), which we split into two mutually exclusive groups for each repeated experiment. In each of the repeated experiment, each of the 312 SA2s will be assigned to either the training, or testing group, but will not appear simultaneously in both training and testing data; samples of varying sizes are then taken from either the training or testing SA2s to fill either training or testing data.

The process of splitting of data into training and testing data, model calibration, and validation in the testing data is repeated many times, to ensure the results obtained are not due to chance. In each repetition, each model is fed identical training data for estimation. In the next step, every model is applied using identical testing data.

Among the base models, both the linear model and machine learning models use the housing prices as the dependent variable, and the set of variables listed in Table 4.1 as explanatory variables. The linear model uses ordinary least squares (OLS) for estimating parameters. We use a single layer neural network structure with 5 neurons for the neural network; the classification tree has a max depth of 30 steps; the random forest and gradient boosting machine use 128 and 650 trees respectively.

4.3.2 Combine Models

Ensemble models combine base models using predefined rules. We test base models and three categories of ensemble models in this hedonic application, namely, simple rules, stacking, and ensemble of ensembles.

The goal of combining different types of base models is to test whether the combined ensemble model can have better performance than the best base model. This research is not about performance optimization of any specific model, so individual model formulations are intended to have 'good', instead of the 'best' performance.

- **Base models** include five types of models: linear model, classification tree, random forest (RF), gradient boosting machine (GBM), and neural network (NN). The base models independently predict the transaction price of houses, using the same set of explanatory variables and the same training data.

- **Ensemble models with simple rules** combine base model predictions as weighted averages. Two weighting schemes are used, namely the simple averages with equal weights, and weighted average that use model performance metrics (RMSE) of each base model in the training data as weights. Predictions from the same type of base model are assigned the same weight.

- **Meta-learner ensemble models (stacking)** use forecasts from the 5 base models as explanatory variables in predicting the transaction price. The training data is further divided into two portions, one used to calibrate the base models, and the other to calibrate the meta-learners. Three meta-learner ensemble models: linear, RF, and

GBM, are trained to combine predictions from base models. A peculiar type of meta-learner uses a RF classifier, and the same set of explanatory variables used by based models, to identify for each residential property, which of the base models is likely the most accurate. For simplicity, the RF classifier is only allowed to choose between two base models: RF and GBM.

- **Ensemble of ensembles** combines different methods of combining base models. Here we use the simple average of the three meta-learner ensemble models for ensemble of ensembles.

4.3.3 Combine Data (Time-series Accessibility Measurements)

In single-model forecasts, the average or median measurement from different data sources is often used to account for uncertainties in the measurement data. Ensemble forecasting provides an alternative to combining data from different sources, by calibrating parallel models that use data from different sources, producing a distribution of possible outcomes. Variations between multiple measurements for the same explanatory variable suggest that these different inputs may produce different paths, and therefore different outputs from the model. In the case of hedonic models, the convenience of transport (as measured by access to jobs) differs at different times of the day depending on traffic conditions and transit schedules.

To account for temporal variations, a common practice among transport professionals is to use time-averaged transit access, and its variation over time (Owen and Levinson, 2015) as explanatory variables in a single-model forecast. In addition to averaging input measurement data, we apply an ensemble approach that uses time-series accessibility measurements in parallel base models, which averages model predictions. The difference is in when the data sources are combined. This ensemble approach in combining data is able to show a range of possible outcomes from parallel models that use different input data, and improves the interpretability of model outputs. For example, the effect of adverse traffic conditions for a short period of time will be preserved in the model output, instead of being lost in the averaging of time-series measurements. This may better reflect human perceptions, as people will perceive travel times based on a particular time of day rather than as time averaged.

The difference in model performance between the two methods: (a) averaging the measurements before modeling, and (b) averaging the model outputs (ensemble method) are compared. The ensemble method has an advantage in attenuating accumulated error, so theoretically should have better performance in long-range forecasts, which is in part responsible for the success of ensemble models in weather forecasting (Blum, 2019). It is not clear whether the ensemble model's advantage in the better handling of accumulated error would be significant in this case, or in other transport applications. In this section we test whether these two different methods would make a noticeable difference in model performance.

Specifically, we compare single-model forecasts with an ensemble method of combining data. On the one hand, the single-model method involves averaging 7 accessibility measurement data, from 07:45 to 08:15, in 5-minute interval, and feeding the time-averaged accessibility into a single model. On the other hand, the ensemble method uses parallel ensemble models (7 parallel models in this study), and calibrates each of the models using one of the time series accessibility data as input; the average of the parallel model outputs

becomes the ensemble forecast. We apply this ensemble method for combining time series transit accessibility data using three base models: linear, RF, and GBM. The standard deviation of transit accessibility are included in both methods. Only transit access to jobs is included as the locational variable. We also tested models using transit accessibility at a single 8:00 am departure time, with no knowledge of transit accessibility at other departure times or temporal variation of transit access. This experiment is repeated multiple times, and each time the parallel ensemble models use the same random starting point as the time-average model, so the only difference will be when data sources are combined (i.e. average measurements vs. average model outputs from different measurements).

4.3.4 Range of Possible Outcomes

In some cases, the range of possible outcomes might be of greater interest than a single number as ensemble model output. In the case of predicting house prices, knowing the extent of price uncertainties would be helpful for purchase decisions.

There are numerous ways that a range of possible model outcomes can be produced, and the choice depends on the purpose of modeling and data availability. To enumerate and experiment with all such methods will be beyond the scope of this research. In this application, we repeatedly take subsamples from training data to calibrate models and to form a distribution of predicted prices for a single house.

The distribution of predictions for house prices comprises ensemble models. Each time, 15,000 cases are taken (all 312 SA2s are mixed in the training data to maximize the amount of information) from the entire dataset, with replacement to form a subsample. Base models and ensemble models are calibrated using this subsample, and the final ensemble model output (ensemble of ensembles) comes from the average of three ensemble models (linear, RF, and GBM meta-learners), so each of the model outputs in the distribution is an ensemble of ensemble model. This process is repeated 50 times, and these models form a distribution, which represents the forecast for a single house. This resulting distribution has three layers of ensemble models, namely the initial meta-learner ensemble models, the 'ensemble of ensemble' models, and an ensemble of 50 outputs.

4.3.5 Elasticity of Housing Price in Model Predictions

The elasticity (Mankiw, 2014) of housing price shows the sensitivity of the price responding to changes in various contributing factors, and is measured by the ratio of the percentage change in housing price, to the percentage change of an explanatory variable. An elasticity of "1" indicates perfect elasticity (i.e. linear relationship between housing price and the explanatory variable), and values below "1" would indicate somewhat inelastic relationships, and vice versa.

Although attributes of a house itself may remain unaltered, its surroundings, and the convenience of transport to other places may change over time, which will affect the housing price. In this research we analyze how changes in the convenience of transport affect housing prices. To avoid complications with multiple variables, we keep only one transport related variable: the "Number of jobs within 45 minutes by transit" from the "Access" category in Table 4.1. We alter this variable by 1% in the testing dataset, and examine how different

models respond to this change. We repeat this process 500 times, each time a 10,000 testing data is used to obtain a distribution of elasticity values.

This paper compares different models in terms of their sensitivity to input variations, by examining the difference in elasticity among different models, and whether the elasticity is positive across different models. A positive elasticity corroborated by different models would make a stronger case that, the convenience of transport has a positive effect on housing prices.

4.4 Results

4.4.1 Combine Models

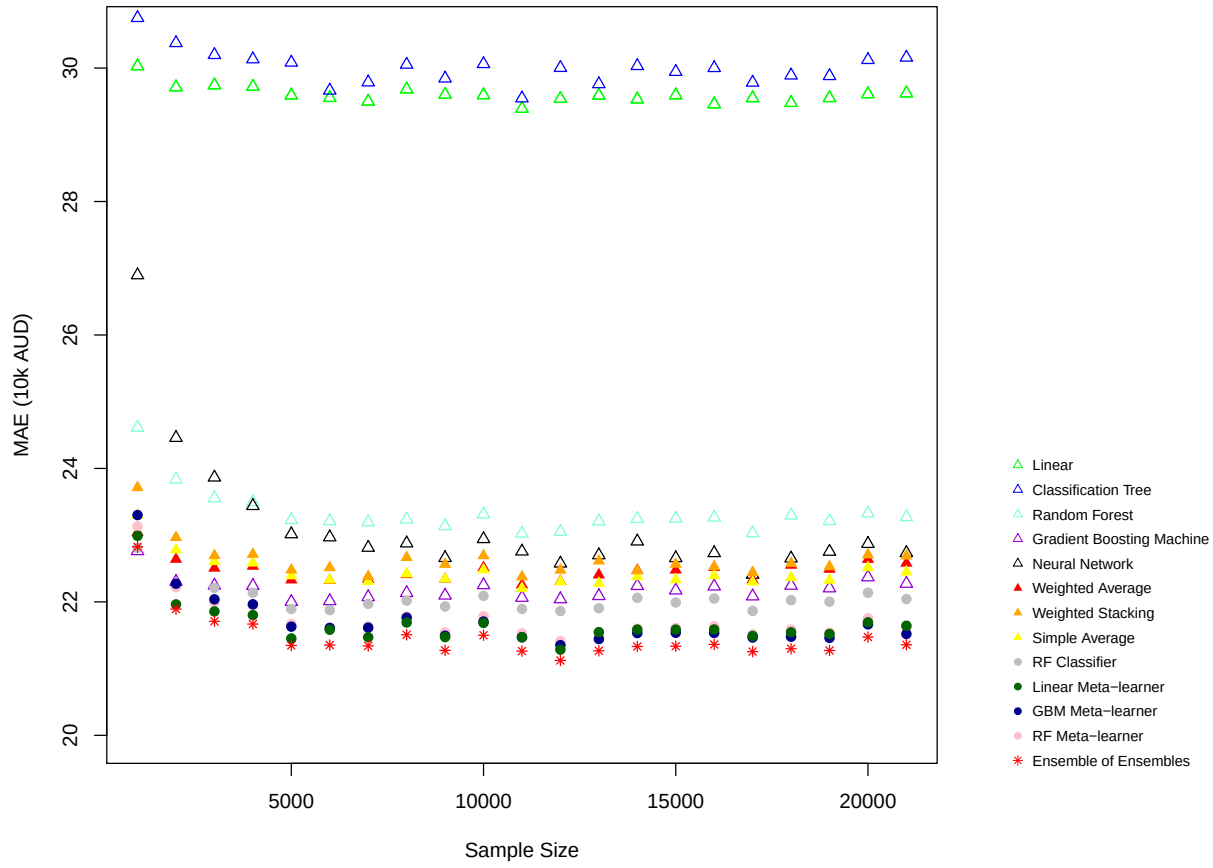
Performance metrics of ensemble models in predicting the house sales price are shown in Figure 4.2 through Figure 4.4.

Base models have different performance in predicting house transaction prices. Among the base models, the GBM has the best accuracy, and its accuracy is also the most stable among base models. The neural network has similar or lower MAE than the RF (Figure 4.2), but also has more large errors, as seen from the higher MSE in neural network predictions (Figure 4.3).

Ensemble models with simple rules, namely simple, weighted average, have slightly lower accuracy than the best performing base model (GBM), and the stability of their model performance is slightly below the GBM (Figure 4.2). However, ensemble models with simple rules did perform better than the remaining base models, with better forecast accuracy, and more stable model performance.

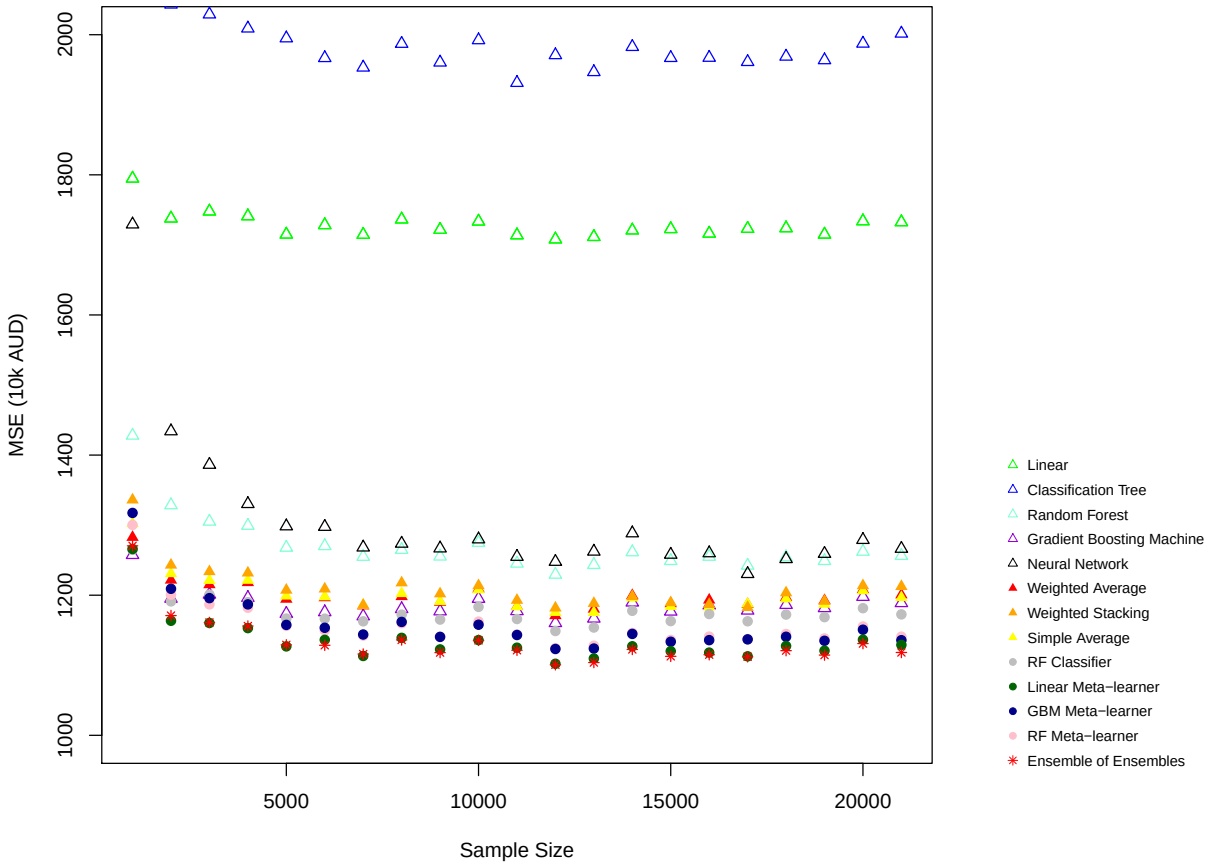
Meta-learner ensemble models (aka. stacking) combining base models are able to improve forecast accuracy and performance stability notably beyond the best base model. The lower MAE and MSE in stacking ensemble models suggest both large and small errors are reduced by stacking; the lower standard deviation of absolute errors shows stacking models have more stable performance than base models.

There is slight difference among the performance of different meta-learners. In Figure 4.2, the linear meta-learner has slightly better performance than the GBM meta-learner and the RF meta-learner. The linear meta-learner is superior in the MSE, and in the standard deviation of absolute errors than the other two meta-learners, with fewer large errors and more stable accuracy performance (Figure 4.3). The RF classifier that is used to select between two base models (Random Forest and GBM) predictions have the lowest performance among stacking models, and have slightly better accuracy than each individual model.



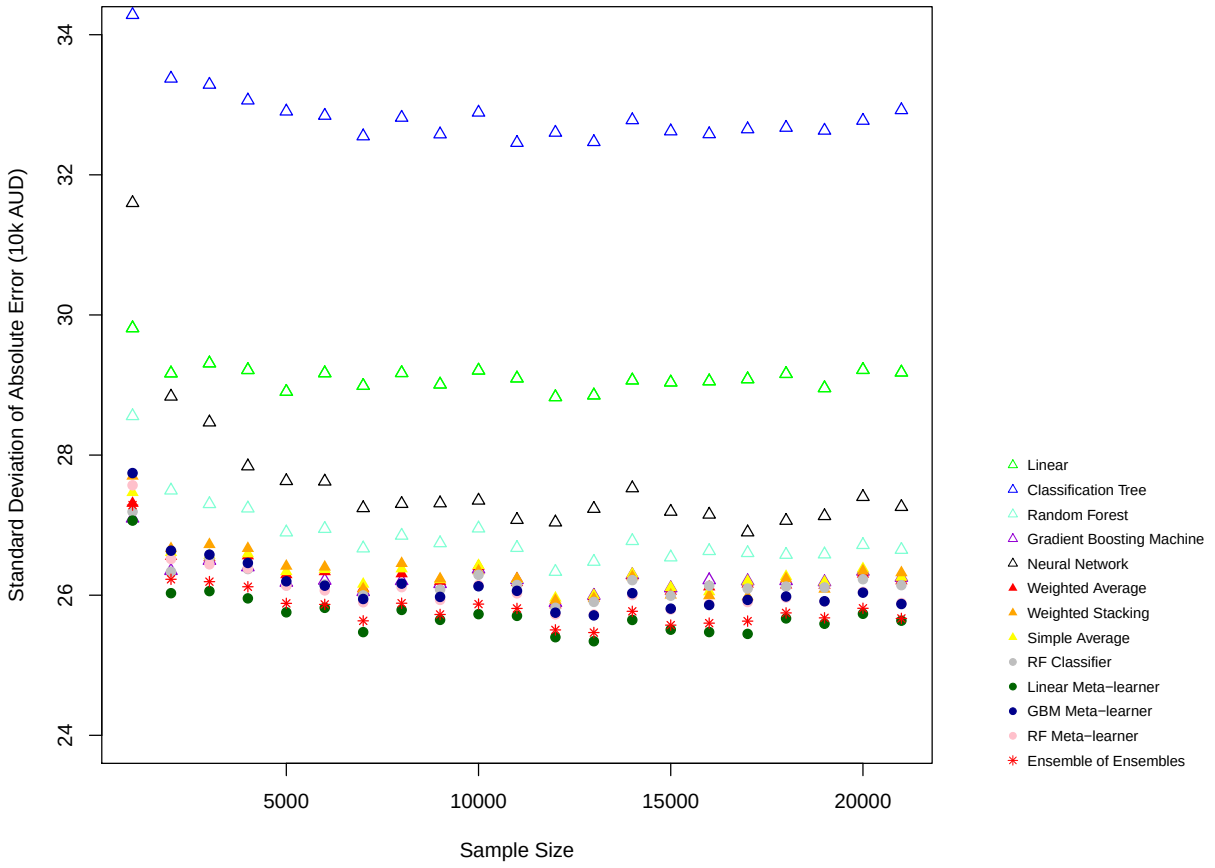
(a) Mean absolute error (b) Models

Figure 4.2: Model performance in predicting house sales price. Mean absolute error of forecast by different models, in testing data. Every dot is the average of 90 experiments, each with 100 testing samples.



(a) Mean square error (b) Models

Figure 4.3: Model performance in predicting house sales price. Mean square error of forecast by different models, in testing data. Every dot is the average of 90 experiments, each with 100 testing samples.



(a) Standard deviation of absolute error (b) Models

Figure 4.4: Model performance in predicting house sales price. Standard deviation of absolute error of forecast by different models, in testing data. Every dot is the average of 90 experiments, each with 100 testing samples.

The ensemble of ensembles model (which is the average of three meta-learners) lowers the MAE of the three meta-learners, but has no effect on the MSE, or the stability of the forecast accuracy. Performance of the ensemble of ensemble model in terms of MSE is similar to that of the linear meta-learner. Ensemble of ensembles also has higher standard deviation of absolute error than the linear meta-learner.

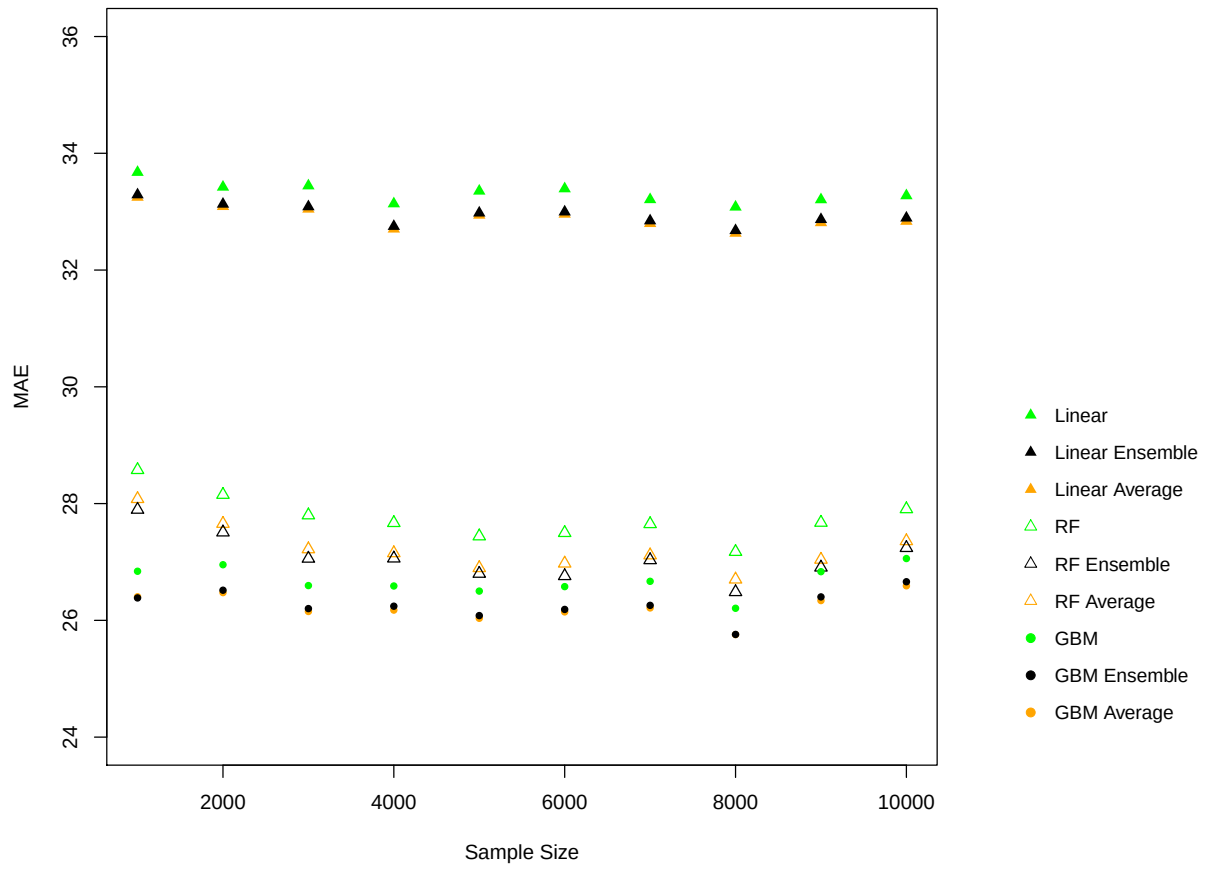
In cases with small training dataset sizes, the meta-learner ensemble models and ensemble of ensembles are not noticeably better than the best single base model. This suggests that the meta-learners require sufficient training data to calibrate and to avoid mistaking noise for signal.

4.4.2 Combine Time Series Accessibility Data

The performance of models using time-averaged transit accessibility and using parallel ensemble models to combine data are shown in Figure 4.5 through Figure 4.7. Models using a single departure time for transit accessibility are used as baselines for comparison.

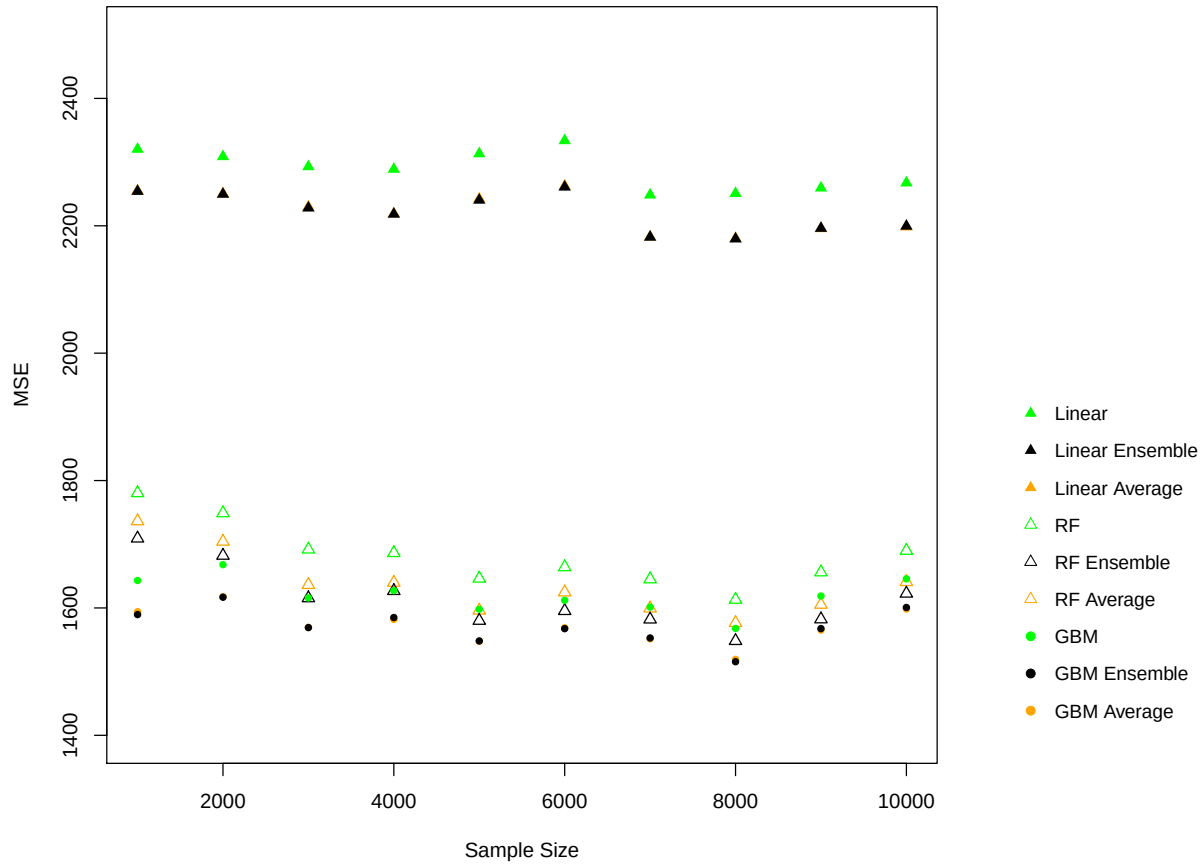
The inclusion of transit accessibility data at different departure times significantly improves model performance beyond the baseline model, which uses a single departure time. This is shown by the difference between green plots and black and orange plots in Figure 4.5 and Figure 4.6.

However, there is only slight difference in combining parallel ensemble models versus using time-averaged accessibility. In RF ensemble models, combining parallel ensemble models (black plots) has a slight advantage over using time-averaged accessibility (orange plots in Figure 4.5 through Figure 4.7). Combining parallel GBM or linear models achieved near identical forecast accuracy and stability as using time-averaged accessibility.



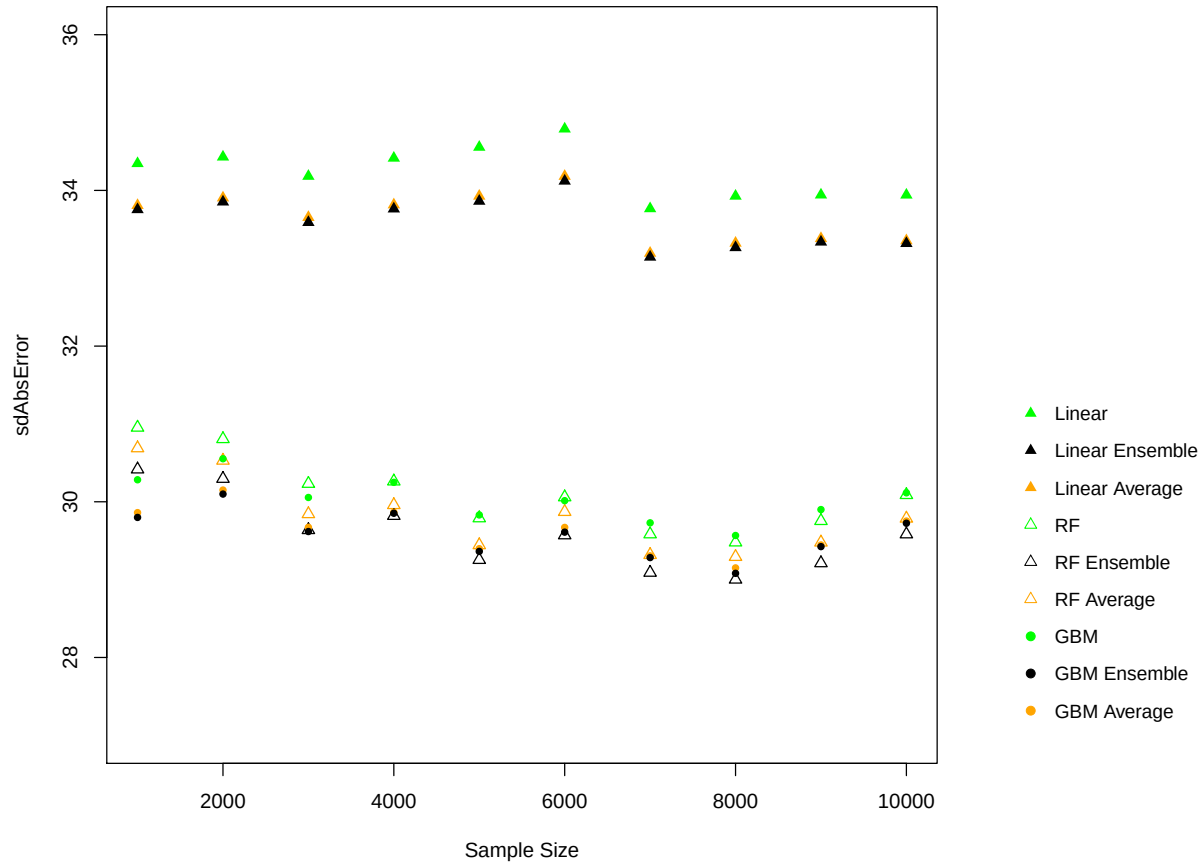
(a) Mean absolute error (b) Models

Figure 4.5: Model performance in predicting house sales price. Mean absolute error of forecast by different models, in testing data. Every dot is the average of 60 experiments, each with 50 testing samples.



(a) Mean square error (b) Models

Figure 4.6: Model performance in predicting house sales price. Mean square error of forecast by different models, in testing data. Every dot is the average of 60 experiments, each with 50 testing samples.



(a) Standard deviation of absolute error (b) Models

Figure 4.7: Model performance in predicting house sales price. Standard deviation of absolute error of forecast by different models, in testing data. Every dot is the average of 60 experiments, each with 50 testing samples.

One advantage of using parallel ensemble models over time-averaged accessibility is in the interpretability of the ensemble model output. A model using time-averaged accessibility variables produces a single number in the output, whereas the composition of the output is unknown. An ensemble model consisting of parallel models can produce a range of possible outcomes; while only the averages of the 7 parallel ensemble models are shown in this section (to compare with the single-model output), a range of 7 numbers are available. In a sense the parallel ensemble models provide predictions that are 'dissected', so the range of possibilities, and the best to worst scenarios (subjective) can be known. The ensemble model predictions are more useful for decision making.

4.4.3 Range of possible outcomes with subsampling

Ensemble forecasting can produce a range of possible model outcomes. Here we present an example of ensemble models producing a range of possible house transaction prices, through ensemble of ensembles and repeated subsampling. There are other methods capable of producing a range of possible model outputs (discussed in section 2.5.3), this section is intended to show the effect of multiple path predictions.

Figure 4.8 illustrates the range of possible prices provided by the ensemble models. The ensemble is made up of 50 parallel ensemble models, each with a 15,000 house transactions size subsample training data. The median of ensemble model predictions (predictions are made before the actual sale of the house) for a particular house is 2.02 million (2019 Australian dollars), and the house sold in 2021 at 2.12 million Australian dollars.

Applying subsampling to ensemble model forecasts is able to present a useful range of possible outcomes, for what would otherwise be a single number in model prediction. The transaction price for any single residential property is inherently volatile, with uncertainties and variabilities in buyers, sellers, and how the property is presented and assessed (behavioral economics (Shiller, 2005)), so a range of possible sales prices are unsurprising. The range of outcomes from ensemble models show what is most likely, and in cases of very low or high prices, indicate how unlikely these outcomes are.

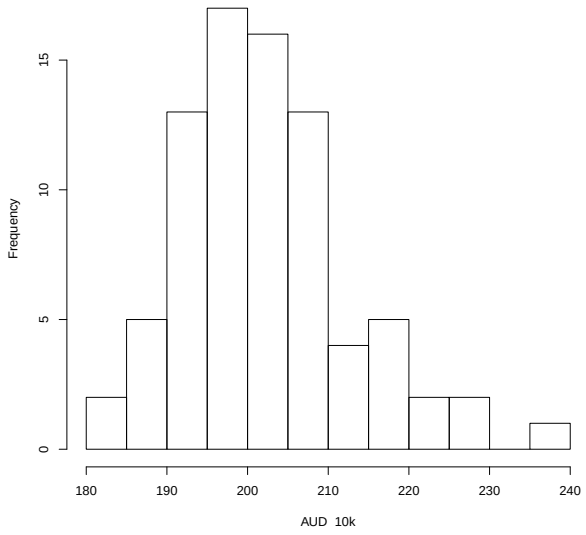


Figure 4.8: Example of parallel ensemble model predictions for a single house, with subsampling. (50 subsamples, each with 15,000 cases)

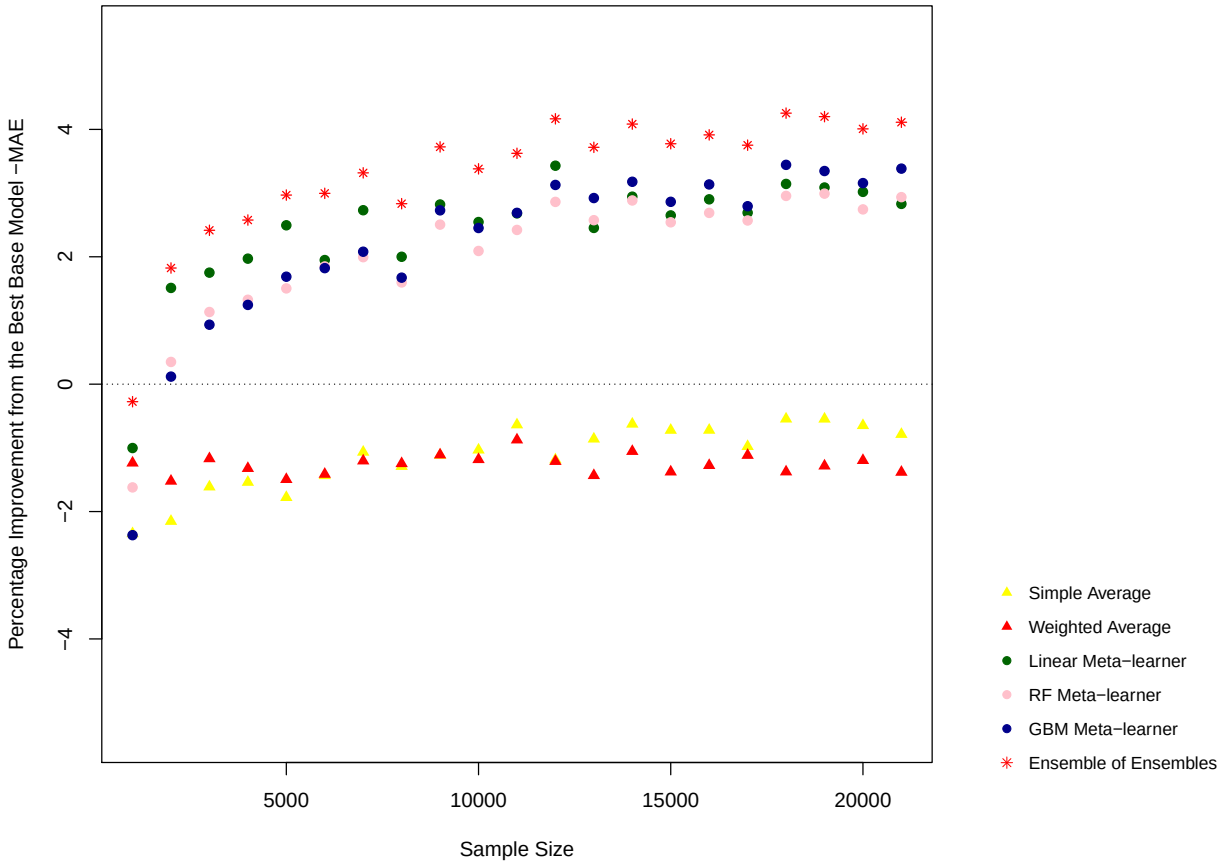
4.5 Base vs. Ensemble Models

The extent of performance improvement from ensemble models are shown in Figure 4.9 and Figure 4.10; percentage improvement in MSE follow identical patterns as MAE, and are presented in Appendix C.

In predicting Sydney house transaction prices, the weighted average, and simple average ensemble models are unable to improve the accuracy upon the best base model. The extent

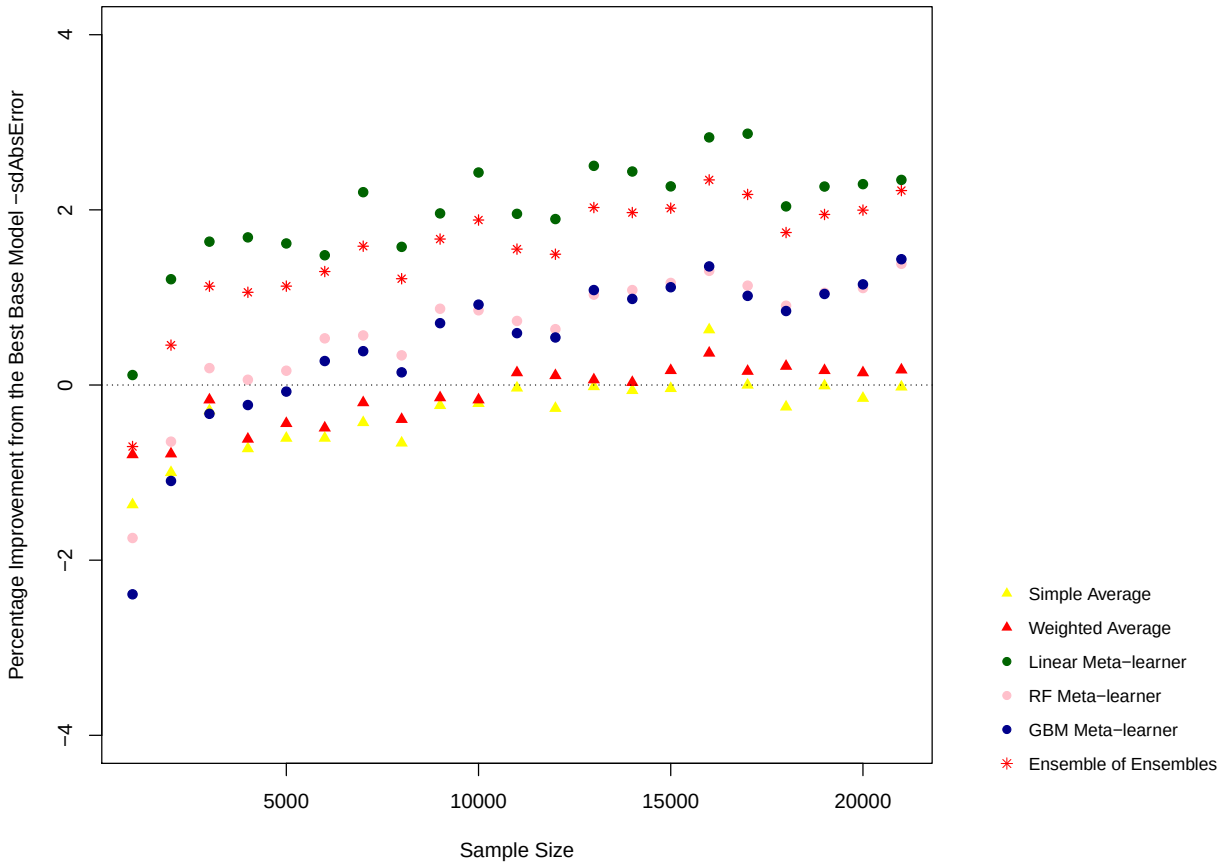
of accuracy improvement from stacking models, and ensemble of ensembles (iteration 1, average of three ensemble models) increases with the size of training data. Ensemble of ensembles model faces diminishing returns; adding additional training data to the models has an increasingly smaller effect, until additional data no long improves forecast accuracy.

Other ensemble of ensembles methods in iteration 1 are tested, namely the linear, RF, GBM as meta-learners combining previous ensemble model outputs, with the addition of the average of these meta-learners in iteration 2 (i.e. an ensemble of ensemble of ensemble model). However, these ensemble of ensembles models do not improve upon the simple average of ensemble models in iteration 1, suggesting meta-learners are beginning to pick up more noise (accidental features) than information. Therefore in this particular case, ensemble of ensembles stops at iteration 1.



(a) Percentage Improvement in MAE (b) Models

Figure 4.9: Performance improvement from the best base model - Sydney Hedonic MAE



(a) Percentage Improvement in SD of Absolute Error (b) Models

Figure 4.10: Performance improvement from the best base model - Sydney Hedonic SD of absolute error

4.5.1 Elasticity of Housing Price

With a 1% increase in the transit accessibility variable, all models (both base models and ensemble models) show positive elasticity values. This is significant because given the difference between models, including different model assumptions, and different methods of pattern recognition, all models arrived at the same conclusion, that the convenience of transport has a positive impact on housing prices.

Different models vary slightly in their sensitivity to changes in the accessibility variable. Elasticity of different models are shown in Table 4.2. The linear model, which is the second least accurate (after the classification tree) is the lease sensitive to changes in the convenience of transport, with an average elasticity of 0.124, meaning a 1% change in transit accessibility produces a 0.124% change in the housing price. The remaining models (both single models and ensemble models) are more sensitive than the linear model, and have similar elasticity values, ranging from 0.150 (simple average) to 0.166 (RF meta-learner). The ensemble of

ensembles model has an elasticity of 0.158.

	Base Models					Ensemble Models				
Model	Linear	CT	RF	GBM	NN	Average	Linear(M)	RF(M)	GBM(M)	Ensemble of Ensembles
Elasticity	0.124	0.180	0.152	0.162	0.166	0.150	0.157	0.166	0.158	0.158

(M): Meta-learners

Table 4.2: Elasticity of Housing Price to Transit Access to Jobs (45 minutes)

4.6 Discussion

This chapter examines the performance of base and different ensemble models in predicting the sales price of houses in Sydney. Predicting housing prices with ensemble models differs from the conventional single-model approach, and better reflects inherent uncertainties in real-world events. Ensemble models not only provide more accurate estimates for house price than the single-model approach, but also internalize, and reflect modeling uncertainties as a range of possible model outputs, making ensemble model outputs more useful as decision support tools.

We find machine learning models generally produce more accurate and more reliable forecasts for the sales prices of houses than linear models (except the classification tree, which is less accurate and not as reliable as the linear regression). Among the base models, the Gradient Boosting Machine (GBM) produces the most stable and accurate forecasts. Comparing the neural network (NN) and random forest (RF), the NN has lower MAE than the RF (with sufficient training data), but produces more large errors, so the MSE and the standard deviation of absolute errors of NN are generally higher than the RF.

Meta-learner ensemble models are able to improve forecast accuracy and reliability beyond the best base model. The linear meta-learner has the best performance among the three meta-learners. The better performance of linear meta-learners might be due to the robustness of linear models. Forecast accuracy (MAE) of ensemble models can be further improved, by combining different methods of combining models (ensemble of ensembles). Future tests are required to determine whether, and to what extent, ensemble of ensembles reduces the chance of large errors or improves the stability of model performance. In this case ensemble of ensembles finished second best, behind a linear meta-learner, on the measure of standard deviation of absolute error, but ahead of all base models and other ensemble techniques.

Meta-learners are calibrated with training data in this chapter, which in part contributed to their good performance from meta-learner ensemble models. In cases with a small training data set available, it may not be possible to calibrate meta-learners, or the inadequately trained meta-learner might have bad performance. Meta-learners are prone to picking up noise for signals with small samples.

With respect to the two different methods of combining measurement data – namely, (a) using the averaged measurement as input in a single model, and (b) using parallel models each with a different measurement data – we find very slight differences in model performance between these two methods. It is possible that this specific application in predicting house prices does not produce large differences in model performance. Future research should explore methods of combining measurement data. We expect applications that are susceptible

to accumulated error, for example, applications with many iterations, would benefit from combining data using parallel ensemble models.

Chapter 5

Predicting For-hire Vehicle (FHV) Trips with Ensemble Models

5.1 Introduction

In this chapter for-hire vehicle (FHV) trips data from Chicago and New York City is used as a case study to examine the performance of ensemble models, and to compare ensemble models against base models. Models are compared in terms of their accuracy, the reliability of model performance, and the usefulness of model output as decision support tools. In order to prevent publication bias (Easterbrook et al., 1991), this chapter will report results from all applications of ensemble models, regardless of whether ensemble models improve model performance.

The advent of for-hire vehicle (FHV), such as Uber and Lyft is a new occurrence, and the share of FHV in all trips (0.5% in 2017 (Federal Highway Administration, 2017)) remains low compared to existing modes of transport, but the FHV continues to grow at a rapid rate, and their numbers have already became significant in some areas. Nearly 10% Americans use FHV in any given month in 2017 (Conway et al., 2018); in New York City in particular, the number of FHVs tripled between 2010 and 2019, to 100,000 (Roberton et al., 2020). At this rate of growth, and with the help of autonomous vehicle technology, the FHV will likely become a significant mode of transport, which might cause conflicts with the existing transport system. Empirical research find FHV to be a significant contributor to traffic congestion (Erhardt et al., 2019), and increases vehicle emission (Roberton et al., 2020). A better understanding of FHV trips is needed.

Predicting travel demand is an important application in transport modeling. Travel demand changes in response to population, demographic, economic or technological changes. Travel demand is measured by trip production, attraction, and traffic flow between places, and are conventionally predicted by four-step transport planning models.

Because of the level of complexity, and lengthy computation time associated with travel models, the single-model approach is prevalent in modeling. Within a four-step model, the number of trips produced (all modes of transport), or attracted to each place is often estimated by a single linear model within the four-step transport planning model; the number of residents and employment within each zone are often included as explanatory variables. The flow between places is based on trip production and attraction, and derived iteratively

through a gravity model (Meyer and Miller, 2001). In models that involve percentages of people using different modes of transport (mode choice), logit models (Ben-Akiva et al., 1985, Meyer and Miller, 2001) are used to predict the likelihood of mode choices, assuming people make decisions to maximize their "utility", which is often a linear function. Travel time, distance, and monetary cost etc. are often elements of the utility function (Yao and Morikawa, 2005).

The trip production and attraction in particular are predicted using linear models, which assume travel demand to increase proportionally with increases in explanatory variables. At an individual level, it is noted that taste variations are often represented with a distribution about an exact value (McFadden, 1974), but the process by which individual or locational characteristics affect trip production and attraction is assumed to be singular; no consideration is given to the possibility that the data generation mechanisms may follow assumptions other than a linear or logit model.

Alternative model assumptions are generally discarded, when predictions made by alternative models appear not as accurate, even when performance differences between models were small. As McCullagh and Nelder (1989) put it: "Data will often point with almost equal emphasis on several possible models, and it is important that the statistician recognize and accept this." Model predictions are inherently probabilistic, so relying on a single model assumption, and presenting model prediction as a singular number is not the best approach.

Our theory of ensemble forecasting suggests that there might be room for improvement in both forecast accuracy and reliability by adopting ensemble models. We focus on for-hire vehicle trips, and test the use of different algorithms as base models, and different ensemble models to combine base models in predicting the trip production, attraction, and flow of for-hire vehicles between places. The goal is to test whether combining different base models will be able to produce forecasts that are more accurate and reliable, than any individual base model. If ensemble models indeed perform better than the single-model approach, then perhaps prevalent modeling approaches in the four-step transport planning models, at least in the prediction for trip production and attraction can be improved, with the adoption of ensemble models.

5.2 Data and Methods

5.2.1 Data

There are some differences between New York City and Chicago that are relevant for comparing model performance. On the data side, the New York City data includes 239 taxi zones, which provides less training data for models predicting trip production and attraction, when compared to 739 census tracts in the Chicago data; ensemble models for New York City would also have less data to calibrate meta-learners. This smaller data size does not pose a significant issue for predicting flows, as the origin-destination matrix of $239^2 = 57,121$ zone pairs is a sufficiently large number.

Chicago

Data for select for-hire vehicle (from ride-hailing companies such as Uber and Lyft) trips are available for the City of Chicago area (totalling 794 census tracts) (Chicago Data Portal, 2019). This dataset includes trip details such as pick-up and drop-off locations, which can be used to tally the number of trip production, attraction within different zones, and the number of trips between zones.

Trips may begin or end outside the City of Chicago area; these external trips constitute a small percentage of total trip numbers, and are not available within the dataset. We extract trips that took place on 26 consecutive Wednesdays in the first half of 2019, and use the average daily trips numbers for models to predict average daily trips.

Explanatory variables include social demographic, and locational factors of a zone. The percentage of non-family households, number of jobs and workers, and the median household income within a zone describe social demographic characteristics of an area; the locational factors can be described by accessibility to jobs. Explanatory variables used in models are listed below.

- 30-minute transit access to jobs
- Percentage of households that are not family units
- Median household income
- Number of workers in the area
- Number of jobs in the area
- Size of the area, in sqrkm
- Road distance between origin and destination zone (in the flow model)

New York City

The New York City for-hire vehicle (FHV, rideshare companies such as Uber and Lyft) trips data comes from the NYC Taxi Limousine Commission (TLC, 2017). The FHV trips data covers the City of New York area, totalling 239 taxi zones.

Trips may begin or end outside the City of New York area; these external trips presumably constitute a small percentage of total trip numbers, and are not available within the dataset. We extract trips that took place on 30 consecutive Wednesdays, beginning in the July 2017, for models to predict average daily trips.

Explanatory variables include social demographic, and locational factors of a zone. The list of explanatory variables used for New York City are the same as the ones used in Chicago, but aggregated to New York City taxi zones, which are generally larger than census tracts uses in Chicago. All taxi zones are divided into two mutually exclusive groups, so data in one group is used to training models, and data in the other group is used for testing model performance.

5.2.2 Models

We compare the performance of base models with different ensemble models that combine base models. The list below shows the category of models used to predict travel demand.

- Base models (Linear, Classification Tree (CT), Random Forest (RF), Gradient Boosting Machine (GBM), Neural Network (NN))
- Ensemble models with simple rules
- Meta-learner ensemble models
- Ensemble of ensembles

Base models include linear model, classification tree, random forest, gradient boosting machine, and neural network. Among the base models, the linear model uses ordinary least squares (OLS) for estimating parameters. We use a single layer neural network structure with 5 neurons for the neural network; the classification tree has a max depth of 30 steps; the random forest and gradient boosting machine use 128 and 650 trees respectively.

We use ensemble models with simple rules (simple, and weighted average), and with meta-learners to combine models. Three types of meta-learners are used, namely the linear, gradient boosting machine, and random forest meta-learners. An ensemble of ensembles method that averages outputs from the three meta-learners is also used as an ensemble model.

5.2.3 Trip Production and Attraction

Models predict the average daily number of trips produced and attracted to each of the 794 zones (census tracts) within the City of Chicago area, and each of the 239 taxi zones within New York City, using demographic and locational data as explanatory variables.

We divide the data into training and testing datasets to evaluate the performance of different models. Generally a zone with many trips previously would continue to have a high trip volume later in time; machine learning models can identify this trend in making predictions and substitute predictions with historical trip numbers, so it would be a mistake to separate training and testing data by time. In this study, the training and testing data are separated spatially, to prevent machine learning models from 'memorizing' specific zones. Any zone in the training dataset will not appear again in the testing dataset. We run repeated experiments to test the performance of models; in each experiment, the trip production and attraction zones are divided into two mutually exclusive groups, data in one group is used to train models, and data in the other group is used to test model performance. Samples are then drawn from the training zones and testing zones to fill either training or testing data.

To evaluate the model performance, we repeat the whole process (including splitting the data into training and testing, model calibration and validation) 400 times to obtain a distribution of model performance metrics.

5.2.4 Flow Model

Models predict the average daily number of trips between any pair of zones, using descriptive statistics from both the origin and destination zones, and the road distance between the two zones.

Generally the number of trips between two zones are reciprocal, as more trips in one direction often suggest a similar amount of trips in the other direction. So it would not be sufficient to divide training and testing data based on trip pairs alone, for example, including one origin-destination pair in the training data does not exclude trips in the other direction from this origin-destination pair in the testing data. In this study, the entire data is divided into training and testing datasets based on unique trip origins. In repeated experiments, each time the trip origin zones are divided into two mutually exclusive groups, data in one group is used to train models, and data in the other group is used to test model performance. Samples are then drawn from the training zones and testing zones to fill either training or testing data.

Training data of various sample sizes are used to calibrate models. For each sample size, the model is calibrated on 90 different training samples, and each time applied to 100 testing samples, to evaluate model performance.

5.3 Results

5.3.1 Trip Production and Attraction

Chicago

Different base and ensemble models have different performance in predicting the average daily number of trips produced from, and attracted to each zone. Figure 5.1 shows the mean absolute error, and Figure 5.2 shows the mean square error of different models in predicting trip production. The stability of model performance, as measured by standard deviation of absolute error is shown in Figure 5.3. Model performance follows identical patterns in predicting trip production and attraction. Figures showing model performance in predicting trip attraction are included in Appendix B.

Among the base models, the linear model has better performance than the classification tree, but has lower performance than other machine learning models. Within the 400 repeated experiments, the linear model has the highest chance to produce large errors.

Ensemble models with simple rules, namely simple average, and weighted average of the base models, improved MAE and MSE beyond the best base model. Figure 5.1 and Figure 5.2 show these two ensemble models (color coded green) to have the fewest cases where the models have large error measures. Within each of 400 repeated experiments, these two ensemble models have the most stable performance accuracy between cases, as shown by the standard deviation of absolute error in Figure 5.3.

Ensemble models with more complex rules, including meta-learners and ensemble of ensembles did not work as well as simple rules ensemble models, and did not significantly improve the mean absolute error of the base models. The linear combination meta-learner and ensemble of ensembles both have lower MSE than the base models. The gradient boost-

ing machine and random forest machine meta-learners provided no noticeable improvement from base models.

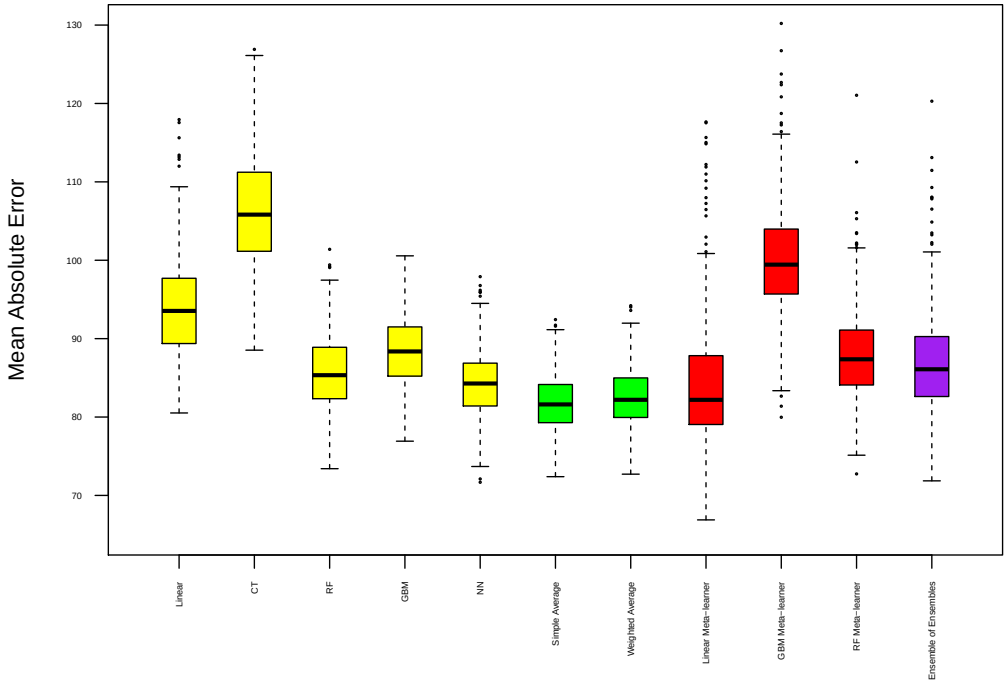


Figure 5.1: Model performance in predicting trip production in Chicago. Distribution of mean absolute error from 400 experiments. Boxes span 25th to 75th percentile, whiskers extend to max/min of the values, excluding outliers (1.5 box heights away from box edges, which are shown as dots)

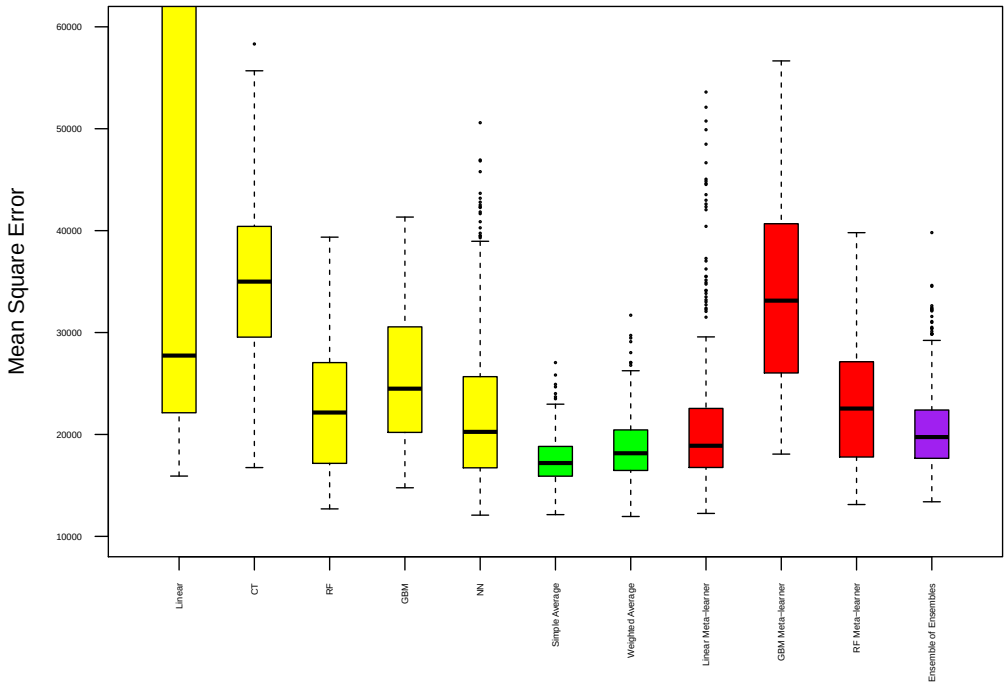


Figure 5.2: Model performance in predicting trip production in Chicago. Distribution of mean square error from 400 experiments. Boxes span 25th to 75th percentile, whiskers extend to max/min of the values, excluding outliers (1.5 box heights away from box edges, which are shown as dots)

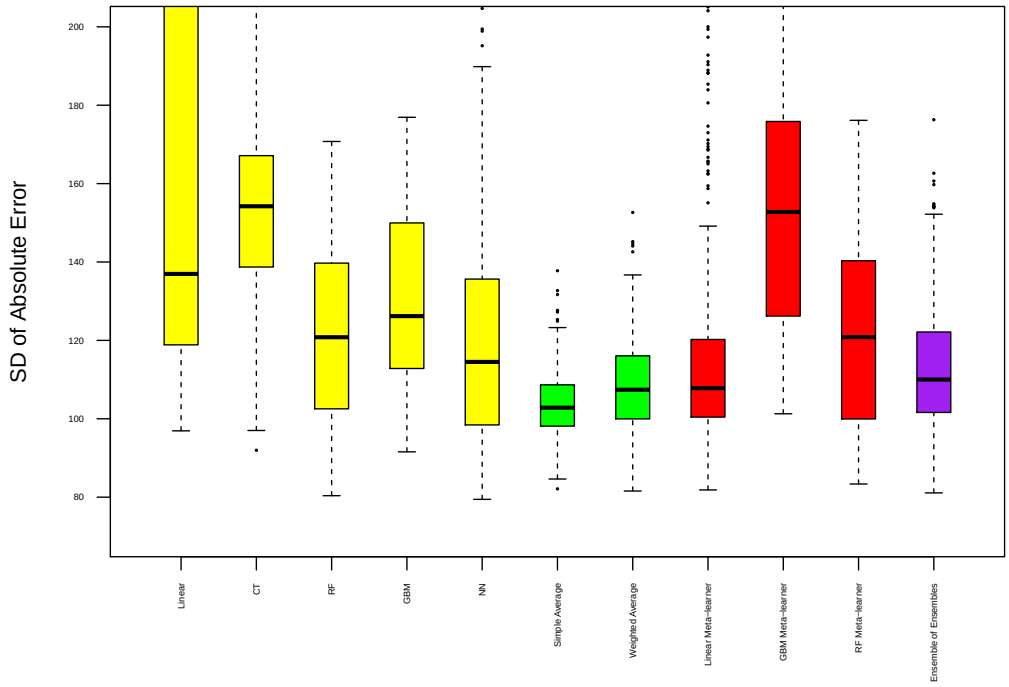


Figure 5.3: Model performance in predicting trip production in Chicago. Distribution of standard deviation of absolute error from 400 experiments. Boxes span 25th to 75th percentile, whiskers extend to max/min of the values, excluding outliers (1.5 box heights away from box edges, which are shown as dots)

New York City

In predicting trip production and attraction in New York City, only the simple average ensemble model is able to improve model performance beyond the best single model (linear). The linear stacking model has similar performance as the best single model.

Model performance in predicting trip production is shown in Figure 5.4 through Figure 5.6. Model performance in predicting trip attraction has identical patter as in predicting trip production, and is shown in Appendix B.

When the classification tree is removed as one of the base models, the simple average of the 4 remaining base models perform better than the best single model (linear) in terms of MAE and MSE, and 'no worse' in the stability of model performance.

Weighted average ensemble models, and RF, GBM stacking models performed worse than best single model. The ensemble of ensembles method, which averages RF, GBM and linear meta-learners also performed worse than the best single model.

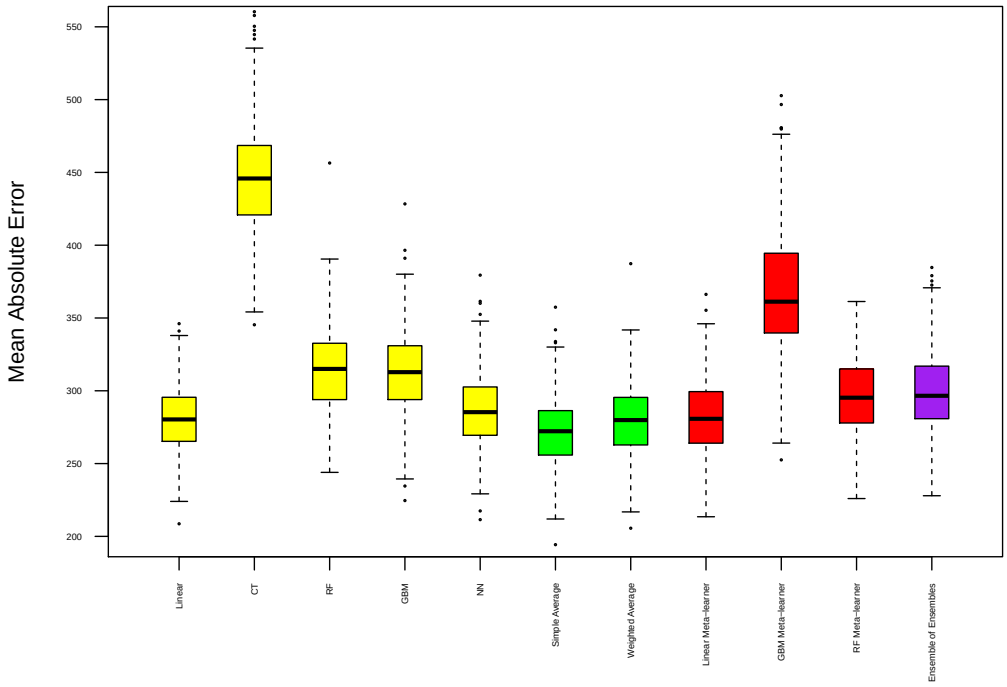


Figure 5.4: Model performance in predicting trip production in NYC. Distribution of mean absolute error from 400 experiments. Boxes span 25th to 75th percentile, whiskers extend to max/min of the values, excluding outliers (1.5 box heights away from box edges, which are shown as dots); base models exclude used in ensemble models classification tree

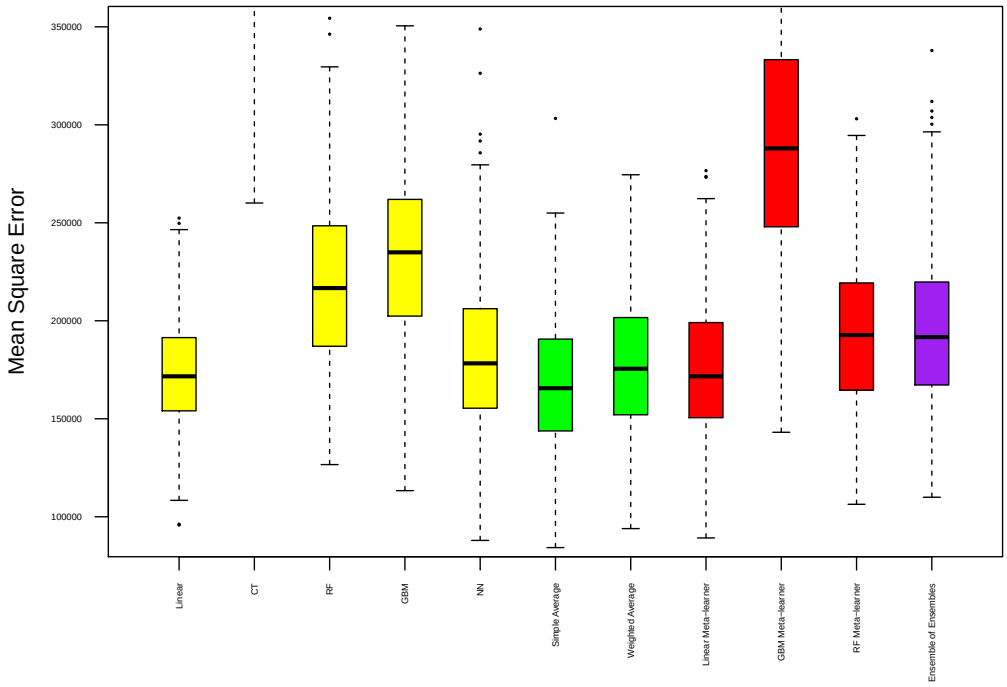


Figure 5.5: Model performance in predicting trip production in NYC. Distribution of mean square error from 400 experiments. Boxes span 25th to 75th percentile, whiskers extend to max/min of the values, excluding outliers (1.5 box heights away from box edges, which are shown as dots); base models used in ensemble models exclude classification tree

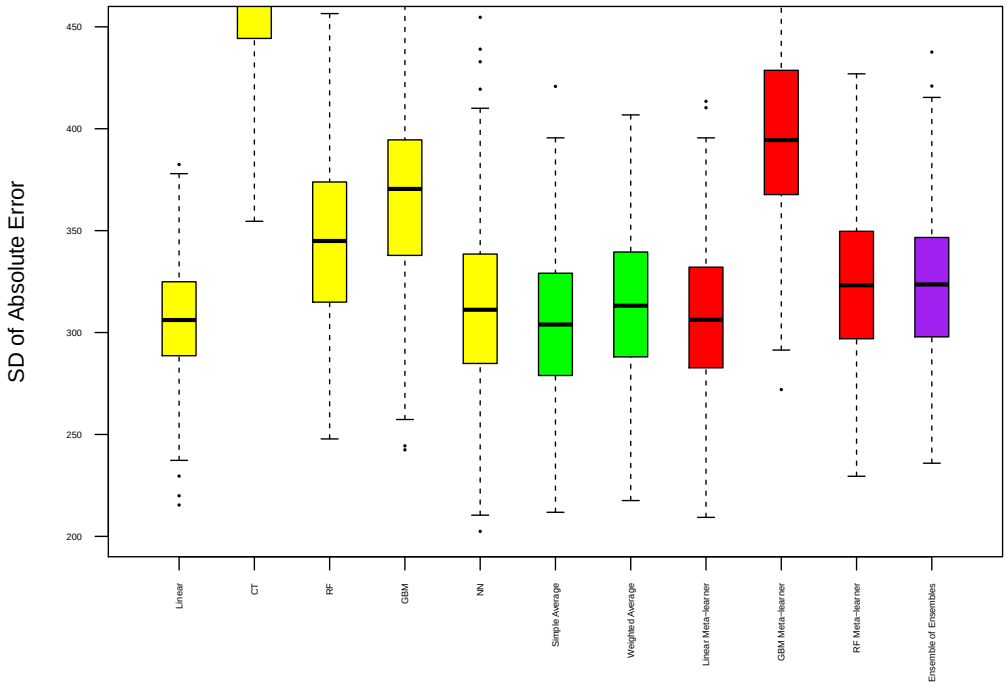


Figure 5.6: Model performance in predicting trip production in NYC. Distribution of SD of absolute error from 400 experiments. Boxes span 25th to 75th percentile, whiskers extend to max/min of the values, excluding outliers (1.5 box heights away from box edges, which are shown as dots; base models used in ensemble models exclude classification tree

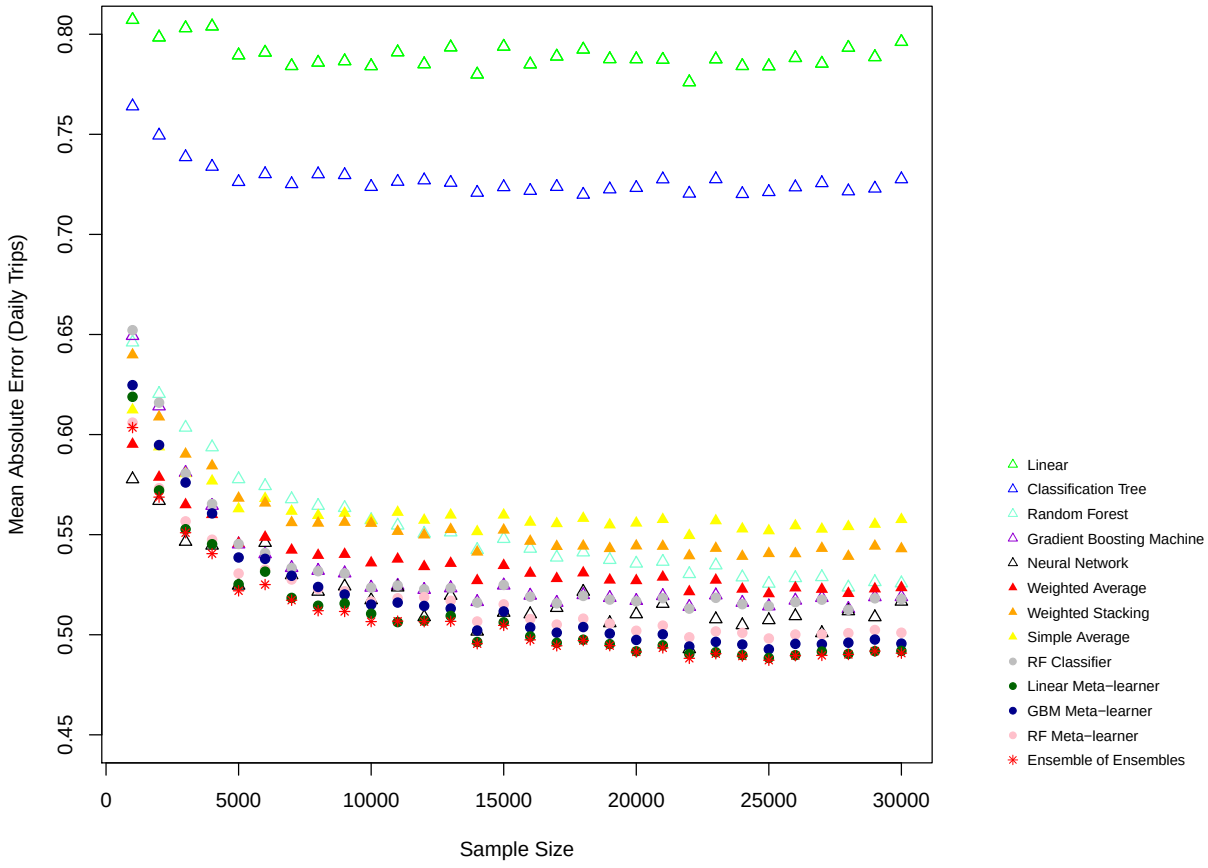
5.3.2 Flow Model

Ensemble models with simple rules provide no improvement from the best single model prediction, but accuracy of these ensemble models are generally similar to the best single model. The neural network model produced low mean absolute error, but has high mean square error, and standard deviation of absolute error, suggesting the presence of large errors in the neural network model. On the other hand, the weighted average ensemble model scored well (although not the best) on all three measures.

Ensemble models with stacking methods improved the mean square error beyond the best base model, suggesting a reduction in large errors. Mean absolute error of stacking models is a slight, but not significant, improvement beyond the best base model.

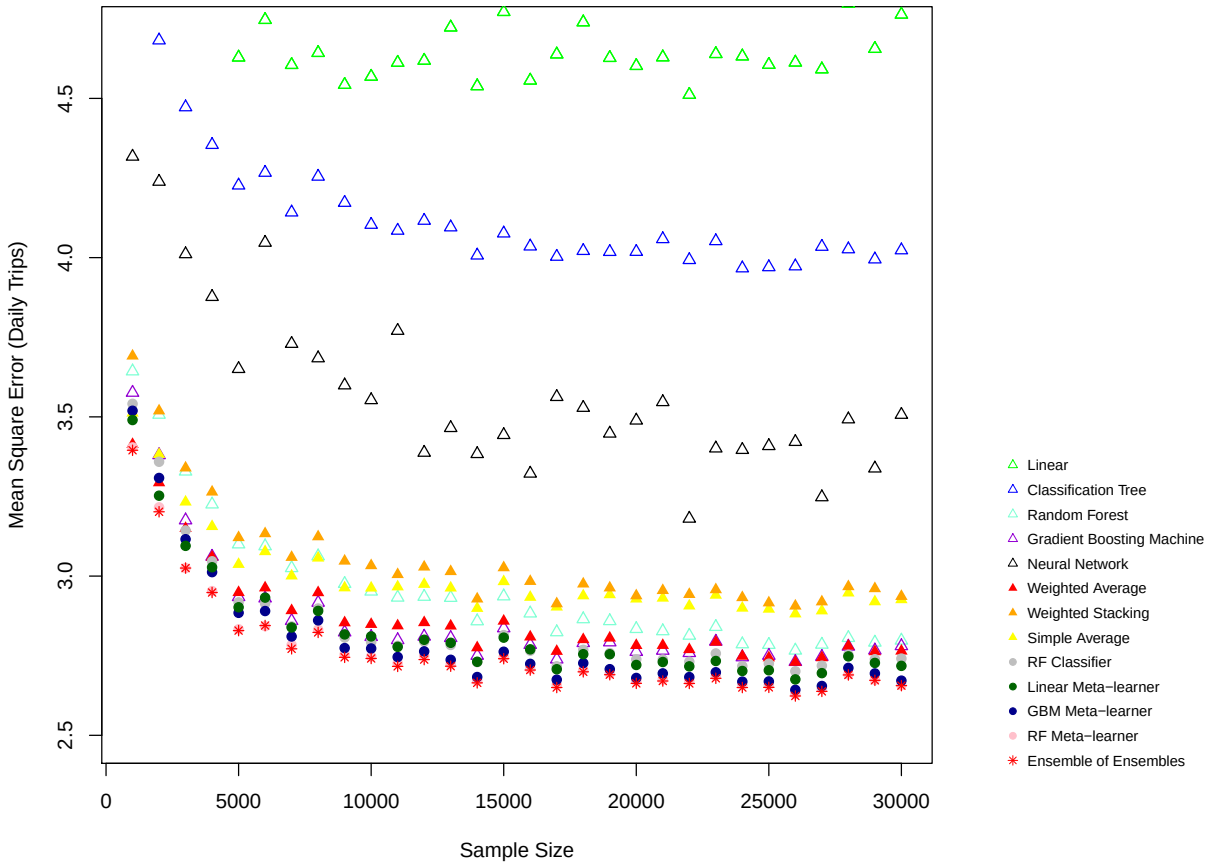
In predicting the flow of for-hire vehicles, the weighted average ensemble model improves forecast accuracy beyond the best single model when the training sample size is small; once the sample size increases, the weighted average becomes less accurate than the best single model prediction.

Chicago



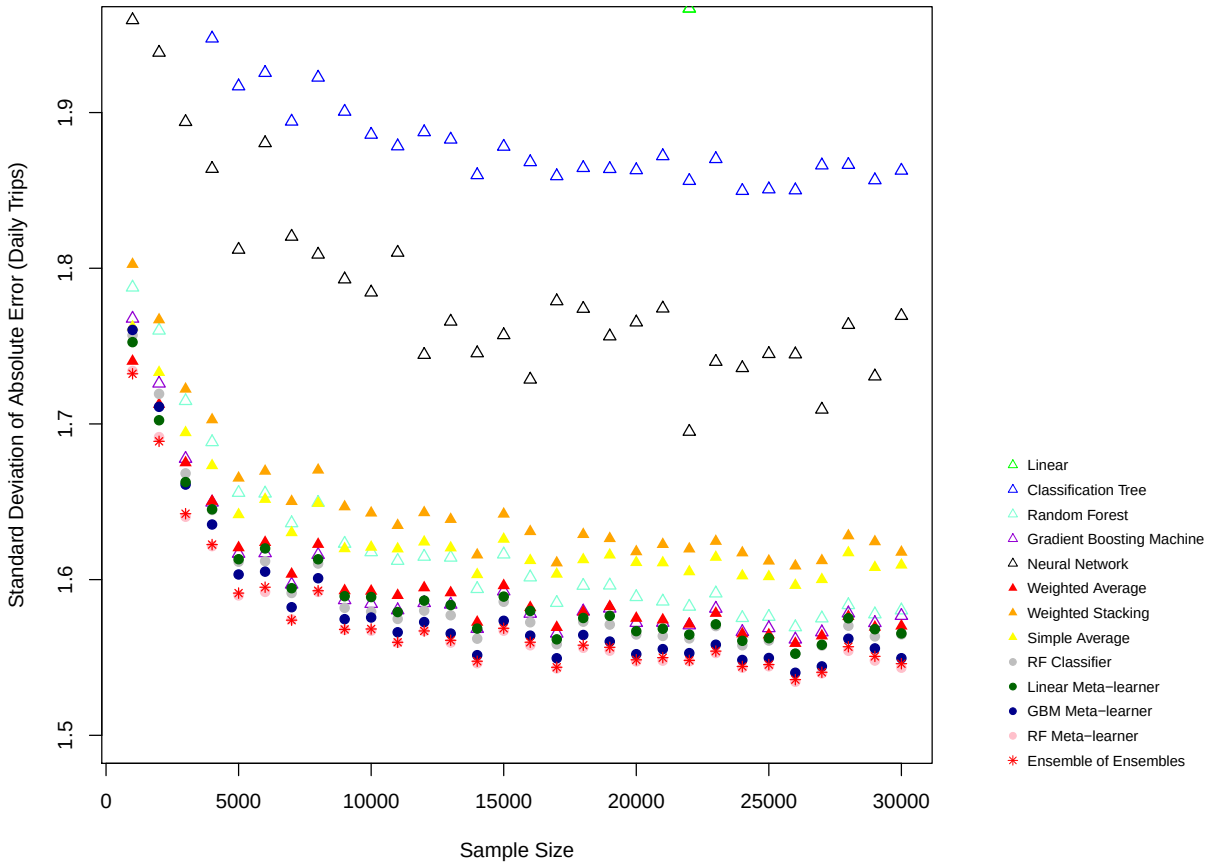
(a) Mean absolute error (b) Models

Figure 5.7: Model performance in predicting FHV flow in Chicago. Mean absolute error of different models, in testing data; every dot is the average of 90 experiments, each with 100 testing samples



(a) Mean square error (b) Models

Figure 5.8: Model performance in predicting FHV flow in Chicago. Mean square error of different models, in testing data; every dot is the average of 90 experiments, each with 100 testing samples



(a) SD of absolute error (b) Models

Figure 5.9: Model performance in predicting FHV flow in Chicago. Standard deviation of absolute error of different models, in testing data; every dot is the average of 90 experiments, each with 100 testing samples

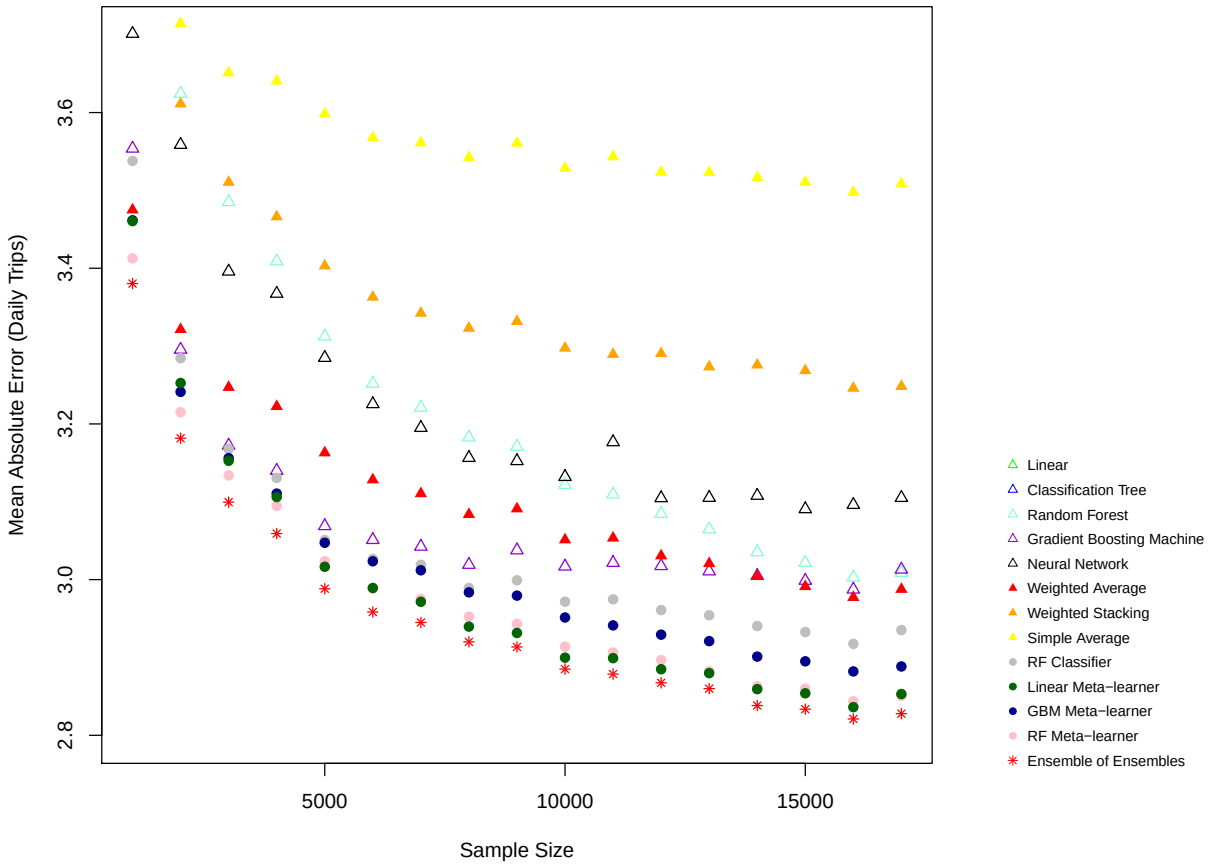
New York City

Model performance in predicting the flow of for-hire vehicles is shown in Figure 5.10 through Figure 5.12.

In Figure 5.10 and Figure 5.11 among the base models, the gradient boosting machines has the best performance in MAE, MSE, and in the standard deviation of absolute errors. The neural network has good performance in the MAE measure, but very high MSE in predicting flow, suggesting many large errors in neural network predictions.

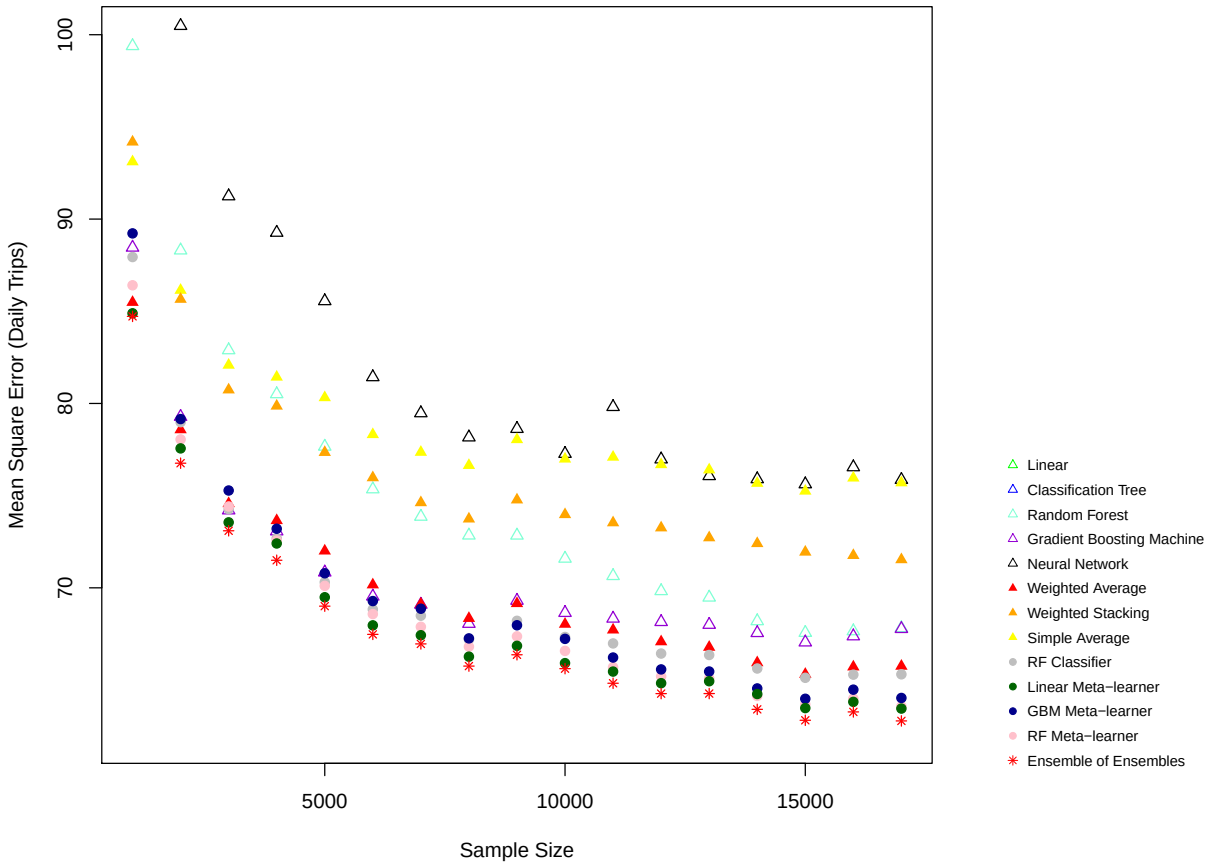
Stacking ensemble models are able to improve model performance beyond the best single model. Among different stacking ensemble models, the linear meta-learner has the best accuracy (MAE and MSE), and produces forecasts with the most stable accuracy.

Ensemble of ensembles, averaging three stacking ensemble models (Linear, RF, GBM meta-learners), improves accuracy beyond the best stacking ensemble models.



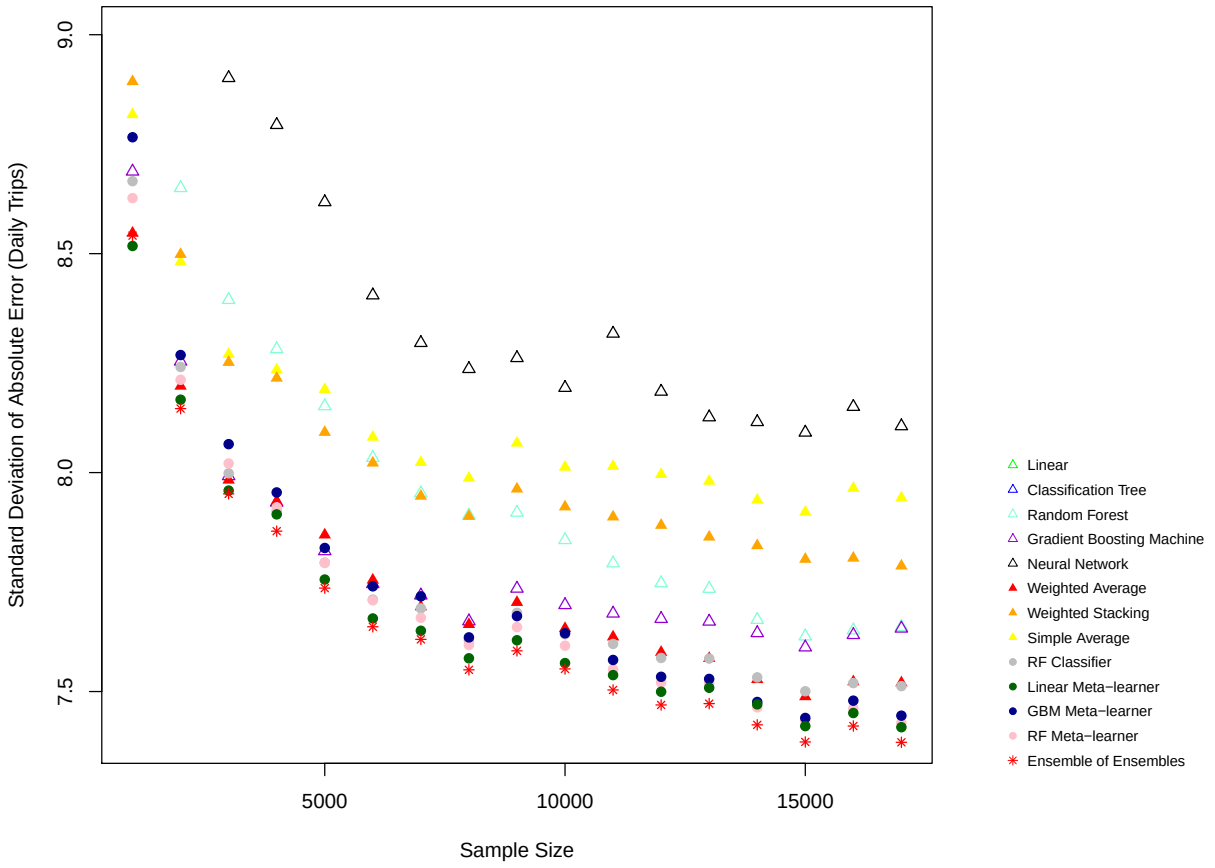
(a) Mean absolute error (b) Models

Figure 5.10: Model performance in predicting FHV flow in NYC. Mean absolute error of different models, in testing data; every dot is the average of 90 experiments, each with 100 testing samples



(a) Mean square error (b) Models

Figure 5.11: Model performance in predicting FHV flow in NYC. Mean square error of different models, in testing data; every dot is the average of 90 experiments, each with 100 testing samples



(a) SD of absolute error (b) Models

Figure 5.12: Model performance in predicting FHV flow in NYC. Standard deviation of absolute error of different models, in testing data; every dot is the average of 90 experiments, each with 100 testing samples

5.4 Base vs. Ensemble Models

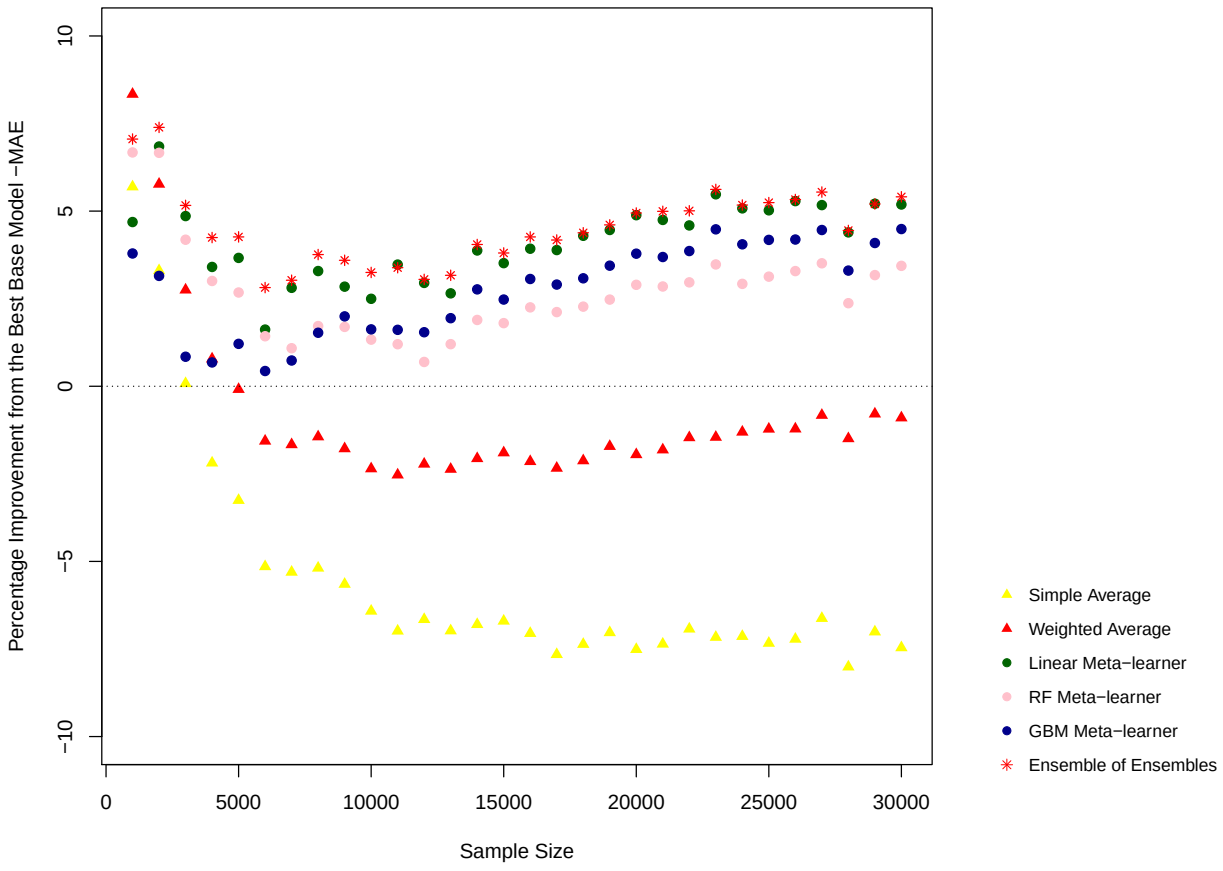
The extent of performance improvement from ensemble models are shown in Figure 5.13 through Figure 5.16; percentage improvement in MSE follow identical patterns as MAE, and are presented in Appendix C.

An initial drop in the performance improvement from stacking ensemble models can be observed in Figure 5.13 for Chicago, and Figure 5.14 for New York City, which shows that at certain levels of training data size, the amount of improvement obtainable from ensemble models is reduced. This can be explained by the difference in how fast base models and meta-learners improve their accuracy: if the best base model rapidly improves itself with more training data, and the meta-learners were to have a slower rate of improvement than the base model, then the gaps between the best base model and the ensemble models will be reduced,

resulting in the noticeable kink in the figures. Diminishing returns set in as performance improvement per unit of extra training data drops, and may eventually disappear, in the base models, and the meta-learners are able to further improve upon base models. For this reason, the extent of accuracy improvement with meta-learners generally increases with the size of training data to a point.

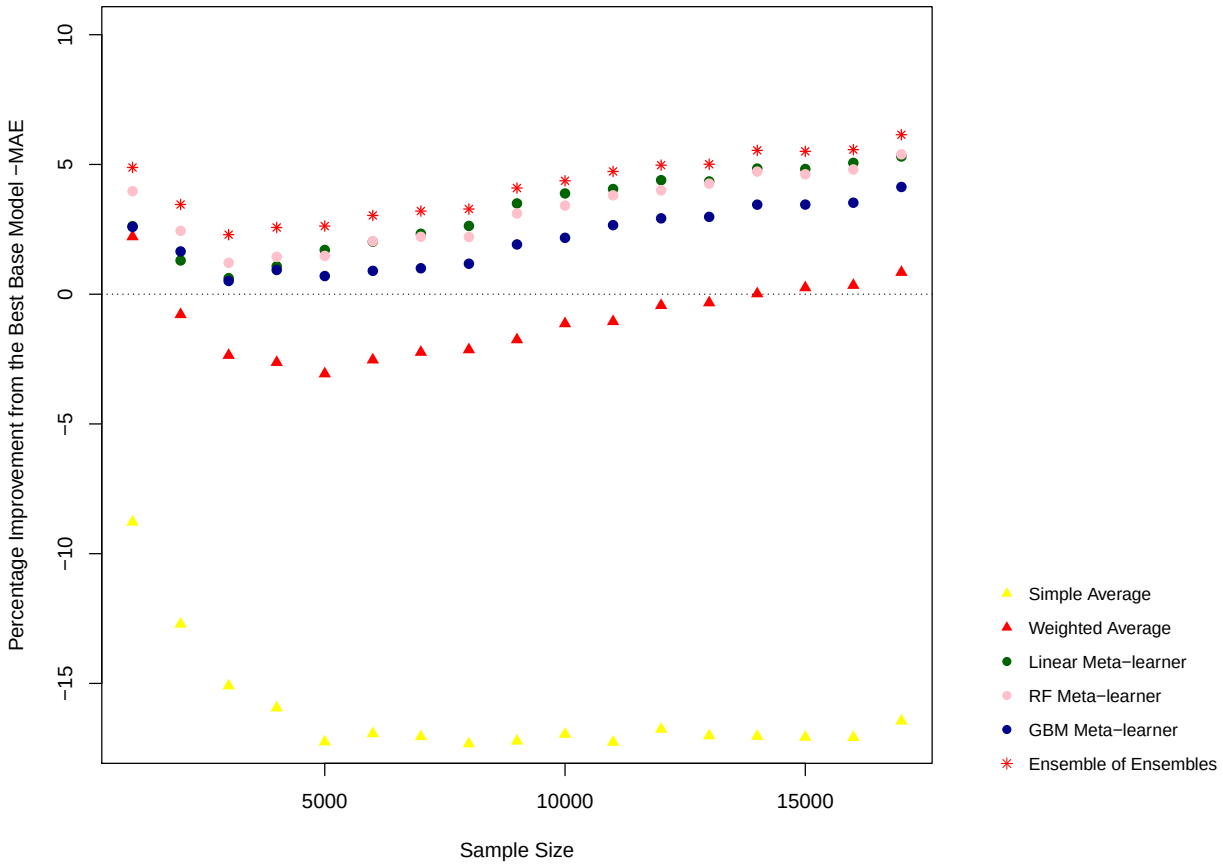
On the other hand, if the meta-learner improves faster than the best base model, then the amount of performance improvement from ensemble models will not diminish with more training data, and the kinks in Figure 5.13 and Figure 5.13 will not appear. This is possibly the case in the hedonic models in Figure 4.9 (Sydney hedonic models) where the amount of performance improvement from ensemble models increases steadily with training data size.

Other ensemble of ensembles methods in iteration 1 are tested, namely the linear, RF, GBM as meta-learners combining previous ensemble model outputs, with the addition of the average of these meta-learners in iteration 2 (i.e. an ensemble of ensemble of ensemble model). The results are identical to the case in Sydney Hedonic models: these ensemble of ensembles models do not improve upon the simple average of ensemble models in iteration 1.



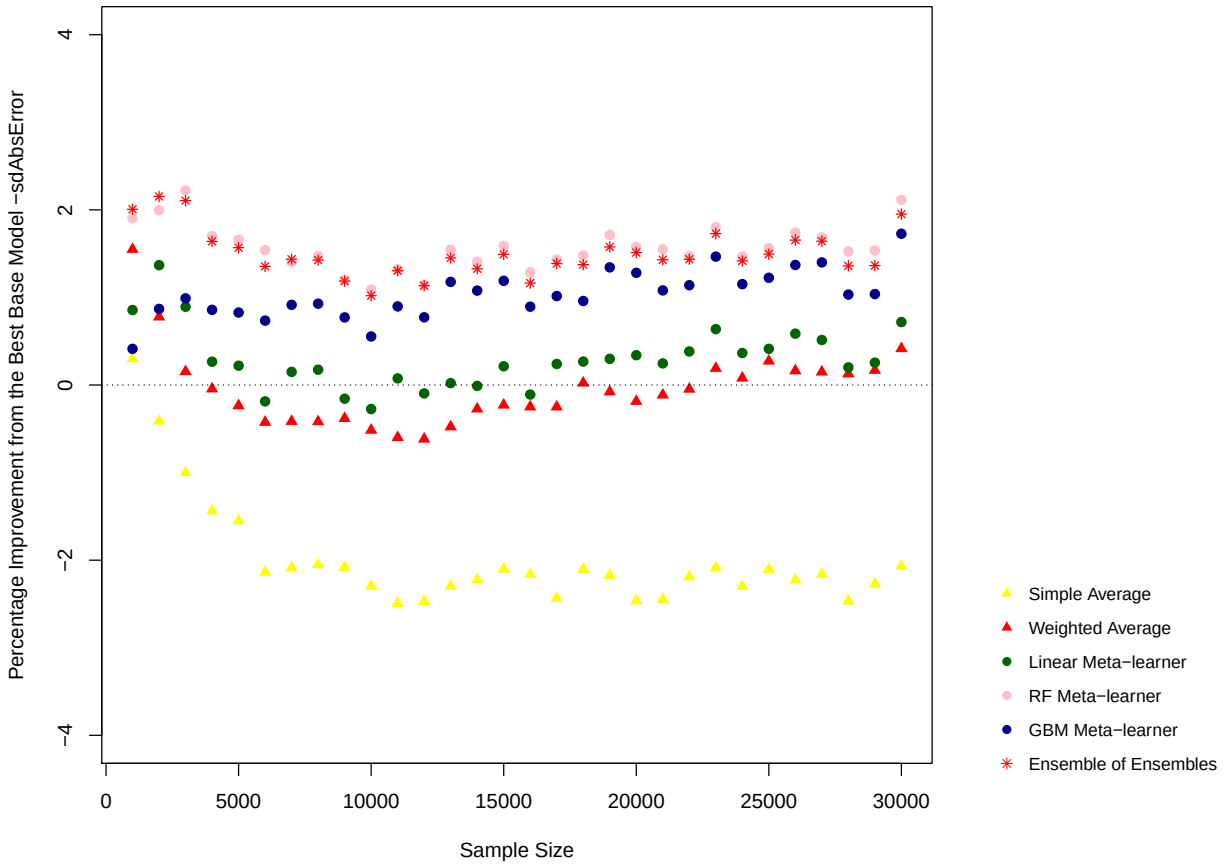
(a) Percentage Improvement in MAE (b) Models

Figure 5.13: Performance improvement from the best base model - Chicago MAE



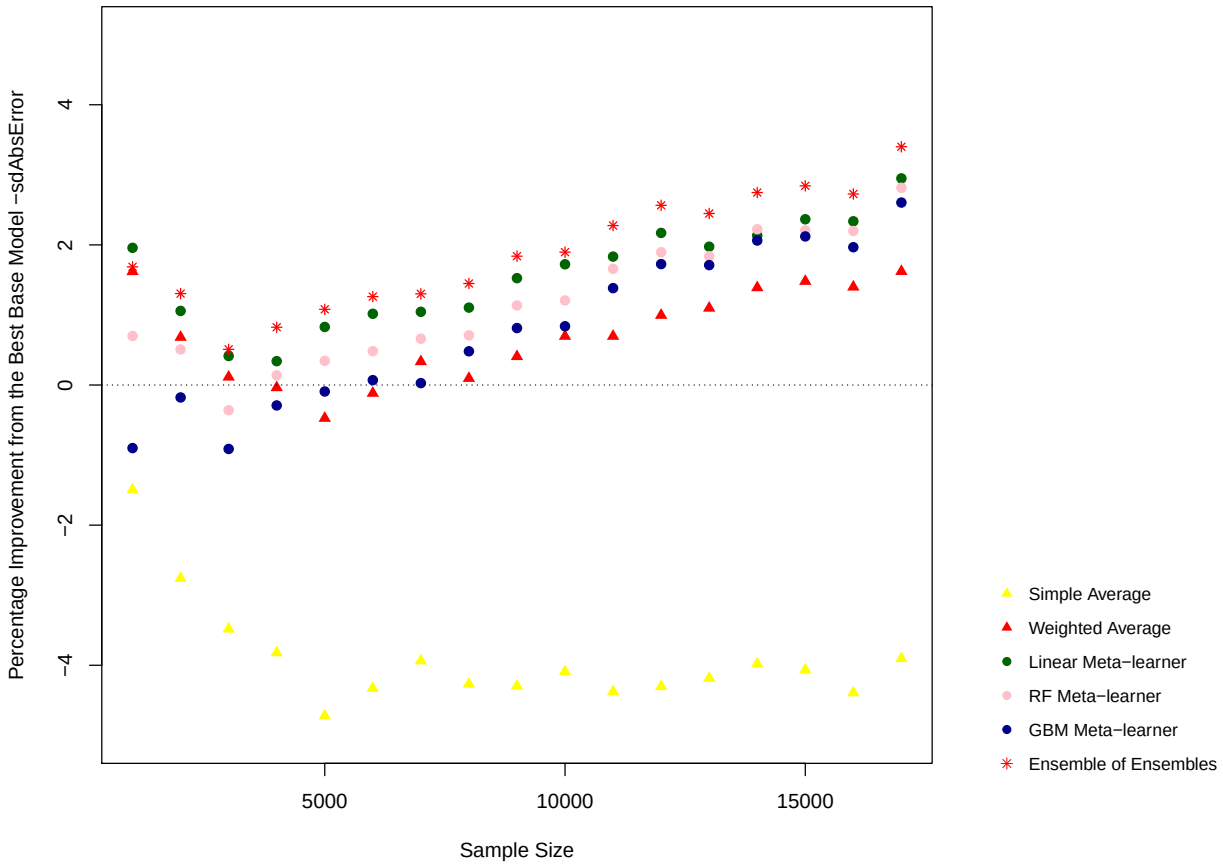
(a) Percentage Improvement in MAE (b) Models

Figure 5.14: Performance improvement from the best base model - NYC MAE



(a) Percentage Improvement in SD of Absolute Error (b) Models

Figure 5.15: Performance improvement from the best base model - Chicago SD of absolute error



(a) Percentage Improvement in SD of Absolute Error (b) Models

Figure 5.16: Performance improvement from the best base model - NYC SD of absolute error

5.5 Discussion

In the case of predicting trip production and attraction, with a small training sample size, we find the simple averages combining rule to outperform other ensemble models and base models, producing both more accurate, and more reliable forecasts. Stacking ensemble rules provide little to no improvement from the best single model. This is possibly because stacking models require sufficient data to be calibrated; with a small sample size, both the base models and meta-learners are not adequately calibrated. We also find the linear meta-learner, and ensemble of ensembles to have robust performance, in that, although providing no significant improvement, these two ensemble models have similar performance to the best single model.

In predicting flow, meta-learner ensemble models are calibrated with sufficient training data, and are able to improve model performance beyond the best single model. In most cases ensemble of ensembles has the best performance. The neural network model has low mean absolute error, but also a significant amount of large errors, which resulted in a high

MSE. So reliance on a single model based on one set of performance measure can be risky.

Ensemble models (especially robust ensemble algorithms such as linear meta-learner and ensemble of ensembles) can generally improve model performance. Ensemble models are also well rounded in performance, providing a good balance between forecast accuracy, large and small errors, and stability of forecast accuracy. In general, ensemble models are either better, or 'no worse', than the best single model forecast. However, discretion is needed in applying ensemble models under different scenarios. In cases without enough data to calibrate models, simpler and robust ensemble models become preferable.

The comparison between ensemble models and the single-model approach shows that, relying on a single model is not the best modeling practice, even when a single model appears to have the best performance; further performance gains can be obtained by combining models with different assumptions or pattern recognition methods.

Chapter 6

Benefits and Costs of Ensemble Forecasting

Ensemble forecasting has a number of benefits over single-model predictions, that make ensemble models particularly useful in a wide range of applications. In this chapter we discuss specific drawbacks with the single-model doctrine, connecting these drawbacks with how the single-model doctrine often fall short of its role as decision support tools. We then draw from our experience with ensemble models in previous chapters, in discussing benefits of ensemble forecasting, and how ensemble models can provide remedy to many problems with the single-model prediction. Costs of using ensemble models are also discussed.

The literature often describe specific advantages of using ensemble models in a specific context. For example, weather forecasts often emphasize the initial measurement error, and ensemble models in weather applications focus almost exclusively on reducing the accumulation of error over time. Economic models, business and political predictions often focus on uncertainties in model assumptions, and combine different assumptions to obtain more robust forecasts; the effect of accumulated error may or may not be incorporated in these models. The different benefits of ensemble forecasting often cross paths with each other, and one ensemble method may have multiple benefits. Different pieces of ensemble forecasting are often not well connected by the literature. Therefore it is necessary to summarize the scope of potential benefits from ensemble forecasting, and string together what ensemble models are good at. The performance of ensemble models, and both the advantages and disadvantages of ensemble models are also summarized in this chapter.

While other chapters generally view ensemble forecasting from a transport perspective, and seek to utilize ensemble models in transport applications, this chapter presents benefits of ensemble forecasting without shackles of any specific discipline. There are different theories as to what benefits can be obtained from ensemble forecasting, and about mechanisms of these benefits.

6.1 Base vs. Ensemble Models Performance

Ensemble models are applied in previous chapters to a range of different transport problems. Both ensemble and single-model predictions are presented as single numbers, in order to

		Training Data Size - Large			Training Data Size - Small	
Ensemble Method	Performance Metric	Sydney Hedonic	Chicago FHV Flow	New York FHV Flow	Chicago FHV Trips P&A	New York FHV † Trips P&A
Simple Average	MAE	⊙0.72	⊙7.51	⊙17.07	⊗3.01	⊗3.12
	MSE	⊙0.62	⊙6.01	⊙12.24	⊗21.82	⊗2.90‡
	Stability	=	⊙2.46	⊙4.07	⊗14.08	=
Weighted Average	MAE	⊙1.37	⊙1.95	⊗0.25‡	⊗2.24	=
	MSE	⊙0.95	⊙0.73†	⊗2.58	⊗16.53	⊙2.77‡
	Stability	⊗0.17‡‡	=	⊗1.48	⊗10.11	⊙2.50†
Linear Stacking	MAE	⊗2.65	⊗4.88	⊗4.82	=	=
	MSE	⊗4.82	⊗1.50	⊗5.32	=	=
	Stability	⊗2.27	⊗0.34†	⊗2.37	⊗2.85‡‡	=
RF Stacking	MAE	⊗2.54	⊗2.90	⊗4.62	⊙4.47	⊙5.36
	MSE	⊗3.46	⊗3.33	⊗5.01	=	⊙11.67
	Stability	⊗1.17	⊗1.58	⊗2.21	=	⊙5.57
GBM Stacking	MAE	⊗2.86	⊗3.79	⊗3.46	⊙18.79	⊙30.94
	MSE	⊗3.67	⊗2.97	⊗4.57	⊙52.11	⊙71.19
	Stability	⊗1.12	⊗1.28	⊗2.12	⊙26.22	⊙29.99
Ensemble of ensembles (Average)	MAE	⊗3.77	⊗4.95	⊗5.51	⊙3.25	⊙6.61
	MSE	⊗5.44	⊗3.59	⊗6.31	⊗7.88	⊙12.28
	Stability	⊗2.02	⊗1.51	⊗2.84	⊗6.35	⊙5.09†

Performance metrics: Mean Absolute Error (MAE) and Mean Square Error (MSE); Lower is better
Model performance measured as percentage improvement(⊗), decline(⊙), no change(=) compared to the best single model(GBM);
Training sample size: 15,000 for hedonic and NYC flow models; 20,000 for Chicago flow models;
Testing sample size: 5,000 each time; repeated 9,000 times to obtain a distribution of performance metrics;
p-value < 0.001 unless marked with daggers. †: $p < 0.02$; ‡: $p < 0.05$; ‡‡: $p < 0.1$;
†: The Classification Tree has especially low performance, and therefore removed from base models.

Table 6.1: Percentage improvement by ensemble forecasting over the best base models (in testing data)

compare different models in this section. The performance of ensemble models relative to base models is shown in Table 6.1. The Gradient Boosting Machine (GBM) is the best performing base model in all cases, and the performance of ensemble models in Table 6.1 are measured as whether the ensemble model improves (⊗) or worsens (⊙) the performance of best base model, and how much (percentage) compared to the GBM. Model performance is measured by Mean Absolute Error (MAE), and Mean Square Error (MSE), so a performance improvement would mean lower MAE and lower MSE. Stability of model performance is measured by the standard deviation of absolute error. The results confirm that, if applied properly, ensemble models can further improve forecast accuracy of the best base model by a notable degree.

Model performance is obtained from repeated experiments, that form a distribution of model performance metrics; the table shows the mean of each performance metric in each application. Among the three performance metrics, the mean absolute error (MAE) measures the average size of errors. The mean square error (MSE) supplements the MAE metric in measuring both the average size, and the presence of large forecast errors.

The performance of simple and weighted average ensemble models have a mixed result. With sufficient training data, stacking ensemble models generally perform better than the simple average and weighted average ensemble models, and provide positive improvements upon the best performing base model. Ensemble of ensembles (iteration 1, average of meta-learners, which combines different ways of combining models) provides further performance enhancement upon ensemble models, and generally provides the best performance among ensemble models. The linear meta-learner is generally the second-best ensemble method and

has good robustness. More complex meta-learners from iteration 1, and simple average from iteration 2 provide no improvement beyond the simple average in iteration 1, suggesting ensemble models picking up more noise than information.

The extent to which ensemble models improve forecasts from the best single model depends on many factors, and the size of the training data is a significant factor. The performance differences between base and ensemble models are statistically significant, and the performance advantage of ensemble models tends to be greater as the training data becomes larger in size. The performance gain of ensemble models eventually levels off, at which point additional training data can no longer improve model performance.

The performance stability of a model is measured by the standard deviation of absolute error, which reflects the degree of variation in the size of forecast errors. Since the model accuracy is measured by the average size of forecast errors, large variations in error sizes would make the model accuracy measure less reliable in individual cases.

The ensemble of ensembles method most consistently improves the stability of model performance across different training sample sizes. Other ensemble methods improve performance stability less consistently, or to a lesser degree. For example, the linear ensemble method improves performance stability more than the ensemble of ensembles method in predicting property prices, but provides little improvement in predicting Chicago FHV flow. The random forest and gradient boosting machine generally require a large training data in order to improve performance stability beyond the best single model.

6.2 Problems with the Single-model Doctrine

We summarize three major problems with the single-model doctrine.

6.2.1 Lack of consideration for multiple paths

The first major problem with the single-model doctrine relates to the exclusion of other possible data generation mechanisms, and methods of data exploration. Real-world events often have multiple possible paths, and that a different data generation process may be at work for different persons, under different circumstances, or at different times. Multiple sources of data measurements might lead to different model outcomes. The single-model doctrine denies the intrinsic complexity within real-world events, and instead bet on the 'most likely' data generation mechanism. And one set of data may have multiple underlying patterns, therefore a single method for identifying patterns may not extract all available information.

The economic common sense of 'don't put all your eggs in one basket' applies to modeling as well, here we rephrase it as 'don't put all your faith in one model'. In cases where the path of real-world event deviates away from the single-model assumption, this can lead to some very inaccurate predictions. There is no 'checks and balances' in single-model predictions, so the failure of a single model will result in the failure of an entire model prediction.

The relative performance of various models are "for these data and the approaches employed", so some models will work better in other datasets, or with different sizes of training data. This makes it more difficult for finding the 'best" single model.

There are applications where knowledge on different possible outcomes can be particularly useful, even when some of the outcomes have a low probability of occurrence. Such applications often have disproportionate consequences for large forecast errors compared to smaller errors, or large penalties for failing to account for low possibility events. For instance, a hurricane evacuation area is based on many possibility paths of the hurricane; in this case, the forecast is interested not only in the most likely path, but in all possible paths that a hurricane might take, so that residents can take appropriate precautions.

6.2.2 Presentation of model input and output as single numbers

The second major problem with the single-model doctrine is the presentation of model output as a single number. This problem is partially derived from the limitation of a single model assumption, which lacks consideration for multiple paths, and allows for only one possible outcome from the model prediction.

Explanatory variables rarely deal in absolutes, and many variables are linked to some possibility of occurring, or have some flexible variations. A single measurement data may not accurately reflect the reality, for example, body temperature varies at different times of a day, to which the 'normal' body temperature measure is just a 'typical' daytime temperature. Variations in the measurement data cause differences in the model outcome, which are not reflected by a model prediction as a single number.

Other sources of modeling uncertainties, including in model assumptions, and from external shocks to the system, etc. also add to the level of uncertainties in the model output. Therefore all that can be obtained from models is a range of possible outcomes. The single model doctrine presenting model output as a single number assumes a level of certainty that does not exist, and does not provide information on the range and variability of possible outcomes. In section 4.4.3 we show the use of ensemble models to predict the housing price as a range of possible outcomes, which may be used as a decision support tool. The real-life housing price is inherently probabilistic depending on the bargaining and interaction between buyers and sellers. The single-model approach of predicting a single value fails in that respect.

6.2.3 Accumulation of error in long-range forecasts

The inability of the single-model doctrine to alleviate accumulated error is a major problem in long-range forecasts. In non-linear systems, variations of the same extent but in different directions have different degrees of impact on models, so averaging measurements still embed errors within the explanatory variables; these errors tend to accumulate over time, and often render long-range forecasts useless.

This problem can be explained to a laymen with an example from a popular team-building game called 'telephone', where participants would form into a straight line, and a message is passed on from one person to the next, and from the start to the end of the line. Each time the message is passed on to a different person, subtle differences are added to the sentence, which accumulates and often make the message unrecognizable in the end.

In section 4.4.2 we experimented with using multiple input data for models, although this produced little difference in model performance in short-term predictions. It is hoped that

in long-range model predictions, the use of ensemble models will make a more significant difference, which is the topic of future research in ensemble models.

With the single-model doctrine, neither other possible outcomes, nor the range of possible variations in the model outcome can be known. This attribute of the single-model doctrine makes long-range forecasting mostly useless. Long-term transport demand forecasts have very low accuracy (Boyce and Williams, 2015).

6.3 Advantages of Ensemble Forecasting

In this section, we summarize major advantages of the ensemble forecasting, based on results from applications of ensemble models conducted in this research, and combined with our own understanding of ensemble models. In applying ensemble models, we find that ensemble models exhibit many desirable properties, that provide remedy to many existing problems in transport modeling.

We believe that transport modeling can benefit from the adoption of ensemble models, which includes higher accuracy, better reliability, and more informative model predictions. Since ensemble forecasting is not a specific method, the idea and awareness of ensemble forecasting also sheds light on the limitations of the single-model doctrine, which shall help practitioners in choosing models, and in interpreting their modeling results. Hence the dissemination of ensemble forecasting into transport modeling would have some positive effects.

6.3.1 Improve forecast accuracy

Accuracy improvement is a major advantage in using ensemble models. In previous chapters, we find that when base models are combined into ensemble models, and with sufficient training data, and properly chosen and executed ensemble models, there is about 4% to 5% improvement in the mean absolute error beyond the best base model. This shows that the act of combining multiple model assumptions improves forecast accuracy.

The accuracy improvement comes with a prerequisite for sufficiently large training data sizes. In all applications tested in this research, ensemble models consistently have good performance when the size of training data is large. With the exception of the linear-meta learner, ensemble models don't perform well in cases with a small training data, and tend to have slightly worse performance than the best base model. The linear meta-learner is a robust ensemble method, that provides similar performance as the best base model when the sample sizes were small.

Improvement in forecast accuracy is a major motivation for the adoption of ensemble forecasts in many cases. The theoretical basis for accuracy improvement stems from two sources, namely the pooling of information, and the pooling of errors.

Pooling information: A philosophical basis for combining forecasts views forecasts as information, and the combining of forecasts as aggregating information (Winkler, 1989). Useful information might be dispersed widely, making it difficult for a single model to capture all information (Tetlock and Gardner, 2016). In this respect, more information about the data generation process can be obtained by combining different models that contain non-overlapping pieces of information, even when these models have lower accuracy than the

best model. So combining different models essentially pools information, that the ensemble models may produce accuracy higher than each individual model.

Pooling errors: In addition to pooling information, the different data sources and models combined in ensemble models also pool errors and biases. Different models are common in extracting patterns within a data set, but differs in how they mistake noise for signals; combining different models filters out noise while preserving signals (Zhang and Ma, 2012). If the signal and the noise were to be visualized as sound waves on a two dimensional graph, and the process of combining models visualized as superimposing different graphs, then the signals from different models will tend to reinforce each other, and the noise (errors) in different models will tend to cancel out each other, because the occurrence of errors are more random by nature.

Although the exact quantitative mechanism through which ensemble models improve forecast accuracy is not clear, and many ensemble models remain a black box, we discuss the some common ensemble methods, and their role in improving forecast accuracy.

Simple Rules

Differences in model assumptions and biases often cause the actual dependent variable to be between different model predictions. Averaging (both weighted and unweighted) different model predictions by simple rules causes the combined forecast to get closer to the actual value, than randomly selecting one model prediction from different base models; this is known as 'bracketing' (Larrick and Soll, 2006). When the 'bracketing' does not occur, the error from the combined forecast and the single model forecast will be of the same magnitude, because the error from the combined forecast will be identical to randomly choosing one model from the ensemble of models (Graefe et al., 2014, Larrick and Soll, 2006).

Different biases and random errors among different base models make the 'bracketing' more likely to occur (Graefe et al., 2014), so it is important to have diversity both in the model formulations and in the data. When theories used in models are more diverse, it has been found that the ensemble of models has lower errors (Batchelor and Dua, 1995).

The rationale for combining models using weights comes from the complexity in the real-world data generation process. A factor that is crucial in one mechanism can be totally irrelevant in another. As a result, when predicting an event, the observed dependent variables are not derived from a single process, but a turbid mixture of data generated by different mechanisms. Each type of model has its unique structure, representing one assumption on the data generation process; in the single-model doctrine, the best model captures the dominant mechanism. By having an ensemble of models that have different formulations, although none of the models are 'true', the combined forecast may better replicate phenomena in the real-world. Weights on different individual models can be viewed as chances that each model is closest to the reality among the selection of models (Winkler, 1989).

Stacking

The stacking ensemble methods can be further divided into two categories, namely the meta-regression, and meta-classification. The main objective of stacking is to improve the overall model accuracy, and the method doesn't guard against large errors, and neither does it aim

to improve model robustness. Although sometimes the stacking method reduces the chance of large errors by providing 'checks and balances' in cases where one of the base models malfunctions.

The stacking ensemble models place more emphasis on pooling information, than on pooling errors. In some cases the stacking models can act similarly as ensemble models with simple rules; the linear meta-learner in particular, provides coefficients to each base model, which is very similar to that of the weights in the weighted average ensemble method, except that these coefficients don't necessarily added up to one. Compared to simple average and weighted average ensemble models, stacking models aim directly at maximizing model accuracy. With sufficient training data, and if the meta-learners were able to be properly calibrated, the resulting stacking models can have good accuracy measure.

Ensemble of ensembles

We term the practice of further combining ensemble models 'ensemble of ensembles'. This recognizes that a single method of combining base models may not be optimum. Ensemble of ensembles pools both information and errors from models that combine base models, so an improvement in forecast accuracy is an expected benefit from ensemble of ensembles models. In the experiments from previous chapters, we did find ensemble of ensembles to improve forecast accuracy beyond ensemble models.

6.3.2 Provide the range of possible outcomes

Ensemble forecasting internalizes uncertainties in models, showing a range of possible outcomes as ensemble model output, instead of a single number. In other words, ensemble models are able to represent the level of uncertainty that is inherent within the system (Murray, 2018). When a forecast is provided as a single number, as in most transport models, it gives a false sense of determinism and infallibility. Real-world events rarely unfold exactly following model assumptions, so it is common for observed values to deviate from model predictions. The range of possible outcomes from ensemble models signals the level of uncertainty in the model forecast, and establishes the range of possibilities from best to worst scenarios, which will be very useful as decision support tools.

The range of possible outcomes can be obtained by a great variety of methods, that are centered on accounting for model uncertainties. Choice of the ensemble method can be customized based on needs, such as calibrating parallel ensemble models using subsampling (bootstrapping), using different data sources or time-series data, by using data with perturbations, or by using different models each with different assumption on the data generation process. Model output resulting from different ensemble methods carry different interpretations, that can be applied in different contexts.

Ensemble models are better at combining contextual information with the likelihood of an event. Model outputs presented as a range of possible outcomes can better incorporate contextual information. Passenger overflows at busy CBD locations are more likely to cause stampedes than in other places; a transport model predicting a 20% chance in travel demand exceeding transit capacity at a busy CBD location may cause a greater concern, than if the model predicts a 60% chance of demand overflow for a remote location. Other examples

include a three-year construction plan, so that segments that are likely to have a high traffic demand in the near terms can be prioritized; in a PM peak period congestion response plan for example, segments where congestion tend to occur, or accident prone areas can also be better identified with ensemble models, and limited resources (both in infrastructure investment and incident response) can be spent at places where it is most likely to be needed. In long-range models, such as a 40-year transportation plan, ensemble models can provide more robust predictions, and a range of possible predictions, so more informed decisions can be made; a divergent pattern in ensemble model outputs is an indication of the lack of predictive ability, in which case the ensemble model does not hide its level of uncertainties.

6.3.3 Informative Model Outputs

Having more informative model outputs is one of the major benefits of ensemble forecasting. Ensemble models can include multiple predictions from using different model assumptions, different data sources, and with perturbed measurement data, that form a range of possible outcomes. Cases with severe disagreement among models can be flagged and further investigated for the underlying causes, and predictions for such cases can be marked with a lower reliability. Ensemble models don't arbitrarily reject any possibility in the modeling process, and can produce one prediction for each uncertain possibility, so the ensemble prediction is presented as a range of paths and possibilities.

The level of uncertainties increases with long-range forecasts, and more possibilities in the modeling process would produce a wider spread of possible model outcomes than short-range forecasts, indicating the ensemble model is becoming less certain of its predictions. In iterative ensemble models, the accumulation of error in long-range forecasts will cause base model predictions (each of which uses one perturbed measurement data) to further diverge, indicating that the ensemble prediction is becoming less reliable. On the other hand, if the accumulated error has little effect on the model, then the ensemble prediction will continue to have a narrow spread, showing that the prediction to still be reliable. Events that are further into the future are subjected to more uncertainties, and are inherently more difficult to predict. The ability of ensemble models to incorporate the level of uncertainties in model outputs reflects the reality of how the world works, and at the same time makes ensemble models a type of 'honest' models, that acknowledges its deficiencies.

Ensemble forecasting is a type of 'honest' models that can provide the range of forecast variations caused by uncertainties, so model users can make informed decisions as to how much the model output can be trusted. If the level of modeling uncertainties only cause a small spread in the model output, such as in short-range predictions, then more confidence can be had in the ensemble prediction; on the other hand, if the amount of uncertainties produces a wide range in the ensemble model output, which is commonly seen in long-range predictions, then it is a signal that the ensemble forecast cannot be trusted. In ensemble models, the level of uncertainties can be obtained for each individual model prediction, which is not possible with the single-model doctrine. Knowing the reliability of model predictions can alleviate the impact of potentially inaccurate model predictions.

Ensemble models can include the range and spread of possibilities, the best and worst case scenarios, and the sensitivity of the result to uncertainties. In some cases, the question of interest may be the worst-case scenario predicted by the models, instead of the usual expected

values. The worst-case scenario carries a significant penalty for high stake events, and will be of concern even if the event has a low probability. For example, by showing multiple possible trajectories of a hurricane (Leutbecher and Palmer, 2008), the evacuation areas, and different degrees of evacuation orders can be based on the likelihood of each possible path. The epidemic growth model can be based on all the known information, including different assumptions on the weather, and the transmission rates, so the predicted trajectory of the epidemic includes the most likely, and also the best and worst case scenarios. Information on the best to worst scenarios helps decisions in preparation for events, and makes

Ensemble forecasting is more useful as decision support tools than single-model predictions. Ensemble model outputs can either be used directly as the final prediction, or further aggregated into a single value. The conventional single-model forecast assumes accurate measurements and model assumptions, and produces a discrete number as the output, giving a false sense of determinism and infallibility.

6.3.4 Reduce the chance and impact of inaccurate predictions, Portfolio Theory

Decisions based on inaccurate predictions can be very costly. Therefore a lower chance of having inaccurate predictions, combined with the ability to identify cases with potentially unreliable predictions can alleviate the impact of inaccurate model predictions. Both attributes are present in ensemble forecasting.

Ensemble models consider different possibilities in the data generation process, combining different model assumptions and data sources, therefore providing 'checks and balances' in cases where one or more of these assumptions went wrong. In applications of ensemble models in this research, we find that ensemble models can provide between 4% to 5% reduction in the mean square error (appendix C), and about 2% reduction in the standard deviation of absolute error, which shows a reduction in large errors, and a more stable model performance.

The previous section 6.3.1 focuses on the accuracy benefit of ensemble forecasting, which means model predictions are 'on average' closer to actual values. But there are cases where the occurrence of a single large forecast error will overshadow the improvement in forecast accuracy. In such cases, it would be desirable to focus on the frequency of occurrence, or the magnitude of large errors, even at the cost of some average forecast accuracy.

A lower chance of large errors is another desirable property of ensemble forecasts, especially in situations where large errors are costly. If the risk of prediction errors were to be measured with the 'hazard and exposure' scale, the hazard (chance) of a large error is low for models with good past performance, but large errors are almost certain to occur in a large enough number of predictions. Models built on assumptions will always produce large errors, and some models do it more often than others. Machine learning algorithms have good accuracy in general, but can produce colossal errors when they malfunction. Therefore applications relying on a single algorithm have a taste of playing Russian roulette, where the penalties for large errors can be significant.

Combining forecasts from different sources also avoids choosing, and depending on a single sources (Silver, 2012), so that the failure of a single forecast does not put the entire prediction at risk, even though the combined forecast does not necessarily have the highest accuracy (Graefe et al., 2014). The presence of other models provides 'checks and balances' in cases

when one of the models went wrong, just as the 'checks and balances' mechanism would normally prevent a single tyrant from taking over a democratic society. The conditional probability ensures that, given one model malfunctions (assuming independent models), the chances of multiple models malfunctioning simultaneously will be lower than any single one of the models, so combining forecasts reduces the risk of extremely bad predictions.

In Portfolio Theory, committing different proportions of assets into different investments translates into the choice between risk, and expected return; the proportions are determined in a way that maximizes return, and minimize the risk (Peters and Adamou, 2018). In transport modeling, the goal of combining different model outputs is to maximize forecast accuracy, while minimizing the chance of large errors.

Ensemble forecasting combining different models and data sources is reminiscent of the idea of the portfolio theory (Markowitz, 1991) in economics. The price of financial instruments (e.g. stocks) depends on both future returns, and are affected by uncertainties in the market, including irrational behavior (Shiller, 2005). Investors value both the 'risk' and 'return' on their investment; so instead of pouring everything into the stock with the best past performance, investors diversify their investment into different companies, and various financial instruments to spread the risk while maintaining an acceptable level of returns (Markowitz, 1991). The investors adhering to portfolio theory may not become rich overnight, but they are less likely to go bankrupt. The collapse of the investment bank Lehman Brothers is an example of the potential cost of the single-model doctrine, which, in chasing a significant amount of profit, had given up diversification and focused on a single type of mortgage backed security (Silver, 2012); the historically good track record of the mortgage market (at that time) was no indication that it would continue to perform well.

The analogy to the 'ensemble forecast' in the stock market, is the market itself. The market is so efficient that, it is almost impossible for individual investors to beat the average return of the stock market over a sustained period (Fama, 1960); yet the risk associated with individual stocks is diluted infinitely by the entire stock market, that the major risk is an overall economic downturn. From the ensemble forecast point of view, a rational investor should avoid managed funds, and simply rely on the collective wisdom of the market, and buy stock indexes.

In the transport context, research by Levinson and Zhu (2013) show that with real-world uncertainties in network conditions, there is generally no single route that is consistently better than others, so for individual travelers, choosing a portfolio of routes is a more rational choice than choosing a single route.

The term 'large errors', and 'average accuracy' are our modeling analogy to the portfolio theory's 'risk' and 'return'. Rational modelers should base their predictions on different models and data sources to reflect uncertainties, just as rational investors hold not just the best performing stock, but different stocks in their portfolios. Even when the average model accuracy is not improved by the ensemble forecast, the chances of large errors are much lower.

The point here, is that the ensemble forecast acts like investors pooling risk, while the single-model doctrine acts like gamblers. In practice, using ensemble forecasts improves accuracy in most cases, and in other cases it maintains the same level of accuracy as the best single model (Armstrong, 2001). The avoidance of large errors makes the ensemble forecasts more useful as decision support tools.

6.3.5 Dilution of error in long-term forecasts, the Chaos Theory

There is a tendency for model predictions to become less accurate when predicting events that are further into the future, which sometimes could render the model completely useless. The model becomes increasingly inaccurate as the time-space of model prediction increases. And this tendency applies to any type of model, and in predicting any kind of events; and the only difference is in the rate of how fast the model becomes useless for future events. This phenomenon is summarized by the Chaos theory (Lorenz, 1995, Thietart and Forgues, 1995), which states that,

- Small differences in the initial system state, if not measured precisely, can amplify over time; and
- For a given initial system state, small changes not captured by the model can accumulate over time, and make the system unpredictable.

Errors in transport models become larger, as the forecast time span increases (Armoogum, 2003). In iterative models where the output of the model becomes its input in the next cycle, single-model predictions accumulate errors in each cycle, even when the events can be completely and deterministically described by the model. The real-world can also be an iterative model (e.g. income-spending cycles), which drifts away from baseline conditions; so even non-iterative models can become increasing inaccurate over time.

In transport demand modeling, the conventional strategic planning (four-stage) model (Meyer and Miller, 2001) is a sequence of four different models that first makes a guess of the travel demand originating from each zone, then distributes that demand between zones, and among different models of transport; that demand is finally assigned to specific links in the road network. In the case of four-stage models, the measurement for the initial state (including the population numbers, or mode share in the base year) can include small variations, and the model assumptions can be faulty in different ways. In a study examining the propagation of the initial misrepresentations in the four-stage model inputs (Zhao and Kockelman, 2002), it is found that errors propagate and amplify across stages, and are counteracted by the final traffic assignment stage; so the prediction for travel demand between locations can be significantly impacted by the amplified measurement errors, but has less effect on the traffic forecast on each road segment. This counterbalance to error propagation is possibly due to the capacity limits of road segments. We might also think that the flow on each link is an ensemble of sorts of flows from each OD pair; route choice and demand are a negative feedback system, more congestion reduces demand, dampening error propagation.

Ensemble forecasting provides the opportunity to account for accumulated error, by actually incorporating uncertainties in the modeling process, including measurements with slight perturbations, and different assumptions (models) on the system mechanism to obtain the possible range of variations in the ensemble forecast. And because signals tend to amplify, and noises tend to cancel each other, the mode of ensemble forecasts tend to become more accurate than single-model predictions.

Ensemble models with slightly different initial conditions and model assumptions were first introduced in meteorology, which has greatly improved the weather forecast accuracy, and since became the standard practice in that field. Ensemble forecasts generally produce

more accurate longer-term predictions than single model predictions in iterative models (Zhu, 2005).

The relevance of ensemble forecasting in alleviating accumulated errors is apparent for transport applications. Modelers using single-model predictions, such as in Markov chains should be aware of the accumulated error over time; the adoption of ensemble models should be able to improve forecast accuracy in iterative or long-term transport models. What is not clear is the extent of improvement for transport models. Weather systems are generally non-linear, and errors would accumulate quickly over-time to render the model useless; the accumulation of error in transport models may not be as fast or significant as in weather models, so the benefit of adopting ensemble models for transport modeling would remain to be explored by future research.

6.3.6 Using Ensemble Models for Analysis

In this research, we focus mainly on the *forecast* role of ensemble models, but ensemble models can also be used for *analysis*, and with lower chance of erroneous conclusions. In Chapter 4 we calculated the sensitivity of the dependent to changes in explanatory variables, and all base and ensemble models returned the same positive sign for the sensitivity. Models in an ensemble can be made up of different assumptions, or different methods of pattern recognition - disagreement between models suggests some of the assumptions might be faulty, or data issues that needs further investigation, or even the event itself may be unpredictable.

On the other hand, consensus among different models in the ensemble can make a stronger case in analyzing relationships. Analysis using a single model may capture some 'accidental' features, and risk arriving at the wrong conclusion; results corroborated by multiple models are more reliable. Different models can 'vote' on the relationship of variables. Since real relationships are more persistent than accidental features, error by one model is more likely to be diluted and corrected by other models.

6.3.7 Computation Time

Computation time is a concern for many applications in modeling. Models that have good performance, but require extensive computation time may not be practical for use, and this is especially the case for real-time applications. Historically, limitations in computation power played a major role in delaying the progress of ensemble models in weather forecasting, where the first operational ensemble model (Palmer, 2019) came two decades after the initial identification of the problem in the 1960s (i.e. the Chaos theory) (Lorenz, 1995). To this day, weather forecasting still requires computer clusters to produce timely predictions. Although ensemble models are more complex, computation time of these ensemble models would not necessarily exceed that of base models.

Models are generally calibrated by computers. The single-thread performance on central processing unit (CPU) has stagnated (relatively speaking) for years after hitting the limits of fundamental physics; the quantum tunneling phenomenon (Tucker et al., 1994) also limits the downsizing of transistors, so heating problem still persists. Manufacturers compensated by adding more cores, which creates opportunities for improving computation time by parallelizing processes.

Ensemble models are ideal for parallel processing in computers. Base models can be calibrated independently and in parallel with other models (Chen et al., 2016); models with different assumptions or using different data can in theory be calibrated and applied to new data in parallel, and only combined toward the end of computation; taking subsamples, and training multiple ensemble models can also occur simultaneously, since each of this is an independent process. The computation time of the entire model is limited by the single base model with the longest computation time, plus the computation time in combining different base models; although an ensemble model would require more computation than any base model, this attribute limits the added time penalty, and the computation time will generally be comparable with single-model predictions. This means that ensemble models are scalable, and any increase in the number of base models, data sources, etc. would not significantly impact computation time.

To reach the same level of forecast accuracy, ensemble models might need less training data for each of the base models, and the combining of different models generally takes less time than training base models. Therefore it is possible for the ensemble models to have lower computation time than single models.

6.4 Disadvantages of Ensemble Forecasting

Ensemble forecasting has many benefits over the single-model doctrine. However, we do recognize that ensemble models have a number of deficiencies, that are mainly associated with inappropriate use and understanding of ensemble methods. A philosophical idea governing sciences has been that 'there is no free lunch'; something must be sacrificed in order to obtain something else. Ensemble models are able to outperform single-model forecasts in many respects, but that inevitably comes at some cost, such as increased model complexity, and lower interpretability. These costs and trade-offs are discussed in this section.

It shall be noted that ensemble forecasting does not absolve modelers from the burden of assessing models themselves, and at times, why the model doesn't work. The divergence of ensemble model predictions can be an indication that more work is need to improve models, or even the event itself is inherently unpredictable. It is still the responsibility of modelers to understand and improve models.

The idea behind ensemble forecasting is simple, yet the use of ensemble models is only recent, and ensemble methods have not been widely adopted in transport modeling. And further to this point, the existence of ensemble forecasting as an alternative to single-model forecasts is not widely known among modelers. So there appears to be some friction in the deployment of ensemble models. In this section we gather our experience with ensemble models, and provide explanation for the slow adoption of ensemble forecasting.

Data size requirement

A major deficiency with ensemble forecasting relates to the size of available training data. Both base models, and meta-learners require sufficient data to calibrate. In situations with a small training data set, the meta-learners may not be sufficiently calibrated to combine base model predictions, and the resulting ensemble forecast may become less accurate than the

best performing base model. This problem can sometimes be alleviated by choosing a robust meta-learner, such as a linear meta-learner.

In this research, ensemble models cannot reliably outperform the best single model when the size of training data is small. Therefore in applications that don't have access to sufficient data, there is little incentive to adopt more complicated ensemble models.

Data collection can be prohibitively expensive in some cases, or even physically impossible for some rare events. The field of weather forecasting is the birthplace of ensemble forecasting, and is in modern times characterized by rich, relatively cheap-to-collect atmospheric measurement data (Blum, 2019); models can be calibrated, and validated frequently by subsequent weather observations. So weather forecasting provides the ideal environment to experiment with, develop, and apply ensemble models. However, this is not the case for many other disciplines.

Many potential applications of ensemble forecasting suffer from insufficient data. Earthquake prediction is notorious for the difficulty in data collection, not only for the low frequency of large earthquakes, but also for the difficulty in getting measurement data before an earthquake. In the transport context, large scale travel surveys are very expensive, and dedicated travel surveys are rare; travel related information are sometimes collected together with national level census data, which limits the level of detail in each survey. It is hoped that new data collection methods, such as from personal smart devices, and more quality data will become available with technological developments, the data barrier in applying ensemble forecasting will be reduced, so that ensemble models will see wider adoption.

Increased complexity, manual efforts and computation cost

Although ensemble forecasting can outperform single-model predictions, ensemble models are invariably more complex than single-model predictions. Efforts will be required in matching specific modeling needs to the right type of ensemble method, and selecting base models; the performance of different ensemble methods will need to be compared, and whether the selected ensemble method produces desired outcome will also need to be tested and confirmed. In contrast, forecasts using a single model require fewer steps, and single-models would generally work out-of-the box.

In some cases, the extent of model performance improvement may not justify the adoption of ensemble models. The extent of accuracy improvement from ensemble models may be trivial in some cases. For applications that value modest accuracy improvements, the additional work, and a higher level of expertise required of ensemble models will likely deter some applications.

Interpretability

Ease of interpretation is a major drawback for ensemble forecasting; presenting and interpreting ensemble model outputs take more effort than using a single model. Model outputs resulting from different ensemble methods will have different meanings and interpretations. For example, an ensemble model using perturbed measurement data accounts for uncertainty from inaccurate measurements, but does not necessarily consider different assumptions on the data generation process. So not only will the explanation of model output become more

complex, the user of ensemble models will also need some technical knowledge in order to fully understand implications of ensemble forecasts.

The 'black box' problem with machine learning algorithms may continue into ensemble models. Many of the base models, and higher level models combining base models are machine learning algorithms, which are not very interpretable; these machine learning algorithms are sometimes referred to as 'black boxes'. The theoretical development for ensemble forecasting is still in its early stages, and more theoretical work would be needed to optimize the performance of ensemble models. Ensemble models produced by the combination of uninterpretable parts further make the ensemble model uninterpretable.

Ensemble forecasting raises the technical requirement for both modelers, and model end-users (e.g. decision makers, general public). Compared to single-models, it is more difficult to explain ensemble model outputs to a technical laymen, or even to people with some technical understanding of modeling. Without a proper understanding of ensemble forecasting, the true value within ensemble model predictions may not be appreciated. And yet ensemble models can produce more intuitive outputs than single-models; the graphic illustration of multiple possible paths and outcomes may be easier to perceive, many people seem to understand probabilities of potential hurricane trajectories when displayed on maps, for instance, so a full understanding of the underlying ensemble forecasting methods may not be required to interpret the output.

By being less interpretable, ensemble forecasting ventures deeper into the modeling black box. The complex and latent relationships between factors within ensemble models might be beyond the limits of human understanding. The history of applied science is rich with examples where techniques are used without understanding; a flashlight, for example, can be used without a full understanding of the nature of photons. Therefore we may recognize that we may not be able to understand everything. When models become more complex, eventually there will come a point where the level of complexity begins to defy understanding; hopefully this point has not been reached, because after this point, only philosophy will provide us with some guidance.

6.5 Applicability of Ensemble Forecasting

As with all methods, ensemble forecasting also has a limited area of applicability, and does not apply to all scenarios, or solve all modeling problems. Ensemble models are unable to account for factors not included in the model, or external shocks that change the course of events. This is exemplified by the 'unknown-unknowns' mentioned in section 1.7, that exogenous factors not included in the model formulation will not automatically be accounted for by ensemble models, and that long-range forecasts are intrinsically less accurate. However, ensemble models can include models that include different sets of explanatory variables, or with different formulation, so would have a higher chance for factors affecting the event outcome to have been included in one of the base models.

Ensemble forecasting is intended for applications with inherent uncertainties, that are otherwise difficult to account for. In completely deterministic processes and with accurate measurement data, such as in pure simulations, or in describing mechanical events, conventional physical models (single-model) will work better than ensemble models. Because

these events are completely deterministic at the scale of interest, therefore models can be judged as simply right or wrong, based on whether these models captures the underlying data generation mechanism.

Chapter 7

Conclusion

Ensemble forecasting, and the single-model doctrine represent two different schools of thoughts in modeling. Transport modeling generally adopts the single-model doctrine, which uses a single assumption for the data generation process, or method of pattern recognition (model). Although the single-model doctrine has a number of problems, the doctrine has well established methodology, and widely accepted standards by practitioners. This uniformity of ideas and practices facilitates communication among practitioners, and is both a blessing, and a curse. In the history of scientific revolutions, there are often different schools of thoughts that both compete with, and complement each other; practitioners are not bound by a single common belief, and were free to conduct independent experiments (Kuhn, 2012). The ensemble of different thoughts, and tolerance to different ideas and scientific beliefs, has been the source of scientific progress.

The track record of transport models under the single-model doctrine has not been very good, especially travel demand forecasts, which have poor accuracy, and are not improving over time (Boyce and Williams, 2015); many cases of inaccurate transport models with detrimental effects have been documented in the literature (Federal Reserve Bank, 2020, New Civil Engineer, 2018, Transport for NSW, 2012). In practices, very little attention has been paid to the effect of uncertainties in modeling (Rasouli and Timmermans, 2012). The single model procedure also produces a gap between model outputs, which are deterministic, and real-world events, which are, to the best of our knowledge, probabilistic.

In this research we present the case for using ensemble forecasting to improve transport modeling, which provides a different approach to modeling, and addresses many problems with the single-model doctrine. Ensemble forecasting is more of a paradigm for modeling than a specific modeling method, it enables modelers to view the real-world events (data generation processes) as having multiple possible paths and causes, recognizing some degree of uncertainty.

Events involving complex interactions and human behavior are inherently complex and chaotic, and not all variables can be observed or measured. This research points out the folly in the misuse of physical models for chaotic real-world events; modeling these complex events will therefore need to internalize uncertainties, which makes ensemble models intrinsically different from Newtonian physical models.

On the philosophical side, the idea of ensemble forecasting conforms with the Occam's Razor (Domingos, 1999), which favors simple solutions over more complex ones. In particu-

lar, the Occam's Razor favors models with simpler, and fewer assumptions, as these models are more likely. The elegance and simplicity of the single-model doctrine is highly valued in mathematics, and in Newtonian physics; ensemble models are more complex than single models, and include many base models, each containing a different set of assumptions. This dilemma that simple models not being as accurate as complex models is pointed out by Breiman et al. (1996), who suggested that the real data generation process is nature's black box, and complex models are merely imitating that nature's black box. To address this dilemma with Occam's Razor, we posit that in an ensemble of different models, different methods for combining models reduce reliance on a single model's assumption on the data generation process, or pattern recognition; and an ensemble of data sources reduces dependence on a single data. Ensemble of ensembles takes one step further in accounting uncertainties in ensemble methods themselves. Ensemble forecasting reduces the sensitivity of models to particular assumptions, so are in fact more elegant, and simpler than the single-model doctrine.

This research concludes that ensemble models generally have better performance than single models, and are well suited for applications in transport modeling. Based on our findings, we conclude that the better performance of ensemble models that are often described in the literature is mostly based on facts, and not due to the publication bias making null results less likely to get published (Easterbrook et al., 1991). However, there are nuances in the application of ensemble methods, including a requirement for sufficient training data, and the applicability to non-deterministic processes. These conclusions are based on findings from this research, and our current understanding, and experience with ensemble forecasting. This research establishes a basic theoretical framework for using ensemble forecasting, and shows some promising results with a number of applications of ensemble models in transport problems. We also discuss the caveats of this research, and potential future research directions.

7.1 Caveats, and Future Research

7.1.1 Greater number and variety of base models

In this research five models are used as base models. Although this small number of base models is able to demonstrate the case for ensemble models, there is likely a lack of variety in model assumptions and pattern recognition methods. Some of the ensemble models, namely the simple average and weighted average ensemble models, may require a larger variety of different models in order to function better.

While the performance of stacking ensemble models consistently exceed the best base models, performance of simple average and weighted average ensemble models are often below the best base model. There is a possibility that, by increasing the number and variety of base models, the performance of simple average and weighted average ensemble models will improve. For stacking ensemble models, however, a larger number of base models presents more uncertainties, and meta-classifiers will select models from a greater variety of options, which might confuse the meta-learners, and require a larger training data to calibrate ensemble models.

To our knowledge, there is currently no consensus on the required, or optimum number of base models for different ensemble methods. Future research may explore the link between the number and variety of base models, and the performance of ensemble models.

7.1.2 Base model from varying model formulations

Machine learning algorithms can have different model formulations. In linear regression, model formulations selected based on highest R^2 are often wrong (Bacon, 1977, Kennedy, 2003, Mayer, 1975); in the same light, machine learning formulations based on model performance may also be wrong.

Future research may test combining machine learning algorithms of the same type but with different formulations (e.g. neural network structure, learning rates, random starting points), or to combine different model formulations with different types of models.

7.1.3 Long-range forecasts using ensemble models

Ensemble models are used routinely in weather forecasting, and other disciplines, to alleviate accumulated errors over time. Long-range transport predictions face similar problems as in weather forecasting, where accumulated errors make model predictions less reliable, and less accurate over time. The efficacy of transport ensemble models in this respect has not been tested.

In section 4.4.2 we compared model performance between two different methods, namely (a) directly combining data, and (b) combining data with ensemble models, in which we find very little difference between these two methods. However, this application is not iterative, and does not involve the accumulation of error over time. It is possible that once errors accumulate over time, differences in model performance between these two methods will begin to materialize.

Future research may apply ensemble forecasting to long-range transport predictions to test its efficacy. The method by which transport models produce long-range predictions may undergo a significant reshuffle, if ensemble models can be proven to alleviate accumulated errors.

7.2 Discussion

The adoption of ensemble forecasts by various fields has been driven mostly by necessity. For instance, decisions based on economic and business models carry significant costs, if the predictions were off the mark by a large margin. Weather forecasts are worthless if not delivered on time, or without sufficient accuracy; there is public pressure for weather forecasters to improve accuracy, and the ensemble model's presentation of weather forecasts as probabilities makes the prediction more useful. Valuable lessons have been learned from the application of ensemble forecasting in these fields. Other disciplines can take in these hard learned lessons and experience, instead of developing models for their fields from scratch.

Most transport-related cases seem to be predictable, as least in the short and mid-term range. Historically there had been doubts as to whether the political events, earthquakes,

and even the weather can ever be predicted. Experience has shown that with sufficient signal from measurements, and a fine-tuned modeling method, the prediction of weather is practical, and the ensemble forecast played a crucial role. Some early failures and stumbles along the way are to be expected, before a fully functioning modeling procedure can be established (Blum, 2019). Historically 'unpredictable' events all include fatal deficiencies in either the measurement data or inadequate validation data; for example, there is insufficient geological measurement data for earthquakes.

Transport forecasts generally have access to sufficient measurement data, and historical data for model calibration and validation, and there are patterns in the behavioral tendency of human beings. Hence the prediction of transport-related events should be intrinsically solvable. However, as a field dealing with human behavior, it may be inherently more difficult that physically-dominated phenomena like weather and earthquakes. The trail of inaccurate transport predictions highlights the current single-model transport modeling doctrine as a bottleneck.

Modeling practices are highly siloed across different disciplines, and people who use forecast models don't often look beyond their disciplinary boundaries for the best practices. The single-model doctrine is deeply rooted in transport modeling. To fully utilize ensemble forecasts in transport would require a more widespread understanding of its theoretical basis, applicability, and methods of implementation in the transport context. This type of understanding is lacking among transport professionals, and the theory and methodology of ensemble modeling is not yet formalized for the transport field. This research fires the opening shot at formalizing the theory and methodology of ensemble forecast in the transport field, and suggests that transport forecasts should catch up with the development in other fields, and embrace the idea of ensemble forecasting.

There is difference in terminology between disciplines, so what is meant by "ensemble" might differ significantly, depending on elements constituting an ensemble. Basic forms of ensemble models, such as the Random Forest, may themselves be termed "ensemble" models, therefore what this research refers to as "ensemble" models, might be termed "ensemble of ensembles" in some other context (Ghosh et al., 2020). These differences in terminology reinforces existing disciplinary barriers, and this research calls for better clarity in ensemble methods in future research, and to avoid the use of vaguely defined terms in titles and abstracts.

We foresee a number of issues in the future adoption of ensemble models, which are discussed in this section.

Adoption of ensemble forecasting by the industry

The wider adoption of ensemble forecasting will likely take an extended period of time. In practice, various constraints ranging from modeling complexity to computation cost will likely hinder the adoption of ensemble forecasting in transport modeling. The acceptance of ensemble forecasting will likely take time, and some push back from the status quo is to be expected.

As a first step, ensemble models can be use in parallel with current conventional modeling approaches. Ensemble models can take advantage of more data sources than conventionally models, so the collection of more data is necessary in order to take the full advantage of

ensemble models. In actual use cases, a stringent peer-review process comparing the performance of ensemble model against conventional models will be needed to identify whether, and where ensemble models have a clear advantage over conventional models. Because transport models can have a relatively long time span (several decades) compared to ensemble models in other disciplines, iterations and improvements to transport ensemble models will also take longer, and the review of model performance may be performed by a different generation of modeler from the ones that initially designed the models; careful documentation of ensemble model procedures will also be needed.

Adoption of ensemble forecasting in general

The adoption of the ensemble approach in transport has taken longer than one might expect. If history provides us with any guidance on this matter, it is that useful new concepts and ideas generally take a long time to enter common knowledge, and to reach wide adoption. For instance, the concept of probability, which is considered common knowledge today, didn't formally enter modern science until the 19^{th} century (Hájek, 2002). And knowledge does not become common knowledge equally for everyone.

While it is not a requirement for the general public or decision makers to use ensemble models, some understanding in the idea of ensemble forecasting will help them better interpret the results, and make more informed decisions.

Appendix A

Dummy Example - Compare Averaging Data with Averaging Models

A.1 The Average Trap

In Table A.1 we show a simplified example of how in non-linear systems, averaging multiple measurements can produce different results, than averaging the model outputs that uses different measurements. Assuming that x represents one measurement for an explanatory variable, and $f(x)$, $g(x)$, $h(x)$ are three separate data generation processes; $f(x)$ is super-linear, $g(x)$ is linear, and $h(x)$ is sub-linear. For ease of understanding, we assume the data generation process to be the monthly investment return on x amount of investment. To predict the average monthly return over three months, averaging the three x will either over or under-estimate the investment return, depending on whether the system is sub or super-linear. The idea is similar to the summation in calculus.

The example shows that averaging measurements does not equate averaging how the different measurements can act on the dependent variable through the model; despite its simple concept, the risk associated with averaging measurements is not well perceived or even mentioned.

Measurement	Measurement (x)	Super-linear $f(x) = x^3$	Linear $g(x) = 3x$	Sub-linear $h(x) = \sqrt{x}$
First	6	216	18	$\sqrt{6}$
Second	7	343	21	$\sqrt{7}$
Third	8	512	24	$\sqrt{8}$
Average of Measurements	7			
Combine data by averaging measurements		343	21	2.646
Combine data by averaging models (ensemble)		357	21	2.641

Table A.1: Comparing the difference between averaging the measurements, and averaging model outputs in a non-linear system

Appendix B

FHV Trip Attraction - Model Performance in Predicting Trip Attraction

B.1 Chicago

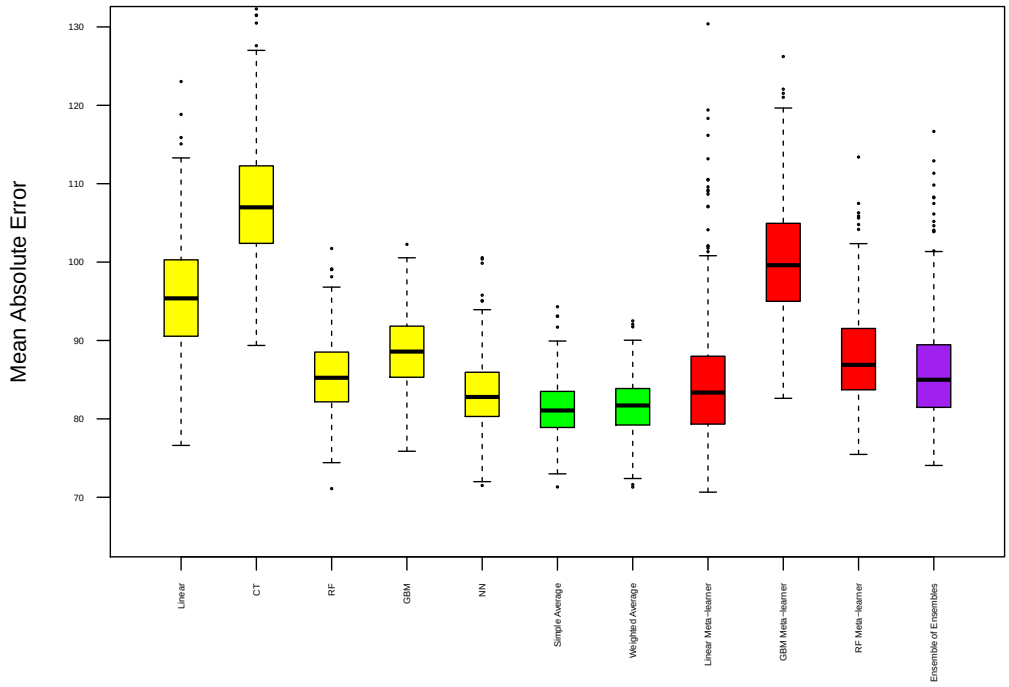


Figure B.1: Model performance in predicting trip attraction in Chicago. Distribution of mean absolute error from 400 experiments. Boxes span 25th to 75th percentile, whiskers extend to max/min of the values, excluding outliers (1.5 box heights away from box edges, which are shown as dots)

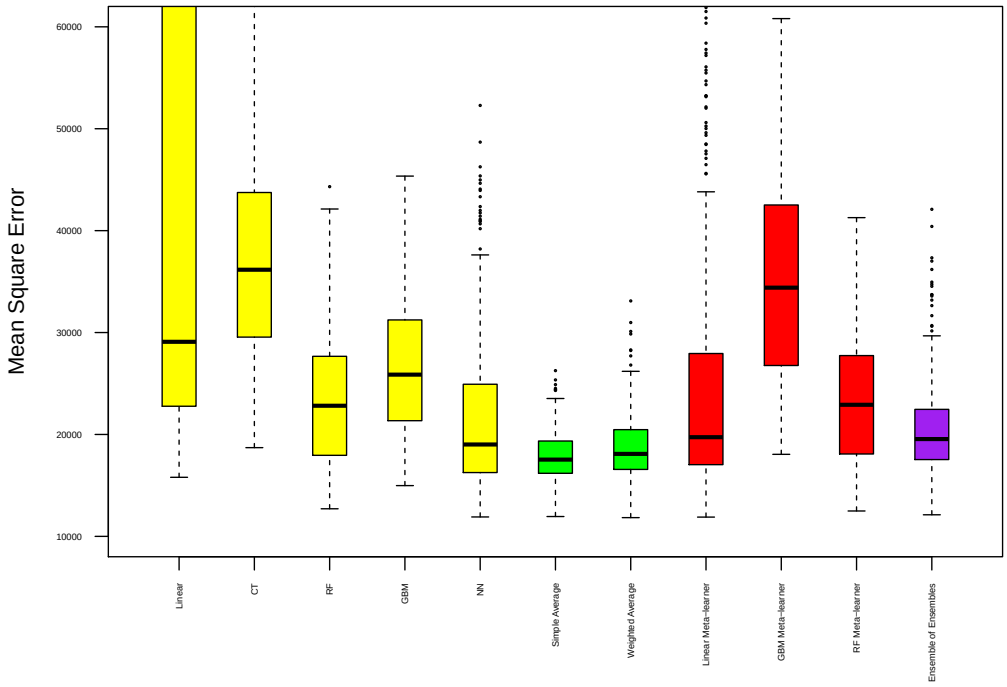


Figure B.2: Model performance in predicting trip attraction in Chicago. Distribution of mean square error from 400 experiments. Boxes span 25th to 75th percentile, whiskers extend to max/min of the values, excluding outliers (1.5 box heights away from box edges, which are shown as dots)

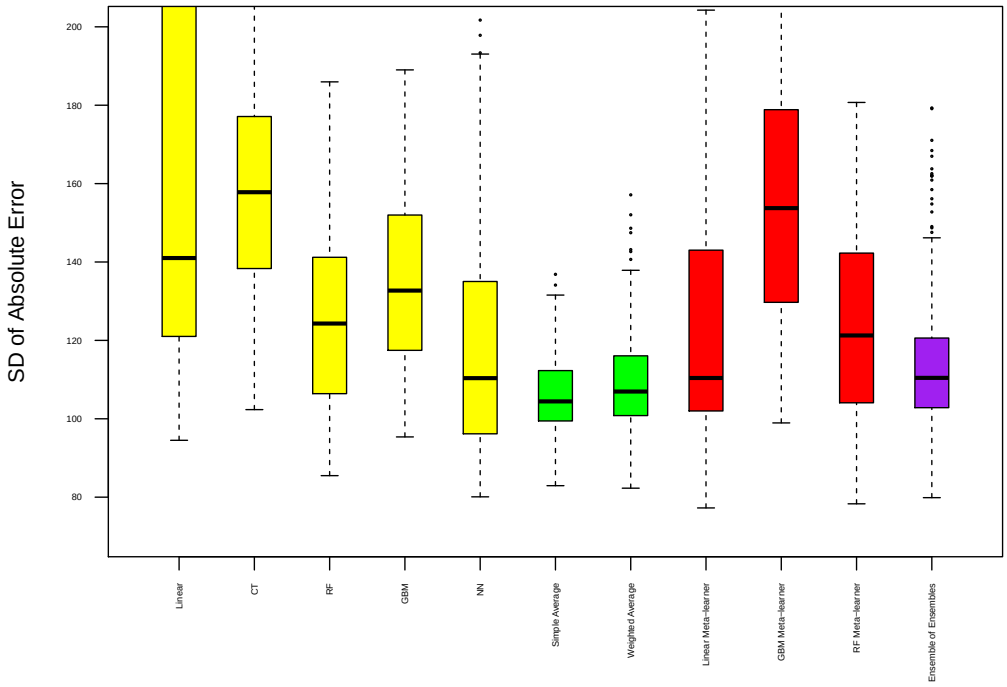


Figure B.3: Model performance in predicting trip attraction in Chicago. Distribution of standard deviation of absolute error from 400 experiments. Boxes span 25th to 75th percentile, whiskers extend to max/min of the values, excluding outliers (1.5 box heights away from box edges, which are shown as dots)

B.2 New York City

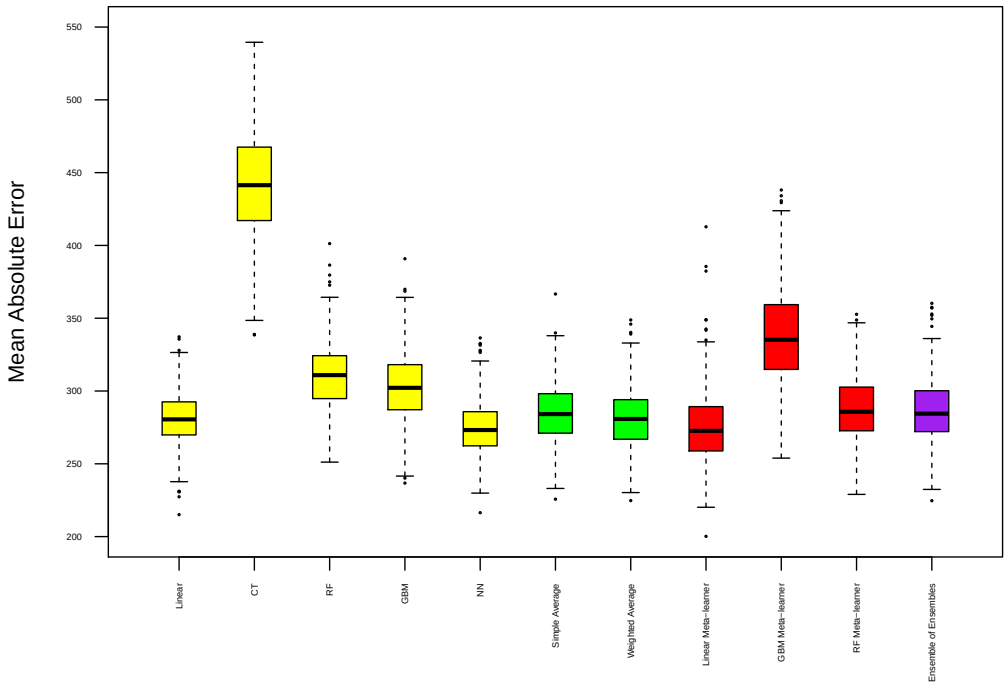


Figure B.4: Model performance in predicting trip attraction in NYC. Distribution of mean absolute error from 400 experiments. Boxes span 25th to 75th percentile, whiskers extend to max/min of the values, excluding outliers (1.5 box heights away from box edges, which are shown as dots); all 5 base models

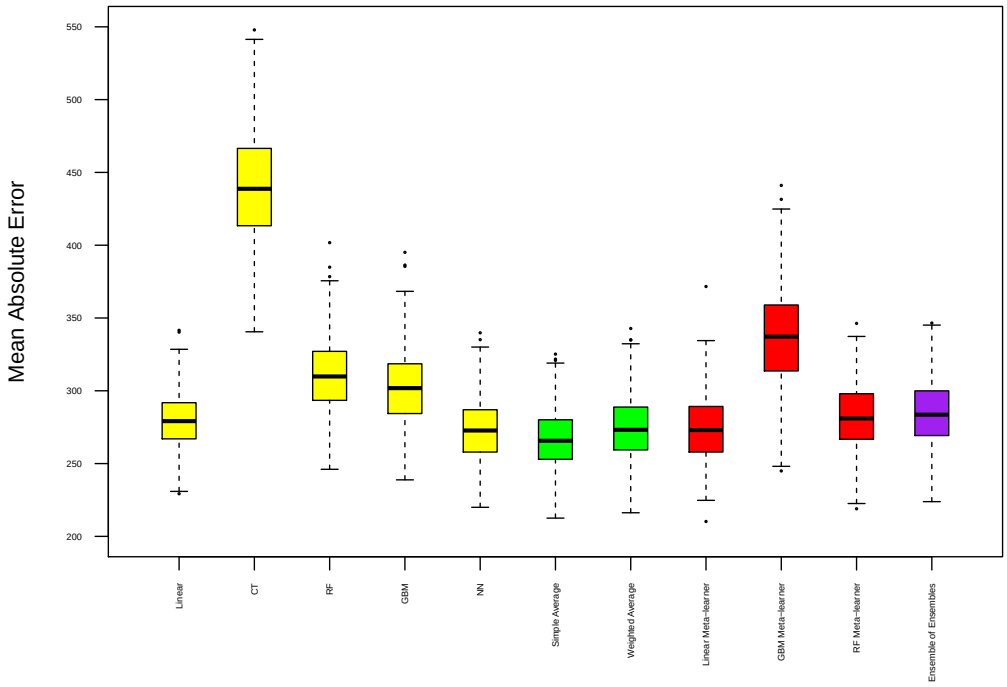


Figure B.5: Model performance in predicting trip attraction in NYC. Distribution of mean absolute error from 400 experiments. Boxes span 25th to 75th percentile, whiskers extend to max/min of the values, excluding outliers (1.5 box heights away from box edges, which are shown as dots); base models exclude classification tree

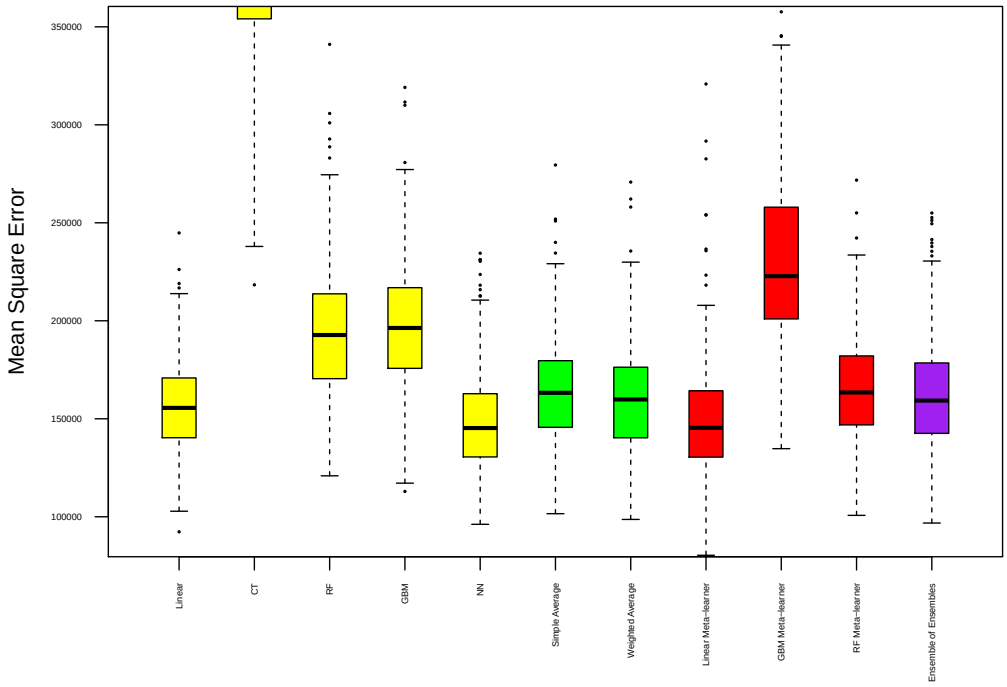


Figure B.6: Model performance in predicting trip attraction in NYC. Distribution of mean square error from 400 experiments. Boxes span 25th to 75th percentile, whiskers extend to max/min of the values, excluding outliers (1.5 box heights away from box edges, which are shown as dots); all 5 base models

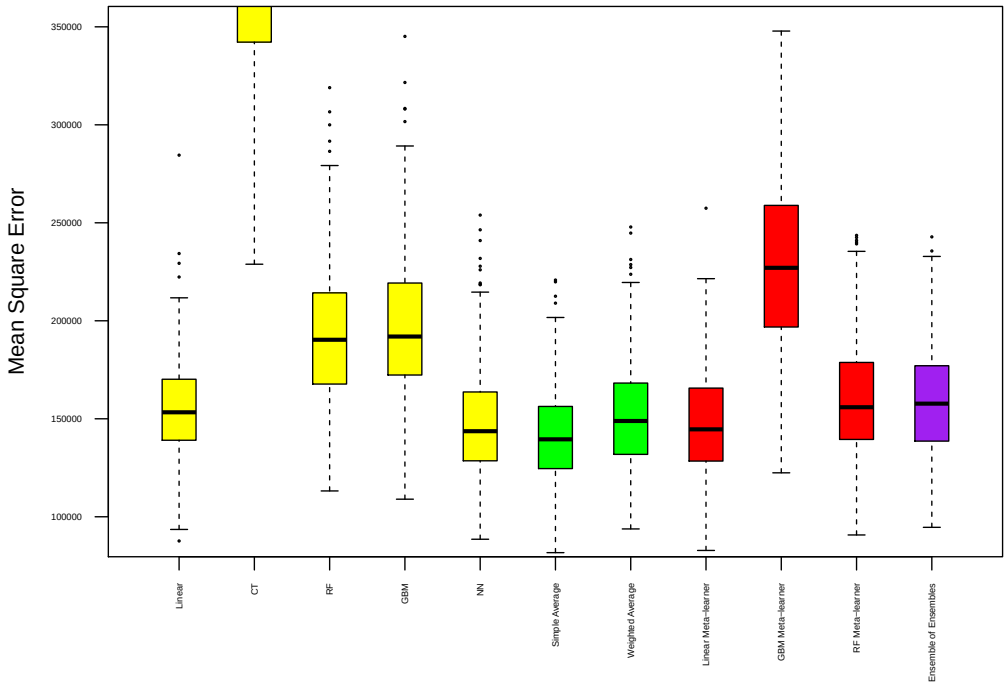


Figure B.7: Model performance in predicting trip attraction in NYC. Distribution of mean square error from 400 experiments. Boxes span 25th to 75th percentile, whiskers extend to max/min of the values, excluding outliers (1.5 box heights away from box edges, which are shown as dots); base models exclude classification tree

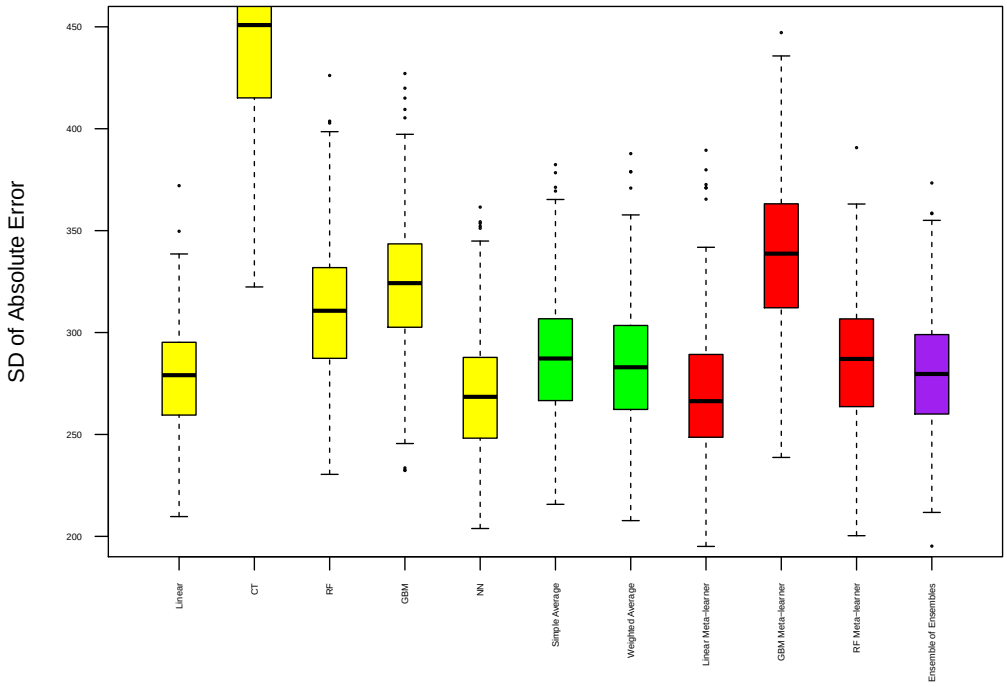


Figure B.8: Model performance in predicting trip attraction in NYC. Distribution of SD of absolute error from 400 experiments. Boxes span 25th to 75th percentile, whiskers extend to max/min of the values, excluding outliers (1.5 box heights away from box edges, which are shown as dots); all 5 base models

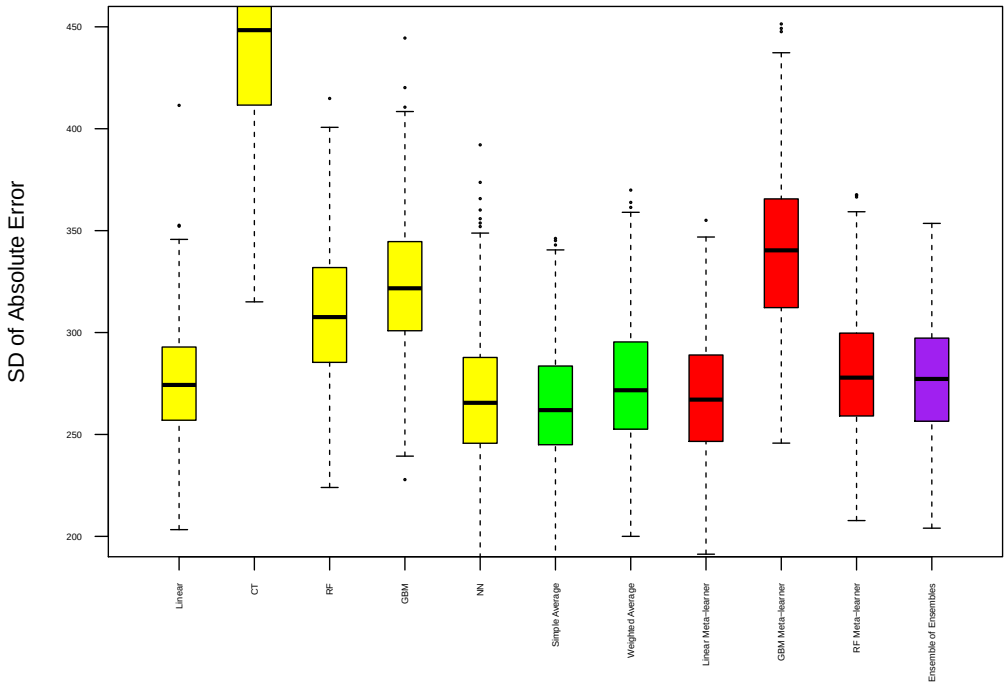
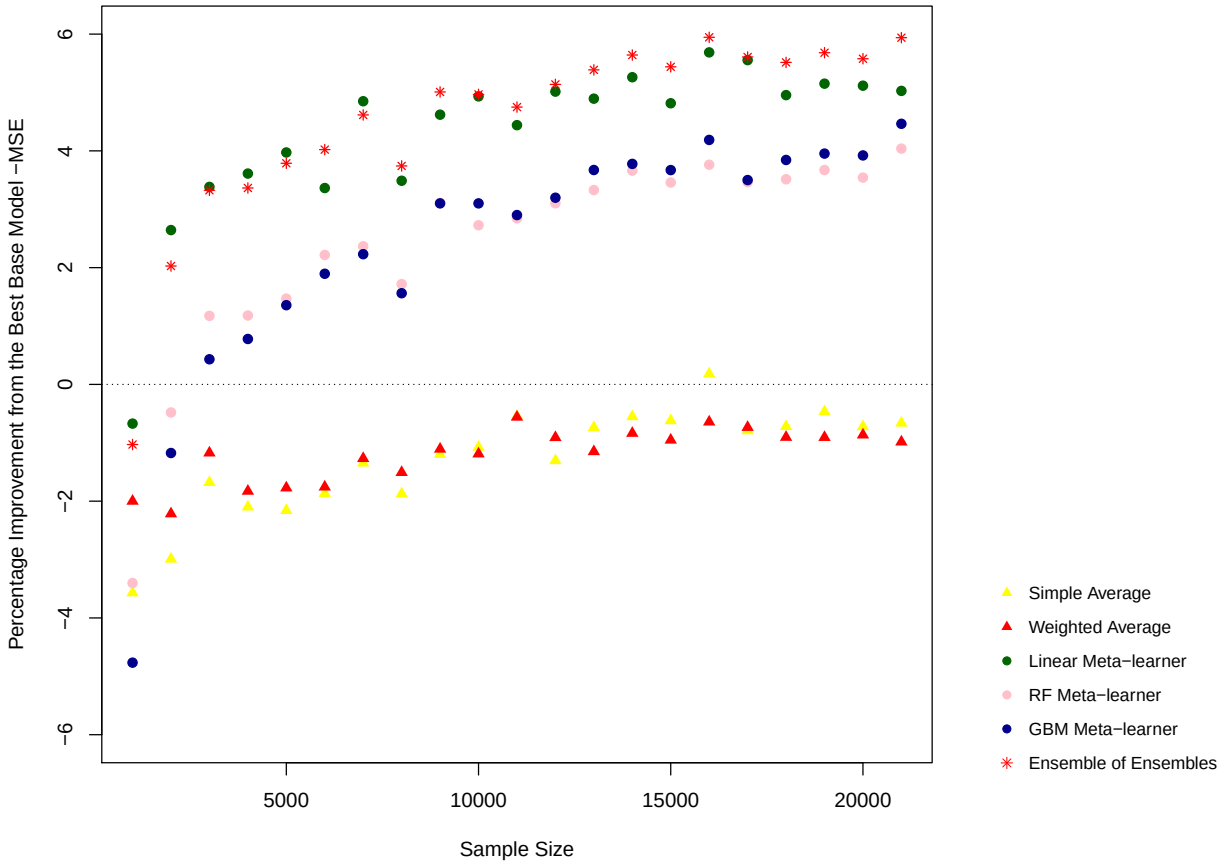


Figure B.9: Model performance in predicting trip attraction in NYC. Distribution of SD of absolute error from 400 experiments. Boxes span 25th to 75th percentile, whiskers extend to max/min of the values, excluding outliers (1.5 box heights away from box edges, which are shown as dots; base models exclude classification tree

Appendix C

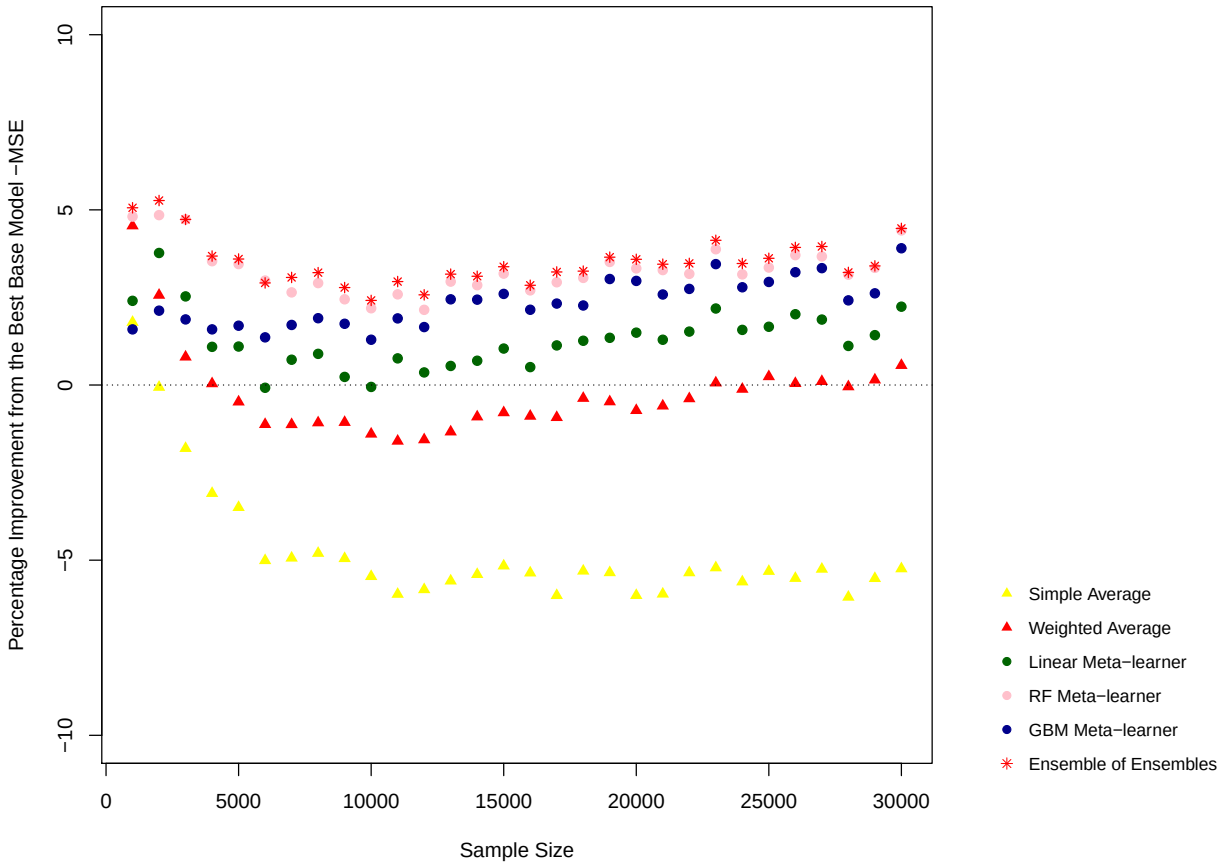
Performance Improvement from Ensemble Models

C.1 Mean Square Error



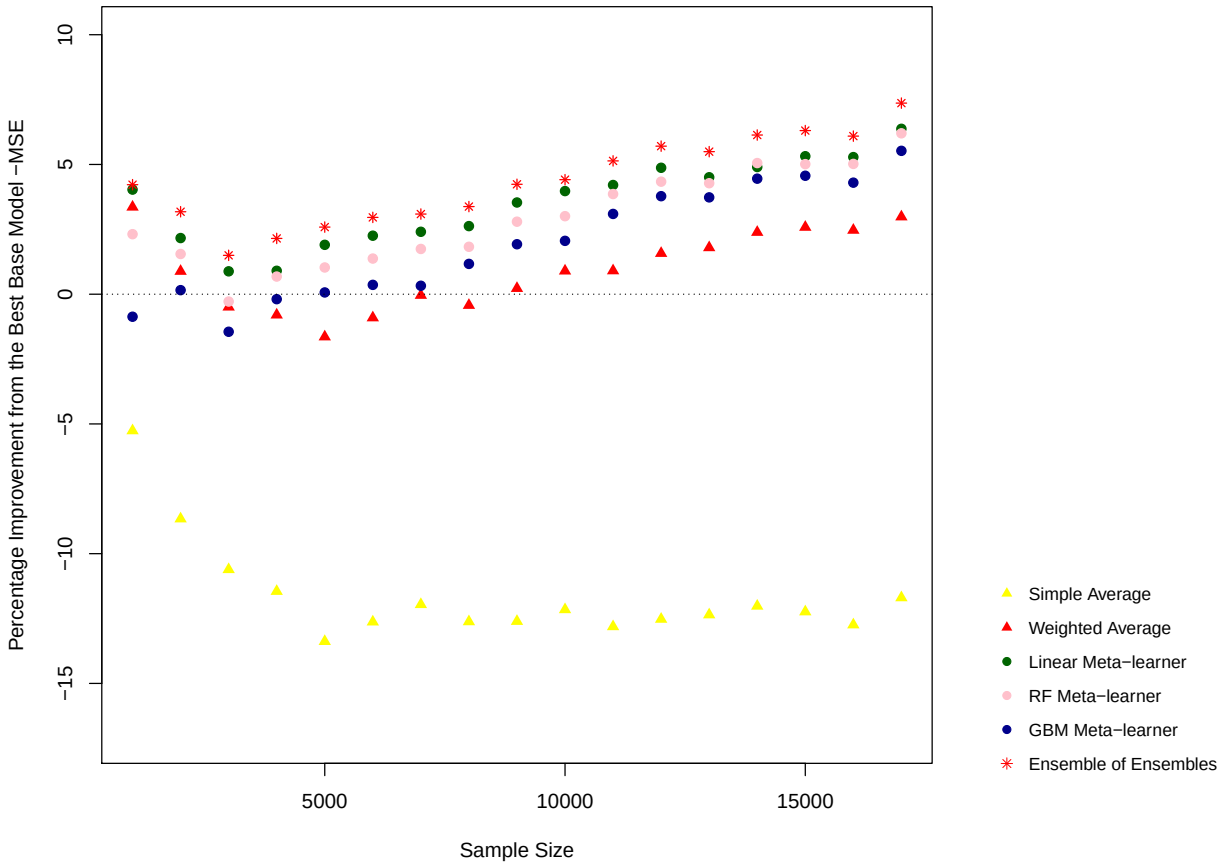
(a) Percentage Improvement in MSE (b) Models

Figure C.1: Performance improvement from the best base model - Sydney Hedonic MSE



(a) Percentage Improvement in MSE (b) Models

Figure C.2: Performance improvement from the best base model - Chicago MSE



(a) Percentage Improvement in MSE (b) Models

Figure C.3: Performance improvement from the best base model - NYC MSE

Bibliography

Alonso, W. et al. (1964). *Location and land use*. Harvard University Press Cambridge, MA.

Anderson, J. L. (1996). A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *Journal of Climate 9*(7), 1518–1530.

Armoogum, J. (2003). Measuring the impact of uncertainty in travel demand modelling with a demographic approach. In *Proceedings of the European Transport Conference (ETC), October 2003, Strasbourg, France*.

Armstrong, J. S. (2001). Combining forecasts. In *Principles of Forecasting*, pp. 417–439. Springer.

Ashton, A. H. and R. H. Ashton (1985). Aggregating subjective forecasts: Some empirical results. *Management Science 31*(12), 1499–1508.

Australian Bureau of Statistics (2016). Working population profile (WPP).

Australian Bureau of Statistics (2017). Sydney transit mode share 2017.

Australian Property Monitors (2019). Residential property transaction data 2017 - 2019.

Bacon, R. W. (1977). Some evidence on the largest squared correlation coefficient from several samples. *Econometrica: Journal of the Econometric Society*, 1997–2001.

Batchelor, R. and P. Dua (1995). Forecaster diversity and the benefits of combining forecasts. *Management Science 41*(1), 68–75.

Bates, J. M. and C. W. Granger (1969). The combination of forecasts. *Journal of the Operational Research Society 20*(4), 451–468.

Ben-Akiva, M. and T. Watanatada (1981). Application of a continuous spatial choice logit model. *Structural Analysis of Discrete Data with Econometric Applications*, 320–343.

Ben-Akiva, M. E., S. R. Lerman, and S. R. Lerman (1985). *Discrete choice analysis: theory and application to travel demand*, Volume 9. MIT Press.

Blum, A. (2019). The weather machine: A journey inside the forecast. Ecco.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association 71*(356), 791–799.

Boyce, D. E. and H. Williams (2015). *Forecasting urban travel: Past, present and future*. Edward Elgar Publishing.

Breiman, L. (1996). Stacked regressions. *Machine Learning 24*(1), 49–64.

Breiman, L. et al. (1996). Heuristics of instability and stabilization in model selection. *Annals of Statistics 24*(6), 2350–2383.

Breiman, L. et al. (2001). Statistical modeling: The two cultures. *Statistical Science 16*(3), 199–231.

Brigham, E. F. (1965). The determinants of residential land values. *Land Economics 41*(4), 325–334.

Chand, N., P. Mishra, C. R. Krishna, E. S. Pilli, and M. C. Govil (2016). A comparative analysis of SVM and its stacking with other classification algorithm for intrusion detection. In *2016 International Conference on Advances in Computing, Communication, & Automation*, pp. 1–6. IEEE.

Chau, K. W. and T. Chin (2003). A critical review of literature on the hedonic price model. *International Journal for Housing Science and Its Applications 27*(2), 145–165.

Chen, J., K. Li, Z. Tang, K. Bilal, S. Yu, C. Weng, and K. Li (2016). A parallel random forest algorithm for big data in a spark cloud computing environment. *IEEE Transactions on Parallel and Distributed Systems 28*(4), 919–933.

Chen, X. M., M. Zahiri, and S. Zhang (2017). Understanding ridesplitting behavior of on-demand ride services: An ensemble learning approach. *Transportation Research Part C: Emerging Technologies 76*, 51–70.

Cheng, L., X. Lai, X. Chen, S. Yang, J. De Vos, and F. Witlox (2019). Applying an ensemble-based model to travel choice behavior in travel demand forecasting under uncertainties. *Transportation Letters*, 1–11.

Chicago Data Portal (2019). *Transportation network providers - Trips*. City of Chicago.

Christensen, H. M., I. Moroz, and T. Palmer (2015). Stochastic and perturbed parameter representations of model uncertainty in convection parameterization. *Journal of the Atmospheric Sciences 72*(6), 2525–2544.

Compass (2019). *Australian roads speed data, Nov. 11 - Nov. 25, 2019*. Compass IoT Pty Limited.

Conway, M. W., D. Salon, and D. A. King (2018). Trends in taxi use and the advent of ridehailing, 1995–2017: Evidence from the US national household travel survey. *Urban Science 2*(3), 79.

Cowgill, B., F. Dell'Acqua, S. Deng, D. Hsu, N. Verma, and A. Chaintreau (2020). Biased programmers? or biased data? A field experiment in operationalizing AI ethics. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pp. 679–681.

Cui, M. and D. Levinson (2021). Shortest paths, travel costs, and traffic. *Environment and Planning B: Urban Analytics and City Science 48*(4), 828–844.

Dalrymple, D. J. (1987). Sales forecasting practices: Results from a United States survey. *International journal of Forecasting 3*(3-4), 379–391.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist 34*(7), 571.

De Jong, G., A. Daly, M. Pieters, S. Miller, R. Plasmeijer, and F. Hofman (2007). Uncertainty in traffic forecasts: Literature review and new results for the netherlands. *Transportation 34*(4), 375–395.

Delen, D., L. Tomak, K. Topuz, and E. Eryarsoy (2017). Investigating injury severity risk factors in automobile crashes with predictive analytics and sensitivity analysis methods. *Journal of Transport & Health 4*, 118–131.

Delle Monache, L., F. A. Eckel, D. L. Rife, B. Nagarajan, and K. Searight (2013). Probabilistic weather prediction with an analog ensemble. *Monthly Weather Review 141*(10), 3498–3516.

Department of Health (2019). Commonwealth hospital declaration.

Domingos, P. (1999). The role of Occam's razor in knowledge discovery. *Data Mining and Knowledge Discovery 3*(4), 409–425.

Easterbrook, P. J., R. Gopalan, J. Berlin, and D. R. Matthews (1991). Publication bias in clinical research. *The Lancet 337*(8746), 867–872.

ECMWF (2013). *The ensemble prediction system.* European Centre for Medium-Range Weather Forecasts.

Elliott, G. (2011). Averaging and the optimal combination of forecasts. *Manuscript, Department of Economics, UCSD.*

Erhardt, G. D., S. Roy, D. Cooper, B. Sana, M. Chen, and J. Castiglione (2019). Do transportation network companies decrease or increase congestion? *Science Advances 5*(5), 2670.

Espey, M. and H. Lopez (2000). The impact of airport noise and proximity on residential property values. *Growth and Change 31*(3), 408–419.

Fama, E. F. (1960). *Efficient market hypothesis.* Ph. D. thesis, University of Chicago, Graduate School of Business.

Fay, D. and J. V. Ringwood (2010). On the influence of weather forecast errors in short-term load forecasting models. *IEEE Transactions on Power Systems 25*(3), 1751–1758.

Federal Highway Administration (2017). 2017 National household travel survey (NHTS).

Federal Reserve Bank (2020). Moving 12-month total vehicle miles traveled.

FHWA (2002). Status of the nation's highways, bridges, and transit: Conditions and performance. Technical report, US Department of Transportation.

Freund, Y., R. E. Schapire, Y. Singer, and M. K. Warmuth (1997). Using and combining predictors that specialize. In *Proceedings of the Twenty-ninth Annual ACM Symposium on Theory of Computing*, pp. 334–343.

Garrison, W. L. and D. M. Levinson (2014). *The transportation experience: policy, planning, and deployment*. Oxford University Press.

Geoscape (2020). Geoscape data for buildings in Sydney.

Ghasemian, A., H. Hosseinmardi, A. Galstyan, E. M. Airoldi, and A. Clauset (2020). Stacking models for nearly optimal link prediction in complex networks. *Proceedings of the National Academy of Sciences*.

Ghosh, I., M. K. Sanyal, and R. Jana (2020). An ensemble of ensembles framework for predictive analytics of commodity market. In *2020 4th International Conference on Computational Intelligence and Networks (CINE)*, pp. 1–6. IEEE.

Gneiting, T., A. Raftery, F. Balabdaoui, and A. Westveld (2004). Verifying probabilistic forecasts: Calibration and sharpness. In *Preprints, 17th Conf. on Probability and Statistics in the Atmospheric Sciences, Seattle, WA, Amer. Meteor. Soc*, Volume 2.

Gneiting, T., A. E. Raftery, A. H. Westveld III, and T. Goldman (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review 133*(5), 1098–1118.

Graefe, A., J. S. Armstrong, R. J. Jones Jr, and A. G. Cuzán (2014). Combining forecasts: An application to elections. *International Journal of Forecasting 30*(1), 43–54.

Graefe, A., H. Küchenhoff, V. Stierle, and B. Riedl (2015). Limitations of ensemble Bayesian model averaging for forecasting social science problems. *International Journal of Forecasting 31*(3), 943–951.

Gupta, O. and R. Raskar (2018). Distributed learning of deep neural network over multiple agents. *Journal of Network and Computer Applications 116*, 1–8.

Gustafsson, N. (2002). Statistical issues in weather forecasting. *Scandinavian Journal of Statistics 29*(2), 219–239.

Haider, M. (2019). Diminishing returns to density and public transit. *Findings*.

Hájek, A. (2002). Interpretations of probability.

Hamill, T. M., J. S. Whitaker, and X. Wei (2004). Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Monthly Weather Review 132*(6), 1434–1447.

Hawkins, J. (2021). *A thousand brains: A new theory of intelligence*. Basic Books.

Heisenberg, W. (1985). Über den anschaulichen inhalt der quantentheoretischen kinematik und mechanik. In *Original Scientific Papers Wissenschaftliche Originalarbeiten*, pp. 478–504. Springer.

Hoehn, J. P., M. C. Berger, and G. C. Blomquist (1987). A hedonic model of interregional wages, rents, and amenity values. *Journal of Regional Science 27*(4), 605–620.

Hong, L. and S. E. Page (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences 101*(46), 16385–16389.

Hoque, J. M., G. Erhardt, and D. e. a. Schmitt (2021). The changing accuracy of traffic forecasts. *Transportation 1*.

Ji, A. and D. Levinson (2020). Injury severity prediction from two-vehicle crash mechanisms with machine learning and ensemble models. *IEEE Open Journal of Intelligent Transportation Systems 1*, 217–226.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kang, H. (1986). Unstable weights in the combination of forecasts. *Management Science 32*(6), 683–695.

Kaplan, R. S. (2009). Conceptual foundations of the balanced scorecard. *Handbooks of Management Accounting Research 3*, 1253–1269.

Kaplan, R. S. and D. P. Norton (1998). Putting the balanced scorecard to work. *The Economic Impact of Knowledge 27*(4), 315–324.

Karlaftis, M. G. and E. I. Vlahogianni (2011). Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. *Transportation Research Part C: Emerging Technologies 19*(3), 387–399.

Kennedy, P. (2003). *A guide to econometrics*. MIT Press.

Konečnỳ, J., B. McMahan, and D. Ramage (2015). Federated optimization: Distributed optimization beyond the datacenter. *arXiv preprint arXiv:1511.03575*.

Kronvang, B., H. Behrendt, H. E. Andersen, B. Arheimer, A. Barr, S. Borgvang, F. Bouraoui, K. Granlund, B. Grizzetti, P. Groenendijk, et al. (2009). Ensemble modelling of nutrient loads and nutrient load partitioning in 17 European catchments. *Journal of Environmental Monitoring 11*(3), 572–583.

Kuhn, T. S. (2012). *The structure of scientific revolutions*. University of Chicago Press.

Lammers, J., J. Crusius, and A. Gast (2020). Correcting misperceptions of exponential coronavirus growth increases support for social distancing. *Proceedings of the National Academy of Sciences 117*(28), 16264–16266.

Landefeld, J. S. and E. P. Seskin (1986). A comparison of anticipatory surveys and econometric models in forecasting US business investment. *Journal of Economic and Social Measurement 14*(1), 77–85.

Larrick, R. P. and J. B. Soll (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science 52*(1), 111–127.

Layton, D. F. and S. T. Lee (2006). Embracing model uncertainty: Strategies for response pooling and model averaging. *Environmental and Resource Economics 34*(1), 51–85.

Legendre, A. (1805). Nouvelles methodes pour la determination des orbites des cometes courcier paris. *Appendice Sur la Mehode des Moindres Quarre's*, 72–80.

Leutbecher, M. and T. N. Palmer (2008). Ensemble forecasting. *Journal of Computational Physics 227*(7), 3515–3539.

Levinson, D., H. Wu, B. Lahoorpoor, H. Rayaprolu, R. Kohan, and B. Haddock (2020). *Liverpool Sustainable Urban Mobility Study*. Transportlab, University of Sydney.

Levinson, D. and S. Zhu (2013). A portfolio theory of route choice. *Transportation Research Part C: Emerging Technologies 35*, 232–243.

Li, L., X. Chen, and L. Zhang (2014). Multimodel ensemble for freeway traffic state estimations. *IEEE Transactions on Intelligent Transportation Systems 15*(3), 1323–1336.

Linstone, H. A., M. Turoff, et al. (1975). *The Delphi method*. Addison-Wesley Reading, MA.

Liu, Y., C. Lyu, A. Khadka, W. Zhang, and Z. Liu (2019). Spatio-temporal ensemble method for car-hailing demand prediction. *IEEE Transactions on Intelligent Transportation Systems*.

Lobo, G. J. (1992). Analysis and comparison of financial analysts', time series, and combined forecasts of annual earnings. *Journal of Business Research 24*(3), 269–280.

Lorenz, E. (2000). The butterfly effect. *World Scientific Series on Nonlinear Science Series A 39*, 91–94.

Lorenz, E. N. (1995). *The essence of chaos*. University of Washington press.

Lucas, P. and L. Van Der Gaag (1991). Principles of expert systems.

Ma, Z., J. Xing, M. Mesbah, and L. Ferreira (2014). Predicting short-term bus passenger demand using a pattern hybrid approach. *Transportation Research Part C: Emerging Technologies 39*, 148–163.

Makridakis, S., E. Spiliotis, and V. Assimakopoulos (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PloS One 13*(3), e0194889.

Mankiw, N. G. (2014). *Principles of macroeconomics*. Cengage Learning.

Margenau, H. (1967). Quantum mechanics, free will, and determinism. *The Journal of Philosophy 64*(21), 714–725.

Markowitz, H. M. (1991). Foundations of portfolio theory. *The Journal of Finance 46*(2), 469–477.

Mayer, T. (1975). Selecting economic hypotheses by goodness of fit. *The Economic Journal 85*(340), 877–883.

McCullagh, P. and J. Nelder (1989). Generalized linear models.

McFadden, D. (1974). The measurement of urban travel demand. *Journal of Public Economics 3*(4), 303–328.

McFadden, D. et al. (1973). Conditional logit analysis of qualitative choice behavior.

McNees, S. K. (1990). The role of judgment in macroeconomic forecasting accuracy. *International Journal of Forecasting 6*(3), 287–299.

Meehl, P. E. (1954). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.

Meier, P. (1953). Variance of a weighted mean. *Biometrics 9*(1), 59–73.

Meyer, M. D. and E. J. Miller (2001). Transportation planning: A decision-oriented approach. *McGraw Hill*.

Mohring, H. (1961). Land values and the measurement of highway benefits. *Journal of Political Economy 69*(3), 236–249.

Mulley, C. and C.-H. Tsai (2017). Impact of bus rapid transit on housing price and accessibility changes in Sydney: A repeat sales approach. *International Journal of Sustainable Transportation 11*(1), 3–10.

Mulley, C., C.-H. P. Tsai, and L. Ma (2018). Does residential property price benefit from light rail in Sydney? *Research in Transportation Economics 67*, 3–10.

Murray, S. A. (2018). The importance of ensemble techniques for operational space weather forecasting. *Space Weather 16*(7), 777–783.

NCAR (2020). *Analog ensemble*. National Center for Atmospheric Research.

Nelson, J. P. (1977). Accessibility and the value of time in commuting. *Southern Economic Journal*, 1321–1329.

New Civil Engineer (2018). Arup settles 1.3 billion traffic forecast lawsuit.

NOAA (2012a). HPC improvement to NCEP models (1-inch day 1 QPF forecast).

NOAA (2012b). HPC pct improvement vs MOS.

Okun, A. (1960). The value of anticipations data in forecasting national product. In *The Quality and Economic Significance of Anticipations Data*, pp. 407–460. Princeton University Press.

Ollinaho, P., S.-J. Lock, M. Leutbecher, P. Bechtold, A. Beljaars, A. Bozzo, R. M. Forbes, T. Haiden, R. J. Hogan, and I. Sandu (2017). Towards process-level representation of model uncertainties: Stochastically perturbed parametrizations in the ECMWF ensemble. *Quarterly Journal of the Royal Meteorological Society 143*(702), 408–422.

Open Source (2020). Open trip planner.

Owen, A. and D. M. Levinson (2015). Modeling the commute mode share of transit using continuous accessibility to jobs. *Transportation Research Part A: Policy and Practice 74*, 110–122.

Palmer, T. (2019). The ECMWF ensemble prediction system: Looking back (more than) 25 years and projecting forward 25 years. *Quarterly Journal of the Royal Meteorological Society 145*, 12–24.

Pavlyuk, D. (2020). Towards ensemble learning of traffic flows' spatiotemporal structure. *Transportation Research Procedia 47*, 361–368.

Penrose, R. and N. D. Mermin (1990). The emperor's new mind: Concerning computers, minds, and the laws of physics.

Perrone, M. P. and L. N. Cooper (1992). When networks disagree: Ensemble methods for hybrid neural networks. Technical report, Brown University Providence RI INST for Brain and Neural Systems.

Peters, O. and A. Adamou (2018). Ergodicity economics. *London Mathematical Laboratory*.

Petropoulos, F. and S. Makridakis (2020). Forecasting the novel coronavirus COVID-19. *PloS One 15*(3), e0231236.

Politis, D. N., J. P. Romano, and M. Wolf (1999). *Subsampling*. Springer Science & Business Media.

Powe, N. A., G. Garrod, and K. Willis (1995). Valuation of urban amenities using an hedonic price model. *Journal of Property Research 12*(2), 137–147.

Qi, Q. and P. H. Kwok (2020). Graph ensemble net and the importance of feature and loss function design for traffic prediction.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski (2005). Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review 133*(5), 1155–1174.

Rasouli, S. and H. Timmermans (2012). Uncertainty in travel demand forecasting models: literature review and research agenda. *Transportation Letters 4*(1), 55–73.

Rasouli, S. and H. J. Timmermans (2014). Using ensembles of decision trees to predict transport mode choice decisions: Effects on predictive success and uncertainty estimates. *European Journal of Transport and Infrastructure Research 14*(4).

Rayaprolu, H. S. and D. M. Levinson (2019). What's access worth? A hedonic pricing approach to valuing cities.

Reid, D. J. (1968). Combining three estimates of gross domestic product. *Economica 35*(140), 431–444.

Ren, L. and Z. Zhao (2002). An optimal neural network and concrete strength modeling. *Advances in Engineering Software 33*(3), 117–130.

Roberton, J., S. Schmidt, and R. Stiles (2020). Emissions from the taxi and For-Hire Vehicle transportation sector in New York City.

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy 82*(1), 34–55.

Rumsfeld, D. (2011). *Known and unknown: A memoir*. Penguin.

Sanger, T. D. (1989). Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural Networks 2*(6), 459–473.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning 5*(2), 197–227.

Schrödinger, E. (1935). Die gegenwärtige situation in der quantenmechanik. *Naturwissenschaften 23*(49), 823–828.

Servizi, V., N. C. Petersen, F. C. Pereira, and O. A. Nielsen (2020). Stop detection for smartphone-based travel surveys using geo-spatial context and artificial neural networks. *Transportation Research Part C: Emerging Technologies 121*, 102834.

Sewell, M. (2008). Ensemble learning. *Research Note 11*(02).

Shahriari, M., G. Cervone, L. Clemente-Harding, and L. Delle Monache (2020). Using the analog ensemble method as a proxy measurement for wind power predictability. *Renewable Energy 146*, 789–801.

Shiller, R. J. (2005). Behavioral economics and institutional innovation. *Southern Economic Journal 72*(2), 269–283.

Silver, N. (2012). *The signal and the noise: Why so many predictions fail–but some don't*. Penguin.

Sirmans, S., D. Macpherson, and E. Zietz (2005). The composition of hedonic pricing models. *Journal of Real Estate Literature 13*(1), 1–44.

Stathopoulos, A., L. Dimitriou, and T. Tsekeris (2008). Fuzzy modeling approach for combined forecasting of urban traffic flow. *Computer-Aided Civil and Infrastructure Engineering 23*(7), 521–535.

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

Sydney Airport (2018). Sydney airport ANEF. https://aircraftnoise.sydneyairport.com.au/wp-content/uploads/2018/07/180824-ANEF-A1-Map-ENDORSED.pdf.

Tan, M.-C., S. C. Wong, J.-M. Xu, Z.-R. Guan, and P. Zhang (2009). An aggregation approach to short-term traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems 10*(1), 60–69.

Tetlock, P. E. and D. Gardner (2016). *Superforecasting: The art and science of prediction*. Random House.

Thietart, R.-A. and B. Forgues (1995). Chaos theory and organization. *Organization Science 6*(1), 19–31.

TLC (2017). *New York City FHV trip record data*. NYC Taxi and Limousine Commission.

Transport for NSW (2012). Strategic travel model standard outputs.

Transport for NSW (2020). All modes historical patronage.

Tselentis, D. I., E. I. Vlahogianni, and M. G. Karlaftis (2014). Improving short-term traffic forecasts: To combine models or not to combine? *IET Intelligent Transport Systems 9*(2), 193–201.

Tucker, J., C. Wang, and P. S. Carney (1994). Silicon field-effect transistor based on quantum tunneling. *Applied Physics Letters 65*(5), 618–620.

Turing, A. (1936). On computable numbers, with an application to the decision problem. *Proceedings of the London Mathematical Society 2*(42), 230–265.

Van Cranenburgh, S., S. Wang, A. Vij, F. Pereira, and J. Walker (2021). Choice modelling in the age of machine learning. *arXiv preprint arXiv:2101.11948*.

Van der Laan, M. J., E. C. Polley, and A. E. Hubbard (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology 6*(1).

Wanas, N., G. Auda, M. S. Kamel, and F. Karray (1998). On the optimal number of hidden nodes in a neural network. In *Conference Proceedings. IEEE Canadian Conference on Electrical and Computer Engineering*, Volume 2, pp. 918–921. IEEE.

Wann, C.-D. and M.-H. Lin (2004). Data fusion methods for accuracy improvement in wireless location systems. In *2004 IEEE Wireless Communications and Networking Conference*, Volume 1, pp. 471–476. IEEE.

Wei, Y. and M.-C. Chen (2012). Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks. *Transportation Research Part C: Emerging Technologies 21*(1), 148–162.

Wichard, J. D. and M. Ogorzalek (2004). Time series prediction with ensemble models. In *2004 IEEE International Joint Conference on Neural Networks*, Volume 2, pp. 1625–1630. IEEE.

Wigner, E. P. (1990). The unreasonable effectiveness of mathematics in the natural sciences. In *Mathematics and Science*, pp. 291–306. World Scientific.

Winkler, R. L. (1989). Combining forecasts: A philosophical basis and some current issues. *International Journal of Forecasting 5*(4), 605–609.

Winkler, R. L. and R. M. Poses (1993). Evaluating and combining physicians' probabilities of survival in an intensive care unit. *Management Science 39*(12), 1526–1543.

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software engineering*, pp. 1–10.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks 5*(2), 241–259.

Wu, H. and D. Levinson (2019). *Access Across Australia*. TransportLab.

Wu, H. and D. Levinson (2020). *Access Across New Zealand*. TransportLab.

Wu, H. and D. Levinson (2021). The ensemble approach to forecasting: A review and synthesis. *Working Paper*.

Wu, S. (2009). Personal communication.

Xiao, Y., J. J. Liu, J. Xiao, Y. Hu, H. Bu, and S. Wang (2015). Application of multiscale analysis-based intelligent ensemble modeling on airport traffic forecast. *Transportation Letters 7*(2), 73–79.

Xu, J., P.-N. Tan, and L. Luo (2014). Orion: Online regularized multi-task regression and its application to ensemble forecasting. In *2014 IEEE International Conference on Data Mining*, pp. 1061–1066. IEEE.

Yao, E. and T. Morikawa (2005). A study of on integrated intercity travel demand model. *Transportation Research Part A: Policy and Practice 39*(4), 367–381.

Yu, L., S. Wang, and K. K. Lai (2005). A novel nonlinear ensemble forecasting model incorporating GLAR and ANN for foreign exchange rates. *Computers & Operations Research 32*(10), 2523–2541.

Zador, P. and D. Levinson (2013). Evaluating perturbation impact on key travel models. https://nexusresearch.files.wordpress.com/2015/03/censdcproject1finalreport.pdf.

Zhang, C. and Y. Ma (2012). *Ensemble machine learning: Methods and applications*. Springer.

Zhang, X., S. T. Waller, and P. Jiang (2020). An ensemble machine learning-based modeling framework for analysis of traffic crash frequency. *Computer-Aided Civil and Infrastructure Engineering 35*(3), 258–276.

Zhang, Y. and A. Haghani (2015). A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies 58*, 308–324.

Zhao, Y. and K. M. Kockelman (2002). The propagation of uncertainty through travel demand models: An exploratory analysis. *The Annals of Regional Science 36*(1), 145–163.

Zhou, L., K. K. Lai, and L. Yu (2010). Least squares support vector machines ensemble models for credit scoring. *Expert Systems with Applications 37*(1), 127–133.

Zhu, Y. (2005). Ensemble forecast: A new approach to uncertainty and predictability. *Advances in Atmospheric Sciences 22*(6), 781–788.