Forecasting Electricity Price Spikes: A Comparative Analysis of GAM and SVM

Romana Shwe

THE UNIVERSITY OF SYDNEY

November 2025

School of Economics

University of Sydney

Thesis submitted in partial fulfilment of the award course requirements of the Bachelor of Economics (Honours)

Supervised by Dr Ye Lu

Statement of Originality

I hereby declare that this submission is my own work and, to the best of my knowledge, it contains no material previously published or written by another person. Nor does it contain any material which has been accepted for the award of any other degree or diploma at the University of Sydney or at any other educational institution, except where due acknowledgment is made in this thesis. Any contributions made to the research by others with whom I have had the benefit of working at the University of Sydney are explicitly acknowledged.

I also declare that the intellectual content of this study is the product of my own work and research, except to the extent that assistance from others in the project's conception and design is acknowledged.

Acknowledgment

My deepest thanks go to my supervisor, Dr Ye Lu, whose steady guidance and thoughtful advice carried me through every step of my thesis journey. I am truly grateful for her time, insight, and encouragement, which have been invaluable in shaping this research. Her kindness and approachable, easygoing nature made my experience as an Honours student far less daunting. I am also grateful to the econometrics staff for providing insightful feedback that helped further improve my thesis. I am forever indebted to my parents and brothers (and my cats), whose unwavering support and love have been my greatest source of motivation throughout my studies. I would also like to thank my dear friends, Saniya, Frida, Firdos, Iggle, Genie, and Zahra, for always being there to make me laugh and keep me sane during this thesis journey.

Above all, Praise be to Allah (SWT).

I acknowledge the use of ChatGPT-5 mini to refine the academic language of my own work and to assist with coding.

Abstract

The Australian National Electricity Market (NEM) exhibits high price volatility, with prices determined every five minutes. Sudden and unpredictable price spikes create substantial uncertainty and financial risk for market participants, including generators, retailers, and households. Forecasting these rare events is challenging because non-spike periods can dominate the data, leading to the issue of data imbalance. This can bias traditional models toward the majority class, reducing their ability to accurately forecast the minority class (price spikes). This thesis addresses this challenge by adapting the Generalised Additive Model (GAM) and Support Vector Machine (SVM) with class weights to improve the prediction of rare spikes. Moreover, model performance is evaluated using imbalance-aware metrics such as the F1-score, G-mean, and the precision–recall area under the curve (PR-AUC). Results indicate that the GAM outperforms the SVM in predicting spikes while maintaining low false-positive rates. Both models balance performance across spike and non-spike classes and are robust to variations in weighting schemes and hyperparameters. These findings demonstrate the potential of GAM and SVM for electricity price spike forecasting (EPSF) in Australia, enabling generators to optimise production, retailers to maximise profits, and households to reduce costs.

Contents

1 Introduction

As of 2023, Australia has been named as having one of the most volatile electricity markets in the world. This volatility is driven by supply side issues where Australia faces natural disasters that weaken transmission lines, resulting in unplanned coal power plant outages. Established in 1998, electricity in Eastern and Southern Australia is traded through the NEM. The NEM consists of New South Wales (NSW), South Australia, Victoria, Tasmania and Queensland. NEM has one of the longest transmission lines, spanning from Queensland to South Australia. These transmission lines allow energy to be exported and imported as needed. Similarly, electricity is transported from generators to retailers via high-voltage transmission lines, and then delivered to households and businesses through lower-voltage distribution networks.

Spot prices, also known as the Regional Reference Price (RRP), represent the real-time price of electricity, updated every five minutes. They are determined by the Australian Energy Market Operator (AEMO) through a real-time dispatch process that matches supply with demand. Generators submit offers every five minutes of every day, specifying the amount of electricity they are willing to generate. As of now, bids can range from -$1,000/MWh to $15,000/MWh. Generators typically determine their bid prices based on their operational costs, the cost of shutting down and restarting generation, and their portfolio positions. In some cases, generators may even offer their capacity at negative prices, as it can be more costly or inefficient to shut down generation units abruptly. These bids are then submitted to AEMO by 12:30 PM on the day prior to dispatch. AEMO then determines the amount of generation required to meet demand for each trading interval and compiles all generator bids into a bid stack, ordered from the lowest to highest price. Generators are subsequently dispatched from the cheapest offers upward until total demand is satisfied. This means that suppliers with the cheapest generation are dispatched first, ahead of those with more expensive generation. The five-minute spot price is then determined as the price of the marginal generator, the most expensive generator required to be dispatched to balance supply and demand (Flow Power, 2021). It is important to note that the five-minute settlement was only introduced on 1 October 2021, replacing the previous 30-minute settlement system. Under the old arrangement, each 30-minute trading interval comprised six five-minute dispatch prices. As a result, if one of these dispatch intervals experienced a price spike, the entire 30-minute average price was also considered a spike. Overall, electricity is purchased by retailers at the realised spot prices and then resold to households and firms at a heavily regulated price.

EPSF provides important advantages to many participants in the electricity market. From a corporate perspective, EPSF is particularly valuable for energy retailers that supply consumers with electricity. Retailers such as AGL, Origin Energy, and Energy Australia often have contracts with consumers to supply energy at a fixed price. However, they are exposed to financial risk when electricity prices spike, as these costs cannot always be passed directly to consumers. To manage this risk, retailers typically purchase hedging contracts which allows them to lock in electricity prices in advance. While these contracts reduce exposure to volatile prices, they come at a significant cost. Access to EPSF allows retailers to strategically minimise financial exposure and mitigate potential losses. For example, if a price spike is anticipated during a certain period, a retailer can selectively enter into a hedging contract for that time, reducing unnecessary expenses.

Generators can also benefit from EPSF, where they can use these forecasts to bid strategically. For example, if a price spike is anticipated, a generator may submit bids to supply more electricity at higher prices. Forecasts also help generators manage their generation schedules efficiently. By expecting price spikes, they can adjust output, plan maintenance, and optimise fuel procurement. These measures allow generators to maximise revenue while minimising operational costs.

EPSF also provides benefits to households, who can use these forecasts to optimise their energy consumption and minimise costs. By knowing when prices are expected to be high, households can shift discretionary electricity usage, such as running appliances or charging electric vehicles, to periods with lower prices, thereby reducing their costs.

To establish whether a spot price is categorised as a price spike, a threshold is needed to be numerically defined. For simplicity, a threshold of $300/MWh is chosen in this research, where any spot price equal or above $300/MWh is considered a price spike. The $300/MWh level is used because it represents the standard strike price of the most heavily traded cap products in the Australian NEM (Shell Energy, 2023). These cap products are basically financial hedging contracts that protect electricity retailers from extreme spot price spikes. In effect, cap contracts operate like insurance against high price volatility: retailers pay generators a premium in exchange for compensation whenever the spot price exceeds $300/MWh. This means that if the spot price is under $300/MWh, retailers pay the usual spot price. However, if the spot price goes over $300/MWh, then retailers are only obliged to pay $300/MWh, and generators have to cover the difference between the actual spot price and $300/MWh. This makes $300/MWh a practical threshold for extreme price events, as it marks the point at which financial protection mechanisms in the market are activated.

Due to the nature of electricity being non-storable and influenced by factors not controlled by market forces, such as weather conditions, transmission line capacity, and government policies, changes in supply and demand can induce short-lived price spikes (Stathakis et al., 2021). Price volatility in electricity can be two orders of magnitude higher than any other commodity or financial asset (Maciejowska et al., 2022), making it highly unpredictable and rare. While forecasting each and every electricity price observation is theoretically possible, in practice it is less feasible due to the extreme volatility of prices and the dominance of routine noise. Thus, this paper focuses on predicting price spikes, which is a more tractable task, providing more actionable information for market participants.

The highly volatile nature of electricity prices is further illustrated in Table 1, which presents the monthly summary statistics for NSW electricity spot prices. Firstly, the standard deviation in most months is very large, with some values exceeding twice the monthly mean. This indicates that electricity prices frequently deviate substantially from their average, reflecting high short-term volatility. We also observe that both the mean and standard deviation vary widely across months; for instance, the mean spot price for May 2024 was \$273.6/MWh, whereas in September 2024 it reached \$57.9/MWh. Such large fluctuations highlight the presence of occasional extreme price spikes that are not captured by simple averages. Secondly, the kurtosis values show that the distribution of electricity prices exhibits heavy tails, particularly at very high price levels in the thousands of dollars per MWh. Together with the high skewness, this confirms that electricity prices are not normally distributed but are instead heavily right-skewed.

Table 1: NSW Monthly Electricity Price Summary Statistics ($/MWh)

Month	Mean	Median	SD	Min	Max	Skewness	Kurtosis
Feb 2024	111.7	65.1	610.9	-809.5	16600.0	23.1	575.0
Mar 2024	70.5	61.8	46.0	-62.4	300.0	1.4	9.6
Apr 2024	89.8	88.0	64.6	-52.3	457.5	1.1	5.4
May 2024	273.6	110.0	1443.0	-46.0	16600.0	10.1	105.0
Jun 2024	152.8	116.8	108.3	-47.0	1117.8	1.4	7.2
Jul 2024	127.2	89.6	270.4	-48.6	17290.0	38.2	2079.8
Aug 2024	174.5	98.2	926.2	-595.5	17500.0	15.1	240.0
Sep 2024	57.9	62.7	265.0	-64.7	17307.0	60.2	3841.5
Oct 2024	77.4	75.0	228.5	-1000.0	13919.0	45.8	2425.8
Nov 2024	219.7	106.5	1207.0	-60.8	17500.0	12.5	162.5
Dec 2024	134.7	87.5	664.1	-1000.0	17500.0	21.3	483.7
Jan 2025	83.3	76.0	426.2	-1000.0	17500.0	35.6	1344.7
Feb 2025	90.3	82.6	364.1	-523.4	17500.0	42.8	1987.7
Mar 2025	90.0	76.8	427.6	-1000.0	17480.0	37.9	1479.3

Similarly, another key characteristic of the electricity price data is its pronounced nonlinearity. This is illustrated in Figure 1, which plots all spot prices in NSW between February 2024 and March 2025. The absence of any clear linear pattern highlights the complex and irregular dynamics in the data, suggesting that the relationships between prices and their potential predictors are highly non-linear.

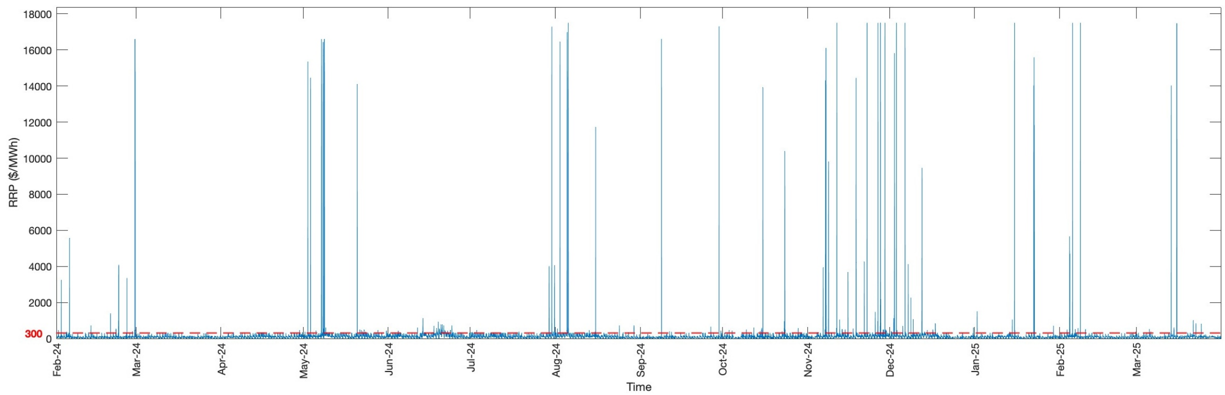


Figure 1: 5 Minute Spot Price in NSW Electricity Market (Feb 24 - Mar 25)

Another important characteristic of Australian electricity price data is its imbalance, where non–price spike events occur far more frequently than price spikes. This imbalance is a well-known feature of EPSF, as spike events are extremely rare. In fact, an event is typically classified as rare when its probability is below 5%, and price spike data from NSW, Australia (February 2024–March 2025) show an average occurrence of only 1.64%. Observing Table 2, we can see how the probability of observing a price spike is very rare, ranging between 0-6.7% out of the entire period. This brings about the issue of imbalanced data, which needs to be accounted for when training and evaluating models to avoid misleading performance results. Moreover, conventional methodologies are typically built on the assumption that each class to be predicted is sufficiently represented in the dataset (Calabrese & Osmetti, 2015). When forecasting price spikes, using such traditional models without adjustment tends to favour the majority class (non–price spikes), resulting in poor detection of the minority class (price spikes).

Table 2: Percentage of Price Spikes in NSW Electricity Market

Feb-24	Mar-24	Apr-24	May-24	Jun-24	Jul-24	Aug-24	Sep-24	Oct-24	Nov-24	Dec-24	Jan-25	Feb-25	Mar-25
2.05%	0.00%	0.32%	2.90%	6.66%	1.40%	1.83%	0.16%	0.49%	3.30%	2.26%	0.45%	0.77%	0.43%

Existing studies on EPSF in Australia give limited attention to the issue of data imbalance, which can lead to misleading conclusions and models that appear to over-perform or underperform. When this issue is acknowledged, it is typically addressed only through adjustments to evaluation metrics. This thesis aims to fill this gap in the Australian EPSF literature by directly tackling data imbalance through model adaptations and the use of imbalance-aware performance measures. Specifically, we compare the performance of an adjusted GAM with that of an adjusted SVM. Both models have been modified to account for the highly imbalanced data via weighting schemes, ensuring that the price spike class is effectively identified and not dominated by the non-price spike class. We also evaluate the robustness of the models to small changes in hyperparameters. This analysis helps determine how sensitive the models are to parameter selection and whether their forecasting performance remains stable under slight variations.

In this paper, we find that the GAM model outperforms the SVM, as it is able to forecast price spikes more accurately while minimising false predictions. The SVM is still able to forecast most price spikes, but this comes at the cost of a higher misclassification rate. Nevertheless, both models effectively address the data imbalance, achieving balanced performance in classifying both price spike and non–price spike events. We also find that imbalance-aware metrics provide a more reliable assessment of model performance compared to traditional metrics such as accuracy. Finally, changes in model parameters is found to have little effect on forecasting performance, particularly regarding the weighting scheme applied during model adjustment.

2 Literature Review

Christensen et al. (2012) highlight that most electricity price forecasting (EPF) models aim to predict spot price movements over time, but traditional approaches often struggle with forecasting rare events like price spikes (Amjady & Keynia, 2010). Electricity prices share many characteristics such as mean reversion, long memory, complex seasonality patterns, price jumps, making rare event forecasting, such as price spikes, particularly challenging (Stathakis et al., 2021). To properly capture these components, Jiang and Hu (2018) point out the seven main EPF models: simulation models, multi-agent models, statistical models, computational intelligence models, deep learning models, hybrid intelligent models and combining forecast models. Based on the chosen models of this paper, we will be focusing on EPSF literature that centres around statistical and computational intelligence models.

In the early stages of EPSF research, traditional parametric approaches were often adopted. Mount et al. (2006) applied a stochastic regime-switching model to the U.S. single-settlement market, extending it by allowing key parameters to depend on time-varying variables. They identified demand and reserve margin as critical drivers of price spikes, arguing that when the reserve margin falls below 20%, spikes are considerably more likely to occur. Using data from 1999–2000, they found that 11 of 13 observed spikes had switching probabilities exceeding 0.5. Importantly, they emphasised that predictive accuracy is highly sensitive to the quality of explanatory variables, highlighting the need for further research on feature selection. The paper's focus on reserve margin as a critical driver of price spikes directly informed this thesis, guiding the choice to incorporate generation reserve as an explanatory variable in forecasting Australian electricity price spikes.

Conventional methods such as logit and probit models have also been applied in the EPSF context. Christensen et al. (2012) incorporate an autoregressive conditional hazard (ACH) model and an ordered probit model, the latter augmented to include exogenous variables. Using a simple threshold of $300/MWh and features such as load and temperature deviations, the study finds that the ACH model outperforms the probit model. The authors attribute the probit model's poor performance to the imbalanced sample, which hinders precise parameter estimation. Although adjustments to parametric models like logit and probit have been introduced in the past to address class imbalance in EPSF, these have been limited. Therefore, this thesis extends the logit framework by incorporating explicit adjustments for data imbalance, aiming to improve predictive accuracy for rare price spike events.

King and Zeng (2001) propose an adjustment made on the Generalised Linear Model (GLM) model through the link function. They note that using logit or probit links on imbalanced data can bias probabilities, often assigning low values to the minority class. A proposed solution is to carry out a weighted exogenous sampling maximum-likelihood estimator. Instead of maximising the simple log-likelihood function, they maximise a weighted version. Weights are assigned to be $w_1 = \frac{\tau}{\bar{y}}$ and $w_0 = \frac{1-\tau}{1-\bar{y}}$, where τ is the fraction of ones in the population and $\bar{y}$ is the fraction of ones in the sample. A key drawback is that most studies, such as EPSF, do not have access to population-level data to calculate τ. Consequently, this thesis adopts a similar approach but applies a weighting scheme that is more practical and easier to implement in the absence of population-level information.

Non-parametric methods have also been introduced in EPSF namely Machine Learning (ML) methods. This stems from how ML models can capture non-linear dynamics and complex patterns from the data. Zhao et al. (2007) presented a paper exploring the use of a data mining-based approach to predict price spikes in Australia's electricity market. Using SVM and probability classifier, the paper uses feature selection to highlight the importance of accurate explanatory variables, drawing from demand, supply, spike history, season/time, net interchange, and dispatchable load. Instead of using traditional classification performance measures, they use spike prediction accuracy and spike prediction confidence. This ensured that heavily imbalanced data in EPSF was accounted for, leading to reliable and accurate interpretations of results. Furthermore, SVM predicted spike occurrences with over 50% accuracy, demonstrating its reliability even on imbalanced data. This underscores the importance of using adjusted, imbalance-aware metrics, which are applied in this thesis to evaluate model performance more reliably.

Similarly, Vu et al. (2021) explored the SVM model with a radial basis function (RBF) kernel and a polynomial kernel on the Australian electricity market. The paper integrated the use of variable thresholds to determine the price spike where features with the highest fisher scores were selected being surplus generation, supply-demand index, demand, and relative demand index. To deal with the significant data imbalance, choice based sampling is employed where resampling is reduced from 100% to 1%, meaning they resample the non-spike data only using 1% of its initial data. Using AUC to effectively compare the two thresholds, variable threshold resulted in better performance with higher true positive rate. The study also found that the RBF kernel outperformed other kernel functions, achieving the highest AUC, which motivates the use of the RBF kernel in this thesis.

Correspondingly, Stathakis et al. (2021) instituted the Multi-class SVM model for EPSF in the German market. The paper follows Byström's (2005) two-step method where Peak-Over-Threshold (POT) is used to distinguish price spikes. Variables (and 24 lags of itself) chosen were logged electricity prices, vertical load (total electricity demand that must be met by conventional power plants), wind in-feed, solar in-feed, total vertical load, total wind in-feed, total solar in-feed and net exports. To address class imbalance, different weights to each class is applied in the SVM, adjusting the penalty for misclassified data points accordingly. Authors also note that simple measures of accuracy tend to benefit the majority class over the minority classes. Hence, precision, recall , F1-Score, and G-mean accuracy scores are employed as more appropriate methods for the imbalanced classification problem. Using a 2 layered neural network model and an extreme gradient boosted machine model as the benchmark models, SVM resulted in having the highest F1-score. Their adjustment on the SVM model ultimately motivates the methodology of this thesis, particularly on the SVM model to account for imbalanced data. It also motivates the use of evaluation metrics such as F1 score and G-mean to assess model performance in Australian EPSF.

While several EPSF studies acknowledge the effect of data imbalance, few directly address it. The following section reviews broader literature on imbalance treatment across domains, forming the foundation for the methodological choices in this thesis.

In binary classification, imbalanced data commonly arises when one class dominates the dataset. This can lead to misleading results, as a classifier may achieve high accuracy simply by predicting the majority class. Few past studies on EPSF explicitly address this issue. In other fields, however, imbalanced data is widely managed through model specification adjustments and evaluation methods tailored to data imbalance. These approaches are clearly relevant to EPSF but remain under-explored in forecasting price spikes. The following section reviews key studies from other disciplines, highlighting strategies that inform the modelling framework of this thesis, with particular focus on comparing resampling and cost-sensitive learning (CSL) methods.

The traditional approach to addressing class imbalance is to force the dataset toward balance. This can be done by undersampling the majority class, although this reduces the amount of data and may distort the original distribution (Chen et al., 2021; Thölke et al., 2023). Alternatively, oversampling the minority class can be employed, but this risks introducing noise and increasing the likelihood of overfitting.

Soh and Yusuf (2019) applied oversampling to predict credit card fraud transactions in Europe, where fraudulent transactions occur far less frequently than legitimate ones. The authors implemented this technique across several machine learning models, including random forest, k-nearest neighbours (KNN), decision tree, and logistic regression. They found that logistic regression achieved the best overall performance, while the other models were prone to overfitting.

Similarly, H. Li and Sun (2012) forecasted firm failure 1–3 years in advance within the tourism industry using an oversampling technique. In addition to this adjustment, they implemented a performance indicator that accounts for total error, false positives, and false negatives. Weights were assigned to each type of error, with false positives weighted three times more heavily than false negatives. Overall, the nearest neighbour support vectors successfully predicted firm failures with accuracy rates around 90%. The study highlights that, alongside data resampling, model adjustments improve classification stability and enhance predictive accuracy for the minority class.

To avoid the data integrity issues as mentioned, the methodologies from this thesis focus on CSL where we apply model adjustments and utilise tuning hyperparameters. Recent studies have explored improving algorithms for imbalanced data by adjusting model parameters and incorporating weights or penalties. For example, Chen et al. (2021) aimed to predict consumer purchasing behaviour for travel services, where purchases of target items occurred far less frequently than other behaviours. Using a feature combination method on a random forest model, the authors applied CSL, increasing the penalty for misclassifying the minority group. This ensured that the model placed greater emphasis on correctly predicting target purchases. The CSL model was then compared with a basic random forest using resampling techniques such as DBFUS (undersampling) and SMOTE (oversampling). Overall, CSL achieved higher F1 scores than the resampling-based methods.

Similarly, Rodríguez Velasco et al. (2023) developed a model to forecast late-stage dropout among graduate students in online programs. The authors emphasised the use of probability threshold adjustments to address class imbalance. They found that modifying thresholds in a random forest model improved robustness by balancing recall and precision, concluding that a base model with adjusted thresholds and tuned hyperparameters can perform effectively without altering the underlying data distribution.

Likewise, Amirshahi and Lahmiri (2024) applied an ensemble learning technique to predict bankruptcy in a highly imbalanced dataset. Ensemble models combine multiple individual models to produce a single, more robust predictive model. The authors compared an ensemble with tuned hyperparameters, optimized via grid search, against the same ensemble applied to an oversampled, balanced dataset. They found that oversampling degraded the performance of the base classifiers, and the ensemble with the original data and tuned hyperparameters achieved the best results.

The choice between resampling and CSL remains a topic of ongoing debate regarding which approach is superior. Weiss et al. (2007) concluded that there was no clear overall winner when comparing the two methods. Importantly, they also note that for datasets with more than 10,000 observations, CSL consistently outperforms sampling-based methods. This finding, along with the literatures discussed, supports the use of CSL in this thesis. Large historical electricity price datasets are employed, allowing model adjustments and hyperparameter tuning to address class imbalance effectively without altering the original data distribution.

3 Methodology

In this section, we introduce binary classification models for forecasting electricity price spikes. We extend these models to explicitly account for the inherent class imbalance in electricity price data by applying model adjustments and using evaluation metrics tailored to imbalanced datasets.

3.1 Generalised Additive Model (GAM)

In the context of binary forecasting, we are interested in the conditional probability of observing a price spike, denoted $p(x)$, where Y is a binary random variable that takes the value 1 if a price spike occurs, and 0 otherwise. Our objective is to model the probability that $Y = 1$ given a set of predictors X, which can be expressed as

$$\Pr(Y = 1 \mid X = x) = p(x),$$

where $X = x$ denotes the specific values of the predictors. This probability $p(x)$ is bounded between 0 and 1, whereas the predictors can take on any real value. To model $p(x)$ in relation to the predictors, $p(x)$ needs to be transformed so it can take values on the entire real line. A GLM addresses this by using a link function $g(\cdot)$ to map bounded probabilities to the real line, expressed as

$$g(p(x)) = x^\top \beta \ , \tag{1}$$

where β denotes the coefficients on the predictors. As a result, we can now use the predictors to model $p(x)$. Furthermore, GLMs extend classical linear regression by accommodating non-normal outcome distributions and different types of response variables, such as the binomial distribution for binary outcomes, which is the focus in EPSF. Using the **logit link**, the standard choice for modeling a binary outcome, the link function is defined as $\log \frac{p(x)}{1-p(x)}$. Substituting this into (1) and solving for $p(x)$ gives the familiar logistic regression model

$$p(x) = \frac{1}{1 + e^{-x'\beta}}.$$

However, the GLM assumes linearity in the predictors, which is often violated in electricity market data. To relax this linearity assumption, we extend the GLM to a GAM, which models the predictors using flexible smooth functions, expressed as

$$g(p(x)) = \beta_0 + f_1(x_{i1}) + \cdots + f_p(x_{ip}),$$

where $f_p(x_{ip}) = \sum_{k=1}^{K_p} \beta_{pk} B_{pk}(x_{ip})$. Here, $B_{pk}(x_{ip})$ denotes the basis functions used to represent the smooth function, β_{pk} are the parameters to be estimated, and K_p controls the flexibility of the smooth function. Smooth functions allow for flexible modelling of non-linear relationships in the data. In this thesis, we estimate $f(\cdot)$ using **cubic splines**, a widely used basis function for constructing smooth functions. Cubic splines are piecewise functions in which the data is divided into intervals, with a cubic polynomial fitted within each interval. These polynomials are joined together to ensure the overall function is smooth and continuous across the entire range of the data.[1]

The coefficients of the basis functions (β) in the GAM is estimated using Penalised Maximum Likelihood;

$$\ell_p(\beta) = \ell(\beta) - \frac{1}{2} \sum_{j=1}^{p} \theta_j \int \left[f_j''(x)\right]^2 dx,$$

where θ_j is a penalisation term to ensure that the smooth functions do not overfit the data. Overall,

[1] For further details on cubic splines, see Mantid Project (2021).

the terms inside the summation measure the 'wiggliness' of the smooth functions, controlling model complexity. Further details in Hastie and Tibshirani (1990) and Wood (2004).

Dealing with Imbalanced Data:

It is important to adjust the GAM framework to account for the class imbalance present in the electricity market data. We implement two adjustments:

1. Adjusting the logit link function as proposed by King and Zeng (2001): Instead of maximising the log-likelihood, maximise the **weighted log-likelihood function**:

$$\ln L_w(\beta \mid y) = w_1 \sum_{y_i=1} \ln\left(\frac{1}{1+e^{-x_i'\beta}}\right) + w_0 \sum_{y_i=0} \ln\left(1 - \frac{1}{1+e^{-x_i'\beta}}\right),$$

   where w_1 is the weight assigned to the positive class (price spikes) and w_0 is the weight assigned to the negative class (non-price spikes). To determine the value of w_1 and w_0, we employ two different weighting schemes to investigate whether forecasting performance is affected. This reflects how robust the forecasting model is to changes in tuning parameters. The first weighting scheme, developed by Shrivastava (2020), is the inverse sample size weighting method. This approach assigns weights to each class inversely proportional to its frequency, where $w_1 = \frac{1}{\text{frequency of } y_1}$ and $w_0 = \frac{1}{\text{frequency of } y_0}$. This method ensures that the minority class is always assigned a higher weighting compared to the majority class. The second weighting scheme involves performing a grid search over the positive class weight, while the weight for the negative class (non–price spike) is fixed at 1 for simplicity. Overall, the aim is to assign a higher weight to the positive class, thereby increasing the GAM's sensitivity to misclassifying the minority class. By penalising errors in the spike category more heavily than those in the non-spike category, the model becomes more attentive to rare spike events.

2. **Adjust probability threshold**: the default threshold of 0.5 can result in poor performance in the presence of severe class imbalance (Khan, 2019). Past research has shown that selecting an optimal threshold, instead of the default 0.5, can enhance performance metrics such as F1-score, accommodating to the imbalanced data (Rodríguez Velasco et al., 2023). Similar to prior studies, the classification threshold is selected to maximise the F1 score, which represents the harmonic mean of precision and recall on the Precision–Recall (PR) curve. This approach balances the trade-off between identifying true price spikes and limiting false detections.

Together, these adjustments enhance price spike prediction, allowing the model to remain sensitive to minority-class observations while avoiding an excessive number of false positives.

3.2 Support Vector Machine (SVM)

SVM is a supervised machine learning algorithm that is best suited for classification tasks. The model seeks to find a hyperplane that separates the two classes by maximising the margin. A hyperplane in an n-dimensional feature space can be represented as:

$$\mathbf{w}^\top \mathbf{X}_i + b = 0,$$

where $\mathbf{w}$ is the weight vector perpendicular to the hyperplane, $\mathbf{X}_i$ is the predictor vector of the i-th data point, and b is the bias term that shifts the hyperplane from the origin. The margin is determined by the support vectors, which are the data points lying closest to the separating hyperplane. These points satisfy the condition that the decision boundary equals either $+1$ or -1. As such, they play a crucial role in defining both the position and orientation of the optimal separating hyperplane.

Classification is based on the sign of the hyperplane where a positive sign indicates the positive class (price spike), while a negative sign indicates the negative class (non-price spike). The concept of the separating hyperplane and the associated margin can be illustrated as shown in Figure 2.

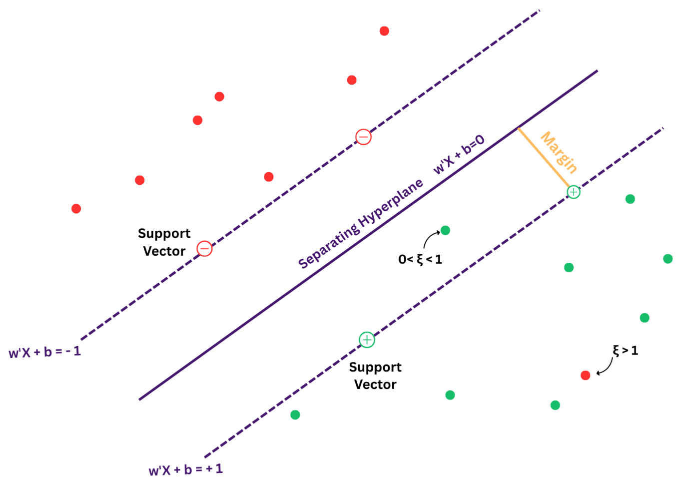


Figure 2: Illustration of a linear SVM model showing the optimal separating hyperplane between two classes

We focus on the soft-margin case, which allows for some misclassifications in order to achieve better generalisation performance on unseen data. This is achieved by introducing slack variables ξ_i, which permit certain data points to lie within the margin or even on the wrong side of the separating hyperplane. This formulation enhances the model's robustness to noise and overlapping classes, leading to improved generalisation compared to the hard-margin SVM.

In the soft-margin case, the SVM optimisation problem can be expressed as

$$\max_{\mathbf{w},b} \; M \quad \text{s.t.} \quad \frac{Y_i(\mathbf{w}^\top \mathbf{X}_i + b)}{\|\mathbf{w}\|} \geq M(1 - \xi_i), \quad \xi_i \geq 0, \quad i = 1, \dots, n, \tag{2}$$

where M denotes the margin and $\|\mathbf{w}\|$ is the weight vector norm. After some rearrangement of (2), an inverse relationship between the margin M and the weight vector norm $\|\mathbf{w}\|$ becomes evident, which allows us to reformulate (2) as a minimisation problem for computational efficiency, formulated as

$$\min_{\mathbf{w},b,\xi} \ \frac{1}{2}\mathbf{w}^\top\mathbf{w} + C\sum_{i=1}^{n}\xi_i \quad \text{s.t.} \quad Y_i(\mathbf{w}^\top\mathbf{X}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \tag{3}$$

where C is a penalisation term introduced on the slack variables. This term controls the trade-off between maximising the margin and allowing classification errors. If C is set too low, the model will tolerate a large number of misclassifications, resulting in a wider margin and potential under-fitting. Conversely, if C is set too high, the model will attempt to perfectly separate the data, leading to a narrower margin and potential over-fitting. Equation (3) is subject to constraints ensuring that each data point lies on or beyond the margin, unless a margin violation is permitted through the corresponding slack variable. The slack variables are further restricted to be non-negative, meaning they can only take values greater than or equal to zero.

Lagrangian is used to solve for (3) where we get the Karush-Kuhn Tucker conditions, defined as

$$\mathbf{w} = \sum_{i=1}^{n}\alpha_i Y_i \mathbf{X}_i \tag{4}$$

$$\alpha_i\big(1 - \xi_i - Y_i(\mathbf{w}^\top\mathbf{X}_i + b)\big) = 0, \tag{5}$$

where after some rearrangements on (5) we get equation $Y_i(\mathbf{w}^\top\mathbf{X}_i + b) = 1 - \xi_i$. From equation (4), it follows that only observations with $\alpha_i \neq 0$ contribute to the weight vector $\mathbf{w}$ and, consequently, influence the orientation of the hyperplane. Equation (5) further characterises which points correspond to support vectors. Specifically, for observations with $\alpha_i \neq 0$, those with a slack variable $\xi_i = 0$ lie exactly on the margin and are correctly classified, while those with $\xi_i > 0$ lie either inside the margin or are misclassified. In both cases, these observations are considered support vectors, as they directly determine the position of the hyperplane.

The convex optimisation problem can be solved as a dual problem, which gives us the final equation,

$$\max_{\boldsymbol{\alpha}} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} Y_i Y_j \alpha_i \alpha_j \mathbf{X}_i^\top \mathbf{X}_j \tag{6}$$

$$\text{s.t.} \quad \sum_{i=1}^{n} \alpha_i Y_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, n.$$

In equation (6), we aim to maximise α_i, the Lagrange multiplier that determines the influence of each observation on the hyperplane. The formulation also ensures that points from different classes that are close together do not disproportionately affect the hyperplane. The objective function is subject to constraints that require the weighted contributions of positive and negative classes to balance, thereby placing the hyperplane centrally between the classes. Additionally, the influence of each observation is limited by the penalisation factor C, ensuring that no single point dominates the optimisation.

Adjusting for Non-linearity

Predictors in the EPSF problem often exhibit non-linear relationships, which makes it difficult to separate the classes using a simple linear hyperplane. To address this, the vectors of predictors are mapped into a higher-dimensional feature space H via a transformation φ, denoted as $\varphi(\mathbf{X}_i)$. Inputting these transformed points in equation (6) gives us

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} Y_i Y_j \alpha_i \alpha_j \varphi(\mathbf{X}_i)^\top \varphi(\mathbf{X}_j), \tag{7}$$

where the inner product $\varphi(\mathbf{X}_i)^\top \varphi(\mathbf{X}_j)$ can be computed directly using a kernel function, without explicitly mapping $\varphi(\mathbf{X}_i)$ to a higher-dimensional space. Formally, the kernel function is defined as

$$K(\mathbf{X}_i, \mathbf{X}_j) = \langle \varphi(\mathbf{X}_i), \varphi(\mathbf{X}_j) \rangle, \qquad \forall\ \mathbf{X}_i, \mathbf{X}_j \in \mathbb{R}^d.$$

Substituting the kernel function into equation (7) gives us the kernelised SVM objective function, represented as

$$\max_{\boldsymbol{\alpha}} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} Y_i Y_j \alpha_i \alpha_j K(\mathbf{X}_i, \mathbf{X}_j) \tag{8}$$

$$\text{s.t.} \quad \sum_{i=1}^{n} \alpha_i Y_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, n.$$

In this study, we employ the Gaussian (RBF) kernel, defined as

$$K(\mathbf{X}_i, \mathbf{X}_j) = \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{X}_i - \mathbf{X}_j\|^2\right),$$

which measures the similarity between two points based on their Euclidean distance. The parameter σ controls the flexibility of the kernel where smaller values allow the decision boundary to adapt more closely to the training data, whereas larger values produce a smoother boundary. The RBF kernel has been widely used in past literature and has been shown to outperform alternative kernels in several studies (Zhao et al. (2007),Vu et al. (2021)).

Dealing with Imbalanced Data:

Imbalanced data issue in EPSF can cause the hyperplane in SVM to be positioned closer to the majority (non-spike) class, reducing the ability to correctly predict the minority class (price spikes) if the imbalance is not explicitly addressed. To ensure that our SVM model accounts for the imbalance data, we follow the approach of Stathakis et al. (2021) where we extend equation (8) by adjusting the constraint on the penalisation parameter C:

$$\max_{\boldsymbol{\alpha}} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j Y_i Y_j K(\mathbf{X}_i, \mathbf{X}_j) \tag{9}$$

$$\text{s.t.} \quad \sum_{i=1}^{n} \alpha_i Y_i = 0, \quad 0 \leq \alpha_i \leq C_i, \quad i = 1, \ldots, n,$$

$$C_i = \begin{cases} C^{+} = C \cdot w_1, & Y_i = +1, \\ C^{-} = C \cdot w_0, & Y_i = -1, \end{cases}$$

where the objective function is now subject to a modified constraint in which the penalisation parameter C is split between the two classes. Assigning a higher weight w_1 to the minority (price spike) class increases the impact of its errors in the optimisation process, making misclassifications in that class more costly than those in the majority class. Consequently, the associated α_i values for price spike observations can reach a larger maximum C_i (as determined by the constraint), allowing these observations to exert greater influence in defining the optimal SVM hyperplane. In general, this class-weighted adjustment ensures that the SVM model places greater emphasis on correctly identifying price spikes, thereby improving its sensitivity to rare events while maintaining an appropriate balance with the majority class.

Similar to the GAM model, the class weights are determined using two approaches. First, inverse sample size weighting is applied, where the weights are set as $w_1 = \frac{1}{\text{frequency of } y_1}$ and $w_0 = \frac{1}{\text{frequency of } y_0}$. This ensures that the minority class receives a higher weight proportional to its rarity in the dataset. Second, a grid search procedure is employed to fine-tune the class weights in combination with other hyperparameters, thereby optimising the model's predictive performance. Overall, the application of the two weighting schemes allows us to assess whether the SVM model's performance is robust to the choice of weighting method and to variations in hyperparameter settings.

4 Model Evaluation

A key emphasis when dealing with imbalanced data lies in the choice of metrics used to evaluate model performance. Using traditional metrics such as accuracy can be misleading in imbalanced datasets, as a model that always predicts the majority class (non-price spikes) may appear to perform well despite failing to detect the minority class.

This section draws on the ideas proposed by Bekkar et al. (2013), who advocate for the use of several metrics specifically adapted to imbalanced datasets, derived from the **confusion matrix** as shown in Table 3.

Table 3: Confusion matrix

	Predicted Positive	**Predicted Negative**
Actual Positive	TP (True Positive)	FN (False Negative)
Actual Negative	FP (False Positive)	TN (True Negative)

From the matrix, we are particularly interested in precision and recall:

Precision: $\frac{TP}{TP+FP}$, how many predicted positives are actually positives. A model with high precision means that it is able to predict most of the price spikes with few false predictions, making the forecast reliable.

Recall: $\frac{TP}{TP+FN}$, measures how many actual positives are correctly predicted by the model. In the context of EPSF, it represents the proportion of actual price spikes that the model successfully identifies. A high recall level indicates that the model successfully identifies most of the actual price spikes in the dataset.

Precision and recall have a trade-off relationship, which can be visualised using a PR curve. In EPSF, we aim to balance this trade-off, ensuring that most price spikes are correctly predicted while minimising false positives. Additionally, we use other evaluation metrics derived from the confusion matrix to assess the GAM and SVM model, ensuring that the assessment conforms to the imbalance present in the data:

F1 Score: the harmonic mean of precision and recall. This ensures that the selected model can accurately predict price spikes with fewer errors, making it a reliable indicator for imbalanced data. This will be our main metric used to evaluate in-sample and out-of-sample model performance as used in Rodríguez Velasco et al. (2023), Chen et al. (2021) and Stathakis et al. (2021).

G-mean: measures the balance between the classifier's performance on the majority and minority class. The G-mean will be low if the model performs poorly on either the non–price spike or price spike class. In particular, if the model fails to identify price spikes, the G-mean will decrease substantially. This metric has been widely adopted in prior literature, including Stathakis et al. (2021) and Bekkar et al. (2013), to evaluate model performance under class imbalance.

Precision Recall Area Under the Curve (PR-AUC): PR curve plots precision against recall across all classification thresholds, with a particular focus on the positive (minority) class. Saito and Rehmsmeier (2015) demonstrated that PR curves are more robust to imbalanced data than other evaluation curves, such as the receiver operating characteristic (ROC) curve, because they emphasise the model's performance on the minority class. The PR curve can be summarised into a single metric, the PR-AUC, which quantifies the overall trade-off between precision and recall. A model with a PR-AUC value of 1 represents a perfect classifier, while a value of 0 indicates no discriminative ability between the classes. The PR-AUC metric also has a baseline corresponding to the proportion of observations belonging to the positive (price spike) class; therefore, a model achieving a PR-AUC higher than this baseline can be considered to perform better than random, providing meaningful predictive value.

For all metrics used, larger values indicate better performance of the GAM and SVM model.

5 Data

For this study, publicly available data from the AEMO, referred to as 'Dispatch Data,' will be utilised. The dispatch data contains information on the operation of the national electricity market, split by region. It includes 5-minute interval data on potential predictors such as total demand, demand forecast, available generation, cleared supply, and others. Our models will focus on the 5-minute data from the NSW region, although the methodology can be extended to other regions. In total, 122,976 observations are used, spanning from February 2024 to March 2025. These observations are split into 80% for model training and 20% for model testing.

Choice of Predictive Variables

Predictors for the models are selected using the Fisher score, as implemented by Vu et al. (2021). The Fisher score identifies features that maximise the separation between classes while minimising the variability of data points within the same class, thereby enhancing class discrimination in the predictive model.

The Fisher score is calculated as:

$$F(c_1, c_2) = \frac{(m_1 - m_2)^2}{s_1^2 + s_2^2},$$

where m_1 and m_2 denote the means of classes one and two, respectively, and s_1^2 and s_2^2 denote the variances of classes one and two, respectively. This metric quantifies the separation between two classes by comparing the squared difference between their means to the sum of their variances, such that features with larger Fisher scores provide greater class discrimination.

Table 4: Fisher scores for selected predictors

Variable	**Fisher Score**
Lagged Price	2.8856
Generation Reserve	2.1226
Total Demand	2.0100
Generation from Renewable Sources	0.8381
Net Interchange	0.4523

From Table 4, lagged price is identified as having the strongest influence on predicting price spikes. This variable has been widely used in EPSF studies, with Eichler et al. (2014) noting that price spikes often occur in clusters, making lagged price a crucial predictor.

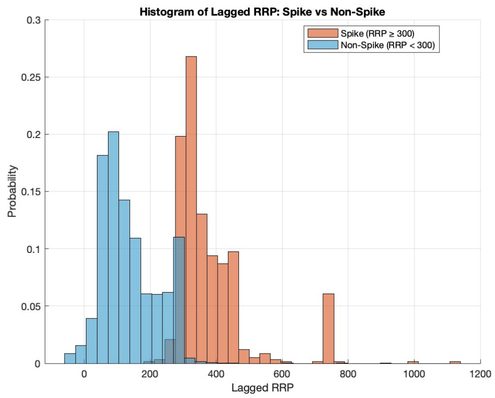
The next most influential variable is generation reserve, which represents the extra generation capacity remaining after demand has been met, available to handle unexpected situations. Past papers have drawn a significant relationship between generation reserve and price spikes where Lu et al. (2005) find that when reserve is less than 20-30% of total demand, a price spike is most likely to occur. Typically, low generation reserve means that when unforeseen events occur, there is little energy backup available, which can lead to price spikes.

We then have total demand as a predictor which is also one of the most reoccurring predictors used in past EPSF papers such as in Mount et al. (2006) , Lu et al. (2005), Zhao et al. (2007) and Vu et al. (2021). Demand plays a major role in determining price spikes because the spot price is set through the dispatch process, where supply and demand intersect. Sharp increases in electricity demand over a short period can exceed the ability of generators to ramp up production quickly, creating an imbalance between supply and demand. This imbalance can result in temporary shortages, which in turn lead to significant spikes in spot prices.

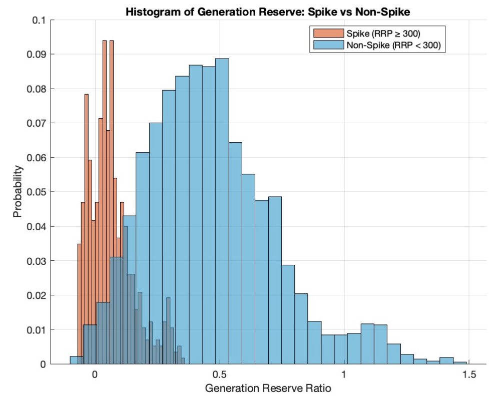
The next predictor is renewable energy (RE), also referred to as unconstrained intermittent generation forecast (UIGF). This includes generation from sources such as solar, wind, and hydroelectric power. It is noteworthy that RE generation in Australia has more than doubled over the past decade (Department of Climate Change & Water, 2025), producing two key effects on electricity prices. First, there is the merit-order effect, whereby increased RE generation tends to lower wholesale spot prices. Second, the intermittency of RE sources introduces greater price volatility, as generation is difficult to adjust in response to fluctuations in demand. Additionally, RE output is heavily influenced by weather patterns and other factors beyond market control, further contributing to market volatility. Including RE generation as a predictor is therefore essential to capture its role in driving price spikes through these volatility effects.

Lastly, we include net interchange as a predictor, noting that Australia has six interconnectors between states. Net interchange is defined as the difference between electricity imports and exports to and from other regions. The interconnector system plays a crucial role during periods of regional constraints, as imported energy can significantly alleviate supply shortages and prevent undersupply. This also implies that price spikes in one region can propagate to others. Furthermore, Abban and Hasan (2021) demonstrate that significant spillover effects exist between regions, particularly from NSW to other states, resulting in volatility transmission. Conversely, physical constraints on interconnectors can limit electricity flows, causing congestion and potentially leading to local price spikes. Including interconnector flows as a predictor is therefore essential to capture their dual role in both transmitting and moderating price spikes.

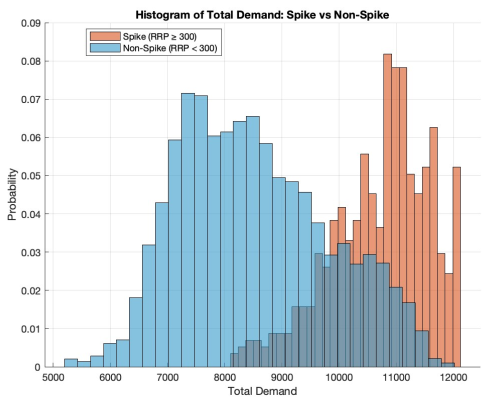
It is important to note that seasonal variables were not included as predictors in the models, as they are already highly correlated with the chosen predictors such as total demand, generation reserve, and renewable generation. The histograms in Figure 3 furthermore illustrates how each predictor relates to price spikes. Panels (a) and (c) show that higher lagged electricity prices and increased demand are associated with a greater likelihood of spikes. In contrast, lower generation reserves, reduced renewable generation, and low net interchange (panels b, d, and e) also tend to coincide with price spikes.



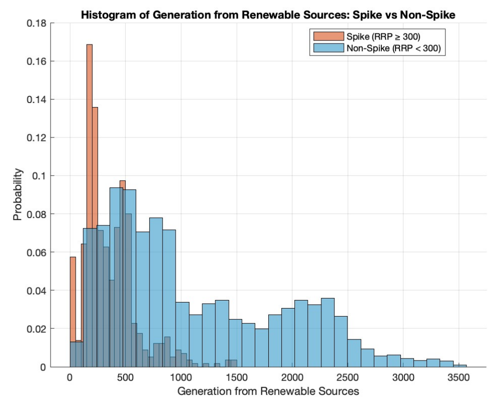
(a) Lagged RRP



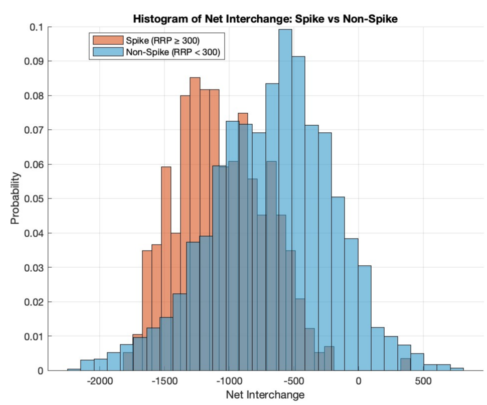
(b) Generation Reserve



(c) Total Demand



(d) Renewable Generation



(e) Net Interchange

Figure 3: Histograms of predictors used in the model

6 Results

6.1 Pseudo Out-of-Sample Model Selection

In this section, we estimate the predictive models used to forecast electricity price spikes. This is achieved using an out-of-sample forecasting approach known as pseudo out-of-sample testing. The ultimate goal is to replicate the real-time forecasting challenges encountered in practical EPSF. In this process, the dataset is divided into two subsets: 80% is allocated for training, where the models learn the relationships between the predictors and electricity price spikes, and 20% is reserved for testing. The test set is not used during model estimation and instead provides an unbiased assessment of the model's ability to predict unseen data. We will be comparing the GAM and SVM model based on its performance on this test data. Pseudo out-of-sample training method has been employed in many other forecasting sectors, including EPSF such as in Li et al. (2024), Rünstler et al. (2009), Stathakis et al. (2021), and Christensen et al. (2012).

We implement a rolling windows approach within the pseudo-forecasting framework, which ensures that the model is not biased toward any single sample and helps mitigate overfitting. A key advantage of the rolling windows method is that it systematically discards older data, retaining only the most recent observations. This is particularly beneficial in EPSF, where high-frequency data are abundant and older observations may have limited relevance for forecasting current price spikes.

The rolling windows approach is applied to the 80% training dataset to select the optimal lag length and hyperparameters for both the GAM and SVM models, as illustrated in Figures 4 and 5. A training window of three months and a validation set of one month are used, with the window rolled forward by one month at each step. These window lengths were chosen to ensure that each segment contains a sufficient number of both price spike and non-spike observations, which is critical given the rarity of price spikes. The combination of lags and hyperparameters that produces the highest average F1 score across all windows is selected, ensuring that the resulting models can accurately predict price spikes while minimising false alarms.

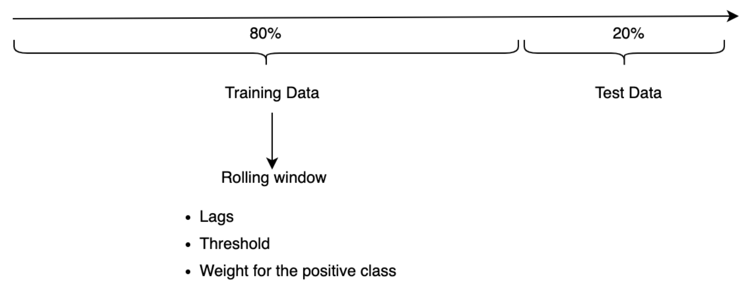


Figure 4: GAM training using an 80/20 train-test split with rolling-window parameter selection

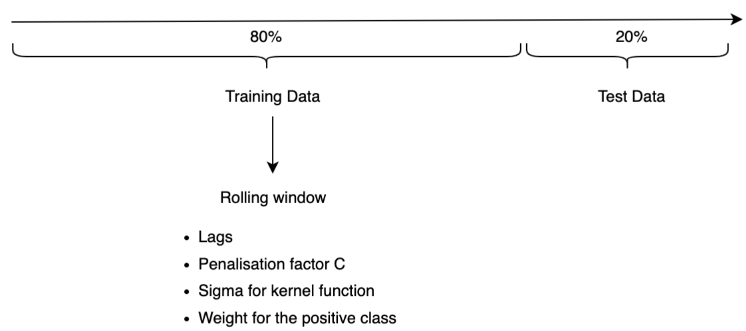


Figure 5: SVM training using an 80/20 train-test split with rolling-window parameter selection

6.1.1 GAM

For the GAM, hyperparameter optimisation plays a crucial role in dealing with the imbalanced data, ensuring classification accuracy. The parameters subject to tuning include the decision threshold and, when using grid search, the positive class weight w_1[2]. When inverse class weighting is applied, w_1 is computed directly from the sample proportions and thus excluded from the optimisation process. The subsequent section presents the in-sample results, offering insights into the model's fitting performance before moving to the out-of-sample evaluation.

In-Sample Result:

[2]Recall that two weighting schemes for w_1 are evaluated in this study.

Observing Table 6, we see that using inverse sample weighting results in a positive class weight of 52, whereas performing a grid search yields a weight of 10. Both lag structures for the two types of GAM model are very similar, except for the inverse of samples weighting model having slightly higher lags. An important observation is that when the positive class weight is large (as with inverse sample weighting), the model employs a higher threshold, reflecting the stronger prior emphasis on the positive class. Conversely, a lower positive class weight (as obtained via grid search) is paired with a lower threshold, allowing the model to predict more positive instances. This is intuitive: a large w_1 means the model already strongly favors the positive class, so a higher threshold ensures that not too many positives are predicted. Conversely, when w_1 is low, the threshold is reduced to allow the model to predict more positives.

The in-sample performance demonstrates a good fit, with an approximate F1 score of 0.6. This indicates that the GAM effectively identifies price spikes, achieving a favorable balance between precision and recall. Furthermore, the average F1 scores under both weighting schemes are very similar, suggesting that small changes in hyperparameters or lag selections do not significantly affect in-sample model performance.

Table 6: Selected lags and model parameters for the GAM that maximise the Avg F1 score

	Inverse ($w_1 = 52$)	**Grid** ($w_1 = 10$)
Demand	4	4
Lagged Price	1	1
UIGF	6	1
Generation Reserve	6	4
Net Interchange	24	24
Threshold	0.98	0.89
Avg F1 Score	0.639	0.641

6.1.2 SVM

For the SVM model, the hyperparameters subject to optimisation include the penalisation factor C, which controls the trade-off between margin width and misclassification, and the kernel bandwidth σ, which determines the smoothness of the decision boundary. When grid search is employed for class weighting, the positive class weight w_1 is also treated as a tunable parameter. Under inverse class weighting, w_1 is externally computed from the class proportions and therefore excluded from the tuning process. The following section presents the in-sample results, offering an overview of the model's performance before moving to the out-of-sample evaluation.

In-Sample Result:

Observing Table 7, we find that the weighting values w_1 are identical to those obtained in the GAM model when using both the inverse of samples weighting and grid search weighting methods. The lag structure between the two weighting schemes in the SVM model is also identical. Importantly, the penalisation factor C is larger when w_1 is small. This ensures that, even with a smaller w_1, the model still places sufficient emphasis on the positive class through stronger penalisation. Conversely, when w_1 is large, as in the case of the inverse weighting method, the corresponding C value is smaller, preventing the model from overfitting the positive class. The in-sample F1 score for the SVM is relatively low, indicating that the model has greater difficulty in accurately identifying price spikes. Overall, the average F1 scores across both SVMs using the different weighting schemes are very similar, emphasising the model's robustness to small changes in parameters.

Table 7: Selected lags and model parameters for the SVM model that maximise the Avg F1 score

	Inverse ($w_1 = 52$)	**Grid** ($w_1 = 10$)
Demand	5	5
Lagged Price	5	5
UIGF	5	5
Generation Reserve	25	25
Net Interchange	25	25
C	10	50
σ	100	100
Avg F1 Score	0.241	0.262

6.2 Out-of-Sample Results

After applying the trained models on the test set, performance is summarised in Table 8. Our analysis primarily focuses on metrics that account for class imbalance, as presented in the table, while traditional metrics such as accuracy are reported only as a robustness check.

Table 8: Out-of-sample performance of the GAM and SVM models

Model	F1 Score	PR-AUC	G-Mean	Recall	Precision	Accuracy
GAM (inverse weighting)	0.6	0.453	0.776	0.604	0.596	0.995
GAM (grid search)	0.625	0.479	0.786	0.619	0.632	0.996
SVM (inverse weighting)	0.116	0.150	0.870	0.813	0.062	0.93
SVM (grid search)	0.248	0.148	0.679	0.468	0.169	0.984

6.2.1 GAM

Focusing on the performance of the GAM model, the grid-searched weighted specification yields slightly better results than the inverse-weighted version. The GAM demonstrates relatively high F1 scores under both weighting schemes, achieving 0.600 for the inverse-weighted model and 0.625 for the grid-searched model. These results indicate that the model is able to predict price spikes with a good balance between precision and recall, successfully identifying true spike events while limiting false detections. This is further supported by the recall values of approximately 0.6 across both schemes, suggesting that around 60% of actual spike occurrences are correctly identified. Furthermore, precision is also around 0.6 for both weighting schemes in GAM, which means that of all predicted price spikes, 60% are actually price spikes.

In terms of the performance on the price spike across all possible threshold, GAM gets PR-AUC of 0.453 for inverse weighting and 0.479 for grid-searched weighting. This indicates that the model captures a substantial share of true spikes while maintaining reasonable precision, highlighting its effectiveness in dealing with the highly imbalanced data. The PR-AUC has a baseline of 0.0057, representing random performance given the class imbalance. Since both GAM models achieve PR-AUC values well above this baseline, it is evident that the GAM model performs substantially better than random, providing a reliable prediction of price spikes.

Similarly, the G-mean is around 0.8 for both weighting schemes meaning that the model was successfully able to balance the prediction of both non-price spikes and price spikes. To confirm that the chosen evaluation metrics appropriately account for the class imbalance issue, the accuracy metric is also reported as a robustness check. The inverse-weighted GAM achieves an accuracy of 0.995, while the grid-searched weighted GAM records an accuracy of 0.996. If model performance were assessed solely based on accuracy, it would misleadingly suggest that the GAM performs almost perfectly. However, this interpretation is inaccurate, as high accuracy in highly imbalanced data primarily reflects the model's ability to predict the majority class rather than its effectiveness in identifying actual spike events.

6.2.2 SVM

Examining the SVM results from Table 8, the grid-searched weighted SVM performs slightly better in terms of the F1 score, achieving a value of 0.248. Overall, the F1 scores for both weighting schemes remain relatively low, indicating that the SVM struggles to achieve a strong balance between precision and recall. This is further evident when examining the individual metrics. The inverse-weighted SVM achieves a recall of 0.813, meaning that approximately 81% of price spikes are correctly identified, while the grid-searched weighted SVM attains a recall of 0.468, correctly identifying around 47% of spike events. Although these recall values suggest that the SVM is able to capture a substantial share of true spikes, this comes at the cost of very low precision. The inverse-weighted SVM records a precision of just 0.062, indicating that only 6.2% of predicted spikes are actual spikes, while the grid-searched weighted SVM achieves a slightly higher precision of 16.9%. These results imply that while the SVM is capable of detecting many true spike events, it also produces a large number of false spike predictions, limiting its overall predictive reliability.

Furthermore, the PR-AUC for the SVM model in both weighting schemes is relatively low, indicating its limited ability to balance recall and precision across all threshold levels. Nevertheless, the PR-AUC values of 0.15 and 0.148 are substantially higher than the baseline of 0.0057, demonstrating that the SVM model still performs better than random in predicting price spikes. In contrast, the G-mean indicates relatively strong performance, suggesting that the model can maintain a reasonable balance between correctly predicting non-price spikes and price spikes, accounting for the data imbalance.

Observing the accuracy as a robustness check, the inverse-weighted SVM achieves an accuracy of 0.93, while the grid-searched weighted SVM records an accuracy of 0.98. These results additionally highlight how misleading traditional performance metrics can be when applied to imbalanced data. Although such high accuracy values may suggest that the SVM performs exceptionally well, they primarily reflect the model's success in predicting the majority (non-spike) class rather than its ability to correctly identify actual price spikes.

6.2.3 Model Comparison

Comparing the GAM and SVM models, it is evident that the GAM outperforms the SVM model. The GAM achieves substantially higher F1 scores than the SVM, a pattern that is also reflected in the PR-AUC values. This indicates that the GAM is better able to balance the accurate prediction of true price spikes while minimising false positives. The performance difference is further highlighted by the confusion matrices shown in Figure 6. The SVM model produces a large number of false positives: 1,697 (inverse weighted) and 320 observations (grid-searched weighted), compared to the GAM, which records only 57 (inverse weighted) and 50 (grid-searched weighted) false positives. Although the inverse weighted SVM predicts slightly more true positives than the overall GAM, the high number of false positives undermines its reliability as a forecasting model.

The consistently high accuracy observed in both the GAM and SVM models reinforces the fact that using metrics that do not account for class imbalance can lead to misleading conclusions. In this dataset, non-price spikes account for more than 95% of observations, so a model that predominantly predicts non-spike events will appear to perform well, resulting in artificially high accuracy values. Traditional metrics such as accuracy fail to reflect the impact of misclassifications, including false positives and false negatives, and therefore can give an overly optimistic impression of model performance.

$$\begin{bmatrix} TP = 84 & FP = 57 \\ TN = 24260 & FN = 55 \end{bmatrix}$$

GAM: Inverse weighted

$$\begin{bmatrix} TP = 113 & FP = 1697 \\ TN = 22619 & FN = 26 \end{bmatrix}$$

SVM: Inverse weighted

$$\begin{bmatrix} TP = 84 & FP = 50 \\ TN = 24267 & FN = 53 \end{bmatrix}$$

GAM: Grid-searched weighted

$$\begin{bmatrix} TP = 65 & FP = 320 \\ TN = 23996 & FN = 74 \end{bmatrix}$$

SVM: Grid-searched weighted

Figure 6: Confusion matrices of the GAM and SVM models

6.3 Price Spike Clusters

It has been well-documented in literature that electricity price spikes tend to occur in clusters (Eichler et al., 2014). Understanding whether forecasting models can accurately predict the onset of these clusters is crucial for market participants. In this context, we assess whether the GAM and SVM model were able to successfully identify the beginning of price spike clusters.

Focusing on the grid-searched GAM and SVM models, we identify periods of price spike clusters. **Note:** In all figures presented in this section (Figures 7–10), the predicted spikes (purple circles) are consistently plotted slightly to the right of the actual spikes (black circles). This offset is applied for visual clarity only, facilitating distinction between observed outcomes and model predictions.

Figures 7 and 8 show price spike clusters on 13th January 2025 between 17:40 and 18:55. Figure 7 illustrates the GAM model's performance, which fails to predict the initial spike but successfully captures subsequent spikes and accurately forecasts the end of the cluster. Similarly, Figure 8 presents the SVM model's performance during the same period. While the SVM model fails to capture the initial spike, it correctly predicts the following spikes. In contrast to the GAM model, the SVM model exhibits a tendency to over-predict, continuing to signal spikes even during periods when no spikes occur.

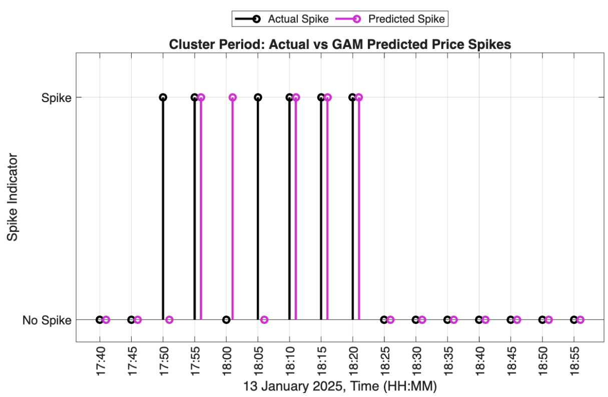


Figure 7: Actual vs **GAM** Predicted Price Spikes

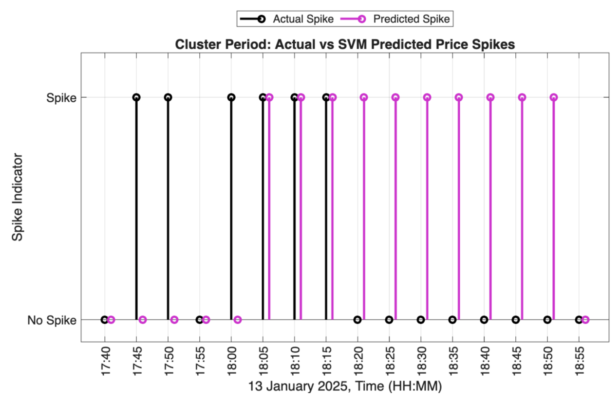


Figure 8: Actual vs **SVM** Predicted Price Spikes

Furthermore, Figures 9 and 10 present a price spike cluster observed on the 22nd of January between 18:15 and 19:55. Figure 9 illustrates the performance of the grid-searched weighted GAM, which fails to detect the first two price spikes but successfully predicts the subsequent ones. Similarly, Figure 10 shows the results of the grid-searched weighted SVM model, which also misses the first two spikes but correctly identifies the third. However, towards the end of the period, the SVM produces multiple false-positive spike predictions, even though no spikes occur. Similarly, comparing figure 9 and 10, we can clearly highlight how GAM can predict most of the price spikes compared to the SVM model.

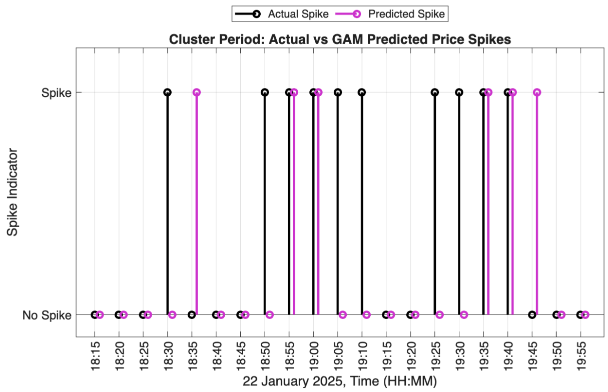


Figure 9: Actual vs **GAM** Predicted Price Spikes

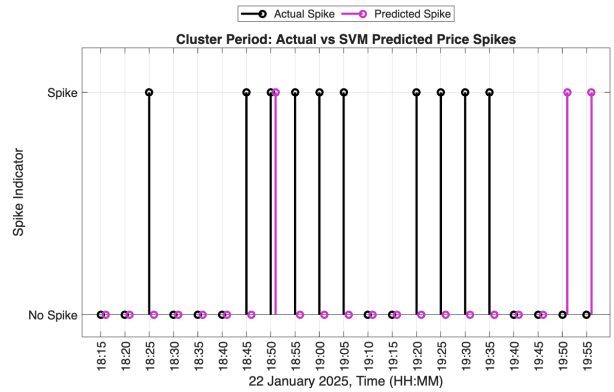


Figure 10: Actual vs **SVM** Predicted Price Spikes

Overall, the results confirm that price spikes tend to occur in clusters, and both the GAM and SVM model face challenges in predicting the initial spike within a cluster. The GAM model performs comparatively better, as it is able to capture most of the subsequent spikes and more accurately identify the start and end of the cluster. Once a price spike cluster begins, both models are able to adjust and successfully predict the following spikes.

6.4 Limitations

A clear limitation in EPSF is the large volume of data that must be processed, which leads to substantial computational cost. Models must handle thousands of high-frequency observations, often with multiple features and complex interactions. Estimating non-linear relationships in the GAM or tuning hyperparameters in the SVM requires repeated evaluations across the dataset, which is time-consuming. In particular, the SVM model demands more computational resources than the GAM due to the greater number of hyperparameters and the complexity of the kernel matrix. These computational constraints also limited the extent of model tuning. While testing additional lag selections and a wider range of hyperparameter combinations might have provided marginal improvements, the high memory and processing requirements made this impractical. Consequently, a predefined set of lag values and hyperparameters was selected for analysis. Moreover, the rare occurrence of price spikes within the large volume of non-spike data necessitates careful handling of class imbalance, further increasing computational demands. Taken together, these factors can restrict the feasibility of real-time forecasting and limit the ability to identify the optimal GAM and SVM models, including the best combination of lags and hyperparameters.

Another important limitation is the time available for analysis, which restricted the exploration of alternative approaches for handling imbalanced data. With additional time, methods such as resampling or more extensive evaluation of imbalance-correction techniques could have been implemented. For instance, resampling methods, as discussed in literature, could be compared to the class weighting and CSL approaches used in this study, allowing for a more comprehensive understanding of their relative effectiveness in addressing class imbalance.

Moreover, electricity price spike forecasts produced in this thesis are less suitable for long-term forecasting, as the models rely on recent lags of predictive values. However, EPSF is primarily intended to provide market participants, such as retailers and generators, with actionable information for short-term decisions. Forecasting spikes within the next 30–60 minutes is particularly valuable in the high-frequency, rapidly fluctuating electricity market, where participants continuously adjust bidding strategies, generation schedules, and hedging positions in near real-time, reducing the relevance of long-term forecasts.

7 Conclusion

Accurate forecasting of electricity price spikes in the Australian NEM is critical due to the high frequency of price changes and the substantial uncertainty faced by market participants such as generators, retailers, and households. Our analysis demonstrates that the GAM consistently provides more accurate spike predictions than the SVM model, particularly by limiting false positives.

The SVM model is still able to forecast a substantial number of price spikes, but it does so at the cost of higher false positives. Nevertheless, both models demonstrate robustness to variations in class weighting schemes and, consequently, to changes in hyperparameters. This stability is important in practice, as it ensures consistent forecasting even under different weighting strategies. Both models also maintain a balanced performance between price spike and non-spike classes, reflecting the effectiveness of class weight adjustments and imbalance-aware evaluation metrics in dealing with the imbalanced data. Despite the rarity of price spikes, both GAM and SVM models provide predictive value, performing substantially better than random guessing. Although predicting the onset of clustered price spikes remains challenging, both models successfully capture subsequent spikes once a cluster has begun. The GAM model is particularly effective at detecting spikes within the cluster and accurately identifying its end, a task that proves more difficult for the SVM model.

Overall, the results demonstrate that GAM and SVM models can be effectively utilised to predict price spikes in Australia's NEM. This contributes to a more stable electricity market and provides generators, retailers, and households, with greater certainty in their operational and strategic decision-making.

8 Appendix

Inverse relationship of M and w in SVM

Rearranging (2) we get,

$$\max_{w,b} \; M \quad \text{s.t} \quad Y_i\left(\left\langle \underbrace{\frac{w'}{\|w\|M}}_{\tilde{w}}, X_i \right\rangle + \underbrace{\frac{b}{\|w\|M}}_{\tilde{b}}\right) \geq 1 - \xi_i$$

This shows that $M = \frac{1}{\|\tilde{w}\|}$

Solving Lagrange

$$L(w, b, \xi, \alpha, \beta) = \frac{1}{2}w'w + \frac{C}{n}\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i\left(1 - \xi_i - Y_i(w'X_i + b)\right) + \sum_{i=1}^{n}\beta_i(-\xi_i)$$

Karush-Kuhn Tucker Conditions:

1. F.O.C of the lagrange:

$$\nabla_w L = w - \sum_{i=1}^{n}\alpha_i Y_i X_i = 0$$

$$\Rightarrow w = \sum_{i=1}^{n}\alpha_i Y_i X_i$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{n}\alpha_i Y_i = 0$$

$$\frac{\partial L}{\partial \xi_i} = \frac{c}{n} - \alpha_i - \beta_i = 0 \quad \text{for all } i$$

2. Primal Constraints:

$$y_i(w'x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

3. Dual Constraints:

$$\alpha_i \geq 0, \quad \beta_i \geq 0$$

4. Complementary Slackness Condition:

$$\alpha_i(1 - \xi_i - y_i(w'x_i + b)) = 0 \tag{9}$$

$$\beta_i \xi_i = 0$$

The convex optimisation problem can then be solved as a dual problem

$$\rightarrow \max_{\alpha,\beta \geq 0} \min_{w,b,\xi} L(w, b, \xi, \alpha, \beta)$$

Using FOCs we can minimise the lagrangian as;

$$\min_{w,b,\xi} L(w, b, \alpha, \xi) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} w^\top w, \quad \text{where } w = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$$

$$= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j Y_i Y_j X_i' X_j$$

Observe that β is no longer in the optimisation, hence we can just solve for α;

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} Y_i Y_j \alpha_i \alpha_j X_i' X_j$$

$$\text{s.t.} \quad \sum_{i=1}^{n} \alpha_i Y_i = 0 \quad \text{and} \quad 0 \leq \alpha_i \leq \frac{C}{n}, \quad i = 1, \ldots, n$$

Support Vector Condition

From the complementary slackness condition (9), we can find the condition of the support vector points. We know that α has to be non-zero to impact the hyperplane. Hence, (9) only satisfies when $(1 - \xi_i - y_i(w'x_i + b)) = 0$:

$$
\begin{aligned}
1 - \xi_i - y_i(w'x_i + b) &= 0 \\
y_i(w'x_i + b) &= 1 - \xi_i
\end{aligned}
$$

References

Abban, A. R., & Hasan, M. Z. (2021). Solar energy penetration and volatility transmission to electricity markets—an australian perspective. *Economic Analysis and Policy*, *69*, 434–449.

Amirshahi, B., & Lahmiri, S. (2024). Bankruptcy prediction using optimal ensemble models under balanced and imbalanced data. *Expert Systems*, *41*(8), e13599.

Amjady, N., & Keynia, F. (2010). Electricity market price spike analysis by a hybrid data model and feature selection technique. *Electric Power Systems Research*, *80*(3), 318–327. https://www.sciencedirect.com/science/article/pii/S0378779609002260?

Bekkar, M., Djemaa, H. K., & Alitouche, T. A. (2013). Evaluation measures for models assessment over imbalanced data sets. *Journal of Information Engineering and Applications*, *3*(10). https://eva.fing.edu.uy/pluginfile.php/69453/mod_resource/content/1/7633-10048-1-PB.pdf

Calabrese, R., & Osmetti, S. A. (2015). Improving forecast of binary rare events data: A gam-based approach. *Journal of Forecasting*, *34*(3). https://onlinelibrary.wiley.com/doi/full/10.1002/for.2335

Chen, S.-x., Wang, X.-k., Zhang, H.-y., & Wang, J.-q. (2021). Customer purchase prediction from the perspective of imbalanced data: A machine learning framework based on factorization machine. *Expert Systems with Applications*, *173*, 114756.

Christensen, T. M., Hurn, A. S., & Lindsay, K. A. (2012). Forecasting spikes in electricity prices. *International Journal of Forecasting*, *28*(2), 400–411. https://www.sciencedirect.com/science/article/pii/S0169207011000550?

Department of Climate Change, t. E., Energy, & Water. (2025). Renewables: Australian energy statistics [Accessed: 25 October 2025].

Eichler, M., Grothe, O., Manner, H., & Tuerk, D. (2014). Models for short-term forecasting of spike occurrences in australian electricity markets: A comparative study. *Journal of Energy Markets*, *7*(1), 245–266. https://static.uni-graz.at/fileadmin/_Persoenliche_Webseite/manner_hans/Publikationen/EichlerGrotheMannerTuerk2013_-_JEM.pdf

Flow Power. (2021). *The nem, bid stacks and five-minute settlement* [Accessed: 28 October 2025]. https://flowpower.com.au/the-nem-bid-stacks-and-five-minute-settlement/

Hastie, T., & Tibshirani, R. (1990). Exploring the nature of covariate effects in the proportional hazards model. *Biometrics*, 1005–1016. https://www.jstor.org/stable/2532444?

Jiang, L., & Hu, G. (2018). A review on short-term electricity price forecasting techniques for energy markets. *2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, 937–944. https://ieeexplore.ieee.org/abstract/document/8581312?

Khan, A. (2019). Imbalanced classification: Advanced algorithms. https://www.kaggle.com/code/ashrafkhan94/imbalanced-classification-advanced-algorithms

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, *9*(2), 137–163. https://www.cambridge.org/core/journals/political-analysis/article/logistic-regression-in-rare-events-data/1E09F0F36F89DF12A823130FDF0DA462

Li, Gong, X., Ge, F., & Huang, J. (2024). Forecasting stock volatility using pseudo-out-of-sample information. *International Review of Economics & Finance*, *90*, 123–135.

Li, H., & Sun, J. (2012). Forecasting business failure: The use of nearest-neighbour support vectors and correcting imbalanced samples–evidence from the chinese hotel industry. *Tourism Management*, *33*(3), 622–634.

Lu, X., Dong, Z. Y., & Li, X. (2005). Electricity market price spike forecast with data mining techniques. *Electric power systems research*, *73*(1), 19–29.

Maciejowska, K., Uniejewski, B., & Weron, R. (2022). Forecasting electricity prices. *arXiv preprint arXiv:2204.11735*. https://arxiv.org/abs/2204.11735

Mantid Project. (2021). *Cubic spline fit function* [Accessed: 31 October 2025]. https://docs.mantidproject.org/v6.0.0/fitting/fitfunctions/CubicSpline.html

Mount, T. D., Ning, Y., & Cai, X. (2006). Predicting price spikes in electricity markets using a regime-switching model with time-varying parameters. *Energy Economics*, *28*(1), 62–80. https://www.sciencedirect.com/science/article/pii/S0140988305000897?

Rodríguez Velasco, C. L., García Villena, E., Brito Ballester, J., Durántez Prados, F. Á., Silva Alvarado, E. R., & Crespo Álvarez, J. (2023). Forecasting of post-graduate students' late dropout based on the optimal probability threshold adjustment technique for imbalanced data. *International Journal of Emerging Technologies in Learning (iJET)*, *18*(04), 120–155.

Rünstler, G., Barhoumi, K., Benk, S., Cristadoro, R., Den Reijer, A., Jakaitiene, A., Jelonek, P., Rua, A., Ruth, K., & Van Nieuwenhuyze, C. (2009). Short-term forecasting of gdp using large datasets: A pseudo real-time forecast evaluation exercise. *Journal of forecasting*, *28*(7), 595–611.

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, *10*(3), e0118432.

Shell Energy. (2023, September). $300 CapContracts – WholesaleEnergyEducationSeries$ [Accessed: 2025-10-29]. https://shellenergy.com.au/energy-insights/300-cap-contracts-wholesale-energy-education-series/

Shrivastava, I. (2020). *Handling class imbalance by introducing sample weighting in the loss function* [Accessed: 2025-05-30]. https://medium.com/gumgum-tech/handling-class-imbalance-by-introducing-sample-weighting-in-the-loss-function-3bdebd8203b4

Soh, W. W., & Yusuf, R. M. (2019). Predicting credit card fraud on a imbalanced data. *International Journal of Data Science and Advanced Analytics*, *1*(1), 12–17.

Stathakis, E., Papadimitriou, T., & Gogas, P. (2021). Forecasting price spikes in electricity markets. *Review of Economic Analysis*, *13*(1), 65–87. https://openjournals.uwaterloo.ca/index.php/rofea/article/view/1822

Thölke, P., Mantilla-Ramos, Y.-J., Abdelhedi, H., Maschke, C., Dehgan, A., Harel, Y., Kemtur, A., Berrada, L. M., Sahraoui, M., Young, T., et al. (2023). Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data. *NeuroImage*, *277*, 120253.

Vu, D. H., Muttaqi, K. M., Agalgaonkar, A. P., & Bouzerdoum, A. (2021). A multi-feature based approach incorporating variable thresholds for detecting price spikes in the national electricity market of australia. *IEEE Access*, *9*, 13960–13969. https://ieeexplore.ieee.org/abstract/document/9321354

Weiss, G. M., McCarthy, K., & Zabar, B. (2007). Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs? *Dmin*, *7*(35-41), 24.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, *99*(467), 673–686. https://doi.org/10.1198/016214504000000980

Zhao, J. H., Dong, Z. Y., Li, X., & Wong, K. P. (2007). A framework for electricity price spike analysis with advanced data mining methods. *IEEE Transactions on Power Systems*, *22*(1), 376–385. https://ieeexplore.ieee.org/abstract/document/4077152?