Vrije Universiteit Amsterdam

Heerema Marine Contractors

Master Thesis

Factor Analysis on Wind Energy Sector

Author: Jeroen Dittmar (2664505)

1st supervisor: Anil Yaman
daily supervisor: Daan Kok/Piet Wessels (Heerema Marine Contractors)
2nd reader: René Bekker

October 22, 2025

Abstract

Since the Paris Agreement came into effect in 2016, 195 countries aim to combat climate change. One of the key strategies to address climate change is the usage of renewable energy sources. Offshore wind is a promising solution due to its scalability and employment opportunities. For wind energy companies, it can be useful to predict future sentiment for investment opportunities. This research aims to classify up/down movements in the wind energy sector for various forecast horizons. Five machine learning models, consisting of Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF), Extreme Gradient Boosting (XGB) and Light Gradient Boosting (LGB) are used for the prediction of future movements. Next, SHapley Additive exPlanations (SHAP) values are computed to show important features in the models' decisions. In addition to various technical indicators, uncertainty measures Economic Policy Uncertainty (EPU) and volatility index (VIX), and commodities Rare Earth and Strategic Metals etf (REMX) and oil are included in the data. This study fills the gap in forecasting wind energy stock price movements and explaining the models' outcomes with SHAP values, using two setups: whole dataset and sliding window. For the whole dataset, SVM showed the most promising outcome for longterm predictions, achieving the best scores on a variety of evaluation metrics. The top ranked feature for this model is REMX, oil being fourth bottom and EPU and VIX in the middle-bracket. For the sliding window method, the gradient boosting models (XGB and LGB) showed better results relative to the other methods, but this may be due to the class imbalance and the streak effect.

Contents

1

Introduction

Since the Paris Agreement in 2016, 196 countries have made the promise of reducing emissions to combat climate change. Almost a decade later, US President Trump signed an executive order to withdraw the United States from the agreement for the second time. Geopolitical crises, such as the Russian invasion of Ukraine, cause energy prices to skyrocket in Europe. These events showed Europe its dependence on other countries for their energy demands and demonstrated how quickly green deals can be canceled. Currently, there is an urgent need for stable energy resources while working toward being climate neutral in 2050. The opportunities for the European Union are to adopt more clean energy in the future and in the meantime acquire new fossil energy sources. There are many ways of producing green energy, such as building nuclear power plants or wind turbines on land. This comes with public opposition, raising concerns about safety issues, wildlife habitat loss, and horizon cluttering. A solution could be to build offshore wind turbines. The North Sea Energy Cooperation aims to create an integrated energy system by 2050 (1). Many countries surrounding the North Sea will be connected to a shared electrical grid to ensure an optimal green energy distribution and to be less dependent on other countries outside the EU.

Although onshore wind installations are less complex and costly compared to offshore turbines, there are several reasons to invest in offshore wind turbines. Esteban et al. (2011) (2) explains that the wind speed is greater and more uniform at sea, resulting in longer battery life of the generators and a higher power output. As there is more free space in the sea, this leads to greater wind farms and thus more energy produced. WindEurope, the leading wind energy association in Europe, adds other benefits to the use of wind power (3). They report that 370,000 jobs are in the European wind industry in 2023, which stayed stable during Covid-19 and the energy crises. Wind energy has saved 139 million tons of CO2 in Europe in 2023 alone. To put this number in perspective, we would have

to plant 4.6 billion trees to absorb that much CO2. It is equivalent to 30.2 million gasoline cars, which is nearly all cars in the UK. The European wind energy industry paid €2.3 billion in 2023, destined for real estate, environmental taxes, and contributions to local developments. Investments in this sector have the potential to accelerate economic growth, drive innovation, and create a sustainable environment.

In the world of finance, many models are used to predict future values based on historical data. Autoregressive Integrated Moving Average (ARIMA) is used for time series forecasting, for instance, future stock prices and economic indicators. The Merton Jump Diffusion model accounts for sudden big price movements, which can occur after the publication of earnings reports or other big financial news. Another example of pricing financial instruments is the Black-Scholes model. This model is used as a framework for pricing derivatives, namely European options. Until recently, machine learning (ML) has been found in financial systems. It can be used in several fields, such as fraud detection, credit scoring, and risk management. The first application of ML was Frank Rosenblatt's perceptron algorithm in 1957. Following its introduction, the technology has made great progress and is now used in a variety of applications, including image recognition, natural language processing, and sentiment analysis. Malkiel et al. (1973)(4) discusses that it is almost impossible to consistently outperform markets and suggests passive index investing rather than going to fund managers or performing technical/financial analysis. He suggests that all public information is already priced in and fundamental/technical analysis cannot be used to achieve higher gains. Fama and French (1970) (5) add that empirically, financial markets align surprisingly well with EMH. Not everyone agrees with this and are questioning whether randomness and EMH hold. Some notable drawbacks and criticisms are: market makers and liquidity providers influence prices, irrational behavior by retail investors, calendar-based anomalies like the January effect, and well-known investors (e.g., Warren Buffet) that have consistently beaten the market.

For our research, it is interesting to see whether we can challenge the random walk theory/EMH and predict future stock prices in the wind energy sector using ML. Furthermore, it is useful to see which factors influence this, which can be vital for policymakers and investors to make well-informed decisions. This study compares several ML models, such as Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF), Extreme Gradient Boosting (XGB) and Light Gradient Boosting (LGB) for classifying up-down movements for different time horizons. To make this classification, technical indicators, volatility indexes, and commodities are used. To explain the outcomes, SHapley Additive exPlanations (SHAP) values are used to turn "black-box" models into interpretable "glass-box" models. This will answer the following research question and sub question:

1. *To what extent can future up-down movements be predicted in the wind energy sector using Machine Learning?*

2. *What are the important factors that contribute to the prediction of the ML models?*

This research is relevant for Heerema Marine Contractors (HMC), as it is active in the energy sector, more specifically in the offshore industry. HMC gains insights into which factors play an important role on a quantitative level, based on historic data. Furthermore, the models incorporate and merge several external datasets to classify the sentiment surrounding the wind energy market. This is handy for the business planners of HMC, as it can guide them to see which projects to take on soon or postpone due to future sentiment surrounding wind energy.

The remainder of the paper is structured as follows. Section 2 details the related work, situating our study among other research. Following this, Section 3 gives an overview of the data used in our analysis, including collection, preprocessing and feature engineering. Section 4 provides a detailed description of the methods applied, and the results are outlined in Section 5. Lastly, Sections 6 and 7 describe the conclusion and discussion, summarizing the main insights and offering suggestions for future research.

2

Literature Review

This chapter provides an overview of the literature relevant for this thesis. In Section 2.1 we discuss the history of forecasting stock prices, highlighting two sides of the random walk theory. We look at several cases where Time Series and ML was used to predict stock/commodity prices in Section 2.2. In Section 2.3 we discuss external influences that impact stock prices. Finally, we outline the contributions of this research in Section 2.4.

2.1 Trends and predictability in stock markets

There are many key participants active in the capital markets, such as investment banks, fund managers, retail investors. Many players active in the stock market try to outperform the market and go to great lengths to achieve this. One example is market makers that trade on various financial instruments, and where latency reduction is critical to maintain a competitive advantage. To profit from price discrepancies, they implement state-of-the-art algorithms and make use of fiber cables to minimize data transmission delays. Asset managers that are in charge of their clients capital are pushed to produce higher returns than the market average. One strategy that can be used is factor investing, which is an investment approach that tries to get higher returns using several characteristics, for example, momentum, size and value. Factor investing is popular since Fama and French (1992) (6) came up with the 3-factor model that includes market risk, the size of firms, and book-to-market. Technical analysis is popular in the field of cryptocurrencies, where various educational programs offer courses on how to trade digital currencies using various chart patterns. A Random Walk Down Wall Street, written by Burton Malkiel (1973) (4), describes that it is impossible to beat the market and it is a waste of time to perform

a technical / fundamental analysis. Random walk theory falls in the category "Strong Form", which means that stock prices reflect public and private information.
On the other side of the spectrum are researchers that reject the random walk theory. Lo et al. (1988) (7) provide evidence that stocks do not follow random walks by using a simple specification test based on variance estimators. They find that the weekly first-order autocorrelation is +30%, which implies that past returns have forecasting abilities. Jegadeesh and Titman (1993) (8) build on the three-factor model and state that there is a momentum factor. Levy (1967) (9) produced a trading rule in his paper, which found that if a stock has performed well in the last 26 weeks, then it is highly likely to perform well in the next 26 weeks. This research was met with criticism in the paper conducted by Jensen and Benington (1970) (10), saying that if many trading rules are investigated, one might eventually work. *'Likewise, given enough computer time, we are sure that we can find a mechanical trading rule which "works" on a table of random numbers—provided of course that we are allowed to test the rule on the same table of numbers which we used to discover the rule.'*(10)

Figure 2.1: The three components of EMH (11)

2.2 Forecasts using Time Series and ML methods

Predicting stock prices has been done many times in the past using various techniques and including several economic factors. Forecasting prices in the stock market remains a formidable task, where Cao et al. (2019) (12) conclude that the underlying changes are nonlinear and non-stationary.

Amilon (2003) (13) proposed a modified Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model to better predict the volatility behavior of low-priced stocks. Bernard et al. (2008) (14) compares GARCH, jump-diffusion and mean-reverting models to forecast the price of the aluminum spot and futures for different time intervals. One of the most extensively studied statistical analysis models is ARIMA to predict future values. Khan and Alghulaiakh (2020) (15) show that a custom ARIMA(1,1,33) outperformed an auto ARIMA when predicting Netflix stock.

One of the first and influential papers to predict stock returns using ML dates back to the late 1980s. White (1988) (16) applied a neural network to forecast daily IBM stock returns and showed that Neural Networks are able to capture nonlinear patterns better than linear models, such as ARIMA. Leung et al. (2000) (17) used a probabilistic neural network and showed better performance than statistical models, such as logit, probit, and discriminant analysis. Park and Shin (2013) (18) investigate the prediction of stock prices on the Korean Stock Exchange. They used a complex interrelation network that consists of multiple economic factors. These included the exchange rate, oil price, technical indicators, and target variables representing upward or downward movements. The researchers used a seven-tuple vector that contains various technical indicators. This approach captures time-series data, gives similarity scores among factors, and eases the application for ML methods. Their semi-supervised learning (SSL) performed better than SVM, artificial neural network (ANN), and Buy-and-Hold (BH).

2.3 External influences on stock markets

Black & Scholes (1973) (19) introduced the option pricing model for pricing derivatives. They provided a standardized method for the valuation of options and were awarded the Nobel Prize in 1997. The Chicago Board Options Exchange launched the CBOE Volatility Index with the ticker symbol VIX, which is based on the work of Black-Scholes. The formula is as follows:

$$VIX = \sqrt{\frac{2e^{r\tau}}{\tau}\left(\int_0^F \frac{P(K)}{K^2}dK + \int_F^\infty \frac{C(K)}{K^2}dK\right)} \tag{2.1}$$

,where:

- τ is the average days in a month (30 days)

- r is the risk-free rate
- F is the 30-day forward price
- K is the strike price
- P(K) is the put price for the strike K
- C(K) is the call price for the strike K

VIX is a measure of the expected market volatility of the S&P 500, and is sometimes called the fear index. Carr (2017) (20) gives multiple arguments for why the VIX can be called the fear gauge. One being that if the Black Scholes model holds, the VIX construction has more capital in ATM and OTM puts than in ATM and OTM calls. The puts reflect fear, while the calls reflect greed, so then VIX is more about fear than greed.
VIX peaks in uncertain times, such as the Global Financial Crisis, COVID, and Trump's tariff war. This is consistent with Bloom (2009) (21), who shows that uncertainty spikes, such as financial crises and geopolitical events, lead to a sharp increase in uncertainty measures and economic slowdowns as companies wait for investments and hiring.
Blair et al. (2001) (22) used VIX for forecasting S&P100 volatility and discovered that VIX captures almost all relevant information. Whaley (2000) (23) concludes that the relationship between stock market returns and changes in the VIX is asymmetric. When the VIX rises, the stock market tends to fall sharply. In contrast, when the VIX declines, the stock market's gains are usually more modest than one might expect.
The Global Economic Policy Uncertainty (GEPU) is an index created in 2016 and is fabricated using multiple components(24). Baker et al. (2016) (25) include the largest newspapers from each country and form an index of articles that discuss economic policy uncertainty (EPU). For the US cases, the second component is the future tax rules set to expire in the coming 10 years. The third and final component is the dispersion of professional forecasters from the Federal Reserve Bank of Philadelphia. Baker et al. (2016) (25) explore the link between the EPU and the volatility of the stock market and conclude that there is a strong connection between the two. However, there was no evidence that EPU can be used to predict future returns. Arouri et al. (2016) (26) contradicts the previous mentioned article and shows that an increase in policy uncertainty significantly reduces the returns of US stocks and that this outcome is more pronounced during extreme volatile times. Another study on Gross Domestic Product (GDP) was conducted in several emerging market economies. The main findings of the work of Gupta et al. (2020) (27) are that,

with an increase in the uncertainty in the US, GDP drops more in developing countries than in the United States.

Rare earth elements (REE) are vital for the green energy transition and other modern high-end technologies. The defense sector has several applications, from radars, jet engines, to night vision devices, in which rare earth elements are needed. For the green sector, wind turbines use generators that contain strong rare earth magnets. China is a global superpower when it comes to producing and consumption of these minerals written by Fernandez (2017) (28). They have almost complete control over the REE market and export quotas/restrictions could have disruptive consequences for other sectors/countries, as stated by the US Geological Survey (29). Morrison and Tang (2012) (30) discuss that in 2010, China blocked almost all shipment of rare earth minerals to Japan for two months after a feud between the two countries. This revealed to industrialized nations its dependence on their supply of REEs of China. Proelss et al. (2020) (31) try to forecast Rare Earth Elements volatilities. A simple Autoregressive fractionally integrated moving average (ARFIMA) showed superior accuracy than other models, such as autoregressive moving-average (ARMA) and GARCH models, and demonstrated that long-term memory models are capable of better capturing the volatility behavior of REE. Henriques and Sadorsky (2023) (32) forecast rare earth stock price directions using various ML methods. Their main findings are that technical indicators tend to be more important than volatility features. Furthermore, the accuracy also tends to improve after increasing the forecast horizon.

2.4 Contributions

There is abundant research done on the forecasting of stock returns using various quantitative methods, such as time series and ML models. There is a lack of literature that tries to forecast wind energy stock price directions with the above mentioned factors using ML. Furthermore, there is a research gap that tries to see if technical indicators, volatility characteristics and commodities have predictive power for the wind energy sector. We also examine the relationship between the rare earth metal sector and the wind energy sector. We compare accuracy scores for different forecast horizons and turn black-box results into explainable glass-box models. Lastly, this research contributes to the literature by presenting the results of the models in two setups: whole dataset and sliding window.

3

Data

This chapter begins with a description of how the data is collected in Section 3.1, followed by handling of the data in Section 3.2, including cleaning and imputing new values. Section 3.3 gives an explanation of the new derived features, and finally Section 3.4 details the data analysis.

3.1 Data Collection

Historical price data is acquired using the London Stock Exchange Group (LSEG) database and GoogleFinance for various commodities and equities. The VIX is downloaded from the website of the Chicago Board Options Exchange(33). The EPU index is retrieved from the data published by Baker et al.(24). The period runs from October, 2010 until May, 2025, meaning well over 14 years of daily historical data.

3.2 Preprocessing

Before new features are derived, the data must be preprocessed. This is done by cleaning, which involves handling missing data, removing duplicates, and outliers. There are no trading days on weekends and holidays, so these rows are removed. In other studies, the data can be user generated, or acquired via sensors, resulting in inaccuracies or measurement errors in the data. For example, the end-user sets a very high number as their age, or machines can break down and report faulty data. In capital markets, companies sometimes do stock splits, resulting in sudden jumps in stock price. The First Trust Global Wind Energy etf (FAN) did not have stock splits, and therefore removing outliers was not necessary. All separate data sources are merged into one data set and transformed using a

Table 3.1: 1-day horizon

Date	Close	Label
28-10-2010	10.46	1
29-10-2010	10.47	0
01-11-2010	10.34	1
02-11-2010	10.61	1
03-11-2010	10.66	1
(...)	(...)	(...)
02-05-2025	15.86	1

Table 3.2: 2-day horizon

Date	Close	Label
28-10-2010	10.46	0
29-10-2010	10.47	1
01-11-2010	10.34	1
02-11-2010	10.61	1
03-11-2010	10.66	0
(...)	(...)	(...)
01-05-2025	15.69	1

scalar. This is done to reduce runtime for all models. The data set is enriched by deriving new features, such as technical indicators, which are explained in the next section.

3.3 Feature Engineering

To be able to retrieve higher accuracy, new features are added to the original data to feed the ML models. These consist of technical indicators (TI) and are in Table 3.3. The intermediate steps for the technical indicators are calculated in Table 7.3. It was decided to get a variety of TI's. One category is trend indicators, where the moving average (MA), moving average convergence divergence (MACD) and average directional index (ADX) fall into. Some momentum indicators calculated are relative strength index (RSI), Stochastic Oscillator (STOCH_OSC) and Williams %R (Will_R). The volatility indicators derived are the Bollinger Bands (BB) and the Average True Range (ATR). Finally, On Balance Volume (OBV) falls into the volume indicator category. The technical indicators are computed using industry standards for the number of days.
To predict directional movements, we label the target variable (y) as 0 or 1. For a downward movement, we use 0, while 1 represents an upward movement. The models are tested on different time horizons. Table 3.1 shows an example of a 1-day forecast horizon and Table 3.2 shows the same data but for a 2-day forecast horizon.

3.4 Data Analysis

Figure 3.1 shows the time series of the different asset classes. The equities show very different price developments, but they all have one thing in common: the prices rise sharply

Table 3.3: Technical indicators

Name	**Description**	**# days**
Moving Average	$MA_t = \frac{\sum_{i=1}^{t} p_i}{t}$	[20,50,100]
Williams R	$\%R = \frac{High_t - Close}{High_t - Low_t} \times -100$	14
Bollinger Bands	$BB_t = MA_t \pm 2\sigma_t$	20
On-Balance Volume	$\mathrm{OBV}_t = \mathrm{OBV}_{t-1} + \begin{cases} \mathrm{Vol}_t, & \text{if } \mathrm{Close}_t > \mathrm{Close}_{t-1} \\ 0, & \text{if } \mathrm{Close}_t = \mathrm{Close}_{t-1} \\ -\mathrm{Vol}_t, & \text{if } \mathrm{Close}_t < \mathrm{Close}_{t-1} \end{cases}$	1
Average True Range	$ATR = \mathrm{MA}_t(\mathrm{TR})$	14
Average Directional Index	$ADX = \mathrm{MA}_t(DX)$	14
Relative Strength Index	$RSI = 100 - \frac{100}{1+RS}$	
Moving Average Convergence Divergence	$MACD = EMA_s - EMA_l$	[12,26]
Stochastic Oscillator	$\%D = \frac{Close - Low_t}{High_t - Low_t}$	14

after the outbreak of Covid-19 in January 2020. After a while, they consolidate and show a downward trend until recently.

Figure 3.2 shows the percentage of up days for different time horizons. For the 1-day forecast horizon, the percentage up days lies just below 50%. This means that in over half of the cases, the next days price turns out to be lower than today's price. On the other hand, for the 20 day forecast horizon, the percentage of up days is 53%, which means that in well over half of the matter, the price in 20 days is higher than today's price.

The plot does not tell the full story, as a stock price could have many more up days

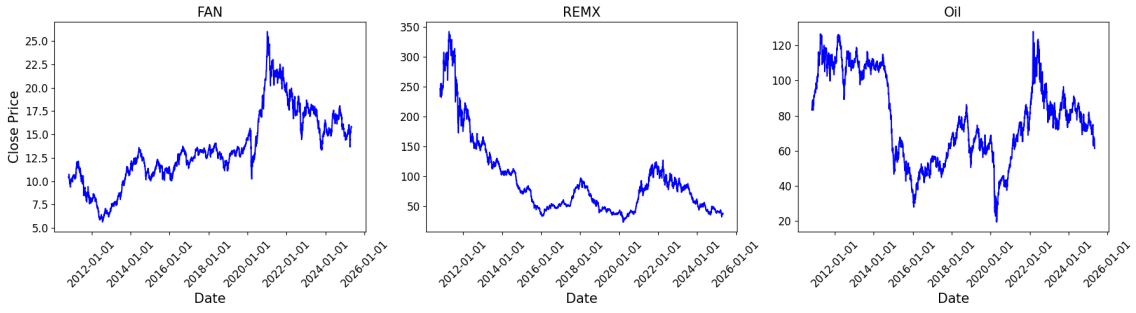


Figure 3.1: Close prices of FAN, REMX and Oil

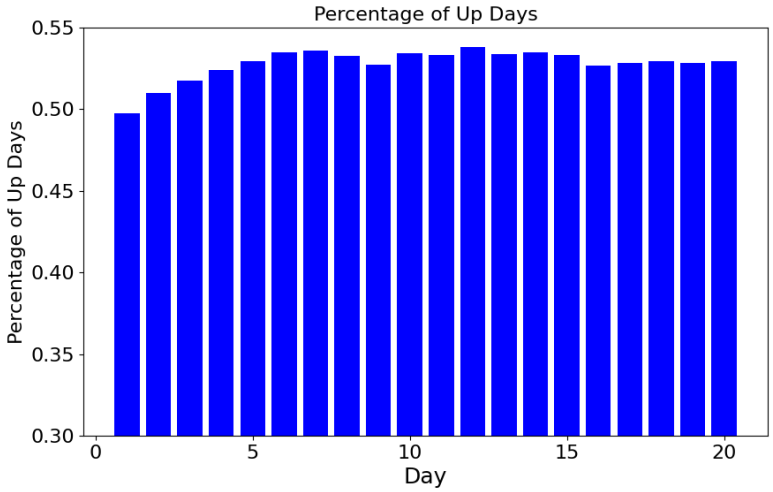


Figure 3.2: % of updays for different time horizons FAN_etf

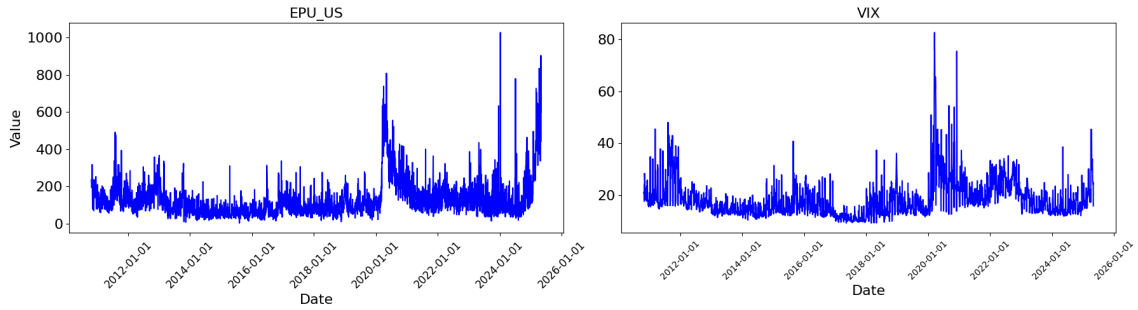


Figure 3.3: Values of EPU and VIX

than down days, but the price could plummet more in down days than the stock could rise in up days. Nevertheless, ensuring a balanced data set enhances the performance of ML models when classifying upward or downward movements. Figure 3.3 shows the time series of the uncertainty measures EPU and VIX. The EPU index consists of three different components. The first is the sentiment of the newspaper about policy-related economic uncertainty. For the United States, the second component is the number of federal tax code provisions set to expire and disagreement among economic forecasters. The third and final part of the EPU index is based on the Federal Reserve Bank of Philadelphia's Survey of Professional Forecasters. The VIX is explained in full detail in Section 2.3. The box plots for the uncertainty measures are in Figure 3.4.

Both uncertainty measures surged in early 2020 with the outbreak of Covid-19. However,

Table 3.4: Summary statistics

	FAN	REMX	Oil	EPU	VIX
count	3649	3649	3649	3649	3649
mean	13.45	94.33	78.20	127.92	18.23
std	3.92	64.35	24.56	97.82	7.07
min	5.62	24.12	19.33	3.32	9.15
25%	11.01	48.90	58.51	68.55	13.52
50%	12.94	78.30	75.94	100.39	16.37
75%	15.88	110.03	102.77	150.79	20.80
max	26.02	342.12	127.98	1026.38	82.69
CV	0.29	0.68	0.31	0.76	0.39

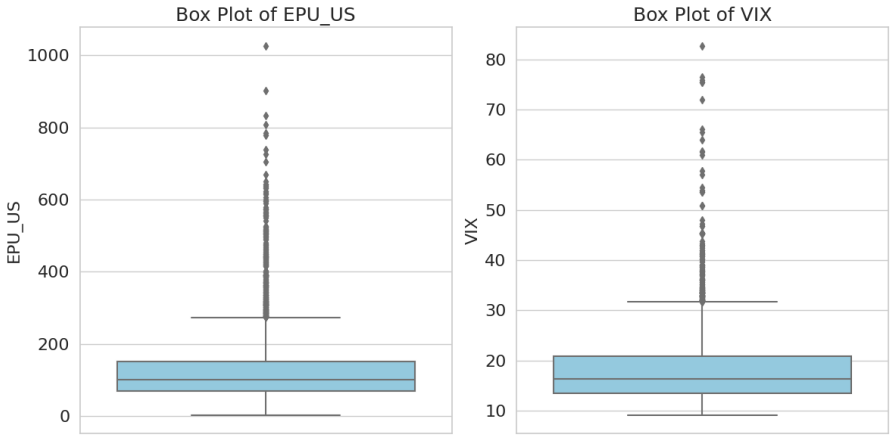


Figure 3.4: Boxplots for EPU and VIX

while VIX spiked again in the beginning of the Russia-Ukraine conflict, EPU showed a modest reaction. This is likely because the US economy is less affected than the US stock market, which often responds more sharply to geopolitical events. We can see that both uncertainty measures increase as the presidential campaign makes progress and Trump is settled as president. Looking at Table 3.4, EPU and VIX show very different values for the different summary statistics, with a coefficient of variation of 0.76 for EPU and 0.39 for VIX. Interestingly, Figure 3.4 tells us that the boxplots are relatively similar in shape.

4

Methodology

This chapter describes in detail all models used for predicting up-down movements. We take a look at some well-known ML models such as LR, SVM, and RF. On top of that, more recent additions to the world of ML (XGBoost and LightGBM) are detailed in Section 4.1. SHAP values are introduced and its use case is explained in Section 4.2. Section 4.3 describes the splitting of the data into training and test sets and feature scaling. Furthermore, Section 4.4 entails the evaluation metrics. Lastly, Section 4.5 describes the tuning of the hyperparameters for each model.

4.1 Models

Logistic Regression (LR) is widely used to forecast for binary classification purposes. Common examples are classifying spam/ham in phishing mails, or fraud detection in the banking world. The formula for the predicted output is as follows:

$$\widehat{y} = \ln(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \tag{4.1}$$

Here, β is the estimated regression coefficients, x are the explanatory variables and p is the probability of the event happening. Logistic regression is often praised for its simplicity and its interpretability. The model comes with a number of assumptions, which we list below. First, logistic regression requires independent observations and there must not be too much multicollinearity, which means that independent variables should not be too highly correlated. Lastly, logistic regression requires a large sample size for predictions.
Our data is a time series of historical prices, and new features are derived using past data. It is therefore difficult to guarantee independence among observations. Commodity prices

have shown little correlation among each other in the period 1982-2004 in the paper of Erb et al (2006) (34). In contrast, Tang et al. (2012) (35) shows that after 2004 there is an increase in price comovements between different commodities.

Support Vector Machines (SVM) can be used for both regression and classification tasks in ML models. This method works by maximizing the margin, through the optimal separating hyperplane. For linearly separable data, SVM separates the data points by a straight line. The general decision boundary can be written as:

$$w^T x + b = 0 \tag{4.2}$$

Here, w is the weight vector, b is the bias term and x is the list of explanatory variables If Equation 4.2 is bigger than zero we assign it class 1, and we assign it class 0 if it less than zero. For nonlinear data, SVM uses a technique called the "Kernel trick", originating from the research of Boser et al. (1992) (36). This allows the SVM to handle non-linear relationships. Some advantages of SVM are that it is effective in high-dimensional spaces, it is memory-efficient, and different kernels can be applied for different needs. The disadvantages of using this method is that it requires careful fine-tuning of parameters, training time is slow when the data is not scaled, and interpretability can be an issue when dealing with non-linear kernels.

Random Forest (RF) is an ensemble of decision trees that is commonly used for regression and classification tasks. This method was first described by Ho (1995) (37) and was later extended by Breiman (2001) (38). The latter introduced bagging and shows higher accuracy than individual decision trees using a majority vote for classification. The following equation shows the majority vote of the RF model:

$$\widehat{y} = \text{mode}(h_1(x), h_2(x), ..., h_k(x)) \tag{4.3}$$

The advantage of using RF models over decision trees is that it is less prone to overfitting due to the majority vote, it has higher accuracy, and it can handle data with many features.

eXtreme Gradient Boosting (XGB) is similar to the RF model, as it combines multiple decision trees, but differs in the way it constructs new trees. While Random Forests builds new decision trees using bagging, XGBoost builds trees sequentially and learns from previous decision trees. Chen and Guestrin (2016) (39) made it publicily available and is

popular in the community of ML competitions and data scientists, as it is a quick algorithm that works efficiently and shows great performance even on limited computational resources. The main disadvantage of XGBoost, just like other multiple decision tree algorithms, is that it is less interpretable

Light Gradient-Boosting Machine (LGB) was created by Microsoft as an alternative to the multiple decision tree algorithms, developed by Ke et al. (2017) (40). The two new techniques: Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) can handle a great number of data points and a large number of features. GOSS randomly drops data points with smaller gradients, while keeping instances with large gradients. EFB reduces the number of effective features and bundles them. The algorithm is able to run fast while maintaining high accuracy. Whereas XGB searches the best split on sorted features values, LGB uses a histogram-based approach. LGB grows trees by choosing the best leaf, while XGB does this level-wise. This can lead to deeper and more accurate trees, but has the potential risk of overfitting on small datasets.

4.2 SHAP Values

SHapley **A**dditive ex**P**lanations (SHAP) is a way of showing feature importance in ML. It can be used for several types of models, such as regression, classification, and deep-learning. It originates from cooperative game theory and was first introduced by Lundberg and Lee (2017) (41). For many business cases, it is often preferable to be able to interpret and explain model's outcomes. Often linear models are opted for, such as Linear Regression and Support Vector Machines, for their interpretability and simplicity, despite lower performance.

For a player i is in a set of F players, the Shapley value is as follows:

$$\phi_i(f,x) = \sum_{\substack{S \subseteq F \setminus \{i\} \\ \text{bounded}}} \frac{|S|!(|F|-|S|-1)!}{|F|!}[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)] \tag{4.4}$$

Here, S is the subset of features that does not include feature i. $f_S(x_S)$ is the prediction of the model using set S, and x_S is the feature value in input x.

4.3 Data split and scaling

The data consists of 3651 records and need to be split for training and testing purposes. The decision was made to use two different setups, namely whole dataset and sliding window. The whole dataset setup uses 80% of the data to train the data, the next 10% are used to tweak the hyperparameters, and the final 10% of unseen data are tested. The drawback of using this framework is that it can only be done once. Therefore, we also apply the sliding window with multiple simulations for robustness. The setup of the sliding window is simple: the first 600 data points are used as training data, and the next day is the test data. After this, we shift the whole setup one day further until there are no more days left. The hyperparameters in the sliding window are chosen that occurred the most in the whole data setup. In other studies, timeseries data were randomly shuffled for ML, often leading to high performance. Most likely, the models are dealing with data leakage, more specifically temporal leakage. The result is that future information is included in the training set, which is not available at prediction time. The models are not generalizable and the evaluation metrics are overinflated. This research respects the time order and does not shuffle the data at random.
The drawback of using the simulation setup with 600 data points is that it could happen that the FAN etf has momentum in either going up or down in value. Across the time frame, the percentage of updays in Figure 3.2 is close to 50%, which means a well balanced dataset. Zooming in a small interval could result in an imbalanced dataset, but having many simulations has the benefit of more robust results than running the model one time. After the features were derived, such as the technical indicators, they were scaled to help the model run more efficiently and with less runtime. This scalar subtracts the mean to each feature, so that the new mean becomes 0. Furthermore, the Z-score is divided by the standard deviation, so it is scaled to unit variance. The formula is as follows:

$$z = \frac{x - \mu}{\sigma} \tag{4.5}$$

Here, x is the original feature value, μ is the mean, and σ is the standard deviation.

4.4 Evaluation metrics

Assessing a model's performance can be done via several evaluation scores. The confusion matrix in Table 4.1 shows that a true positive (TP) is when the model correctly classifies an upward movement, and a true negative (TN) is when the downward movement is correctly

Table 4.1: Confusion matrix for up-down movements

Actual / Predicted	Down	Up
Down	True Negative (TN)	False Positive (FP)
Up	False Negative (FN)	True Positive (TP)

identified. On the other hand, misclassified up days that are actually down movements is a false positive (FP), while a false negative (FN) shows an up movement mislabeled as a down day.

This research uses several evaluation metrics to show the performance of the classification models, specifically accuracy, recall, precision, and F1 score. In addition to these measures, the Receiver Operating Characteristic curve (ROC) and the Area Under the Curve (AUC) are used to further assess the performance of the models.

Accuracy is the proportion of correctly predicted instances out of the total number of observations.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.6}$$

In the ideal scenario, we would have zero false positives/negatives, achieving an accuracy of 100%. For imbalanced datasets, accuracy alone is not enough, and we consider the performance across the other metrics.

Recall measures the proportion of all actual positives that were correctly classified as positives.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.7}$$

This score answers the question: what fraction of actual up movements are correctly classified? In an imbalanced data set, recall is more useful as it measures the ability to correctly classify positives. A high recall means that the model rarely misses bullish signals for the wind energy sector.

Precision measures the proportion of all the model's positive classification that are actually positive.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.8}$$

This score answers the question: out of all the times the model predicted an up movement, how many were actually up? This metric is used when it is very important for positive predictions to be accurate.

F1 score is the harmonic mean of precision and recall. This metric balances both precision and recall and is especially useful for imbalanced datasets.

$$\text{F1 score} = 2 \cdot \frac{recall \cdot precision}{recall + precision} \tag{4.9}$$

This score answers the question: how balanced is the model, and is it able to correctly predict up movements and avoid false alarms? A high F1 Score means that the model has accurate up movement predictions and is sensitive to actual up movements.

The ROC curve is a combination of the false positive rate (FPR) and true positive rate (TPR) calculated at all possible thresholds. The AUC is the probability that the model ranks a random up movement higher than a random down movement. An AUC score of 0.5, alongside a ROC curve that is a diagonal line tell us that the model does not perform better than random guesses or coin flips. In the perfect world, the model has a square ROC curve and an AUC score of 1.0, meaning that it perfectly distinguishes between up and down movements.

4.5 Hyperparameter optimization

For each of the five models, hyperparameter tuning is done to achieve the best model performance to predict up-down movements. This process takes place before the final training/testing phase and can significantly increase accuracy and generalizability. The procedure for finding the optimal hyperparameters can be done in several ways. Grid Search is the most exhaustive as it tests all possible combinations. Random Search is a more advanced method, as it randomly samples hyperparameter values, which can reduce runtime in large search spaces. Another common method is Bayesian Optimization, a sophisticated procedure that uses previous results in a smart way to choose future sets of hyperparameters. This process is continued until a stopping criterion is met. This method balances exploration and exploitation well for finding the best hyperparameters. This research opts for grid search, as it is simple, repeatable, and every combination is explored.

Table 4.2: Hyperparameters and ranges

Hyperparameter	**Symbol**	**Scope**
Penalty	P	[L1, L2]
Inverse of regularization strength	C	[0.01,1,2,10]
Solver	S	[liblinear, saga]
Kernel	K	[linear, rbf]
Trees	T	[100,200,300,400]
Depth	D	[5,10,15,20]

5

Results

This chapter presents the results of the various ML models used for classifying up and down movements. Section 5.1 presents the results of the five models when we use the whole time frame for the classification. Section 5.2 evaluates the results in which the sliding window is applied to the ML Models. For this section the short-, mid- and long-term cases are evaluated. This means that the 1, 10 and 20 day time horizons are discussed in more detail. The tables and figures that contain the 5 and 10 day time horizon can be found in the Appendix

5.1 Whole Dataset

5.1.1 Evaluation Metrics

The results obtained for each model come from training the data on 80%, tuning the next 10%, and then testing on 10% unseen data, illustrated in Figure 5.1
The hyperparameter tuning is explained in detail in Section 4.5. For the LR and SVM models, the hyperparameters C, penalty, solver and kernel are tuned to maximize accuracy. For the tree models, namely RF, XGB, and LGB, the number of trees and tree depth are selected to achieve the highest accuracy. There are many more hyperparameters that can be optimized, but for simplicity and generalizability, these are not explored.

Base models Figure 5.2 shows the accuracy, recall, precision and F1 for LR and SVM. The x-axis denotes the forecast horizon, running from 1 to 20 days, which corresponds to nearly a full month. There seems to be an upward trend when increasing the time horizon across the performance measures. The recall and F1-score for the LR model peak at 5 days, drop down as we increase the time horizon, and later plateaus. We see the

Figure 5.1: 80% training, 10% tuning 10% testing for the FAN etf

same pattern for the SVM, where the recall and F1-score stagnate after the 5-day mark. Despite fluctuations, the precision of the SVM shows an upward trend as the time horizon increases.

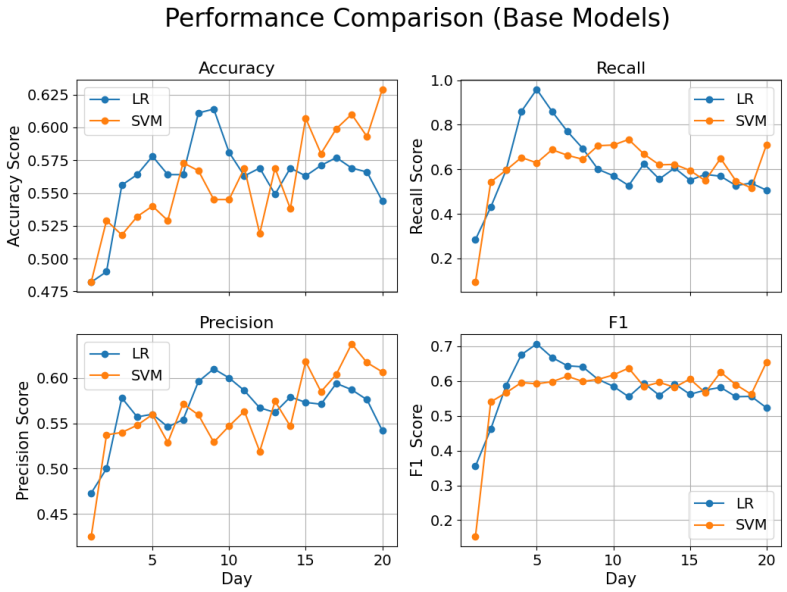
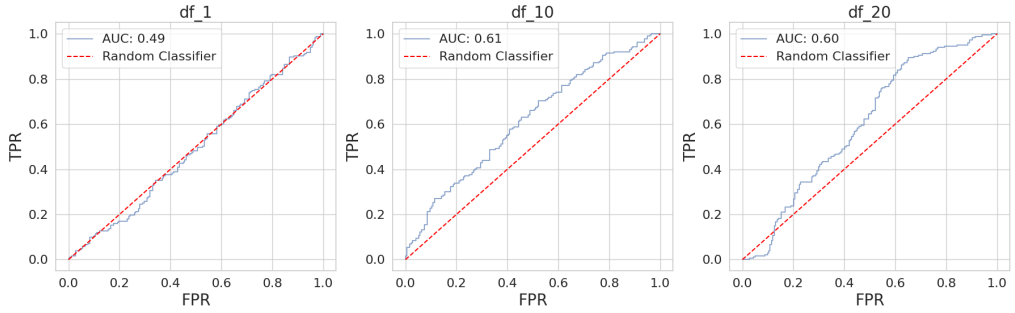
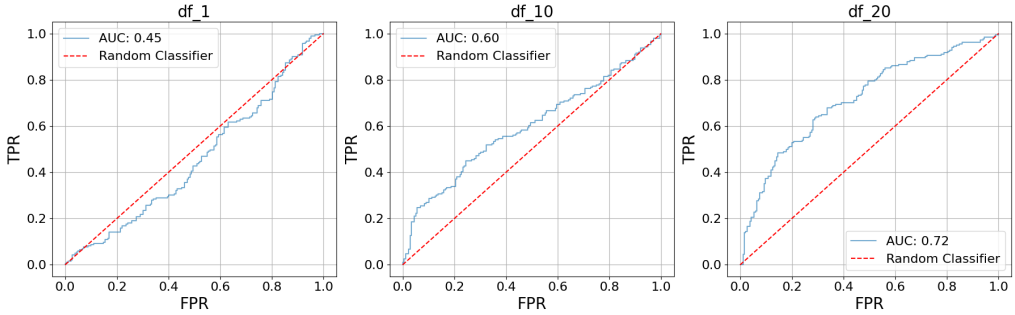


Figure 5.2: Base Models Whole Dataset



(a) LR



(b) SVM

Figure 5.3: ROC Base Models

Figure 5.3 shows ROC curves and the AUC for the base models. In the case of the LR, the model struggles to predict the 1-day ahead outcomes. The result is that the ROC curve closely follows the Random Classifier (RC), which means that the model performs no better than random guessing. The model shows improvements in performance for the mid- and long term. The 10-day AUC climbs to 0.61 and is always above the diagonal line. The 20-day AUC is similar (0.60), and is almost always above the RC.

For the SVM, the 1-day ROC curve is below the RC, with a low AUC (0.45). The opposite occurs in the mid- and long term. The 10-day and 20-day ROC curves are always above the diagonal line. The 10-day AUC has a similar score (0.60) in comparison to that of the Logistic Regression. The big improvement happens for the 20-days ahead predictions. The 20-day ROC curve is closer to hugging the top left corner, indicating a TPR and a low FPR across the thresholds. The 20-day AUC (0.72) is the highest score across the timeframes. On top of that, the 20-day SVM AUC score is the highest compared to the other ML models.

Tree models Figure 5.4 shows the evaluation metrics for the RF, XGB and LGB models. The x-axis again denotes the forecast horizon, ranging from 1 to 20 days, and the y-axis shows the corresponding scores.

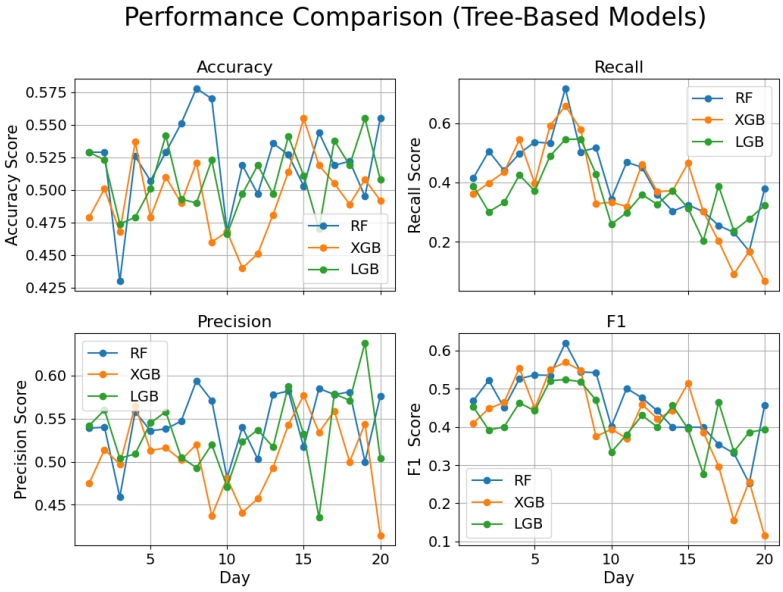
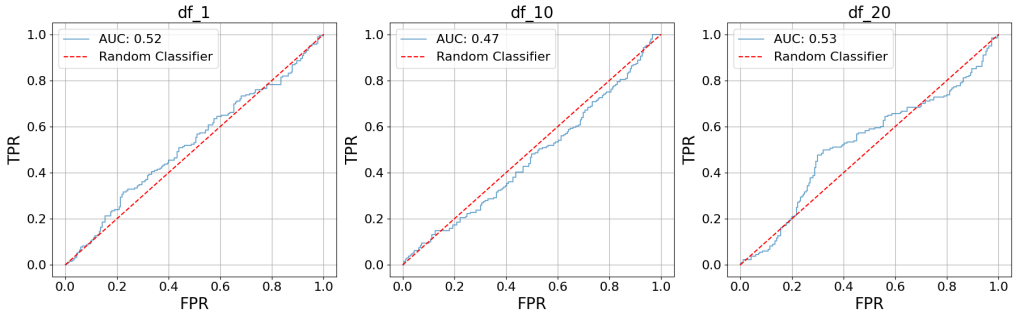
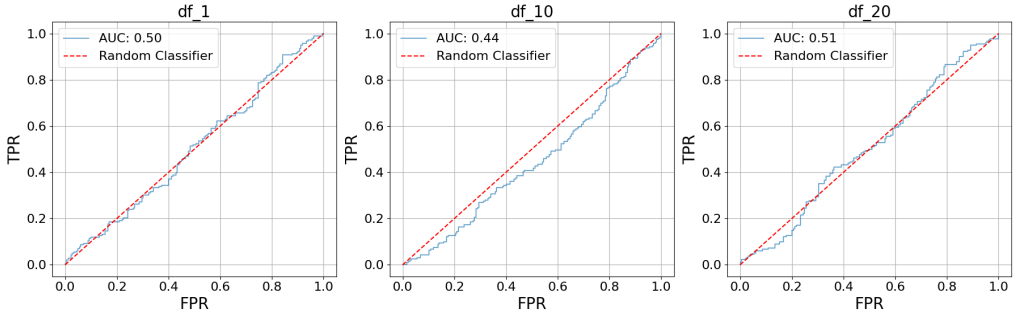


Figure 5.4: Tree Models Whole Dataset

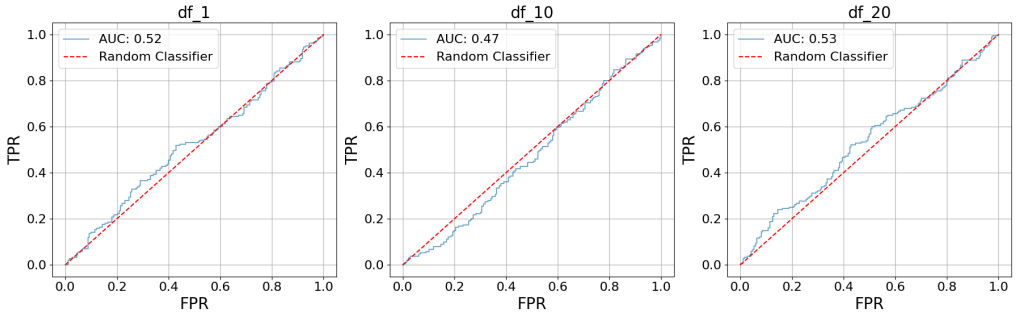
It is hard to see an upward trend for the accuracy and it fluctuates considerably as the time horizon increases in Figure 5.4. The highest accuracy is achieved by the RF for the 8 day forecasts. The recall worsens for all tree models as the time horizon increases. This volatility is likely due to overfitting the given features and shows a lack of robustness. The XGB f1 score shows a downward trend, with a low F1 (0.115). The same can be seen for the RF model, but surprisingly shows a sudden jump for the 20 day F1 score (0.456).



(a) RF



(b) XGB



(c) LGB

Figure 5.5: ROC Tree Models

Figure 5.5 shows the ROC curves and AUC for the tree models. Although very close, the 1-day ROC curve is above the diagonal line for the RF with an AUC of 0.52. The AUC and ROC curve worsen for the medium term, and is below the diagonal line. The long term horizon AUC (0.53) is slightly better than the random chance model, and shows a modest lift above the baseline. The 1-day XGB ROC closely follows the diagonal line, which is reflected in the AUC (0.50). The 10-day ROC curve fails to rise above the chance threshold and has an AUC of 0.44. The 20-day XGB mirrors random chance with an AUC

Of 0.51. Following Figure 5.5c, the LGB model shows a pattern similar to the other tree models. Again, the 1-day ROC curve hugs the baseline with an AUC of 0.52. The 10 day AUC (0.47) is similar to that of the RF model, with a ROC curve slightly below the diagonal line. The 20 day LGB AUC score has the same value as the RF, but shows a more consistent ROC curve that is almost always above the diagonal line.

Table 5.1: Performance whole dataset (1-day)

Model	Accuracy	Recall	Precision	F1-score
LR	0.482	0.284	0.473	0.355
SVM	0.482	0.093	0.425	0.153
RF	0.529	0.415	0.539	0.469
XGB	0.479	0.361	0.475	0.410
LGB	0.529	0.388	0.542	0.452

Table 5.1 shows that RF and LGB have the highest accuracy (0.529) for very small time horizons. RF slightly outperforms LGB when it comes to recall and F1-score. The base models exhibit poor performance, with SVM showing notable outliers for the recall (0.093) and f1 (0.153).

Table 5.2: Performance whole dataset (10-day)

Model	Accuracy	Recall	Precision	F1-score
LR	0.581	0.571	0.600	0.585
SVM	0.545	0.709	0.547	0.618
RF	0.468	0.344	0.481	0.401
XGB	0.468	0.333	0.481	0.394
LGB	0.466	0.259	0.471	0.334

Table 5.2 shows the mid-term case, and here the LR and SVM have the highest accuracies, respectively, 0.581 and 0.578. Moreover, the LR and SVM demonstrate superior performance relative to the other models in various evaluation criteria. Comparing Tables 5.1 and 5.2, the RF and LGB models are performing worse for the mid term case than for the short term horizon across all evaluation metrics. The most pronounced improvement is that of SVM's recall (0.709) and F1 score (0.618).

For the long term period in Table 5.3, again the LR and SVM show better performance than the tree models among the evaluation measures. In particular, XGB shows weak performance, with a recall of 0.067 and an F1 score of 0.115 in Table 5.3. Looking at Tables

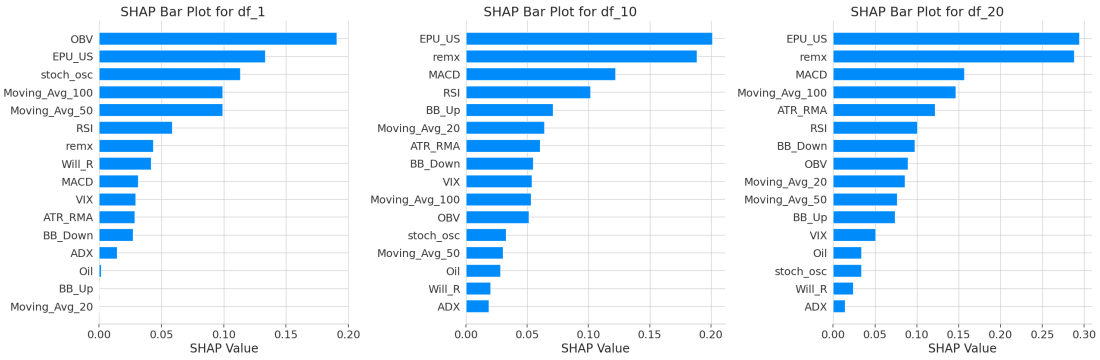
Table 5.3: Performance whole dataset (20-day)

Model	Accuracy	Recall	Precision	F1-score
LR	0.544	0.506	0.542	0.523
SVM	0.629	0.711	0.607	0.655
RF	0.555	0.378	0.576	0.456
XGB	0.492	0.067	0.414	0.115
LGB	0.508	0.322	0.504	0.393

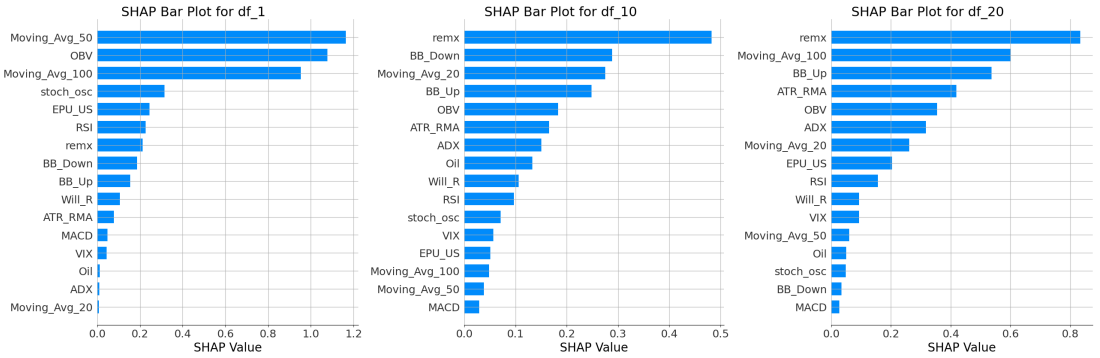
5.2 and 5.3, SVM is again improving in every aspect when it comes to the performance metrics. SVM has the best accuracy across all other time horizons, but is also the best method among the other ML models.

5.1.2 Feature Importance

Base models Figure 5.6 shows the SHAP values for Logistic Regression and Support Vector Machines. The x-axis denotes the SHAP value, and the y-axis all features ranked on importance. In the case of LR, the top three features for the short time horizon are the OnBalanceVolume indicator, EPU and stochastic oscillator. For both the mid-term and long term, the EPU, REMX and MACD are important. ADX and Will_R are at the bottom of the SHAP feature contributions for the mid- and long term. Interestingly, oil is one of the bottom features for predicting directional movements for all forecast time horizons.



(a) LR



(b) SVM

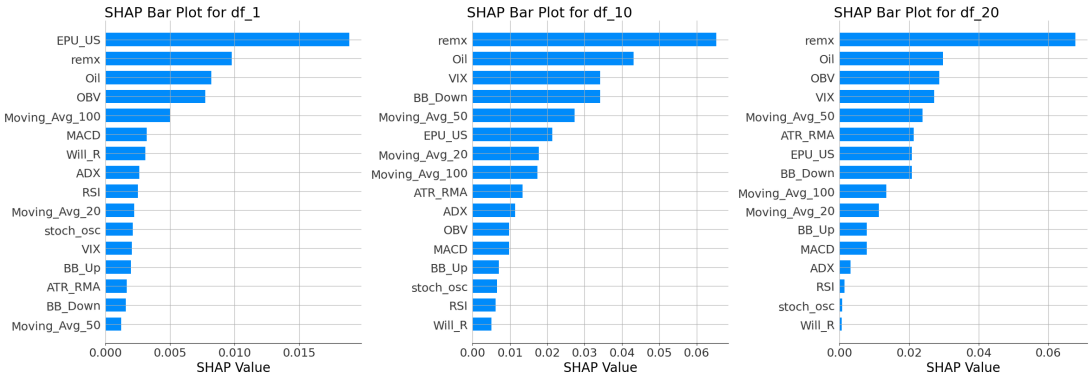
Figure 5.6: Feature Importance Base Models

Looking at Figure 5.6b the 50-MA, OBV and 100-MA are the most important features for the short term. Like the LR, REMX is a top three feature for the mid- and long-term. Furthermore, Oil is ranked third bottom for the 20-day forecast. What stands out, is that the EPU can be found in the midfield for the SVM, whereas it was the top feature for the LR for the long term.

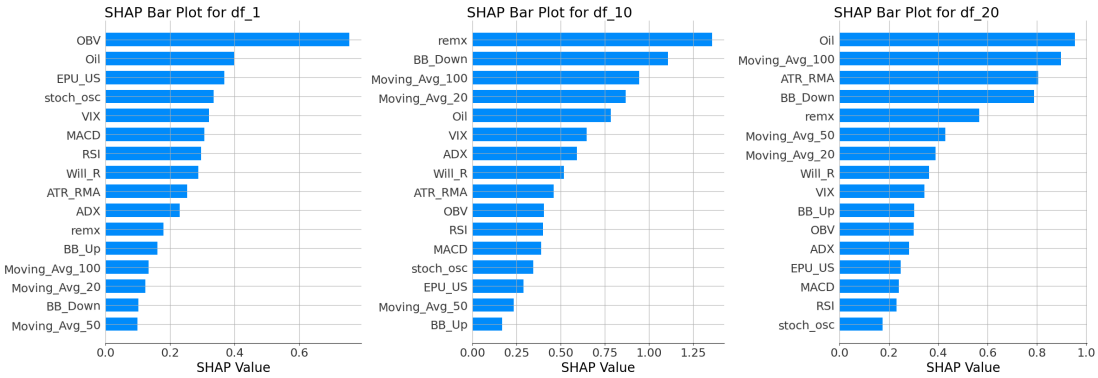
Tree models Figure 5.7 shows the importance of features for the tree models. RF highly ranks the extra added features like EPU, REMX and oil for the short time horizon. For the midterm, the REMX, Oil and VIX are the top three features, again with very low SHAP values. In the long term, the top three is slightly different, now consisting of REMX, oil and OBV.

XGB, just as LR and SVM, highly rates OBV for 1-day predictions. Oil and EPU complete the top three as the most important features. REMX is the main feature, together with BB and 100-MA for the mid-term. Surprisingly, the ATR makes a first appearance as a top 3 feature for the 20-day forecast, with oil and 100-MA at the first and second positions.

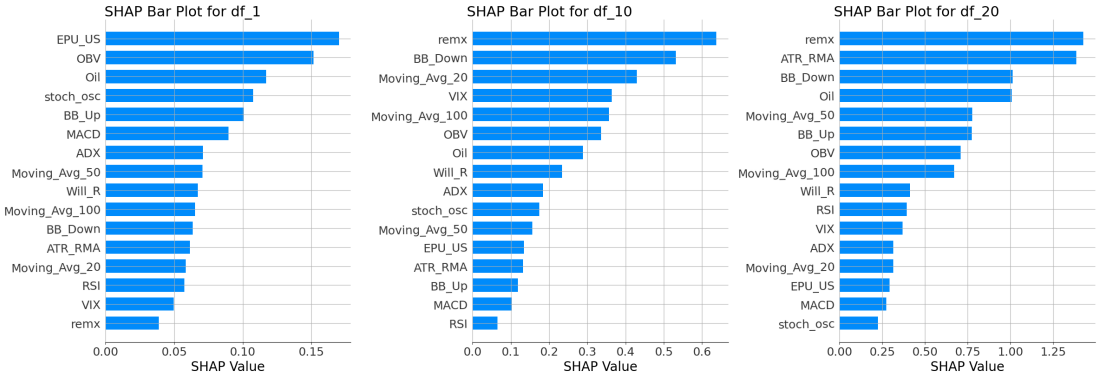
LGB has the same top three as the XGB model for the short-time horizon, but here the order is a bit different. Number one is EPU, followed by OBV and oil. REMX is the bottom predictor for the 1-day horizon, but jumps to the number 1 spot in the mid-term horizon, where BB down and 20-MA complete the top 3. VIX is the second bottom predictor for the short term forecasts, but also makes a jump to fourth place in the mid term. Like the other boosting method in Figure 5.7b, the ATR is included in the top three for the 20-day horizon. REMX remains the top predictor, and in the third position is BB down.



(a) RF



(b) XGB



(c) LGB

Figure 5.7: Feature Importance Tree Models

5.2 Sliding Window

5.2.1 Evaluation Metrics

The results for each model of the sliding window come from training the data on the first 600 data points and testing it on the next case. The 600 instances are chosen because they amount to roughly 2.5 years, which seems like a reasonably long period for training the data. The hyperparameters are chosen via a majority vote following Section 5.1.

Base models Figure 5.8 shows the accuracy, recall, precision and F1 score for the LR and SVM models. Again, the x-axis illustrates the forecast horizon, ranging from 1 to 20 days. One key difference compared to Figure 5.2 with the whole dataset setup, is that the upward trend is clearer and more pronounced. Both models start at around 50% accuracy for the 1-day horizon and rapidly rise as the time increases. The accuracy of the LR only goes to 66.2%, whereas the 20-day SVM's accuracy is 83.3%. Across the evaluation metrics, SVM dominates and is in almost every case the better model. Interestingly, the recall and F1 score of LR top out after 5 days. In contrast, the recall and F1 score of SVM show no signs of plateauing, reaching a score of 86% and 84.8% for the 20-day forecast horizon, respectively.

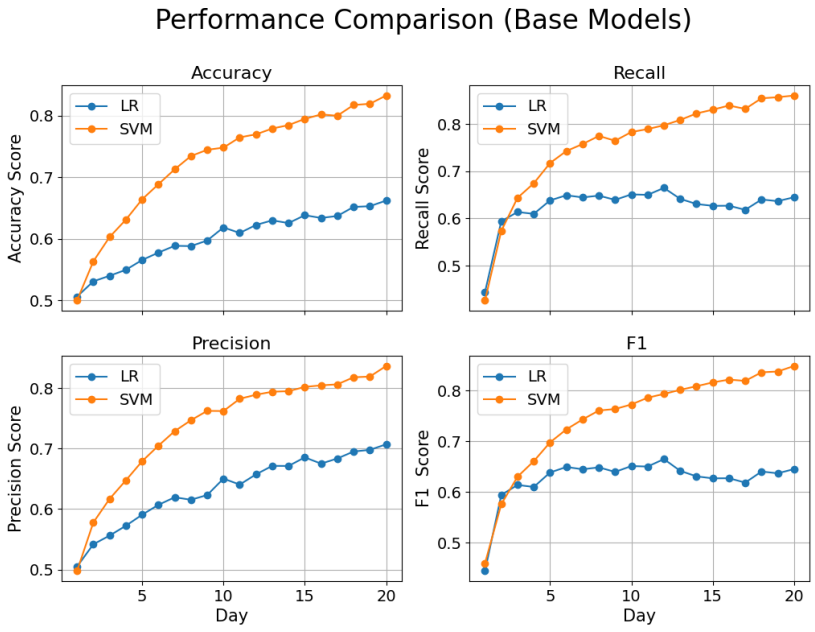
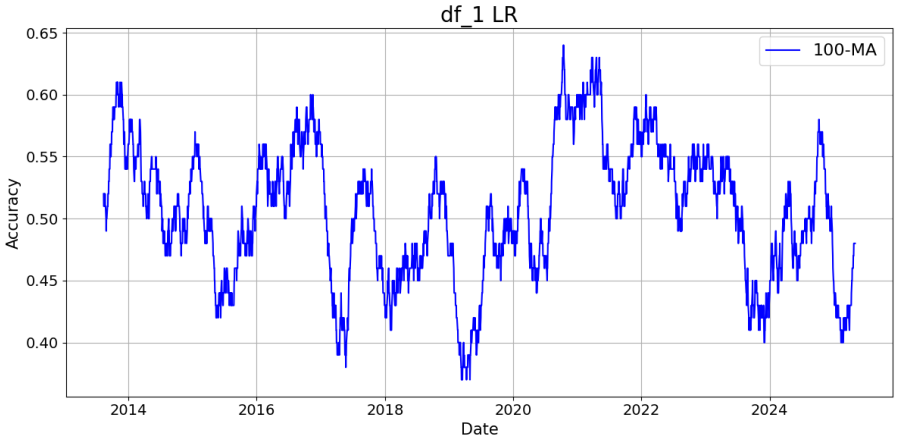
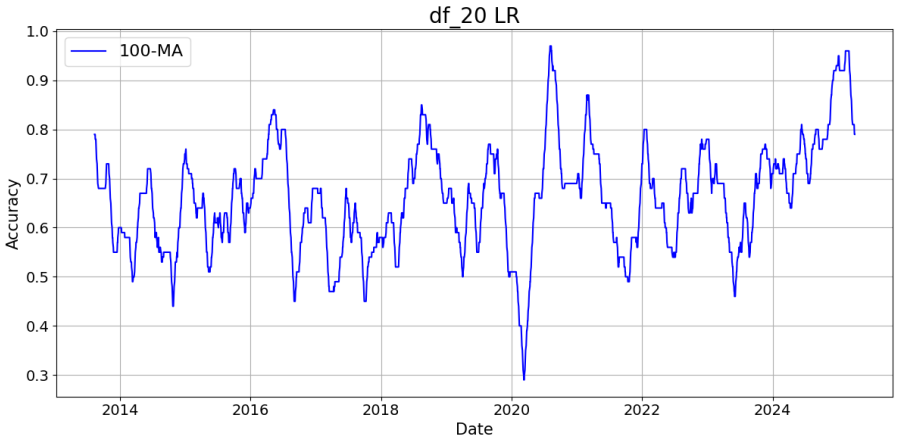


Figure 5.8: Base Models Sliding Window

Figure 5.9 shows the 100-MA accuracy for the 1 day and 20 day forecast horizon for the LR model. The 1-day smoothed accuracy falls in the range of 37% and 64%, whereas the smoothed accuracy is between 29% and 97% for the 20 day ahead predictions. Figure 5.9b shows that the model underperformed when Covid started. Except for this unusual time period, the smoothed 20-day accuracy is almost never below 50%.

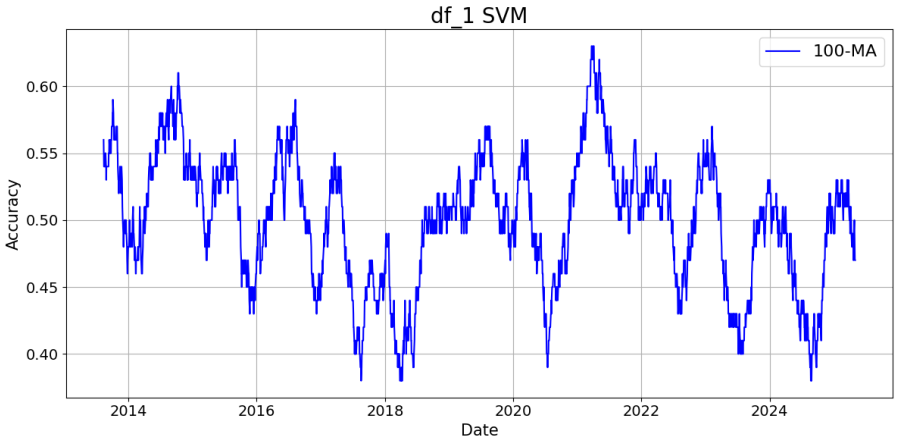


(a) 1-day

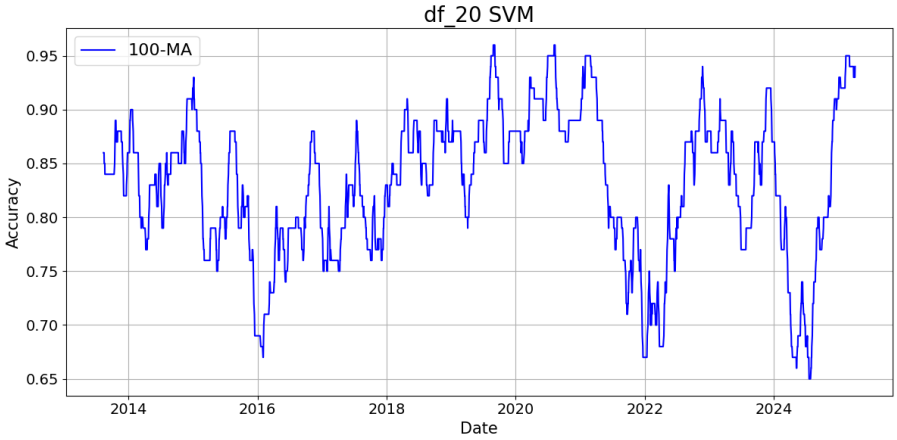


(b) 20-day

Figure 5.9: LR Smoothed Accuracy



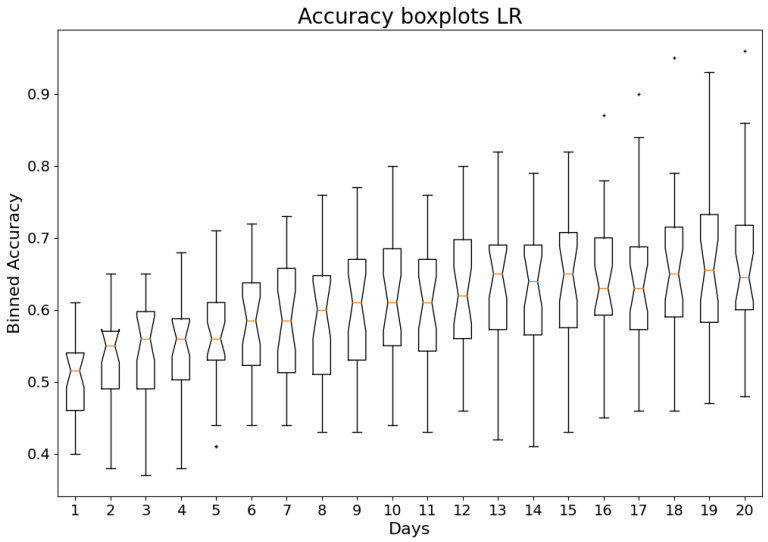
(a) 1-day



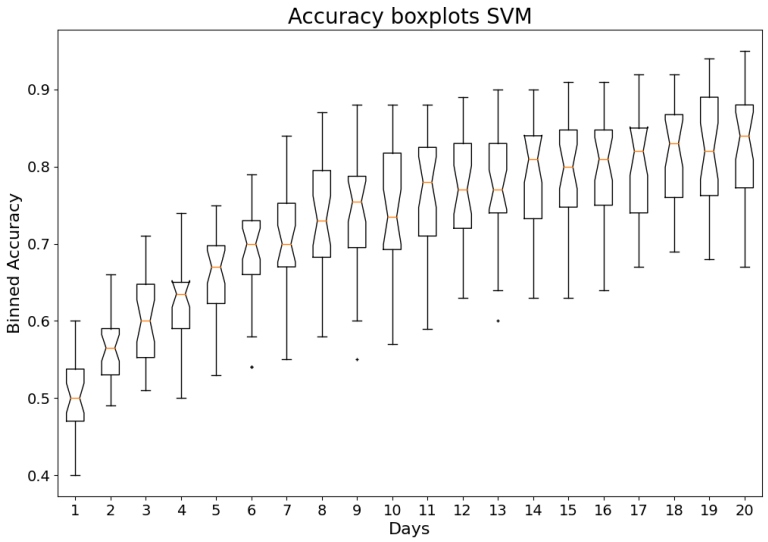
(b) 20-day

Figure 5.10: SVM Smoothed Accuracy

Figure 5.10 illustrates the 100 MA accuracy of the SVM model for forecast horizons of 1 day and 20 days. The 1-day smoothed accuracy falls in the range of 38% and 63%, whereas the smoothed accuracy is between 65% and 96% for the 20 day ahead predictions. Figure 5.10b shows that the model had a difficult time predicting the up-down movements during the armed confrontation between Russia and Ukraine. Compared to the other base model, the 20-day smoothed accuracy never falls below 50%. Figure 5.11 shows the accuracy box plots for the base models. We can see from Figure 5.11a that as we increase the average accuracy improves, but this does come at the cost of a bigger spread. Figure 5.11b shows us that the accuracy increases for larger time horizons for the SVM, but the spread does not widen as much as the LR model.



(a) LR



(b) SVM

Figure 5.11: Base Model accuracy boxplots

Tree models Figure 5.12 shows the accuracy, recall, precision and F1 score for the RF, XGB, and LGB models. Across the evaluation metrics, we can see rapid growth as the forecast period increases. The three models do not differ much for very short forecast periods. The RF model performs slightly worse across the performance metrics compared

to the other tree models when looking at bigger time horizons.

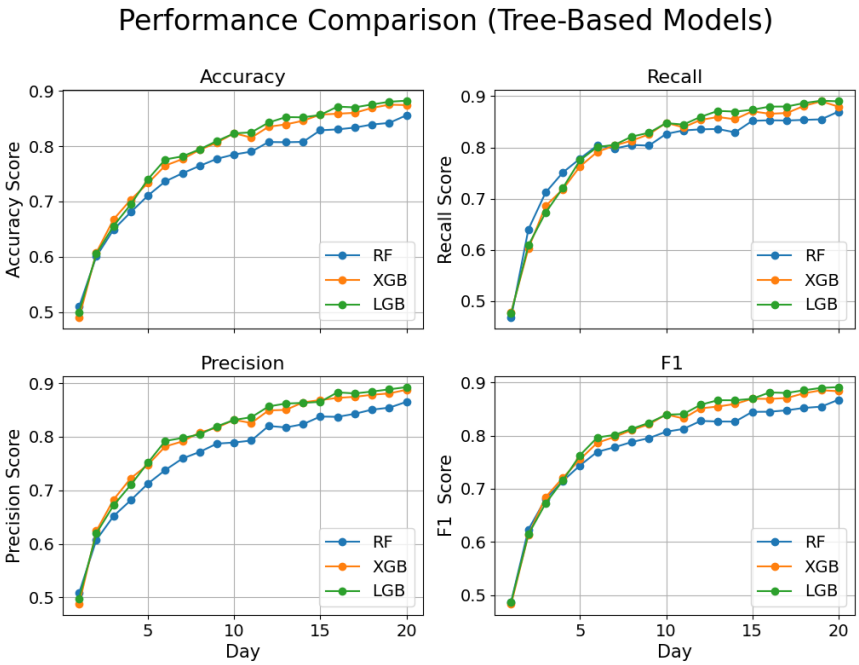
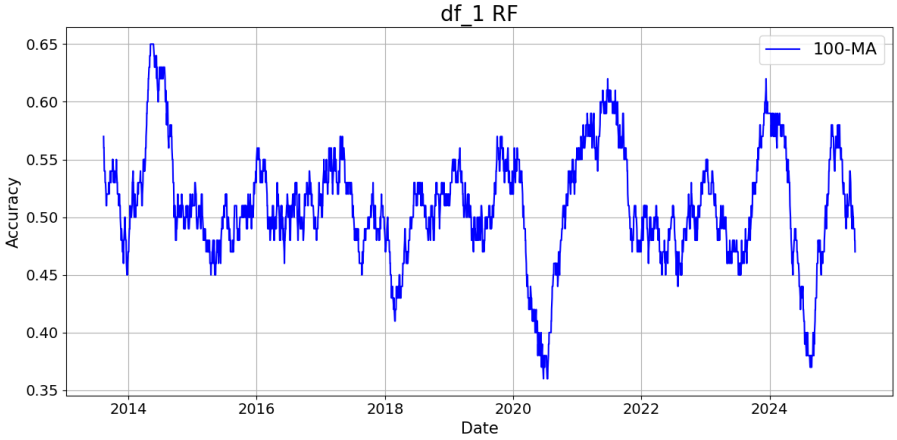
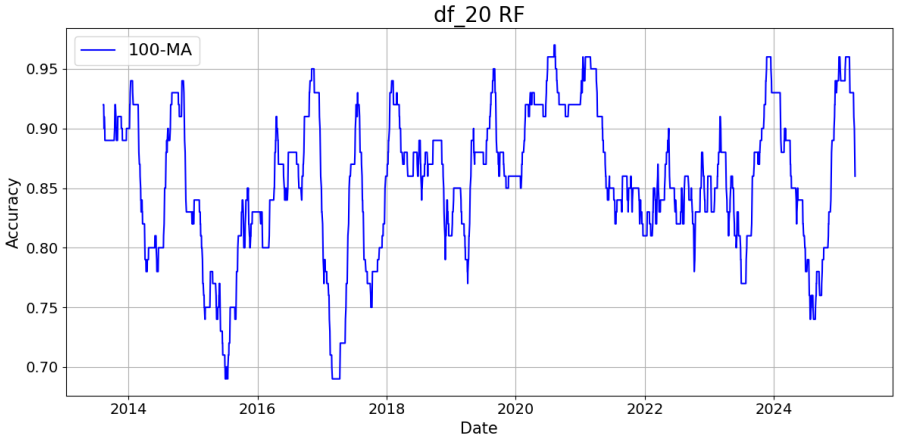


Figure 5.12: Tree Models Sliding Window

Figure 5.13 shows the 100-MA accuracy for the 1 day and 20 day forecast horizon of the RF model. The 1-day smoothed accuracy ranges from 36% to 65%, whereas the smoothed accuracy is between 69% and 97% for the 20 day ahead predictions. Figure 5.13a shows that the model's ability declined when time predicting the 1 day up-down movements when Covid started.



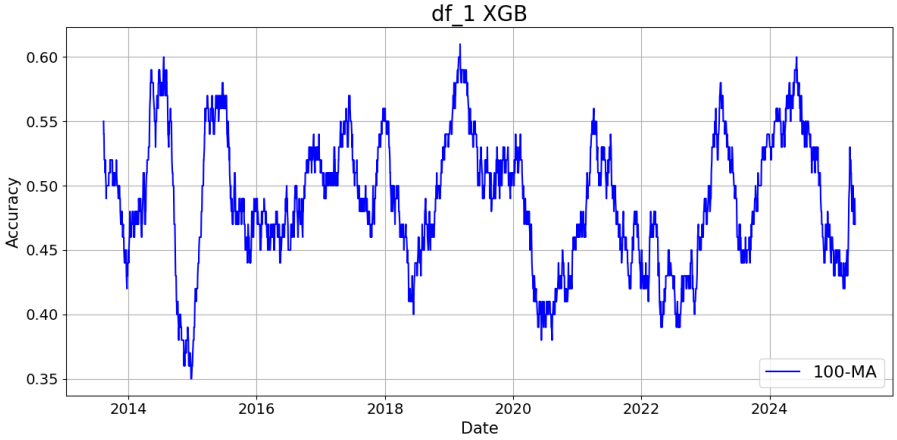
(a) 1-day



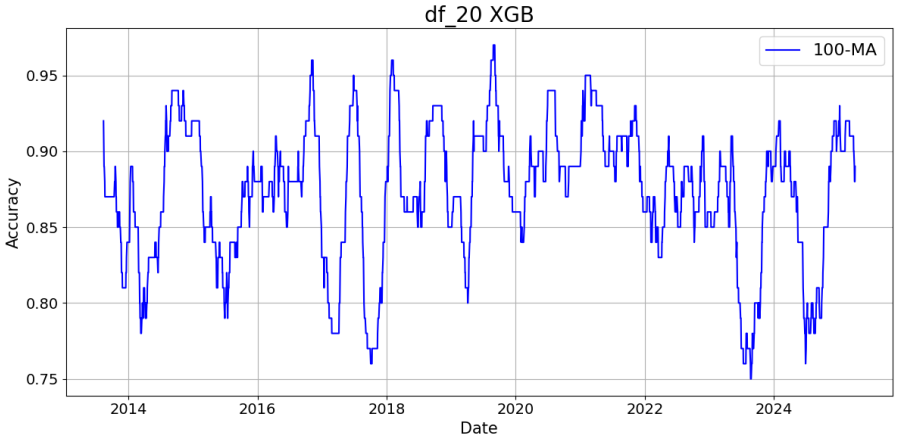
(b) 20-day

Figure 5.13: RF Smoothed Accuracy

Figure 5.14 shows the smoothed accuracy for the short- and long term forecast of the XGB model. The 1 day accuracy is in the bracket of 35% and 61%, whereas the 20 day accuracy falls in the bracket of 75% and 97%. XGB seems to underperform when predicting one day ahead predictions at the turn of the year 2015 and the beginning of Covid.



(a) 1-day

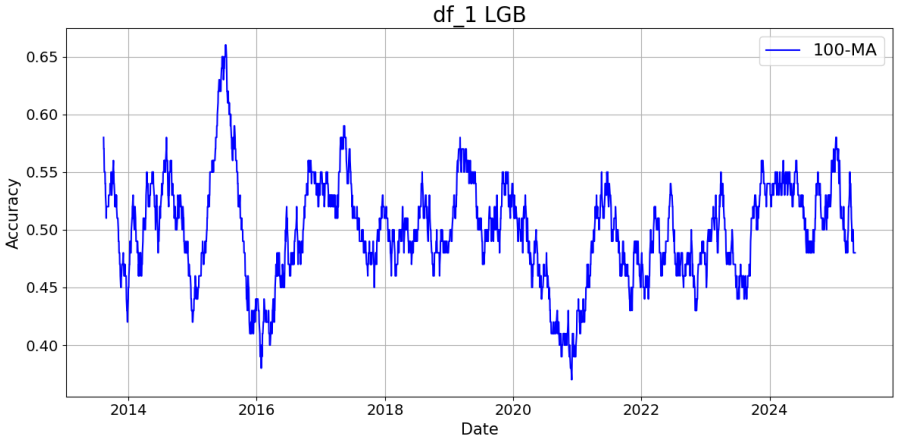


(b) 20-day

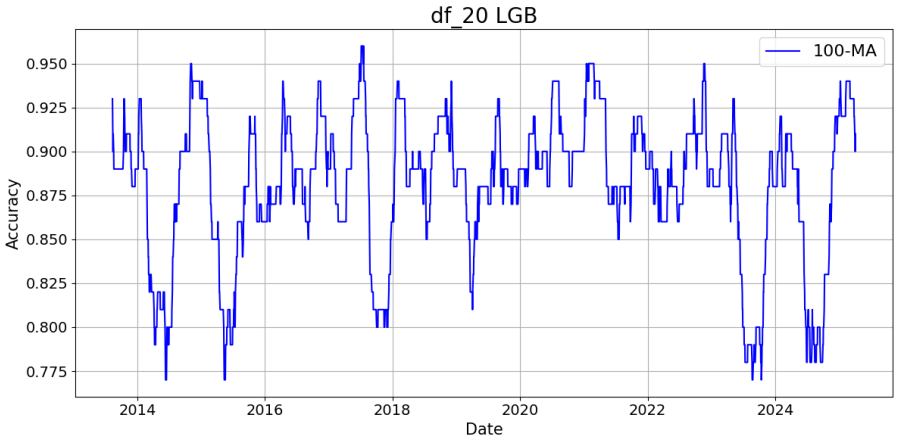
Figure 5.14: XGB smoothed Accuracy

Figure 5.15 shows the 100-MA accuracy for the short- and long term of the LGB model. The 1 day accuracy is in the spectrum of 37% and 66%, quite similar to that of XGB. The 20 day accuracy has a minimum bound of 77% and an upper bound of 96%. LGB showed the weakest performance during the turn of the years 2016 and 2021.

Figures [5.16,5.17,5.18] show the accuracy box plots for the tree models. The three models show a similar pattern, namely, the median of the 1-day forecast is close to 50% and rapidly increases to 90%. Upon visual inspection, the spread in the RF models is greater than the gradient boosting methods. The accuracy boxplots of the LR model stand out negatively in comparison with the tree models, as the median of the accuracy improves marginally, while the spread gets bigger.



(a) 1-day



(b) 20-day

Figure 5.15: LGB smoothed Accuracy

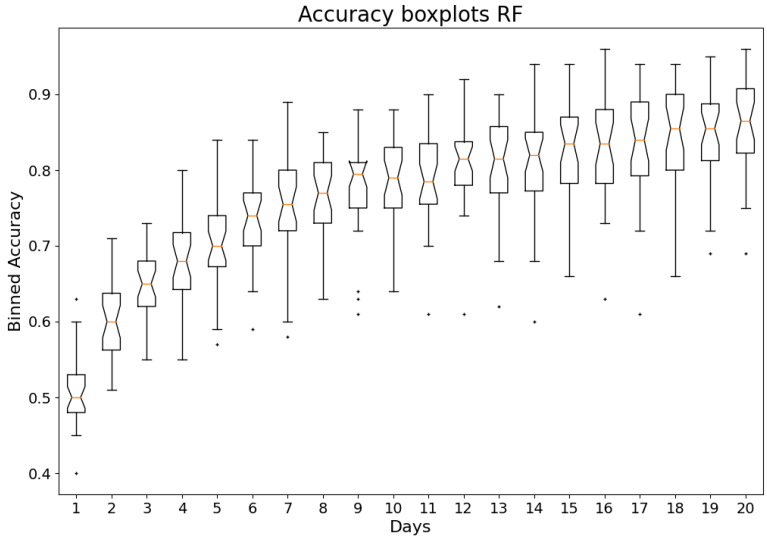


Figure 5.16: Accuracy Boxplot: RF

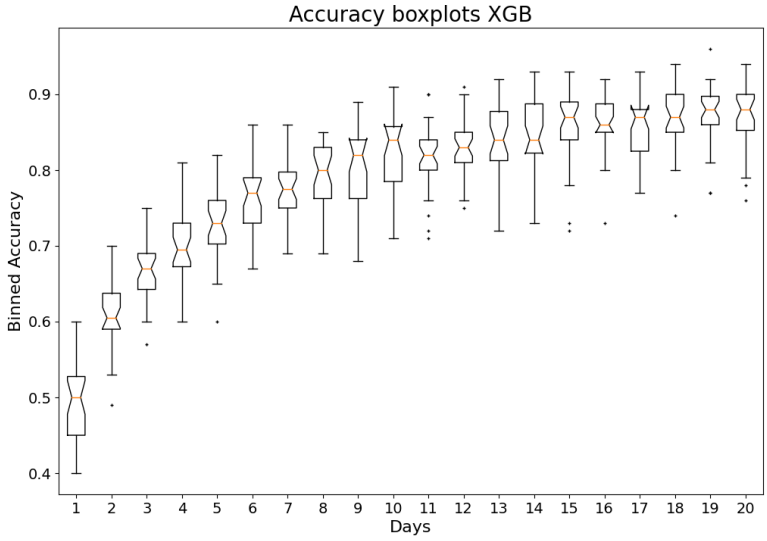


Figure 5.17: Accuracy Boxplot: XGB

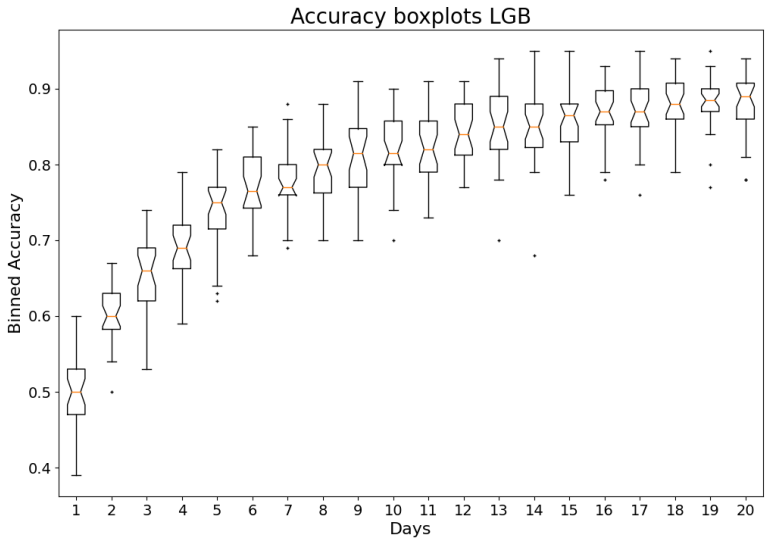


Figure 5.18: Accuracy Boxplot: LGB

Table 5.4 shows that all models are closely matched, with similar scores for evaluation metrics. All accuracies are around 50%, indicating that the models are not better than a random guessing method.

Table 5.4: Performance sliding window (1-day)

Model	Accuracy	Recall	Precision	F1-score
LR	0.506	0.444	0.505	0.444
SVM	0.500	0.427	0.498	0.459
RF	0.510	0.467	0.509	0.487
XGB	0.490	0.478	0.488	0.483
LGB	0.500	0.476	0.498	0.487

Table 5.5 shows notable improvements, especially for the gradient boosting algorithms. Both XGB and LGB have an accuracy of 82.3%, and demonstrated strong performance across the remaining evaluation metrics. LR lags behind, with an accurracy of 61.8%
Table 5.6 again shows notable improvements, with LGB achieving the highest accuracy of 88.2%, closely followed by the other tree models. SVM also has a respectable accuracy score of 83.3%. LR is again the worst performer among the 5 ML models, with an accuracy of 66.2%.

Table 5.5: Performance sliding window (10-day)

Model	Accuracy	Recall	Precision	F1-score
LR	0.618	0.651	0.650	0.651
SVM	0.748	0.783	0.762	0.772
RF	0.785	0.826	0.789	0.807
XGB	0.823	0.848	0.832	0.840
LGB	0.823	0.847	0.831	0.839

Table 5.6: Performance sliding window (20-day)

Model	Accuracy	Recall	Precision	F1-score
LR	0.662	0.645	0.707	0.645
SVM	0.833	0.860	0.836	0.848
RF	0.856	0.870	0.865	0.868
XGB	0.874	0.880	0.888	0.884
LGB	0.882	0.890	0.893	0.891

6

Conclusion

The purpose of this study is to analyze the wind energy sector up-down movements for various forecast horizons using ML. In addition to this, we want to know which features contribute the most when classifying these movements. Five ML models were explored and SHAP values are used to explain the model's decisions. The results were divided in two setups, namely whole dataset and sliding window. Furthermore, the models are divided in two categories: base models (LR, SVM) and tree models (RF, XGB, LGB). Before answering the research question and sub question, a brief summary is given.
The research question that arises is as follows:

To what extent can future up-down movements be predicted in the wind energy sector using Machine Learning?

Whole Dataset
The results following Section 5.1 showed that for the short time horizon (1 day), the tree models RF and LGB were slightly better. The evaluation metrics remained suboptimal, achieving no higher accuracy score than 0.529. This goes to show that predicting short term price movements in this sector is quite difficult. The base models LR and SVM show higher performance in comparison with the tree models for the 10 days ahead predictions. In long term predictions (20 days), the SVM showed superior performance across all other ML models for all evaluation metrics. The SVM's ROC curve hugs the top left corner the most, and the corresponding AUC score (0.72) is the highest among all models and across all forecast horizons.

Sliding Window

This method used around 2.5 years of training data to predict the next case. After this, we slide this setup 1 day, resulting in more than 3000 simulations. Again, for the 1 day predictions, all models showed performance, with an accuracy comparable to random guessing. The real difference occurs when the forecast horizons are expanded. The tree models, in particular the gradient boosting methods showed the highest performance for the mid term (10 days). In the long term (20 days), the LGB took first place, closely followed by XGB and RF, with accuracies nearing 90%.

Projects in the wind energy sector typically extend over several months or even years. Applying a 20 day forecast horizon for the ML models is most likely a too short period for assessing the sentiment of the wind energy industry. Using the whole dataset setup, SVM achieved the highest performance in predicting long term movements, with an accuracy of 0.629, F1 score of 0.655, and an AUC of 0.72. Despite outperforming others, the metrics are below the commonly accepted benchmarks to be considered as an effective and reliable model.

Having discussed the whole dataset setup, we now focus on the sliding window method. The gradient boosting methods showed the highest scores among the ML models, with an accuracy of 0.882 and F1 score of 0.891 for the LGB for longterm predictions. However, caution is required when evaluating the model outputs. As a result of increasing the time horizon, the class imbalance increases as well. Furthermore, as the forecast horizon increases, the streak of the FAN etf also gets larger. Here, a streak refers to a consecutive sequence of periods where the etf moves in the same direction, either upward or downward, without reversing. Most likely, the ML models suffer from this class imbalance and exploit the "streak effect" matter when using the relatively short time frames in the sliding window setup.

In addition, the sub question is formulated as follows:

What are the important factors that contribute to the prediction of the ML models?

Following the results from the five ML models, we want to see which features are important when predicting up-down movements. To answer this question, we compute SHAP values to turn the black box models into glass box models. In addition to the technical indicators,

we added the uncertainty measures and the commodities oil and REMX to see if these have predicting power. The SHAP values were calculated using the whole dataset setup. For the short term, RF and LGB were the highest performing models, and in both models, the economic policy uncertainty (EPU) of the US is the highest rank based on Figure 5.7. Interestingly, whereas REMX is the 2nd highest feature for the RF, it is the bottom predictor for the LGB model. With both models achieving an accuracy of 0.529, it is difficult to rely on these models and say that the EPU is the main predictor.

Now we shift our attention to the SHAP values of the mid term forecast horizon. LR has the highest accuracy, followed by the SVM, respectively 0.581 and 0.545. Both models agree that REMX is in the top 3, but for the rest, the algorithms do not agree on the feature importance rankings. This makes it very hard to make a strong case to pinpoint the main features for the 10 day predictions.

Lastly, we consider the case where the models predict longterm movements in the wind energy sector. As answered in the research question, the SVM is the highest performing model across the evaluation metrics for the whole dataset setup. Here, the top feature is REMX, followed by MA_100 and BB_Up. The uncertainty measures EPU and VIX are in the middle-pack. Surprisingly, oil is the fourth bottom feature for the 20 day forecast horizon. In almost all cases, REMX is in the top three, suggesting that REMX is a leading indicator, though validation experiments are required to be certain.

7

Discussion

7.1 Implications

This study has several implications after the results in Section 5. For researchers in the wind energy sector, the following recommendations are highlighted to improve current financial models.

Firstly, the choice of data is very important. This research used publicly available financial information, namely the FAN etf, the REMX etf, oil price, etc. The FAN etf consists of several companies active in the wind energy sector. The list includes wind turbine manufacturers, wind park developers, and wind power operators. The benefit of using this method is that the data is smoother and more predictable, making it easier for ML models to detect patterns. The downside is that this dataset can be too general, as you can have multiple branches within the wind energy sector.

The REMX etf closely follows companies that are active in mining Rare Earth Minerals, spanning different countries, such as China, Chile and Australia. This does not necessarily mean that the price of minerals is reflected in that basket of mineral companies. A suggestion would be to download data for various rare earth indexes to use as a feature in the models. These can be a proxy for geopolitical tensions and can help forecast the wind energy sector sentiment.

Secondly, keeping the time order intact when dealing with timeseries data is crucial. Some other researchers fall into the trap of shuffling data points around. The consequence is that the models suffer from temporal leakage. Performance on several evaluation metrics are overinflated, and the SHAP values give the wrong feature importance. Without rigorous validation, an overreliance on the models may lead to the situation that stakeholders

are wrongly informed about the future. If models predict strong growth incorrectly, for instance, it could prompt misinformed investment strategies.

7.2 Recommendations and Future Work

In this Section, we go over the limitations of the study and some recommendations for future research.
This study forecasts the FAN etf for varying time frames, namely the 1-day forecast, up to 20-day ahead predictions. This analysis could be beneficial for financial traders and portfolio managers that are active in this branch and work with approximately the same timeframe. Wind energy companies work on a project-by-project basis, spanning multiple months and sometimes even years. This analysis could be extended to incorporate forecasts for multiple months. In this way, the models forecast on a longer time horizon and SHAP shows which features are important according to this new time frame.
In addition, this research only included numerical data for the forecasts. In many other studies, news articles are included in predictive models. In that case, sentiment analysis is performed to assess the mood of the message. The piece of text could fall into three categories, namely positive, negative, or neutral. News articles about wind energy companies could be scraped to be used in addition to existing numerical features. Training could be done manually, by an expert that can label whether a news article is positive or negative for the wind energy sector. This can be labor intensive, and a different approach could be to use X or Reddit datasets to train ML models for sentiment analysis.

Furthermore, the simulation setup with a sliding window could result in imbalanced datasets. This effect is amplified by increasing the forecasting time period. The two main ways of handling this matter are as follows. The first solution is to increase the number of data points in each simulation. In this way, the chance of having imbalanced datasets is reduced. The alternative solution is to leave out the simulation setup and immediately do the classification on the entire dataset. This ensures a relatively balanced dataset, and additionally has the advantage of being more interpretable.
The latest suggestion for future research is to explore other methods than ML models, such as Autoregressive (AR), ARIMA, and Seasonal Autoregressive Integrated Moving Average eXogenous (SARIMAX). The latter model has great potential, as it can detect seasonal patterns, and the addition of exogenous factors makes it so that external variables can be

added. External variables that influence the time series, such as commodities and inflation, can be included and explored to see if there is a connection.

References

[1] European Commission. **The North Seas Energy Cooperation**. `https://energy.ec.europa.eu/topics/infrastructure/high-level-groups/north-seas-energy-cooperation_en`, 2025. Accessed: 21 May 2025. 1

[2] M Dolores Esteban, J Javier Diez, Jose S López, and Vicente Negro. **Why offshore wind energy?** *Renewable energy*, **36**(2):444–450, 2011. 1

[3] Pierre Tardieu Giuseppe Costanzo, Guy Brindley. **Wind energy in Europe - 2024 Statistics and the outlook for 2025-2030**. Technical report, WindEurope, 2024. 1

[4] Judith Malkiel Burton G. Malkiel and Brendan Curry. *A Random Walk Down Wall Street*. W. W. Norton & Company, 1973. 2, 4

[5] Eugene F Fama. **Efficient capital markets**. *Journal of finance*, **25**(2):383–417, 1970. 2

[6] Fama Eugene and Kenneth French. **The cross-section of expected stock returns**. *Journal of finance*, **47**(2):427–465, 1992. 4

[7] Andrew W Lo and A Craig MacKinlay. **Stock market prices do not follow random walks: Evidence from a simple specification test**. *The review of financial studies*, **1**(1):41–66, 1988. 5

[8] Narasimhan Jegadeesh and Sheridan Titman. **Returns to buying winners and selling losers: Implications for stock market efficiency**. *The Journal of finance*, **48**(1):65–91, 1993. 5

[9] Robert A Levy. **Relative strength as a criterion for investment selection**. *The Journal of finance*, **22**(4):595–610, 1967. 5

[10] Michael C Jensen and George A Benington. **Random walks and technical theories: Some additional evidence**. *The Journal of finance*, **25**(2):469–482, 1970. 5

[11] Becca Cattlin. **What is the efficient market hypothesis (EMH)?** `https://www.ig.com/en-ch/trading-strategies/what-is-the-efficient-market-hypothesis--emh---191217`, 2019. Accessed: 17 December 2019. 5

[12] Jian Cao, Zhi Li, and Jian Li. **Financial time series forecasting model based on CEEMDAN and LSTM**. *Physica A: Statistical mechanics and its applications*, **519**:127–139, 2019. 5

[13] Henrik Amilon. **GARCH estimation and discrete stock prices: an application to low-priced Australian stocks**. *economics letters*, **81**(2):215–222, 2003. 6

[14] Jean-Thomas Bernard, Lynda Khalaf, Maral Kichian, and Sebastien McMahon. **Forecasting commodity prices: GARCH, jumps, and mean reversion**. *Journal of Forecasting*, **27**(4):279–291, 2008. 6

[15] Shakir Khan and Hela Alghulaiakh. **ARIMA model for accurate time series stocks forecasting**. *International Journal of Advanced Computer Science and Applications*, **11**(7), 2020. 6

[16] Halbert White. **Economic prediction using neural networks: The case of IBM daily stock returns**. In *ICNN*, **2**, pages 451–458, 1988. 6

[17] Mark T Leung, Hazem Daouk, and An-Sing Chen. **Forecasting stock indices: a comparison of classification and level estimation models**. *International Journal of forecasting*, **16**(2):173–190, 2000. 6

[18] Kanghee Park and Hyunjung Shin. **Stock price prediction based on a complex interrelation network of economic factors**. *Engineering Applications of Artificial Intelligence*, **26**(5-6):1550–1561, 2013. 6

[19] Fischer Black and Myron Scholes. **The pricing of options and corporate liabilities**. *Journal of political economy*, **81**(3):637–654, 1973. 6

[20] Peter Carr. **Why is VIX a fear gauge?** *Risk and Decision Analysis*, **6**(2):179–185, 2017. 7

[21] Nicholas Bloom. **The impact of uncertainty shocks**. *econometrica*, **77**(3):623–685, 2009. 7

[22] Bevan J Blair, Ser-Huang Poon, and Stephen J Taylor. **Forecasting S&P 100 volatility: the incremental information content of implied volatilities and high-frequency index returns**. *Journal of econometrics*, **105**(1):5–26, 2001. 7

[23] Robert E Whaley. **The investor fear gauge**. *Journal of portfolio management*, **26**(3):12, 2000. 7

[24] Scott R. Baker, Nicholas Bloom, and Steven J. Davis. **Daily Trade Policy Uncertainty (7-day Moving Average)**. `https://policyuncertainty.com/`. Accessed: 2025-06-24. 7, 9

[25] Scott R Baker, Nicholas Bloom, and Steven J Davis. **Measuring economic policy uncertainty**. *The quarterly journal of economics*, **131**(4):1593–1636, 2016. 7

[26] Mohamed Arouri, Christophe Estay, Christophe Rault, and David Roubaud. **Economic policy uncertainty and stock markets: Long-run evidence from the US**. *Finance Research Letters*, **18**:136–141, 2016. 7

[27] Rangan Gupta, Godwin Olasehinde-Williams, and Mark E Wohar. **The impact of US uncertainty shocks on a panel of advanced and emerging market economies**. *The Journal of International Trade & Economic Development*, **29**(6):711–721, 2020. 7

[28] Viviana Fernandez. **Rare-earth elements market: A historical and financial perspective**. *Resources Policy*, **53**:26–45, 2017. 8

[29] U.S. Geological Survey. **The Rare-Earth Elements—Vital to Modern Technologies and Lifestyles**. Technical Report FS-2014-3078, U.S. Geological Survey, Mineral Resources Program, 2014. Fact Sheet 2014–3078. 8

[30] Wayne M Morrison and Rachel Tang. **China's rare earth industry and export regime: economic and trade implications for the United States**, 2012. 8

[31] Juliane Proelss, Denis Schweizer, and Volker Seiler. **The economic importance of rare earth elements volatility forecasts**. *International Review of Financial Analysis*, **71**:101316, 2020. 8

[32] Irene Henriques and Perry Sadorsky. **Forecasting rare earth stock prices with machine learning**. *Resources Policy*, **86**:104248, 2023. 8

[33] Cboe Global Markets. **Cboe Volatility Index (VIX)**. `https://www.cboe.com/tradable_products/vix/`. Accessed: 2025-06-24. 9

[34] Claude B Erb and Campbell R Harvey. **The strategic and tactical value of commodity futures**. *Financial Analysts Journal*, **62**(2):69–97, 2006. 15

[35] Ke Tang and Wei Xiong. **Index investment and the financialization of commodities**. *Financial Analysts Journal*, **68**(6):54–74, 2012. 15

[36] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. **A training algorithm for optimal margin classifiers**. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992. 15

[37] Tin Kam Ho. **Random decision forests**. In *Proceedings of 3rd international conference on document analysis and recognition*, **1**, pages 278–282. IEEE, 1995. 15

[38] Leo Breiman. **Random forests**. *Machine learning*, **45**:5–32, 2001. 15

[39] Tianqi Chen and Carlos Guestrin. **Xgboost: A scalable tree boosting system**. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016. 15

[40] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. **Lightgbm: A highly efficient gradient boosting decision tree**. *Advances in neural information processing systems*, **30**, 2017. 16

[41] Scott M Lundberg and Su-In Lee. **A unified approach to interpreting model predictions**. *Advances in neural information processing systems*, **30**, 2017. 16

Appendix

Table 7.1: Holdings of FAN_etf

Company	**Ticker**	**Weight (%)**
Vestas Wind Systems A/S	VWS.DC	7.40
Northland Power Inc.	NPI.CN	7.19
Orsted A/S	ORSTED.DC	7.16
EDP Renovaveis SA	EDPR.PL	6.35
Nordex SE	NDX1.GY	5.32
China Longyuan Power Group Corp Ltd (Class H)	916.HK	4.53
Boralex Inc.	BLX.CN	4.15
Century Iron & Steel Industrial Co., Ltd.	9958.TT	3.03
Siemens Energy AG	ENR.GY	2.78
ERG SpA	ERG.IM	2.69
Iberdrola S.A.	IBE.SM	2.39
RWE AG	RWE.GY	2.31
Enel SpA	ENEL.IM	2.29
GE Vernova Inc.	GEV	2.29
Terna Energy SA	TENERGY.GA	2.27
Engie S.A.	ENGI.FP	2.22
BKW AG	BKW.SW	2.16
Acciona, S.A.	ANA.SM	2.03
Hitachi, Ltd.	6501.JP	2.01
Arcosa, Inc.	ACA	1.94
Prysmian SpA	PRY.IM	1.89
Owens Corning	OC	1.86
Toray Industries, Inc.	3402.JP	1.86
NextEra Energy, Inc.	NEE	1.85
Alliant Energy Corporation	LNT	1.84
AB SKF (Class B)	SKFB.SS	1.78
The Timken Company	TKR	1.59
Hexcel Corporation	HXL	1.52
CS Wind Corp.	112610.KS	1.48

Company	Ticker	Weight (%)
Renew Energy Global Plc (Class A)	RNW	1.31
China Datang Corp Renewable Power Co., Ltd. (Class H)	1798.HK	1.27
Goldwind Science & Technology Co., Ltd. (Class H)	2208.HK	1.27
Clearway Energy, Inc. (Class C)	CWEN	1.19
Serena Energia SA	SRNA3.BZ	0.97
PNE AG	PNE3.GY	0.84
Concord New Energy Group Ltd	182.HK	0.77
Energiekontor AG	EKT.GY	0.74
Enlight Renewable Energy Ltd.	ENLT.IT	0.74
Fugro N.V.	FUR.NA	0.59
China Suntien Green Energy Corp Ltd (Class H)	956.HK	0.45
American Superconductor Corporation	AMSC	0.44
Energix-Renewable Energies Ltd.	ENRG.IT	0.34
Galata Wind Enerji A.S.	GWIND.TI	0.24

Table 7.2: Holdings of REMX_etf

Company	Ticker	Weight (%)
China Northern Rare Earth Group High-Tech	600111 C1	8.30
Albemarle Corp	ALB US	7.20
Lynas Rare Earths Ltd	LYC AU	6.94
Sociedad Quimica Y Minera De Chile SA	SQM US	6.62
MP Materials Corp	MP US	6.34
Pilbara Minerals Ltd	PLS AU	5.64
Shenghe Resources Holding Co Ltd	600392 C1	5.49
Jinduicheng Molybdenum Co Ltd	601958 C1	5.46
AMG Critical Materials NV	AMG NA	5.44
Liontown Resources Ltd	LTR AU	4.97
Xiamen Tungsten Co Ltd	600549 C1	4.94
Baoji Titanium Industry Co Ltd	600456 C1	4.89
Ganfeng Lithium Group Co Ltd	1772 HK	4.28
Lithium Americas Corp	LAC US	4.24
Eramet SA	ERA FP	3.87
Iluka Resources Ltd	ILU AU	3.67

Company	Ticker	Weight (%)
Vulcan Energy Resources Ltd	VUL AU	3.15
Tronox Holdings PLC	TROX US	3.01
Sigma Lithium Corp	SGML US	2.32
Lithium Americas Argentina Corp	LAR US	2.23
Patriot Battery Metals Inc	PMT CN	2.01
AVZ Minerals Ltd	AVZ AU	0.68

Table 7.3: Supporting metrics

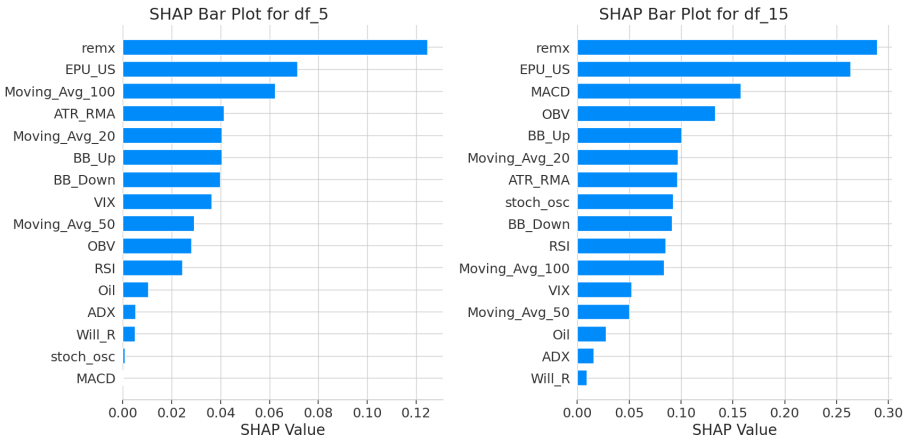
Name	**Description**	**# days**
True Range	$TR = \max \begin{cases} \text{High} - \text{Low} \\ \lvert\text{High} - \text{Close}\rvert \\ \lvert\text{Low} - \text{Close}\rvert \end{cases}$	
Average gain	$\text{Avg gain} = \frac{\sum_{i=1}^{t} \text{gain}_i}{t}$	14
Average loss	$\text{Avg loss} = \frac{\sum_{i=1}^{t} \text{loss}_i}{t}$	14
Relative Strength	$RS = \frac{\text{avg gain}}{\text{avg loss}}$	
Directional Movement$_+$	$DM_+ = \text{current high} - \text{previous high}$	
Directional Movement$_-$	$DM_- = \text{previous low} - \text{current low}$	
Smooth DM$_{+/-}$	$\text{smooth}DM_{+/-} = \sum_{i=1}^{t} DM - \frac{\sum_{i=1}^{t} DM}{t} + \text{current DM}$	14
Directional Indicator$_+$	$DI_+ = 100 \times \frac{\text{Smoothed}DM_+}{\text{ATR}}$	14
Directional Indicator$_-$	$DI_- = 100 \times \frac{\text{Smoothed}DM_-}{\text{ATR}}$	14
Directional Indicator index	$DX = 100 \times (\frac{\lvert DI_+ - DI_-\rvert}{\lvert DI_+ + DI_-\rvert})$	14
Average DX (ADX)	$ADX = \frac{\text{Prior ADX} \times (t-1) + \text{curr DX}}{t}$	14

Table 7.4: Performance whole dataset (5-day)

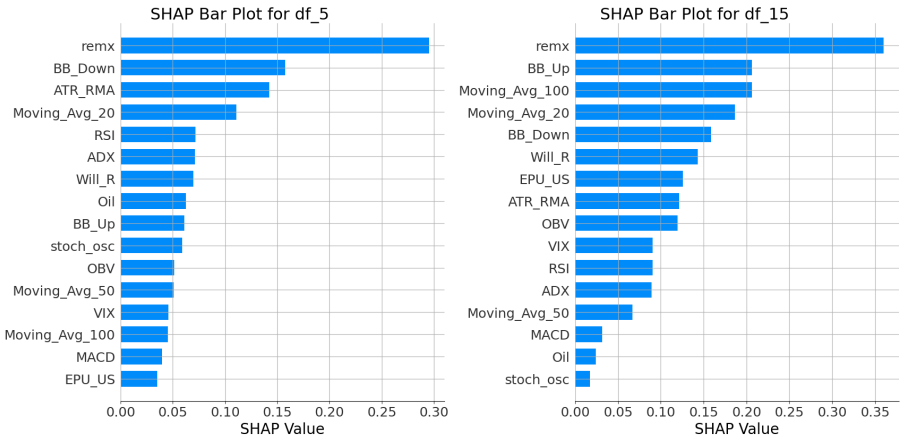
Model	**Accuracy**	**Recall**	**Precision**	**F1-score**
LR	0.578	0.959	0.560	0.707
SVM	0.540	0.629	0.560	0.592
RF	0.507	0.536	0.536	0.536
XGB	0.479	0.397	0.513	0.448
LGB	0.501	0.371	0.545	0.442

Table 7.5: Performance whole dataset (15-day)

Model	**Accuracy**	**Recall**	**Precision**	**F1-score**
LR	0.563	0.551	0.573	0.562
SVM	0.607	0.595	0.618	0.606
RF	0.529	0.415	0.539	0.469
XGB	0.479	0.361	0.475	0.410
LGB	0.529	0.388	0.542	0.452

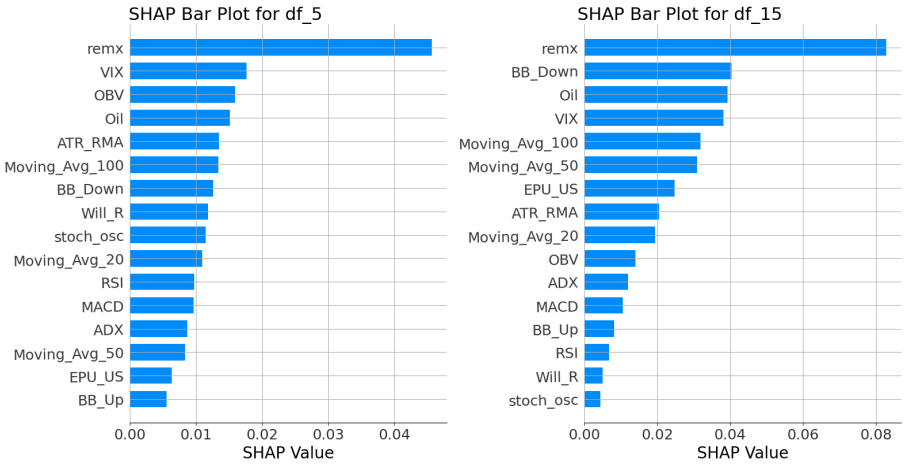


(a) LR

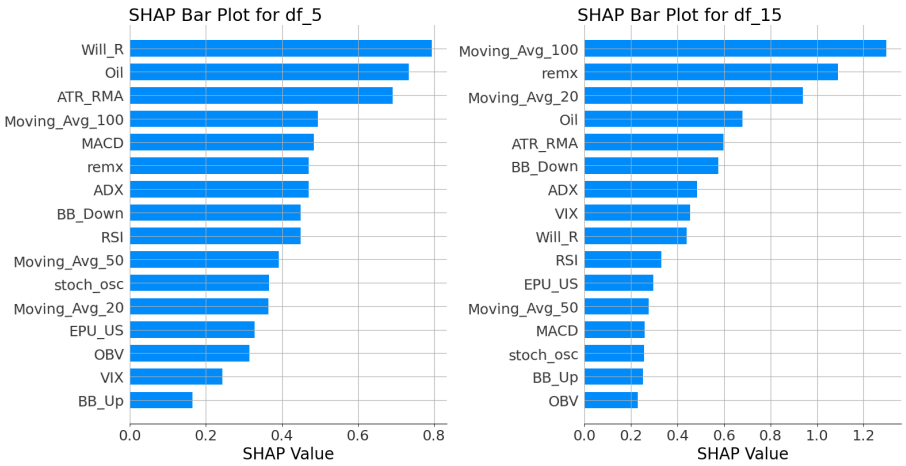


(b) SVM

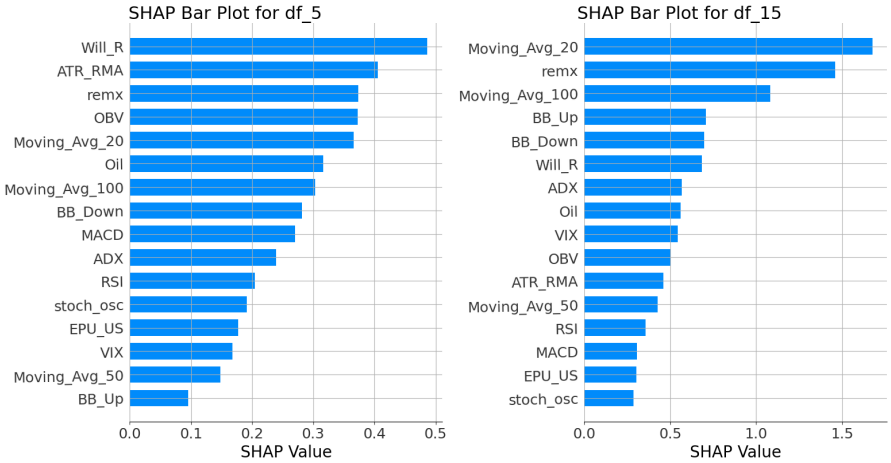
Figure 7.1: Feature Importance Base Models (2)



(a) RF

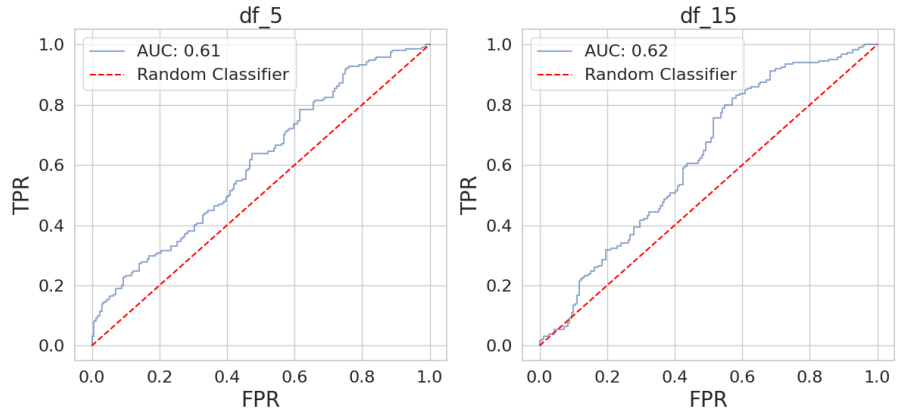


(b) XGB

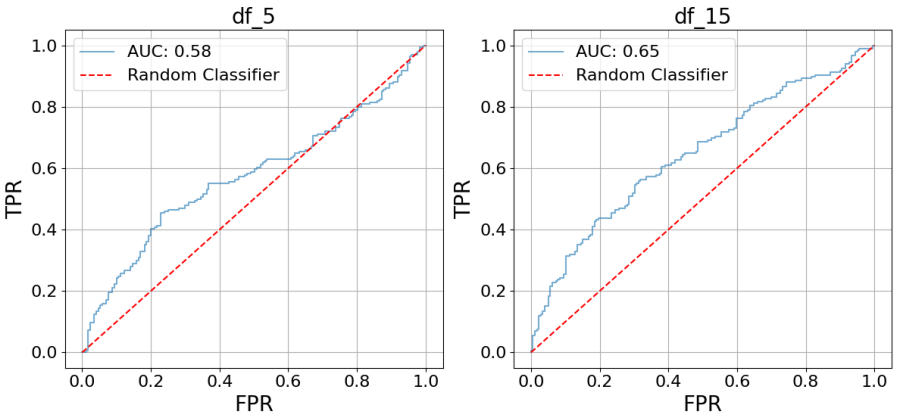


(c) LGB

Figure 7.2: Feature Importance Tree Models (2)

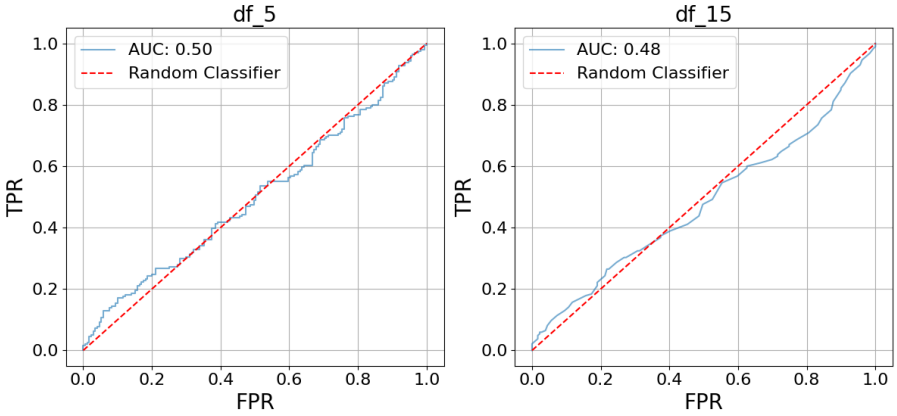


(a) LR

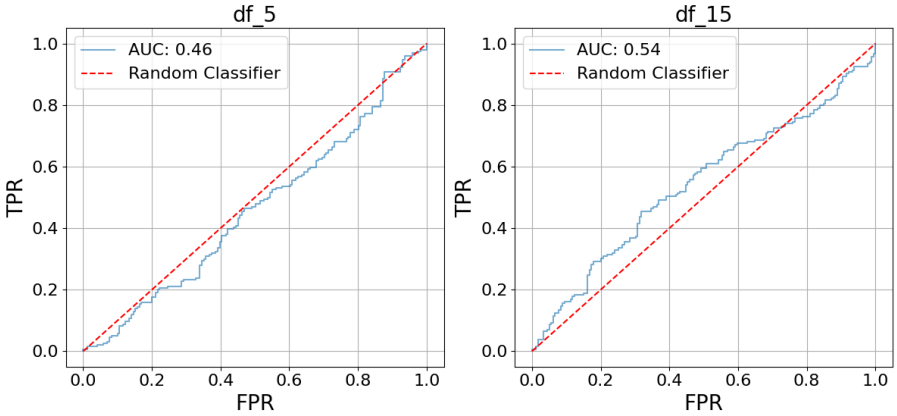


(b) SVM

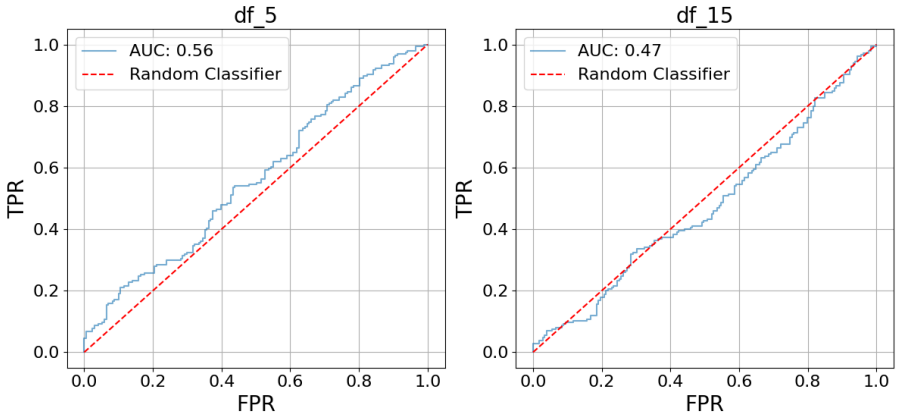
Figure 7.3: ROC Base Models (2)



(a) RF



(b) XGB



(c) LGB

Figure 7.4: ROC Tree Models (2)

Table 7.6: Performance sliding window (5-day)

Model	Accuracy	Recall	Precision	F1-score
LR	0.566	0.638	0.591	0.638
SVM	0.664	0.717	0.679	0.698
RF	0.711	0.778	0.713	0.744
XGB	0.733	0.763	0.748	0.755
LGB	0.740	0.775	0.752	0.763

Table 7.7: Performance sliding window (15-day)

Model	Accuracy	Recall	Precision	F1-score
LR	0.638	0.627	0.685	0.627
SVM	0.795	0.830	0.802	0.816
RF	0.829	0.852	0.838	0.845
XGB	0.857	0.871	0.868	0.870
LGB	0.856	0.874	0.865	0.869