The Art of Interpretation: Unleashing the Potential for Insightful Predictions in FX Data

44480, 47369, 47565, 47976

Department of Statistics, 2022/2023

Submitted for the Master of Science, London School of Economics and Political Science

Abstract

This research employs a rigorous, multi-phase process to analyse the interpretability of machine learning models in classifying the direction of the next day's return of 17 major Foreign Exchange (FX) pairs. We evaluate various black box models, including Long Short-Term Memory (LSTM) and eXtreme Gradient Boosting (XGBoost), and the interpretability methods Submodular Pick - Local Interpretable Model-Agnostic Explanations (SP-LIME) and Permutation Feature Importance (PFI). We use historical FX market data and 20 technical indicators obtained through feature selection utilising the Sequential Forward Floating Selection (SFFS) algorithm with a Linear Discriminant Analysis (LDA) model. Performance is measured based on accuracy, balanced accuracy, and F1 score.

In the first analysis, this study identifies a best-performing prediction model by comparing baseline and complex models. XGBoost outperformed all the other models.

The second analysis identified the best-performing interpretability method; by utilising information on which five variables to retain in subsequent prediction, the interpretability method, which provided the most valuable information, could be ranked as the best-performing interpretability method. The PFI model came out on top.

Subsequently, this study attempts to interpret FX market classification results to showcase the usefulness of interpretation in this paradigm of the financial sector − subjecting the best-performing prediction model, XGBoost, to further interpretability evaluation using the top-performing interpretability method, PFI. This study analyses trading performance using Profit and Loss graphs and monthly Sharpe Ratios to identify distinct periods of best and worst performance. The relative trading performance and interpretation of the results give the reader a better understanding of the XGBoost model's predictions. This can potentially create more trust in the model and aid trading and risk management strategies. Momentum-based technical indicators were most often classified as the driving features behind the classifications of the XGBoost model in the best-performing periods. However, dissecting the results showcases the multifaceted nature of FX markets. Notably, differences arise in the significance of Trend, Momentum, and Volatility-based indicators between USD and EUR-based FX markets. Additionally, variations in the importance of these indicators' downward trends, in contrast to their upward movements, are evident.

This comprehensive approach highlights the balance between model accuracy, interpretability, and trading performance, providing insights into the complex dynamics of FX markets.

Contents

Acknowledgements

First, we would like to thank Blaz Zlicar and Joshua Loftus, who are both responsible for guiding this Capstone Project. They constantly kept in touch, provided feedback and, most importantly, supplied us with the support to keep improving and challenge ourselves. Furthermore, we would like to thank the London School of Economics and Political Science and its professors for providing us with the opportunity to gain knowledge in a safe environment and be able to produce this final results.

Executive Summary

The advent and proliferation of machine learning (ML) models have transformed numerous fields, including finance. Their predictive capabilities offer compelling benefits, especially in foreign exchange (FX) markets. However, despite their impressive potential, these models often serve as 'black boxes', providing little to no insight into their decision-making processes. The absence of interpretability can impede the trustworthiness of these models, ultimately hindering their complete utilisation. This capstone project seeks to bridge this gap by examining the interpretability of ML models over determining the factors responsible for the movement of FX rates across 17 major markets with EUR and USD as the base currencies. More specifically, this study utilises complex classification models such as Long Short-Term Memory (LSTM) and eXtreme Gradient Boosting (XGBoost) in combination with the interpretation methods Submodular Pick - Local Interpretable Model-Agnostic Explanation (SP-LIME), Permutation Feature Importance (PFI), and the feature importance measure of XGBoost to provide more accessible explanations to the stakeholders through a more straightforward representation of complex models.

The application of these classification models over FX markets is non-trivial due to these markets' complex, volatile, and unpredictable nature. These attributes make identifying sustainable, accurate, and interpretable predictive models challenging. To address this, we implement a robust, multi-stage methodology focusing on model selection, interpretability evaluation, and, finally, utilising the potential of interpretation methods to conduct a trading performance analysis on the binary trading strategy in terms of best and worst-performing periods.

In the model selection phase, we train different models on historical FX rates represented by the 20 most essential features obtained through the Sequential Forward Floating Selection (SFFS) feature selection method and evaluate their predictive performance based on accuracy, balanced accuracy, and F1 score to get a holistic understanding. XGBoost demonstrates commendable performance, showcasing potential as a powerful tool for FX market predictions.

The interpretability evaluation phase involves applying SP-LIME, PFI, and the tree-based feature importance measure of XGBoost as interpretation methods and comparing their performance. By utilising information on which five variables to retain in subsequent prediction, the interpretability method, which provides the most valuable information, is ranked as the best-performing one. Here, PFI consistently performs well and often ranks among the best performers.

Finally, we conduct a trading performance analysis to understand the practical implications of the model's predictions. We gain valuable insights into the risk-adjusted return and worst-case scenarios by evaluating the Profit and Loss curves and monthly Sharpe Ratios of the binary trading strategy based on the model's predictions. We aim to extend our understanding beyond static metrics, focusing on real-world trading applications that reflect the complex realities of financial markets. By understanding the nature of the components which make up the datasets of the various currency pairs, this study tries to identify crucial segments for determining the success and failure of the trading strategies. Identifying peak

and vulnerable months of trading performance, coupled with interpretation methods, further elucidates the model's practical application and showcases the insightfulness of utilising interpretation methods within FX markets.

The results illustrate the distinct characteristics of each FX market, where groups of various Trend, Momentum, and Volatility-based variables drive the different markets. However, taking a further look at the results highlights the importance of mostly Momentum-based technical indicators in accurate prediction, supporting trading performance. More specifically, the results suggest that a dominant Trend indicator primarily drives EUR-based FX markets, while USD-based FX markets are mainly driven by an indicator capturing the Volatilitiy of the market. Another thing that stands out is that for making accurate predictions and thus increasing trading performance, the market's downward momentum, movement and volatility are more important than its upward counterparts.

The practical implications of this work provide a framework for understanding the intricate balance between model accuracy, interpretability, and trading performance. This research offers insights that could potentially enhance the trust in models, benefitting market prediction and risk management strategies to contribute to more informed decision-making processes often regarded as essential to the stakeholders in FX markets.

1. Introduction

Since the collapse of the fixed exchange rate system, the foreign exchange (FX) market has undergone significant changes, garnering considerable interest from both academia and industry. Accurate forecasting of the FX market is crucial for international investments, global trade, and economic competitiveness (Melvin and Taylor, 2009). Moreover, FX rates play a significant role in asset pricing, risk assessment, and guiding policymakers in conducting monetary policy. Despite the complexity of the FX market and the limited availability of global volume data, fluctuations in FX rates significantly impact multinational corporations, countries, and the value of imported and exported goods. With a daily traded volume of USD 5.1 trillion, navigating the FX market is no easy task.

The field of finance has long leveraged statistical methods and econometric models to anticipate FX rate fluctuations. However, these traditional techniques often struggle with the markets' complexity and nonlinear characteristics. Thus, the rise of machine learning (ML) has offered new hope for tackling these challenges. ML's capability to learn from and make predictions based on extensive datasets, encompassing various factors influencing currency prices, makes it an ideal candidate for FX market analysis.

However, despite their promise of prediction accuracy, many ML techniques have also raised critical issues, primarily around their interpretability. Hence, interpretability is becoming a topic of pressing interest. In many high-stakes domains, including finance, a model's transparency is vital for its adoption. Without this, the risks associated with 'black box' decision-making − in which the reasons behind a model's predictions are opaque − can outweigh the benefits of accuracy. Notably, users need to trust a model, understand its errors, and know how to improve it, which is difficult, if not impossible, without interpretability.

While the application of ML techniques in financial forecasting has been well-documented, less research has explored their interpretability, particularly in the context of the FX market. This gap extends to a comparative analysis of different interpretability techniques and their usefulness within FX markets. Numerous methods have been proposed, but how they measure up against each other and can be of use within FX market prediction is an area ripe for investigation.

Our research aims to address these issues by employing a multi-phased approach. First, we seek to evaluate the performance of ML techniques in predicting FX market movements, building on the existing body of knowledge while incorporating recent advancements. However, the core of our research revolves around interpretability. We undertake a comprehensive comparison of various interpretability methods, aiming to uncover their strengths and limitations when applied to the predictions following the best-performing complex classification model. Hereafter, this study illustrates the usefulness of interpretation, using a trading performance breakdown in better and worse-performing trading periods to identify the drivers of these respective periods − potentially creating more trust in the complex classification model, enhancing trading and risk management strategies.

We apply ML models to forecast the direction of the next day's FX rate return − 'up' or 'down' − for 17 major pairs with USD and EUR as base currencies. While we use well-established traditional classification models like Linear and Quadratic Discriminant Analysis,

Logistic Regression, and Support Vector Machine, we also turn to more complex models like Random Forest (RF), Long Short-Term Memory (LSTM) networks, and eXtreme Gradient Boosting (XGBoost). These more complex models have demonstrated exceptional performance in various time-series forecasting tasks, as they can capture temporal dependencies (S. Ahmed et al., 2020).

Two popular interpretability models, Submodular Pick - Local Interpretable Model-Agnostic Explanation (SP-LIME) and Permutation Feature Importance (PFI), have garnered significant recognition and demonstrated their effectiveness in diverse fields, such as healthcare (ElShawi et al., 2021; Stiglic et al., 2020a). However, their application to financial data, particularly in the context of FX rate forecasting, remains largely unexplored. Consequently, we will apply these interpretability methods and evaluate them based on their ability to shed light on model decisions and their computational efficiency.

As mentioned, in addition to comparing various interpretability and ML models, our study introduces a unique third component. Building on the findings of our analysis, we identify the best-performing ML model and the most insightful interpretability method. We then employ this optimal ML model in a binary trading strategy within the FX markets to pinpoint the best and worst-performing trading months based on monthly Sharpe Ratios. More importantly, we delve into the 'why' behind these periods using our optimal interpretability method to unravel the intricate factors that dictate these respective trading months − potentially providing valuable insights for financial decision-makers.

This study classified the XGBoost and PFI models as the best-performing ML model and interpretation method, respectively. Moreover, when utilising these two models in the final analysis, this study found meaningful insights into the drivers of the various FX markets. Although most markets display a unique character, this study represented multiple findings. This study highlighted the primary importance of Momentum-based indicators in driving the best-performing trading months. Moreover, it displayed differences between drivers of USD and EUR-based FX markets. A feature that primarily drove the best-performing months of USD-based markets was measuring the market's Volatility. In contrast, for EUR-based markets, this was identified to estimate the market's Trend. Moreover, this study identified various features important in driving specific markets and found that the downward Trend, Momentum, and Volatility were more informative for accurate prediction and, hence, trading performance than its upward counterparts.

The remainder of this paper is organised as follows. The second section covers the literature review. Section three covers the rationale, complexities, and potential approaches evident in comparing interpretation models. In section four, this paper discusses all aspects of the FX data utilised. Subsequently, section five presents a complete picture of the study design. Moreover, sections six, seven, and eight cover the methodology, including the machine learning models, interpretability methods and evaluation metrics. The results are presented in section nine. Finally, the conclusions and suggestions for further research are drawn in section ten.

2. Literature Review

2.1 Machine Learning in Financial Market Predictions

The challenge of predicting financial markets has long been tackled with various statistical and econometric models (Alexander, 2001). The advent of machine learning (ML) brought a paradigm shift, allowing for high-dimensional data processing and discovery of complex, non-linear relationships (Henrique, Sobreiro, and Kimura, 2019). Early explorations intersected simpler models such as linear and logistic regression and decision trees with finance (Ryll and Seidens, 2019). As the potential of ML unfolded, researchers explored advanced models, including neural networks and support vector machines (Sheta, S. E. M. Ahmed, and Faris, 2015; Kurani et al., 2023). Contrary to one might expect, these more complex models are not necessarily superior to linear time series models, even when the data are financial, seasonal and non-linear (JingTao and Tan, 2001).

Eventually, the advancements in neural networks led to the adoption of Long Short-Term Memory (LSTM) networks, known for their ability to learn long-term dependencies in data sequences (Lipton, Berkowitz, and Elkan, 2015; Yu et al., 2019). This ability gives LSTMs a distinct advantage in financial predictions, even as they get compared with Convolutional Neural Networks (CNNs), which also demonstrate effectiveness in structured financial time-series data (Sayavong, 2019; Tsai, 2018). However, the computational complexity and the training expense of LSTM models remain significant limitations (Rasheed et al., 2020).

On the other hand, financial data, characterised by inherent noise and volatility, often benefits more from robust, simpler models that can effectively navigate this complexity. This is where ensemble methods like XGBoost stand out (Gumelar et al., 2020). Operating within the gradient boosting framework, XGBoost uses decision trees as basic building units and incorporates regularisation measures to prevent overfitting (T. Chen et al., 2015).

In a recent study (Paliari, Karanikola, and Kotsiantis, 2021), LSTM showed a slightly superior performance over XGBoost in forecasting stock prices, on the basis of MSE, RSME and MAPE. In addition, (R. Kumar, P. Kumar, and Y. Kumar, 2021) found that a stacked LSTM model surpassed XGBoost in stock market prediction, exhibiting an 11.04% improvement in RMSE and 20.88% improvement in MAPE.

2.2 Machine Learning in FX Market Predictions

The FX market, vital in global finance and economy, is a dynamic and non-stationary field with continuous fluctuations driven by economic events, geopolitical instability, and varying interest rates (Los, 2000). Understanding this complexity is essential for insights into economic indicators, policy impacts, and risk management strategies (Melvin and Taylor, 2009).

Building on the strengths of ML models, researchers have applied LSTM networks and XGBoost in the realm of FX market predictions. LSTM demonstrated a commendable performance in forecasting the directional movement of FX data when combined with technical and macroeconomic indicators (Yıldırım, Toroslu, and Fiore, 2021). Islam, Sholahuddin, and Abdullah (2021) found that XGBoost was particularly effective in forecasting USD to Rupiah exchange rates, exhibiting smaller RMSE values when applied to new datasets and yielding

MAPE values below 10%, indicative of excellent forecasting capabilities. However, while these models are extensively researched individually in the context of FX market predictions, there is a critical gap in the literature: a comparative study examining the performance of LSTM and XGBoost in FX rate predictions. Our study aims to bridge this gap by answering the research question of which predictive model best classifies the directional movement of major FX markets.

2.3 Data and Features

Selecting appropriate features is crucial for FX rate prediction. The literature suggests various features for this task. The commonly used features include the opening, (adjusted) closing, highest, and lowest values of the rates (Yetis, Kaplan, and Jamshidi, 2014). In addition, technical indicators such as Bollinger Bands, Relative Strength Index, Stochastic Oscillator, Money Flow Index, and Moving Average Convergence/Divergence are often employed (Borovkova and Tsiamas, 2019).

The selection of features is not merely a process of aggregation but requires a careful, critical evaluation of their predictive power. For instance, Picasso et al. (2019) used Simple Moving Average, Exponential Moving Average, Average True Range, and Williams Indicators for stock price prediction, but their effectiveness in FX market predictions is not yet fully explored. Similarly, Peng et al. (2021) compiled a list of technical indicators used in financial prediction studies with ML models, suggesting the potential for further investigation and experimentation in the context of FX markets.

Furthermore, feature selection algorithms, such as Sequential Forward Floating Selection (SFFS), play a significant role in identifying the most predictive features (Marcano-Cedeño et al., 2010; Ververidis and Kotropoulos, 2005). SFFS has been lauded for its simplicity, efficiency, and its exceptional performance in technical indicators feature selection (Peng et al., 2021).

2.4 The Trade-off in Model Accuracy and Interpretability

Pursuing better accuracy and classification power often comes at the cost that these models are less interpretable — introducing a trade-off. Neural Networks and Gradient Boosting Decision Trees tend to pivot more weight towards accuracy and raise the problem of interpretability (Zhang et al., 2018). The rising potential of Neural Networks and Gradient Boosting Decision Trees cause it to affect more important decisions (Fan et al. (2021); Guo, Fu, and Sollazzo (2022); Barton et al. (2022)), which makes comprehending the mechanics behind these predictions more crucial than ever. As Molnar (2020) mentions, it is essential to ask why a model made a specific conclusion besides asking what the model predicts. Hence, besides having models targeting the 'what', a well-reasoned analysis should also highlight models explaining the 'why'. Many researchers have been exploring this domain, and various models and approaches have been introduced (Frye, Rowat, and Feige (2020), Stiglic et al. (2020b)).

More specifically, A. Fisher, Rudin, and Dominici (2019a) introduced a model-agnostic variation on the permutation feature importance, first introduced specifically for random forests by Breiman (2001a). The Permutation Feature Importance (PFI) algorithm permutates a particular variable and compares the model error with the model error of not

permutating any variables — therefore, obtaining a global picture of feature importance.

Moreover, Ribeiro, Singh, and Guestrin (2016a) presented an algorithm called Local Interpretable Model-Agnostic Explanations (LIME), constructed by understanding the predictions locally as it incorporates one specific prediction only. With the introduction of Submodular Pick - LIME (SP-LIME), which carefully selects a few instances from all the possible predictions, research designed an instrument to capture information about the entire dataset and provide a global interpretation of a model based on the LIME algorithm.

Both these algorithms have proven attractive for researchers to interpret predictions concerning financial time series. Gite et al. (2021) utilised the LIME algorithm to gain insights into the vital drivers of Indian Stock Market prices, and Riff, Parthiban, and Zhe (2022) used the LIME algorithm on credit risk data and highlighted its importance. Moreover, D. Wu, Wang, and S. Wu (2022) applied the LIME algorithm to decide which model produced the most reliable forecasts in predicting the trend of China's 379 stock market indices by industry. Furthermore, concerning PFI, Varma and Engineer (2020) used PFI as a benchmark to compare with other feature importance measures based on multiple datasets. Schelthoff et al. (2022) applied PFI to gain insights into the feature importance concerning waiting time predictions.

The built-in feature importance mechanism of XGBoost, which ranks features based on their splitting frequency across all decision trees, provides valuable insights into model decisions (T. Chen et al., 2015). This is a critical aspect highlighted in the work of Li et al. (2021), who applied it to FX price prediction. Their research highlighted that XGBoost's feature importance ranking substantially improved model performance. Additionally, stacking LSTM with XGBoost enhanced predictive accuracy for FX prices, underscoring the merits of merging advanced models with interpretability-focused methods. Vuong et al. (2022) work demonstrated that the integration of LSTM and XGBoost models in forecasting stock prices provided superior results, thus reaffirming the potency of these combined methodologies in financial market predictions.

2.5 Comparison of Interpretability Methods

Interpretability methods in ML have been evaluated and compared along several dimensions including the interpretability-accuracy trade-off, computational efficiency, and suitability to specific models or data (Doshi-Velez and Kim, 2017). Interpretability methods such as LIME and SHAP provide a balance between local and global interpretability, but differ in computational demands (Lundberg and Lee, 2017). While PFI is easy to implement, it may have limitations when dealing with feature correlation (Altmann et al., 2010). Moreover, XGBoost's feature importance, which relies on internal model structures, can occasionally be inconsistent (Strobl et al., 2007).

Methods have been proposed for comparing these interpretability techniques. One common approach is to evaluate how well the interpretability method can mimic the predictions of the original model, usually in terms of accuracy or fidelity (Ribeiro, Singh, and Guestrin, 2016b). Another approach involves comparing the stability of the interpretation, i.e., how consistent the explanations are when slight changes are made to the input data (Doshi-Velez and Kim, 2017).

Recent literature (Schmidt and Biessmann, 2019) has urged for quantitative measures

in interpretability assessment, considering aspects such as feature relevance and model complexity. The authors underscored the role of user trust in interpretability, indicating that interpretability comparisons should account for user trust, alongside accuracy and efficiency.

The comparative study conducted by Chan, H. Kong, and Guanqing (2022) on faithfulness metrics, which measure the accuracy of explanations provided by interpretability methods, further enhances our understanding of this field. The authors compared several metrics, including input-output mutual information, infidelity, and sensitivity, to their application in methods like LIME and SHAP. This research emphasises the variability of an explanation method's faithfulness depending on the metric used, highlighting the necessity of careful metric selection and usage when evaluating interpretability methods.

The notion of context dependence in interpretability is evident in the approach introduced by Laugel et al. (2018), where they developed comparison-based inverse classification to evaluate interpretability. This method compares two instances and explains the difference in the model's predictions for these instances, providing instance-specific explanations rather than global ones. This work highlights the importance of comparative approaches in interpretability evaluations, suggesting that the effectiveness of an interpretability method may depend on its application context.

Complementing this view, Carvalho, Pereira, and Cardoso (2019) emphasised the absence of a 'one-size-fits-all' interpretability method in their survey. They proposed a combination of interpretability techniques to provide a comprehensive understanding of model decisions, reinforcing the importance of a comparative approach to determine the most suitable interpretability methods for specific tasks.

Collectively, these studies provide invaluable insights into the multifaceted landscape of interpretability methods, illuminating their unique strengths, weaknesses, and optimal applications. Our research aims to build upon these findings. It addresses the fundamental research question of which interpretability method within the predictive FX market landscape can be identified as the top performer.

2.6 Contribution to Literature

Our study contributes to the literature in several significant ways. Firstly, while many works have applied ML techniques for financial forecasting, our research specifically emphasises interpretability in ML, which remains less explored in the FX market context.

Secondly, we carry out an extensive comparative analysis of three well-recognised interpretability methods: SP-LIME, PFI, and XGBoost's feature importance. Previous studies have not thoroughly investigated the application and performance of these methods in the context of FX market prediction. We fill this gap, shedding light on their relative strengths, weaknesses, and suitability for interpretation.

Lastly, using our top-performing machine learning model, we implement a binary trading strategy to identify best and worst months for FX rate forecasts. Leveraging the optimal interpretability method, we unearth the reasons for these trends. This study not only provides key insights for accurate forecasting but also bridges the academic-industry gap for financial experts.

3. The Focal Points Evident in the Comparison of Interpretability Models

3.1 Rationale Behind Investigating Interpretability Methods

In an age where complex algorithms guide a vast array of decisions, the need for interpretability is increasingly pronounced. This importance is magnified in areas such as the foreign exchange (FX) market, where financial decisions grounded on these predictions bear profound consequences. Models that are accurate but non-transparent pose multiple challenges. For instance, they could harbour inadvertent biases, their outputs may be hard to trust, and regulatory issues could hinder their deployment.

We aim to bridge this gap between accuracy and interpretability by examining interpretability methods. These methods seek to 'translate' the mechanics of the black box models, shedding light on how they make predictions. This transparency not only bolsters trust but can also aid in identifying potential errors or biases, leading to more reliable models and better-informed decision-making.

3.2 The Complexities of Comparing Interpretability Methods

Establishing a framework for comparing interpretability methods presents a labyrinth of challenges. Interpretability methods are not universally applicable. An approach providing profound insights for one model might fail to do so for another. There are various ways in which researchers try to subdivide the domain of interpretation methods. For example, Mahya and Fürnkranz (2023) divide them into directly learning interpretable methods and posthoc explanation methods, whereas Molnar (2020) focuses on the difference between feature importance and feature effects. Moreover, Linardatos, Papastefanopoulos, and Kotsiantis (2020) subdivides the interpretation methods based on its focus points − namely, black box explanation, white box creation, promoting fairness, and, finally, sensitivity analysis. However, all the papers have in common that they discuss that there is no right or wrong answer − there are simply different perspectives − highlighting the complexity of interpretation method comparison.

This study could have focused on interpretability methods following from one of the existing frameworks. However, as this would add little to the existing literature, this study decided to elaborate on the main complexities when comparing a diverse set and different types of interpretation models. More specifically, this study focuses on the Submodular Pick - Locally Interpretable Model Explanation (SP-LIME), Permutating Feature Importance (PFI), and eXtreme Gradient Boosting's (XGBoost's) feature importance methods. Furthermore, this section provides the readers with reasoning for the methodology chosen for the comparative analysis of this diverse set of interpretation methods.

When this study first tried to compare the LIME algorithm with PFI and XGBoost's feature importance method, one main concern was the interpretation methods' global and instance-specific (local) behaviour. Local interpretation methods explain individual predictions, while global methods assemble the average model behaviour. The LIME algorithm explains a model's instance-specific behaviour, while PFI and XGBoost's feature importance

method explain a model's global behaviour. This study is interested in this global, average model behaviour. Therefore, this study needed to modify the LIME algorithm to make it of use to this study, enabling the comparison to the PFI model and XGBoost's feature importance measure. There are several ways to do this, such as the various ways to average the instance-specific explanations discussed in Mahya and Fürnkranz (2023) (e.g. GlocalX (Setzu et al., 2021)). However, the many use cases of SP-LIME (Ahern et al. (2019);Pang and Y. Kong (2022);Ribeiro, Singh, and Guestrin (2016a)) induce this study to implement this specific algorithm.

Another vital distinction between interpretability methods is the presence of an interpretable surrogate model. In this study, the SP-LIME algorithm uses such a surrogate model in which it tries to fit a linear, locally faithful model to gain interpretation. Research often classifies the number one concern for such methods as the level at which this interpretable surrogate model reflects the behaviour of the complex model. This study refers to this concept as the surrogate model's 'faithfulness', which follows the reasoning of Mahya and Fürnkranz (2023), where they refer to this as the 'fidelity' of the surrogate model − a concept often used when comparing interpretation models which utilise such a surrogate. However, the fact that this study deals with surrogate-based and non-surrogate interpretation techniques complicates things and makes this approach unfeasible.

Furthermore, the interpretability of a method is often tightly intertwined with the specificities of the model it's applied to − introducing the discrepancy between model-specific and model-agnostic interpretation methods. This study starts with a comparison between prediction models based on their performance in predicting the upward or downward movement on the next day's return of major FX pairs and continues with the best-performing model. Therefore, deciding upon utilising a particular model-specific interpretation method could not have been anticipated beforehand. Eventually, XGBoost was identified as the best-performing classification model, allowing this study to use the tree-based feature importance method of XGBoost alongside the comparison with SP-LIME and PFI.

3.3 Chosen Methodology to Compare Interpretability Methods

The previous sections explained some of the main complexities that arise when comparing different types of interpretation methods. After careful consideration, this section elaborates on the methodology chosen to employ when comparing the diverse set of interpretation methods: SP-LIME, PFI, and the model-specific feature importance method accompanying XGBoost.

This study adopts a multi-phased approach, primarily focusing on interpretation models and their practical relevance within the context of FX markets. Hence, this study decided to select a methodology for comparing the respective interpretation methods that aligns with the rest of the study, which seeks to ascertain the potential utility of these interpretation methods in FX market prediction and trading.

More specifically, the chosen methodology starts with a model including 20 features. Each interpretability technique then chooses which five features to obtain for subsequent prediction. This study ranked the interpretation method associated with the best-performing classification model in the subsequent prediction as the most insightful and, therefore, best-performing interpretation method.

This study is well aware of the assumptions and speculations it's making, such as the use of prediction accuracy, although indirectly, as a metric for interpretation performance. However, as explained, it aligns with the subsequent analysis showcasing how an interpretation method could be used within FX market prediction and trading. Therefore, after incorporating the complexities of comparing this diverse set of interpretation models, this study believes this to be the best way to employ this comparative analysis on the respective interpretation methods.

4. Data

This section thoroughly analyses all aspects of the data used in this study, encompassing data collection, pre-processing, and construction of the technical indicators. Moreover, in order to be able to replicate this study's results let us emphasise that the primary programming environment for this research is Python, version 3.10.12. All code was executed within the Google Colaboratory (Colab) environment. This study recommends using this or a similar environment when running the code, which can be found in this paper's GitHub repository, to ensure consistency in reproducing the results.

This study compares various interpretability models and focuses on their ability to select the essential features to predict the next day's return direction for 17 major FX pairs. Besides considering the USD/EUR FX pair, this study considers the FX pairs that have USD or EUR as their base currency and JPY, GBP, CHF, CAD, NZD, SEK, NOK, and DKK as quote currencies.

The timeframe considered for this study ranges from January 2005 up until the end of 2019, where the period of 2005-2017 is utilised for training and validation, with the remaining data serving the purpose of testing.

We start our analysis by fetching data using the in-built Yahoo Finance API. The various in-built features obtained through the API include Open, High, Low, and Adjusted Closing values for all the foreign exchanges.

4.1 Data Engineering

To obtain as much information from the data as possible, we increase the number of features in our model by adding various technical indicators. These indicators comprise trend, momentum, and volatility information about the FX rates. Figure 4.1 indicates the entire table of technical indicators.

Open	High	Low
Adj Close	Parabolic SAR (Parabolic_SAR)	Coppock Curve (Coppock_Curve)
Exponential Moving Average (EMA)	Commodity Channel Index (CCI)	Bollinger High Band (BB_HB)
Aroon Up (AROON_UP)	Aroon Oscillator (AROONOSC)	Average Directional Index (ADX)
Linear Regression (LINREG)	Accelerator Low (ACC_LOW)	Accelerator Mid (ACC_MID)
Chandelier Long (CHANDELIER_LONG)	Hull Moving Average (HMA)	Kaufman's Adaptive Moving Average (KAMA)
Typical Price (Typical_Price)	Relative Strength Index (RSI)	Stochastic Oscillator (SO)
Ulcer Index (ULCERINDEX)	Weighted Moving Average (WMA)	Momentum (MOM)
Midprice (MIDPRICE)	Normalised Average True Range (NATR)	Triple Exponential Moving Average (TEMA)
Percent B (PERCENT_B)	Average True Range (ATR)	Keltner Channels Upper (KC_UPPER)
Vortex Negative (VORTEX_NEG)	Vortex Positive (VORTEX_POS)	Moving Average Convergence/Divergence (MACD)
Bollinger Low Band (BB_LB)	Bollinger Band Moving Average (BB_MAVG)	Detrended Price Oscillator (DPO)
Chande Momentum Oscillator (CMO)	Double Exponential Moving Average (DEMA)	Midpoint (MIDPOINT)
Accelerator Up (ACC_UP)	Choppiness Index (CHOP)	Relative Vigor Index (RVGI)
Mass Index (MI)	Moving Standard Deviation (MSD)	Triple Exponential Average (TRIX)
Simple Moving Average (SMA)	Williams Percentage Range (WI)	Price Rate of Change (ROC)
Directional Movement Index (DX)	Triangular Moving Average (TRIMA)	Aroon Down (AROON_DOWN)
Psychological Line (PSL)	Bias (BIAS)	Relative Volatility Index (RVI)
Keltner Channels Lower (KC_LOWER)	Bollinger Band Width (BBWIDTH)	Chandelier Short (CHANDELIER_SHORT)
Percentage Price Oscillator (PPO)	Absolute Price Oscillator (APO)	Donchian Channel (DO_UP)

Figure 4.1: Table of Technical Indicators

Furthermore, the variables are normalised and lagged to ensure we incorporate the time element and make the data suitable for modelling as we determine the direction of the next day's FX rate returns. By considering the information based on several days ago rather

than a single day, we allow for the delay in which past data can provide information on the direction of the return several days later.

4.2 Exploratory Data Analysis: Descriptive Statistic

Name	Mean	Std	Skewness	Kurtosis	K-S test (p-val)	Durbin-Watson Statistic
USD/EUR	1.7624	0.2046	1.2533	1.6810	0.0000	0.0001
USD/JPY	1.0321	0.1205	0.7601	-0.4064	0.0000	0.0000
USD/GBP	1.3097	0.2015	0.4056	-1.4077	0.0000	0.0000
USD/CHF	0.7954	0.0775	0.1389	-0.9148	0.0000	0.0001
USD/NZD	0.6422	0.0894	0.2026	-0.7835	0.0000	0.0000
USD/CAD	8.4991	0.7540	0.4665	-1.0582	0.0000	0.0000
USD/SEK	9.4866	0.6170	0.7268	0.0151	0.0000	0.0000
USD/DKK	1.3932	0.1443	0.8413	1.6225	0.0000	0.0001
USD/NOK	1.1535	0.1283	0.1573	-1.4074	0.0000	0.0000
EUR/JPY	7.4514	0.0097	-0.1133	-0.1781	0.0000	0.0000
EUR/GBP	7.5595	0.9901	0.4090	-0.9868	0.0000	0.0001
EUR/CHF	103.5230	13.2516	-0.5642	-0.7939	0.0000	0.0001
EUR/NZD	5.9237	0.5792	0.1438	-0.9267	0.0000	0.0000
EUR/CAD	130.5750	16.0596	0.2542	-0.1779	0.0000	0.0001
EUR/SEK	1.4510	0.0909	-0.0323	-0.3789	0.0700	0.0000
EUR/DKK	0.8074	0.0781	-0.5166	-1.0890	0.0000	0.0000
EUR/NOK	6.7997	1.1656	0.5138	-1.2184	0.0000	0.0001

Figure 4.2: Descriptive Statistics of Adjusted Closing Prices per FX Market

Figure 4.2 dives deeper into the descriptive statistics of the adjusted closing prices. This study employs the Kolmogorov-Smirnov Test to test for normality. The resultingly small p-values lead us to infer that there is enough evidence to reject the null hypothesis saying that the FX rates are normally distributed. Moreover, this statement is supported by the nonzero skewness values and the respective kurtoses smaller than three. In particular, the EUR/JPY, EUR/CHF, EUR/SEK, and EUR/DKK markets stand out due to their negative skewness values − implying that for these FX markets, the decrease in probabilities is less steep for lower values. We conclude by utilising the Durbin-Watson Statistic to test the presence of autocorrelation in the time series. A value close to two would suggest no autocorrelation. However, our findings yielded values rounding to zero, indicating the presence of positive autocorrelation, which can be of use in the technical analysis of this study.

4.3 Feature Selection: Sequential Forward Floating Selection (SFFS)

Selecting relevant features from abundant FX market data is vital in our model development. To streamline this, we employ the Sequential Forward Floating Selection (SFFS) algorithm, an enhancement of the Heuristic search method (P. Pudil, 1994).

The SFFS method iteratively adds or removes features to optimise the Criterion function $J(S)$, which records the predictive accuracy of our model captured by the subset of features represented by S.

The SFFS method starts by taking a random single feature. Subsequently, the prediction accuracy resulting from the Linear Discriminant Analysis (LDA) model incorporating this single feature is compared with the accuracy resulting from other LDA models made up of subsets of features obtained by adding or removing a specific feature. Finally, the model and respective subset of features resulting in the highest increase in accuracy are retained. This process is performed iteratively until the number of features in the subset is as required.

This dynamic process mitigates overfitting, yielding a more robust model and ensuring that the selected features have demonstrated performance in making accurate predictions.

5. Study Design

This study employs a multi-phased approach to examine the interpretability of a best-performing complex classification model in foreign exchange (FX) markets and the usefulness of interpretation within FX market prediction and trading. The research methodology can be broken down into three distinct phases:

5.1 Selecting the Best Machine Learning Model

In the first phase, this study identifies the most accurate predictive model. A set of classification models, including complex models Long-Short Term Memory (LSTM), eXtreme Gradient Boosting (XGBoost), and Random Forest (RF) together with the baseline models Support Vector Machines (SVM), Logistic Regression (LR), and Linear and Quadratic Discriminant Analysis (LDA and QDA) are trained on historical FX market data. This study measures the performance of each model based on the standard metrics accuracy, balanced accuracy, and F1 Score. Subsequently, this study selects the model that exhibits the best performance for further analysis.

5.2 Selecting the Best Interpretability Method

In the second phase, our focus shifts to interpretability. We employ three interpretation methods: XGBoost's tree-based feature importance, Submodular Pick - Local Interpretable Model-Agnostic Explanations (SP-LIME), and Permutation Feature Importance (PFI) to the predictions following the best-performing machine learning model. We evaluate the performance of these methods based on the accurate information each provides regarding the top five variables necessary for accurate prediction. In this phase, we also evaluate the time complexity, considering the time taken by each method to produce its final result.

5.3 Interpretability Analysis through Economic Performance

This final phase of our research links model interpretability with practical trading performance. We implement a binary trading strategy, guided by the predictions of our top-performing model, and assess its economic implications using the Profit and Loss (PnL) function. To distinguish between the best and worst-performing periods, we employ the Sharpe Ratio, utilising the SOFR (Secured Overnight Financing Rate) as our risk-free rate. This allows us to analyse the model's interpretability under varied market conditions. Subsequently, we apply our chosen interpretability method to these identified periods.

5.3.1 Binary Trading Strategy

Our binary trading strategy hinges on a simple principle: we adopt a long position when our model predicts an uptick and opt for a short position when a downtick is projected − terminating the position after each day and assuming one trades with 100 times leverage or can simply trade these bigger positions. This study is well aware of this trading strategy's simplistic nature and its assumptions. For example, this study accounts no transaction costs to its trades, the trades can always be made at the market opening and executed at the market closing, and they do not move the market as no slippage costs are incorporated. However,

this explicit rule-based strategy lends itself to effective interpretability analysis, ensuring a clear link between the model's predictions and the consequent trading decisions. Given that the primary focus of this paper lies in interpretation, while still being able to discern relative trading performance for the purpose of analysis, a deliberate trade-off is made by prioritising simplicity and clarity over superior overall trading performance.

5.3.2 Performance Metrics

The efficacy of our strategy is determined using two financial metrics: the PnL function and the Sharpe Ratio. The PnL translates abstract model predictions into tangible gains and losses over time. On the other hand, the Sharpe Ratio measures our strategy's performance, considering the risks taken. It helps us identify the best and worst trading periods and gives us insights into how well the model's predictions worked during different market conditions. Higher Sharpe Ratio values imply a more desirable balance between risk and reward.

6. Machine Learning Models

6.1 Linear Discriminant Analysis (LDA)

The prediction of foreign exchange (FX) rate movement can be framed as a binary classification problem, where class 0 represents a downward direction of the next day's return, and class 1 represents an upward direction. To solve this, we utilise Linear Discriminant Analysis (LDA), a renowned dimensionality reduction technique that preserves class discriminatory information (S. Fisher R. A., 1936).

LDA seeks to find a linear combination of features that can effectively characterise or separate these two classes, under the assumption of identical covariance matrices across classes characterised by the features constituting the FX dataset. This assumption leads to a linear decision boundary, providing a straightforward interpretation that aligns well with the inherent trends and movements in the FX market.

6.1.1 Discriminant Function

The discriminant function in LDA provides a score for each class based on the given sample. In the context of our research, the discriminant function for class i ($i = 0, 1$) is defined as:

$$\delta_i(x) = x^T \Sigma^{-1} \mu_i - \frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i + \log P(\omega_i) \tag{6.1}$$

Here, Σ is the shared covariance matrix, μ_i is the mean of class i which is a tuple of averages of various features, and $P(\omega_i)$ denotes the prior probability of class i. By maximising the distance between means and minimising scatter within each class, the discriminant function effectively enhances class separability. The class of an unseen sample is assigned by the class whose discriminant function yields the highest value, thus indicating the predicted direction of the next day's return.

6.2 Quadratic Discriminant Analysis (QDA)

While LDA provides a valuable method for predicting the direction of FX rate movements, it operates under the assumption of identical covariance matrices across classes. To add flexibility in our modeling approach, we also employ Quadratic Discriminant Analysis (QDA), a variant of LDA that allows for unique covariance matrices for each class (Hastie, Tibshirani, and Friedman, 2001).

The ability of QDA to provide a unique covariance matrix for each class allows for a more refined analysis in situations where the variability within each class of FX rate direction is significantly different. By recognising and accounting for these differences, QDA can provide more accurate and nuanced modeling of the FX rate movements.

6.2.1 Quadratic Decision Boundary

The discriminant function for class i ($i = 0, 1$) in QDA is as follows:

$$\delta_i(x) = -\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) - \frac{1}{2} \log |\Sigma_i| + \log P(\omega_i) \tag{6.2}$$

Here, Σ_i represents the covariance matrix for class i, μ_i is the mean of class i, and $P(\omega_i)$ is the prior probability of class i. Unlike LDA, QDA permits a quadratic decision boundary allowing for better separation of data points belonging to different class labels, thanks to the class-specific covariance matrices.

6.3 Logistic Regression (LR)

Logistic Regression (LR) (Cox, 1958) serves as a foundational model in our methodology to predict FX rate trends. Its simplicity, interpretability, and strong statistical foundation make it a suitable choice for our task.

The prediction of 'the direction of next day's return' can be expressed by the LR model given below where X represents the tuple of features associated with our data point and β represent the linear coefficients of the various features representing our FX dataset:

$$\Pr(\mathrm{Y} = 1 \mid \mathrm{X}) = \frac{\exp(\beta_0 + \beta \mathrm{X})}{1 + \exp(\beta_0 + \beta \mathrm{X})} \tag{6.3}$$

6.3.1 Regularised Logistic Regression

Considering the possibility of multicollinearity among our predictors, which could adversely impact our model's performance, we employ Regularised Logistic Regression. Specifically, we use Lasso Logistic Regression (L1 regularisation) to manage multicollinearity and prevent overfitting. This approach effectively performs variable selection, by steering the coefficients of less significant features toward zero.

6.3.2 Cost Function

The cost function for our Regularised Logistic Regression model incorporates both the log-likelihood loss function and an L1 regularisation term. The cost function $J(\beta)$ is:

$$J(\beta) = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(p(Y_i = 1|X_i)) + (1 - y_i) \log(1 - p(Y_i = 1|X_i))] + \lambda \sum_{j=1}^{p} |\beta_j|$$

Here, λ is the regularisation parameter, controlling the strength of the L1 penalty. A larger value of λ implies more regularisation and drives more coefficients toward zero achieving a sparser model, where the direction of the next day's return is dependent on limited features.

6.4 Random Forest (RF)

Random Forest (RF) is an ensemble learning method that constructs multiple decision trees and outputs the class determined by the majority voting among the trees (Breiman, 2001b). We chose RF for its robustness against overfitting and its ability to handle large numbers of features − characteristics which are important in our research, given the complexity and high dimensionality of our dataset.

Given the diverse nature of our dataset, having an ensemble method like RF offers us a more comprehensive perspective by considering multiple possible decision paths. Additionally, RF's technique of randomly subsampling the data and features infuses variety into our model, reducing prediction variance and increasing the model's generalisability to unseen data.

6.4.1 Tree Generation

Within our RF, the decision trees are generated using a bootstrap sample of the training data. A randomly chosen subset of features is considered for splitting at each node. In our case, the maximum depth of each tree, a critical hyperparameter of the RF, is tuned to optimally fit our dataset. We utilise the Gini Index as our split criteria during decision tree generation. It quantifies the impurity within our data subsets. Given L as our training dataset, p_i as the probability of an instance in class i, and k as the number of classes, the Gini Index is calculated as:

$$\text{GINI(L)} = 1 - \sum_{i=1}^{k} (p_i)^2 \tag{6.4}$$

A lower Gini Index indicates a more homogeneous subset, which aids in creating more precise decision trees within our RF.

6.5 Support Vector Machines (SVM)

The primary task of Support Vector Machines (SVMs), invented by Cortes and Vapnik (1995), is to find an optimal hyperplane that maximally separates the observations of different class labels (up and down in our study) represented by different variables associated with the FX markets.

6.5.1 Gaussian SVM

The complexity of FX market behaviour necessitates the use of more flexible and adaptive methods for classification. We employ Gaussian (Radial Basis Function) SVM to effectively model the non-linear structure of our data. In the Gaussian SVM, the training data are mapped to a higher-dimensional feature space, making it possible to construct an optimally separating hyperplane, even for non-linearly separable data.

For the given set of training examples $\{(x_1, y_1), ..., (x_N, y_N)\}$, where $x_i \in R^{20}$ represents an observation from the FX market dataset restricted to 20 selected features and $y_i \in \{0, 1\}$ corresponds to the class labels we aim to predict, we employ the Gaussian kernel:

$$K(x, x') = \exp\left(-\gamma \|x - x'\|^2\right)$$

The hyperparameter γ controls the width of the Gaussian kernel and implicitly determines the complexity of the resulting model. Larger values of γ result in more complex models as they consider fewer surrounding points for the construction of the hyperplane.

Our goal is to solve the following dual form of SVM:

$$\max_{\alpha} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j K(x_i, x_j) \tag{6.5}$$

subject to

$$0 \leq \alpha_i \leq C, \quad \sum_{i=1}^{N} \alpha_i y_i = 0, \quad i = 1, ..., N$$

Once the dual problem is solved, we can classify new examples using the sign of the

decision function:

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b$$

6.5.2 Long Short-Term Memory (LSTM)

The inherently dynamic nature of the FX markets and the continuous temporal flow of information make them complex yet rich areas for predictive modelling. Traditional models often struggle with capturing these complexities due to the non-linear and non-stationary time-series data involved. Recognising these challenges, it is imperative to employ Long Short Term Memory (LSTM), a sophisticated variant of recurrent neural network (RNN) architecture, known for its adeptness in managing long-term dependencies (Hochreiter and Schmidhuber, 1997).

Unlike conventional feedforward neural networks that consider input instances as independent, LSTMs possess an intrinsic memory in the form of a hidden state that propagates information across time steps. This makes LSTM particularly suited to our goal of predicting the next-day FX rate movement, which is inherently reliant on the sequence of preceding movements.

6.5.3 Addressing the Vanishing and Exploding Gradient Problem

Standard RNNs, while useful for sequential data, suffer from the problem of vanishing and exploding gradients, especially when dealing with long sequences. This occurs as the effect of a given input on the hidden layer can essentially vanish or explode, especially over many time steps, causing difficulties in learning long-range correlations. LSTMs overcome this issue by introducing a 'cell state', a mechanism that allows the network to control and manage the flow of information over long sequences. In our FX dataset, where long-term trends and movements can hold predictive power, the cell state of the LSTM plays a significant role.

6.5.4 Leveraging the Unique Structure of LSTM

LSTM's unique structure, as can be seen in Figure 6.1, involves the replacement of usual hidden layers with specialised LSTM cells. These cells are composed of three types of gates: the forget, input, and output gates, which operate in conjunction to process information and determine the cell state and hidden state for each time step.

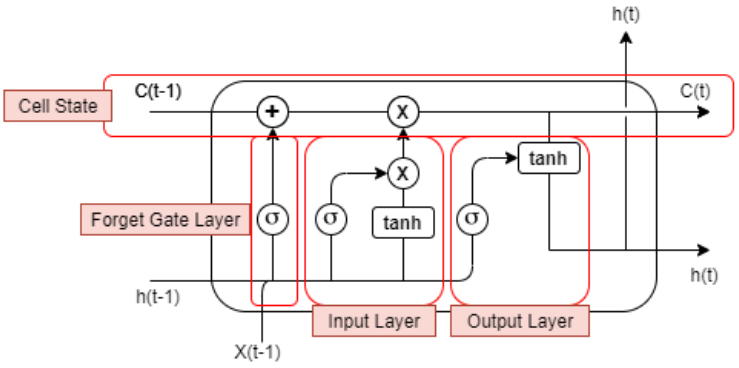


Figure 6.1: LSTM cell

6.5.5 Implications of LSTM's Gates for FX Rate Prediction

The input gate, handling records from our FX dataset, determines how much of the current input should influence the cell state. This gate is crucial as it brings in the most recent FX rate information to update the model considering the liquidity of the FX market.

The forget gate, as the name suggests, allows the model to 'forget' or discard irrelevant or outdated information from the past. This becomes highly relevant in the volatile FX markets where older trends or events may no longer hold predictive power.

The output gate is responsible for generating the final prediction outputted by the LSTM cell. It effectively creates a filtered version of the cell state, which becomes our prediction of the FX rate movement.

Furthermore, to ensure the robustness and accuracy of our model, we apply feature scaling to our dataset before feeding it into the LSTM. This is a crucial pre-processing step that helps in normalizing the input features within a specific range, thereby preventing any single feature from overpowering others. This normalisation strategy also enhances the model's ability to generalise from our training data to unseen data, thus potentially improving its predictive accuracy. In addition, scaling accelerates the model's convergence during the training phase, allowing us to attain a trained model faster.

6.6 eXtreme Gradient Boosting (XGBoost)

Given the complex nature of FX markets, with their high volatility, non-linear relationships, and potential for unexpected exogenous shocks, the need for a robust, accurate, and efficient machine learning algorithm becomes paramount. To address this, we also employ XGBoost, an implementation of gradient-boosting machines specifically designed for performance and computational speed.

The central function that XGBoost optimises is:

$$\Omega(f) = \sum_{i=1}^{N} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{6.6}$$

The differentiable convex loss function l measures the discrepancy between the observed and predicted directional trends of FX rates. $\Omega(f_k)$ is a regularisation term that penalises the complexity of the model.

6.6.1 Tree Boosting

XGBoost's sequential tree building is particularly well-suited to the task of FX rate prediction, where each tree in the model aims to correct the errors of its predecessor. By focusing on reducing misclassifications and minimising the loss of training data, XGBoost is capable of capturing intricate market dynamics over time.

6.6.2 Regularisation

A key advantage of XGBoost, which makes it suitable for our task, is the inclusion of a regularisation term in the objective function. This term mitigates the risk of overfitting by controlling model complexity. As our FX rate data can be noisy and prone to overfitting, this feature of XGBoost becomes particularly valuable.

6.6.3 Sparsity Aware

Given the real-world challenge of handling missing values in financial data, XGBoost's capacity to deal with such values is crucial. Its default strategy searches for the optimal imputation value to reduce loss, ensuring our model's robustness even in the presence of missing market data.

6.7 Hyperparameter Optimisation

All the following models, utilise the validation set, as explained in Section 4, to optimise the hyperparameters. The hyperparameters differ per model and are as follows:

6.7.1 Logistic Regression (LR)

The hyperparameter present in LR is λ, the regularisation parameter controlling the strength of the L1 penalty, which this study tunes by using the MER as the criterion.

6.7.2 Random Forest (RF)

The primary hyperparameters of RF include max_depth (the maximum number of levels considered while constructing the tree), max_features (maximum number of features considered for splitting a node which include the various technical indicators added in the feature space), min_samples_leaf (minimum number of observations allowed in a leaf node), min_samples_split (minimum number of observations placed in a node before the node is split), and n_estimators (number of trees in the forest). These will again be tuned by utilising the misclassification rate as the criterion. RF's performance is usually robust to the exact settings of these hyperparameters, but tuning can further enhance the model's predictive power.

6.7.3 Support Vector Machines (SVMs)

This study obtains the best-separating hyperplane for the various markets by taking into account the misclassification cost by tuning the hyperparameters, such as the adequate neighbourhood size of the RBF kernel γ and the regularisation parameter C.

6.7.4 Long Short-Term Memory (LSTM)

This study optimises the hyperparameters, units (number of neurons in the layers), number of layers, dropout (dropout rate to determine the number of neurons which get activated in the subsequent layers), learning_rate (learning rate to determine how quickly or slowly the model learns from the training data), optimiser (such as tf.keras' optimisers.Adam), present in the LSTM network by considering the MER as the primary criterion.

6.7.5 eXtreme Gradient Boosting (XGBoost)

Hyperparameters leveraged for the XGBoost model in our study include max_depth (controls the maximum amount of partitioning of FX dataset into smaller subsets for achieving pure subsets with only one class labels), learning_rate (the rate at which patterns within our dataset are discovered), min_child_weight (total number of decision trees to be built during the training process), gamma (controls the minimum level of error reduction in the loss function), subsample (fraction of dataset used for building each tree in the training), colsample_bytree (fraction of trees considered for building the trees).

7. Interpretability Methods

7.1 Tree-based Feature Importance Measure of XGBoost

The XGBoost algorithm, beyond its predictive performance, offers a built-in feature importance method which allows some degree of interpretability. This provides us with an initial interpretability mechanism, which although it is global and may fail to capture all the complexities and nuances of the model, serves as a stepping-stone towards deeper interpretation.

In the context of our research, feature importance is gauged based on the contribution of each feature to the model's prediction of the FX market. This is quantified by aggregating the improvement in prediction accuracy that each feature brings across all trees in the model. A feature that frequently results in significant prediction improvement will thus receive a high importance score.

However, this method can sometimes lead to inconsistent or even deceptive results. For instance, consider a feature that is extremely influential under specific market conditions but irrelevant in others. Such a feature might receive a low average importance score, belying its true significance in certain scenarios. This lack of consistency and potential inaccuracy makes it clear that while XGBoost's feature importance provides a helpful initial interpretability layer, it is often insufficient to fully decipher the model's logic.

This inconsistency and potential inaccuracy are why XGBoost models are still often considered 'black boxes' despite this feature importance method. Therefore, although this method provides a useful starting point for interpretability, it is often insufficient for comprehensive interpretability, which is why we make the further comparison with the SP-LIME and PFI models.

7.2 Locally Interpretable Model-Agnostic Explanations (LIME) Algorithm

The Local Interpretability Model-Agnostic Explanations (LIME) Algorithm is designed to help better understand model predictions and potentially build more trust in a model. As the name states, LIME is locally faithful and can be applied to any model as it only utilises its inputs and outputs.

The LIME algorithm tries to fit an interpretable model that is locally faithful to the original model, which classifies the FX rate to go up or down. This study exclusively concentrates on linear models, obtained by performing ridge regression, to restrict the complexity of the locally interpretable model $f : R^p \to R$. This results in $f(\mathbf{x}) = \mathbf{w}\mathbf{x}'$, where $\mathbf{w}$ is a $p \times 1$ vector indicating the weights and p is the number of variables which differs across analyses in this study. Moreover, define $h : R^p \to R$ as the true unknown complex model.

Both f and h output the probability that the input translates to an uptick in the FX returns. By combining all of this, the following minimisation problem can be obtained:

$$\varphi(x) = \arg\min_{\mathbf{w}} \Big\{ \sum_{\mathbf{z} \in Z} \pi_{\mathbf{x}}(\mathbf{z}_i)(h(\mathbf{z}_i) - \mathbf{w}\mathbf{z}_i')^2 + \lambda ||\mathbf{w}||^2 \Big\}. \tag{7.1}$$

Here, $\mathbf{x}$ and $\mathbf{z}_i$ are $p \times 1$ vectors. Moreover, $\mathbf{x}$ is the input vector and Z the set of all perturbed vectors $\mathbf{z}_i$; these perturbed vectors are obtained by sampling, for each feature

separately, from the normal distribution with mean and standard deviation taken from the feature's training data.

Furthermore, $\pi_{\mathbf{x}}(\mathbf{z}_i)$ represents the proximity measure between $\mathbf{z}_i$ and $\mathbf{x}$. This study takes $\pi_{\mathbf{x}}(\mathbf{z}_i) = exp\Big(-\frac{||\mathbf{x}-\mathbf{z}_i||^2}{2\sigma^2}\Big)$, which is the radial basis function (RBF) kernel with σ as the kernel width and symbolises a heatmap of importance around the input vector $\mathbf{x}$.

Before the minimisation problem is optimised, the numerical data is discretised into quartiles to control for two important things; the different ranges of the exploratory variables and the concept of double negatives (i.e. a negative weight for a negative feature is a positive contribution).

Finally, by minimising the objective function, utilising all the quartiles, this study obtains the weights, indicating the LIME explanations for the variables.

7.2.1 Submodular Pick (SP)-LIME

As mentioned before, LIME provides an understanding of the model by locally explaining the individual instances present in the dataset. It cannot provide a global picture, hence it cannot be used to perform comparative analysis between several interpretation models. Thus, this study uses a particular type of LIME method; namely, Submodular Pick (SP)-LIME.

SP-LIME overcomes this problem by carefully selecting a few instances out of all the possible instances which can capture information about the entire dataset and provide a global picture. It leverages the capability of LIME to form human-understandable explanations by constructing an explanation matrix that consists of local-linear representations of all the instances, say X available in the dataset. From the available set of X instances, the PICK step of the algorithm selects $B \in X$ instances that provide non-redundant explanations encompassing the global behaviour of our model. This is done mathematically by maximising the non-redundant coverage with importance function defined as:

$$c(V, W, I) = \sum_{j=1}^{p} \mathbb{1}_{[\exists i \in V: W_{ij} > 0]} I_j, \tag{7.2}$$

and choosing the set of instances in an iterative greedy fashion, in a sense that it adds the instance i to the set V for which there is the highest increase in marginal coverage. Moreover, the W_{ij}'s indicate the weights following from the LIME results and the feature importance is defined as $I_j = \sqrt{\sum_{i=1}^{n} |W_{ij}|}$.

7.3 Permutation Feature Importance (PFI)

To determine the significance of a feature, Permutation Feature Importance (PFI) calculates the loss increment when the values in a feature are randomly permuted.

The method involves randomly permuting the values of a single feature j and measuring the resulting increase in the model's prediction error $e = L(y, f(X))$. Based on this error we calculate the accuracy of the model, $accuracy = 1 - e$. If permuting the values of a feature decreases the accuracy, the feature is considered important, because the model relied on it for accurate predictions. Conversely, if permuting the values of a feature has little effect on the accuracy, the feature is considered unimportant. This technique was introduced by Breiman (2001b) for Random Forests and has since been adapted for the use with other models. A. Fisher, Rudin, and Dominici (2019b) proposed a model-agnostic version of the

feature importance and called it Model Class Reliance (MCR). PFI has the advantage of not requiring the model to be retrained, making it computationally efficient, especially for complex models and large feature spaces.

Thus, only the features which are associated with a positive decrease in the model's accuracy are considered essential for the specific model in consideration while those having negative or zero values with regard to the decrease in model accuracy are considered non-essential.

The PFI algorithm based on A. Fisher, Rudin, and Dominici (2019b) starts with estimating the original model error $e_{orig} = L(y, f(X))$, which is represented as the binary loss function in our specific case. Subsequently, all the features are permutated one by one and their respective errors $e_{perm,j} = L(y, f(X_{perm,j}))$ are estimated. Finally, the permutation feature importance can be calculated as the difference $PFI_j = e_{perm,j} - e_{orig}$.

8. Model Evaluation

This section elaborates on the model evaluation metrics utilised in this study. For the comparative analysis between interpretation methods this study utilises a dual-evaluation method, offering a well-rounded comparison of model efficiency, emphasising both prediction accuracy and computational demands. Hence, this section is subdivided into two subsections, explaining the metrics used for these respective model evaluations.

8.1 Classification Performance Metrics

We employ a range of well-established performance metrics derived from the confusion matrix to evaluate and compare the performance of the different machine learning models and interpretation methods used in this study. The confusion matrix (see Figure 8.1) is a specific table layout that provides a visualisation of the performance of an algorithm, displaying measures such as true positive, false positive, true negative, and false negative predictions of the model.

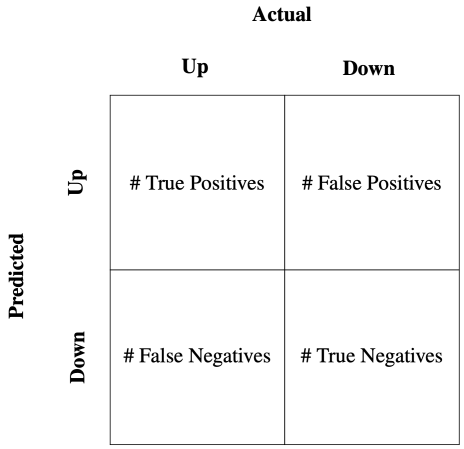


Figure 8.1: Confusion Matrix

From Figure 8.1, we can define:

$$\text{Precision} = \frac{\#\text{ True Positives}}{\#\text{ True Positives} + \#\text{ False Positives}}.$$
$$\text{Recall} = \frac{\#\text{ True Positives}}{\#\text{ True Positives} + \#\text{ False Negatives}}.$$

A high precision indicates that an algorithm produces statistically more relevant than irrelevant results, whereas a high recall suggests that an algorithm returned most of the relevant results. Therefore, one aims for high precision and recall values.

The F1 score is defined as a summary measure of recall and precision with an equal weight in the following manner:

$$F1 = 2\,\frac{\text{Precision}\ \times\ \text{Recall}}{\text{Precision}\ +\ \text{Recall}}.$$

In our study, we utilise our models on the following key metrics:

- Accuracy − measures the ratio of correctly predicted observations to the total number of observations, providing an overall indication of the model's correctness, regardless of the type of errors it may make. This can be misleading in imbalanced datasets.
- Balanced Accuracy − calculates the average of the correct prediction rates for each class individually to account for class imbalance. It is particularly valuable when false positives and false negatives have different consequences, as is often the case in financial predictions.
- F1 Score − combines Precision and Recall into a balanced measure of performance. It ranges from 0 (worst) to 1 (best) and reflects the model's ability to achieve high recall and precision, making it valuable in scenarios with uneven class distribution.

Considering the pre-established conceptual frameworks of this study involving financial data and the intricacies of often imbalanced FX markets, paramount importance is accorded to the balanced accuracy and F1 score metrics during the evaluation process.

8.2 Computational Efficiency

Alongside our core performance metrics, we also gauge the computational efficiency of our employed interpretability methods: XGBoost feature importance, PFI, and SP-LIME.

Computational efficiency, integral in handling large datasets, is often assessed using two factors: time complexity and actual running times. Time complexity, expressed via Big Oh notation, estimates the maximum operations an algorithm may require as data volume varies. For our study, we consider fixed numbers of features, which allows for a standardised comparison of time complexities based on the number of instances.

Running times offer an empirical perspective, indicating the real-world computational costs of each method.

9. Results

9.1 Comparison of the Machine Learning Model

This study starts by comparing the predictive power of the complex less-interpretable models: the Long Short-Term Memory (LSTM) type of Recurrent Neural Network (RNN) and the tree-based methods eXtreme Gradient Boosting (XGBoost) and Random Forest (RF) with the less complicated, more easily interpretable models: SVM with Gaussian Kernel, Logistic Regression, Linear Discriminant Analysis (LDA), and Quadratic Discriminant Analysis (QDA).

	SVM Gaussian Kernel		Logistic Regression		LDA		QDA		Random Forest	
Market	Accuracy	Balanced	Accuracy	balanced	Accuracy	balanced	Accuracy	Balanced	Accuracy	balanced
USD/EUR	0.516667	0.5	0.516667	0.5	0.535897	0.529991	0.492308	0.494395	0.525641	0.522889
USD/JPY	0.485897	0.505691	0.478205	0.5	0.501282	0.505092	0.482051	0.472331	0.483333	0.49058
USD/GBP	0.524359	0.5017	0.523077	0.5	0.517949	0.513599	0.484615	0.492528	0.512821	0.496007
USD/CHF	0.50641	0.516176	0.487179	0.499485	0.497436	0.505821	0.542308	0.535737	0.512821	0.519838
USD/NZD	0.535897	0.504808	0.533333	0.5	0.511538	0.503434	0.50641	0.510474	0.469231	0.475446
USD/CAD	0.491026	0.490906	0.494872	0.5	0.488462	0.489604	0.496154	0.497035	0.505128	0.505234
USD/SEK	0.496154	0.496154	0.496154	0.5	0.494872	0.495651	0.512821	0.512003	0.484615	0.485078
USD/DKK	0.491026	0.511982	0.460256	0.494931	0.484615	0.496378	0.502564	0.487404	0.507692	0.509768
USD/NOK	0.526923	0.527684	0.496154	0.496745	0.521795	0.521108	0.514103	0.51371	0.474359	0.474305
EUR/JPY	0.515385	0.504264	0.515385	0.5	0.469231	0.460831	0.484615	0.489418	0.476923	0.468767
EUR/GBP	0.539744	0.498264	0.54359	0.5	0.524359	0.50799	0.546154	0.505962	0.523077	0.504333
EUR/CHF	0.483333	0.5	0.483333	0.5	0.465385	0.475699	0.516667	0.520279	0.489744	0.494224
EUR/NZD	0.515385	0.499552	0.521795	0.519381	0.512821	0.50998	0.465385	0.480593	0.502564	0.503028
EUR/CAD	0.523077	0.506604	0.521795	0.5	0.521795	0.519261	0.5	0.493001	0.50641	0.506982
EUR/SEK	0.535897	0.529005	0.520513	0.50857	0.537179	0.533652	0.526923	0.513481	0.505128	0.506724
EUR/DKK	0.630769	0.5	0.630769	0.5	0.633333	0.504192	0.497436	0.464134	0.626923	0.50415
EUR/NOK	0.516667	0.514145	0.529487	0.5	0.537179	0.531391	0.502564	0.496731	0.487179	0.470064

	LSTM		XGBoost		True percentage of 0's	True percentage of 1's
Market	Accuracy	Balanced	Accuracy	Balanced		
USD/EUR	0.516667	0.5	0.510256	0.500727	0.483333333	0.516666667
USD/JPY	0.478205	0.475812	0.510256	0.518171	0.521794872	0.478205128
USD/GBP	0.514103	0.515022	0.488462	0.474146	0.476923077	0.523076923
USD/CHF	0.508974	0.51178	0.489744	0.497466	0.530769231	0.469230769
USD/NZD	0.515385	0.521291	0.524359	0.52421	0.466666667	0.533333333
USD/CAD	0.487179	0.487152	0.528205	0.528734	0.505128205	0.494871795
USD/SEK	0.503846	0.5	0.507692	0.508906	0.496153846	0.503846154
USD/DKK	0.507692	0.503106	0.534615	0.537478	0.535897436	0.464102564
USD/NOK	0.487179	0.487523	0.528205	0.529081	0.506410256	0.493589744
EUR/JPY	0.484615	0.5	0.510256	0.50458	0.484615385	0.515384615
EUR/GBP	0.54359	0.5	0.507692	0.499417	0.456410256	0.543589744
EUR/CHF	0.502564	0.494224	0.501282	0.510439	0.516666667	0.483333333
EUR/NZD	0.520513	0.508742	0.487179	0.487163	0.517948718	0.482051282
EUR/CAD	0.501282	0.498933	0.524359	0.522054	0.478205128	0.521794872
EUR/SEK	0.474359	0.5	0.496154	0.494232	0.474358974	0.525641026
EUR/DKK	0.470513	0.478786	0.610256	0.503896	0.369230769	0.630769231
EUR/NOK	0.491026	0.488718	0.517949	0.514901	0.470512821	0.529487179

Figure 9.1: (Balanced) Accuracies of complex and non-complex models

Table 9.1 illustrates that there are some foreign exchange (FX) pairs in which the more complex models perform better than, the less complex ones, but there are also several pairs where that is the other way around. However, there is one clear winner: the tree-based method XGBoost – it serves as an upper boundary for most markets across evaluation measures and is not far off for the others, successfully answering this study's first research question.

The balanced accuracy is the most important as this study wants to account for the potential occurrence of imbalanced upward/downward periods in financial markets. When looking more into this respective column defined in Table 9.1, it can be seen that XGBoost predicts best for almost half of the FX pairs (green); while giving above-average, close to optimal performance for four other FX pairs (orange).

Hence, the XGBoost model is the best model for this study to continue its subsequent

analyses. Research often considers XGBoost less complex and more interpretable than the LSTM type of RNN, as there are some possibilities to perform feature importance on tree-based methods. However, there are also some arguments against using this method of feature importance to interpret the XGBoost model; for example, the lack of consistency and accuracy. Therefore, the complexity of XGBoost, resulting in better predictions of FX pairs compared to the simpler models, still comes at the cost of interpretability. The model is known as a black box — which is precisely the domain this study will uncover further in the subsequent sections of this paper.

9.2 Comparison of the Interpretability Methods

We compare the results of four different XGBoost models to evaluate the predictive power of our interpretability methods. Each model uses five features, with the first model using the top five features selected by the SP-LIME model, the second using the top five features chosen by the PFI model, the third using the top five features selected by the tree-based feature importance measure of XGBoost, and the fourth model using five randomly selected features. These features are chosen from the pool of 20 market-specific features initially obtained through feature selection (SFFS).

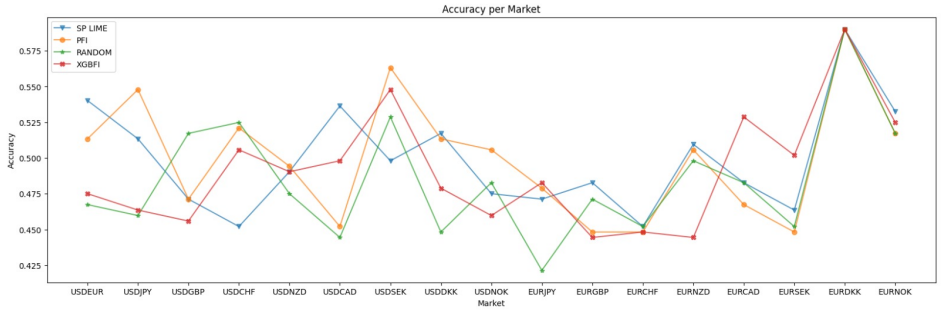
This study compares three key performance metrics − accuracy, balanced accuracy, and the F1 score. To better understand and interpret the results, this study first looks at each model's overall performance and, subsequently, continues by discussing the performance explicitly by market.

9.2.1 Overall Performance of the Interpretation Methods

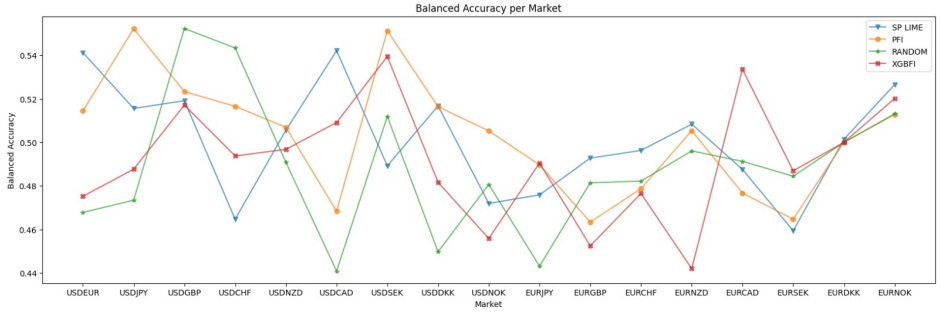
In Table 9.3, green indicates the most informative interpretation method, orange is the runner-up, and red is the worst, based on the different performance metrics. Therefore, a glance at Table 9.3 can already emphasise that the PFI model consistently demonstrated commendable performance across all the markets, particularly given the heightened importance assigned to balanced accuracy and F1 scores due to the often imbalanced nature of financial markets.

Figure 9.2 highlights these classification performance metrics even further − it shows that, for the (balanced) accuracy measure, SP-LIME and PFI primarily occur in the upper bound of the chart across all markets. RANDOM selection serves mainly as the lower bound of the accuracy measures, providing this study with evidence that SP-LIME and PFI provide insightful information about the features. There is less discrepancy in the F1 scores following from the interpretation methods across all the markets. However, Figure 9.3 and Figure 9.2c indicate that PFI is the overall winner on this front. As this study regards the balanced accuracy and F1 score as the most critical measures, PFI is classified as the best performer, answering the second research question of this study − PFI's strong performance suggests it may offer a powerful approach for interpreting machine learning models to classify the direction of the next day's return in FX markets.

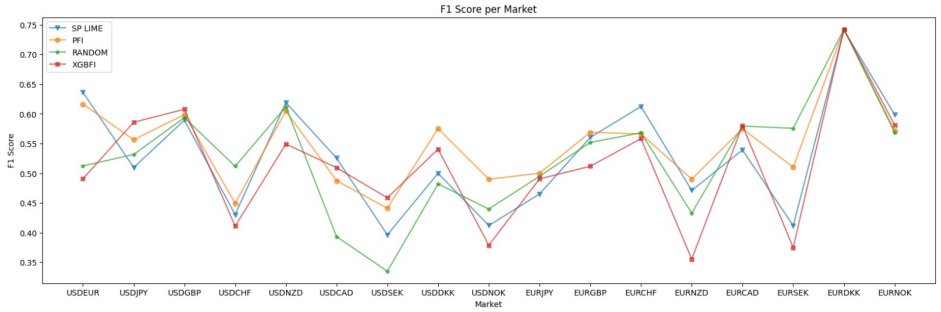
The SP-LIME model exhibits robust performance, especially in the USD/EUR, USD/CAD, EUR/CHF and EUR/NOK markets, indicating that it can also be a viable option for FX market interpretation. However, at the same time, it performed the worst in a considerable number of markets, suggesting some limitations in its adaptability across different market conditions.



(a) Accuracy per Market for all Interpretation Methods



(b) Balanced Accuracy per Market for all Interpretation Methods



(c) F1 Score per Market for all Interpretation Methods

Figure 9.2: Prediction performance following information provided by SP-LIME, PFI, RANDOM and XGBoost models across different markets using Accuracy, Balanced Accuracy, and F1 Score metrics (Line Graph Format)

The tree-based feature importance measure of the XGBoost model did not demonstrate competitive and consistent performance in the FX markets. For most markets, it equals the performance of RANDOM feature selection − providing evidence that this often-considered interpretation tool is not insightful in providing information for accurate classification in FX markets.

In short, SP-LIME showed strong results in some markets, while also performing worst in various others. Moreover, RANDOM selection and XGBoost's feature importance method appear to be the lower limit across all metrics. Finally, PFI consistently demonstrated high-level values in the (balanced) accuracy and F1 score metrics, suggesting its potential as a reliable FX market interpretation tool.

Type	Market	Accuracy	Balanced	F1 Score
XGB	USD/EUR	0.475096	0.47522	0.490706
XGB	USD/JPY	0.463602	0.487662	0.585799
XGB	USD/GBP	0.455939	0.517131	0.607735
XGB	USD/CHF	0.505747	0.493763	0.410959
XGB	USD/NZD	0.490421	0.496794	0.549153
XGB	USD/CAD	0.498084	0.50913	0.509363
XGB	USD/SEK	0.547893	0.53945	0.458716
XGB	USD/DKK	0.478927	0.481673	0.540541
XGB	USD/NOK	0.45977	0.45582	0.378855
XGB	EUR/JPY	0.482759	0.490517	0.490566
XGB	EUR/GBP	0.444444	0.452437	0.511785
XGB	EUR/CHF	0.448276	0.476502	0.558282
XGB	EUR/NZD	0.444444	0.441964	0.355556
XGB	EUR/CAD	0.528736	0.533598	0.580205
XGB	EUR/SEK	0.501916	0.486933	0.375
XGB	EUR/DKK	0.590038	0.5	0.742169
XGB	EUR/NOK	0.524904	0.520176	0.581081

Type	Market	Accuracy	Balanced	F1 Score
PFI	USD/EUR	0.51341	0.514445	0.616314
PFI	USD/JPY	0.547893	0.552214	0.556391
PFI	USD/GBP	0.471264	0.52332	0.598837
PFI	USD/CHF	0.521073	0.51656	0.449339
PFI	USD/NZD	0.494253	0.506941	0.60479
PFI	USD/CAD	0.452107	0.468314	0.487455
PFI	USD/SEK	0.563218	0.551333	0.441176
PFI	USD/DKK	0.51341	0.516389	0.575251
PFI	USD/NOK	0.505747	0.505291	0.490119
PFI	EUR/JPY	0.478927	0.489655	0.5
PFI	EUR/GBP	0.448276	0.463356	0.568862
PFI	EUR/CHF	0.448276	0.478725	0.566265
PFI	EUR/NZD	0.505747	0.505345	0.490119
PFI	EUR/CAD	0.467433	0.47672	0.574924
PFI	EUR/SEK	0.448276	0.46465	0.510204
PFI	EUR/DKK	0.590038	0.5	0.742169
PFI	EUR/NOK	0.517241	0.512824	0.571429

Type	Market	Accuracy	Balanced	F1 Score
RANDOM	USD/EUR	0.467433	0.467792	0.512281
RANDOM	USD/JPY	0.45977	0.473436	0.531561
RANDOM	USD/GBP	0.517241	0.55235	0.593548
RANDOM	USD/CHF	0.524904	0.543382	0.511811
RANDOM	USD/NZD	0.475096	0.490824	0.611898
RANDOM	USD/CAD	0.444444	0.440834	0.393305
RANDOM	USD/SEK	0.528736	0.511941	0.335135
RANDOM	USD/DKK	0.448276	0.449689	0.482014
RANDOM	USD/NOK	0.482759	0.480688	0.439834
RANDOM	EUR/JPY	0.421456	0.443103	0.494983
RANDOM	EUR/GBP	0.471264	0.481428	0.551948
RANDOM	EUR/CHF	0.452107	0.482221	0.567976
RANDOM	EUR/NZD	0.498084	0.496064	0.4329
RANDOM	EUR/CAD	0.482759	0.49127	0.579439
RANDOM	EUR/SEK	0.452107	0.484444	0.575668
RANDOM	EUR/DKK	0.590038	0.5	0.742169
RANDOM	EUR/NOK	0.517241	0.513147	0.568493

Type	Market	Accuracy	Balanced	F1 Score
SP-LIME	USD/EUR	0.54023	0.541251	0.636364
SP-LIME	USD/JPY	0.51341	0.515584	0.509653
SP-LIME	USD/GBP	0.471264	0.519134	0.589286
SP-LIME	USD/CHF	0.452107	0.464668	0.430279
SP-LIME	USD/NZD	0.490421	0.505529	0.618911
SP-LIME	USD/CAD	0.536398	0.542159	0.52549
SP-LIME	USD/SEK	0.498084	0.489179	0.396313
SP-LIME	USD/DKK	0.517241	0.51677	0.5
SP-LIME	USD/NOK	0.475096	0.471958	0.412017
SP-LIME	EUR/JPY	0.471264	0.475862	0.465116
SP-LIME	EUR/GBP	0.482759	0.49276	0.560261
SP-LIME	EUR/CHF	0.452107	0.496296	0.612466
SP-LIME	EUR/NZD	0.509579	0.508371	0.471074
SP-LIME	EUR/CAD	0.482759	0.487566	0.539249
SP-LIME	EUR/SEK	0.463602	0.459375	0.411765
SP-LIME	EUR/DKK	0.590038	0.501426	0.74092
SP-LIME	EUR/NOK	0.532567	0.526559	0.598684

Figure 9.3: Prediction performance following information provided by SP-LIME, PFI, RANDOM and XGBoost models across different markets using Accuracy, Balanced Accuracy, and F1 Score metrics (Table Format)

9.2.2 Market Specific Performance

Figure 9.2 is insightful in comparing each model's performance metrics within individual markets. This study considers balanced accuracy and F1 score as the most crucial evaluation metrics. Therefore, this section will primarily elaborate more on those graphs.

Concerning the F1 scores, it is easy to observe from Figure 9.2c that, although the differences are less present than with the (balanced) accuracy measures, the orange line, indicating the results following the PFI model, serves as the upper boundary for this evaluation measure. Hence, it is relatively easy to conclude that PFI outperforms all other interpretation methods regarding F1 score.

When shifting the focus to balanced accuracy there is a little bit more deviance. The PFI model outperforms all the other models in the USD/JPY, USD/NZD, USD/SEK and USD/NOK markets. In the USD/EUR, USD/CAD, USD/DKK, EUR/GPB, EUR/CHF, EUR/NZD, EUR/DKK, and EUR/NOK markets, SP-LIME outperforms. However, in each of these markets, PFI is close and often classifies as the runner-up. The reverse is less true; when PFI performs well, SP-LIME often ranks worse than XGBoost's feature importance, RANDOM selection, or both − suggesting that PFI more consistently acts as one of the higher performers.

The RANDOM feature selection or XGBoost's feature importance is the best performer in the USD/GBP, USD/CHF, EUR/JPY, EUR/CAD, and EUR/SEK markets. However, PFI takes up the runner-up position in most of these markets, supporting our previously

defined claim. The only exemptions are the EUR/CAD and EUR/SEK markets, for which PFI and SP-LIME both underperform. However, as this study makes many comparisons, expectedly, this phenomenon would occur occasionally.

This section dismantled the overall results per FX market and provided further reasoning for the conclusion drawn in the previous section − PFI predominantly outperforms or belongs to the best performers in comparison to the other interpretation methods in providing insightful information regarding the features for accurately classifying the direction of the next day's return in FX markets.

9.2.3 Computational Efficiency of the Interpretability Methods

To compare XGBoost's feature importance, SP-LIME and PFI based on computational efficiency, we indicate the running times of the algorithms in Figure 9.4. These times cover all 17 markets studied. Therefore, the average per-market runtime is about 0.12 seconds for XGBoost, 0.32 seconds for PFI, and approximately 29.51 seconds for SP-LIME.

Type	Time
XGBoost	2.07
PFI	5.5
SP-LIME	501.59

Figure 9.4: Runtime per Interpretation Method (analysis on all 17 markets)

The feature importance measure provided by the XGBoost model proved to be highly time-efficient, averaging about 0.12 seconds per market. XGBoost's time complexity arises from its part in the model's training process, which includes building multiple decision trees. For each tree, XGBoost scans through all features (F), assessing potential split points for each feature across all data points (N). This process, repeated for every decision tree in the ensemble, leads to a detailed time complexity of approximately $O(N * F)$. As this study fixes F, this simplifies to $O(N)$.

SP-LIME exhibited the longest running time of approximately 29.51 seconds per market due to its more intensive computational demands. SP-LIME involves generating a predetermined number of perturbed samples for each instance and training a simpler model to interpret the complex model's predictions. Its time complexity is $O(N * F^2)$, but again, as this study fixes F, it simplifies to $O(N)$.

PFI, on the other hand, demonstrated a modest increase in computational time over XGBoost, with an average run time of 0.32 seconds per market. PFI permutates each feature individually, measures the subsequent decrease in the model's performance, and incurs a time complexity proportional to the number of instances N and features F ($O(N * F)$). Given that this study fixes F, the time complexity simplifies to $O(N)$.

While PFI has a slightly increased computational cost, its average running time of 0.32 seconds per market was deemed an acceptable compromise, given its superior performance as an interpretability method. Notably, we observed that although there are differences in seconds between the different methods, these are relatively minor. Moreover, we found that all methods had an equal Big Oh time complexity of $O(N)$ due to the fixed number of features − implying that, even for larger samples, the differences in computational time among the methods should remain relatively small.

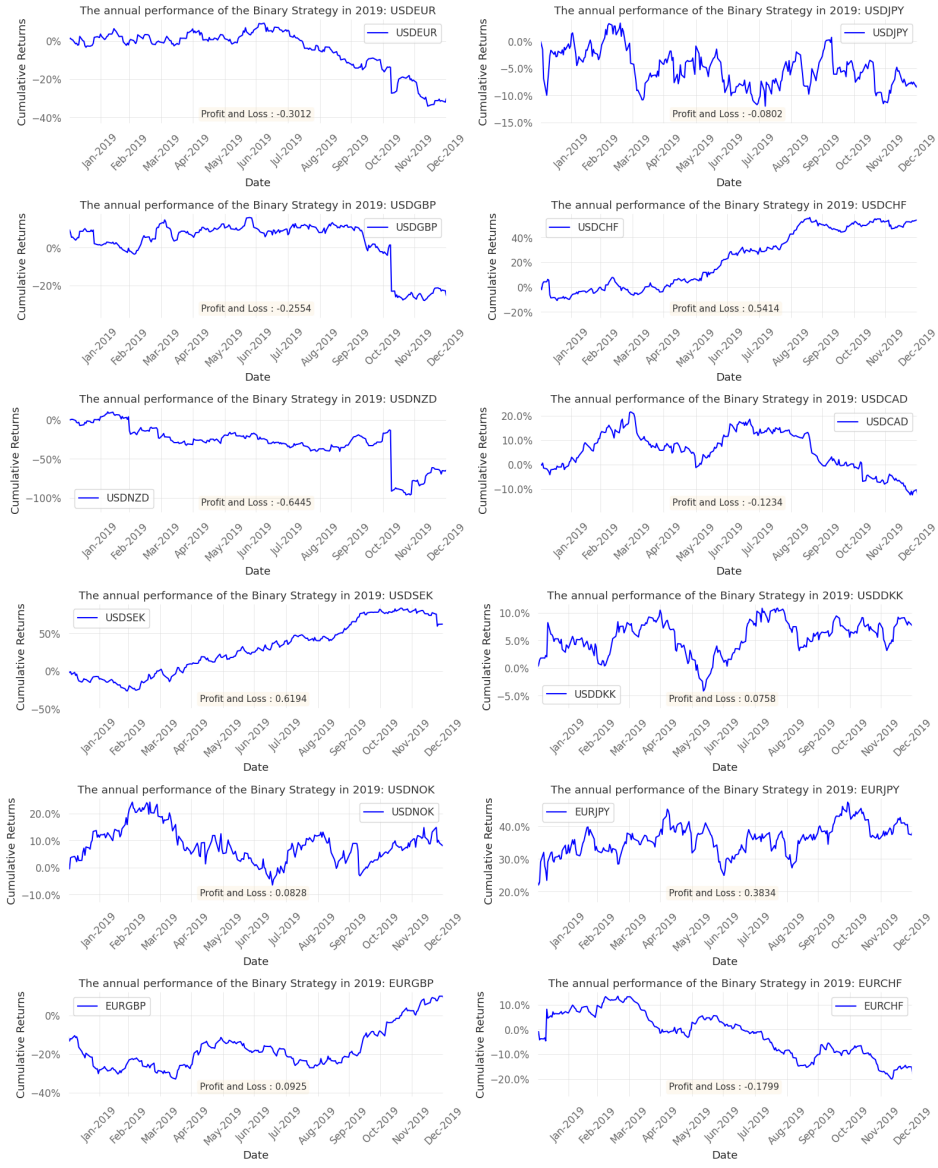
9.3 Interpretability Analysis through Trading Performance

9.3.1 Profit and Loss Analysis

Figure 9.5 presents the binary trading strategy's Profit and Loss (PnL) graphs across different currency pairs based on the classifications following from the XGBoost model, including the 20 market-specific features obtained through Sequential Forward Floating Selection (SFFS). Notable discrepancies can be observed among FX markets − profitable markets following the binary strategy have a slight dominance with nine to eight compared to the losing ones.

The mixed results demonstrate the complexity of FX market prediction and the difficulties faced in consistently achieving profitable trades − this could be due to the countries' differing economic landscapes, different market volatility levels, and many other factors.

It is important to note that while there overall is a slight dominance of profitable markets, there is ample room for improvement in refining the binary trading strategy. However, as the primary objective of this analysis is to demonstrate the usefulness of diverse interpretation methods, the binary trading strategy is preferred − this choice is rooted in the strategy's simplicity, which establishes a direct connection to the model's predictions while still being able to facilitate the identification of relative differences between months of superior and inferior performance, which is the information this study requires.



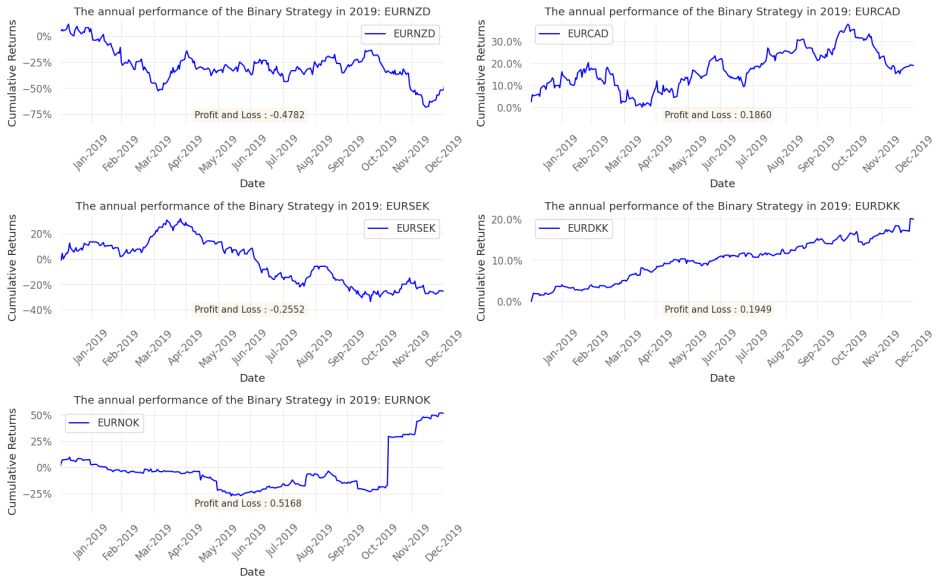


Figure 9.5: The Profit and Loss functions of the Binary Strategy in 2019 for all markets

9.3.2 Identifying Best and Worst-Performing Months

This paper used the Sharpe Ratio as the primary metric to identify the best and worst-performing periods. Figure 9.6 shows that the better (green) and worse-performing (red) months differ across markets. However, January occurs most frequently as the best-performing month, while February and November occur most frequently as the worst-performing month, with February never occurring as the best-performing month.

	USDEUR	USDJPY	USDGBP	USDCHF	USDNZD	USDCAD	USDSEK	USDDKK	USDNOK
2019-01	-0.031451841	-0.081228199	-0.051601064	-0.168184034	-0.016630716	-0.06628908	-0.270816072	-0.031122844	0.154350084
2019-02	-0.272697314	-0.260502058	-0.382193926	0.109812926	-0.243540263	0.226088925	-0.499926736	-0.268990704	0.131181554
2019-03	-0.187549343	-0.182233998	0.29789378	-0.222436292	-0.192324984	0.12106827	0.302798536	0.122233504	-0.102365134
2019-04	0.081486171	-0.179088878	-0.36572126	-0.018898848	-0.273929231	-0.742283885	0.078779302	-0.162889598	-0.423076902
2019-05	-0.308983412	-0.022771615	-0.003968807	0.130241014	0.133231455	-0.316635542	0.092152292	-0.509727986	-0.059828695
2019-06	0.021887775	-0.446158547	-0.144943486	0.290298135	-0.041901442	0.354664248	0.040336623	-0.14301265	-0.237336465
2019-07	-0.165661702	-0.220077573	-0.28233801	0.049739826	-0.529995472	-0.086618623	0.162831896	0.06228191	-0.098880978
2019-08	-0.602296405	0.013030073	-0.093384264	0.279998032	-0.171842653	-0.171558425	-0.056860723	-0.139532963	0.131816031
2019-09	-0.535787953	-0.018937211	-0.032955455	0.104079138	0.126762574	-0.786397439	0.333784835	-0.320734941	-0.277883575
2019-10	-0.207041842	-0.201140879	-0.424429655	-0.044030287	0.093189805	-0.175757779	0.451622165	-0.25464614	-0.074982618
2019-11	-0.113738474	-0.325845839	-0.219253898	-0.106517789	-0.171199668	-0.236415429	-0.155116283	-0.007762142	0.131799652
2019-12	-0.409342463	0.012133488	-0.081257579	-0.037550723	0.160791614	-0.266324158	-0.216500453	-0.165522873	-0.086294573

	EURJPY	EURGBP	EURCHF	EURNZD	EURCAD	EURSEK	EURDKK	EURNOK
2019-01	0.240266085	-0.520057987	0.028990352	-0.028729233	0.264095336	0.134162795	-0.083999757	-0.065694459
2019-02	-0.12770413	0.01773411	-0.03854966	-0.360528188	-0.052849385	-0.496862592	-0.869364656	-0.720475766
2019-03	0.030753405	-0.300998519	-0.04222096	-0.331157149	-0.386397299	0.284290763	-0.57195051	-0.166319489
2019-04	-0.156091946	-0.064373556	-1.972828926	0.30160346	0.038932332	0.136380118	-0.15344703	-0.363525854
2019-05	-0.20626672	0.233737207	-0.005414516	-0.348858183	-0.065152011	-0.642571458	-0.283960824	-0.338195536
2019-06	-0.28338865	-0.554143072	-0.273366417	-0.040680585	0.239828027	-0.242593059	-0.50612403	-0.296576239
2019-07	0.160100986	-0.214316631	-0.611375539	-0.116248826	-0.197290207	-0.474635572	-0.556941677	0.18113769
2019-08	-0.275047955	-0.465324453	-0.509760509	0.125908336	0.123028647	0.105895513	-0.35927196	0.122821026
2019-09	0.092301258	0.023623739	-0.256458465	-0.125987456	-0.290268389	-0.680979717	-0.046843275	-0.439203521
2019-10	0.065858876	0.231393559	-0.110094527	0.009473783	0.237654794	-0.182121343	-0.209489665	-0.179631703
2019-11	-0.409708353	0.287755573	-0.52523933	-0.372027384	-0.463357443	0.211014912	-0.253644079	0.226568702
2019-12	-0.022167118	0.12353107	-0.171931955	0.012685667	-0.278301287	-0.249776952	0.018135067	0.377809626

Figure 9.6: Sharpe Ratio for each month of the Testing Period

The primary occurrence of January as the best-performing period could potentially be explained following the research done by Girardin and Namin (2019). They found that

USD-based FX markets returns were higher in January compared to the rest of the year, introducing this so-called January effect within FX markets − frequently, explained by the concept of tax-loss selling. The presence of higher returns in January, along with an increased frequency of upticks in the next day's return within the FX markets, has the potential to simplify the task for our classification model as this model relies on information derived from trend and momentum-based technical indicators, which in the presence of the January effect will provide the most accurate information to our model.

In the forthcoming sections of this paper, this study applies the PFI method to the XGBoost model and its predictions during the best and worst-performing periods as identified based on the Sharpe Ratio − this will indicate the different features that were important in making that month's predictions enabling this study to gain insights into the drivers in these respective periods.

9.3.3 Applying PFI as an Interpretation Tool to the Best and Worst-performing Months

This section illustrates the usefulness of interpretation models in the complex financial domain − more specifically, FX markets. The previous section indicated the best and worst-performing periods, based on the Sharpe Ratio, following a binary trading strategy. Subsequently, this section elaborates on applying the previously identified best-performing interpretation model, the PFI model, to these periods and hopes to uncover valuable insights.

Figure 9.7 illustrates the results obtained from this analysis. As previously described, the coefficients generated by the PFI model represent the increase in misclassification error resulting from permuting the values of each specific technical indicator. Consequently, the magnitude of a positive coefficient indicates its significance, while the order of these coefficients reflects the relative performance of the technical indicators. Nonpositive coefficients are considered unimportant and therefore not shown in Figure 9.7. Hence, when a specific month for an FX pair does not indicate any technical indicators, none were classified by the PFI model as important.

From Figure 9.7, it becomes evident that seven different months in seven distinct FX markets did not reveal any significant technical indicators. Six of these seven months were classified as the worst-performing month, supporting the reasoning that the XGBoost model had difficulty predicting the direction of that month's next day's returns as no technical indicators seemed to be informative in making predictions.

As could have been expected, the number of essential variables and their positive magnitudes are higher for each best-performing month in comparison to the worst-performing month of a specific FX market. Figure 9.7 displays numerous different technical indicators; at first glance, it's hard to make sense of these results − suggesting the market dynamics for each currency pair exhibit a predominantly unique nature and a diverse set of factors are informative for each FX pair. However, to shed further light on the results and present supplementary insights into addressing the final research question raised in this study, we employ a subdivision of the features extracted from Figure 9.7 and utilise a color-coding scheme to enhance the clarity of the subsequent results.

Figure 9.7: Important Variables in Best and Worst-Performing Periods

Prevalent Technical Indicators

Figure 9.8 summarises statistics of the respective technical indicators occurring in the results of Figure 9.7. The most prominent technical indicators are displayed at the top of Figure 9.8. The technical indicators Directional Movement Index (DX), Commodity Channel Index (CCI), Aroon Down ($AROON\_DOWN$), Relative Vigor Index ($RVGI$), and Percent B ($PERCENT\_B$) occur six times, and the subset of technical indicators Price Rate of Change (ROC), Ucler Index ($UCLERINDEX$), Relative Volatility Index (RVI), and Stochastic Oscillator (SO) occur five times. Besides appearing most often in the results of Figure 9.7, they also have in common that they primarily seem to drive the best month's predictions – with merely once and for DX twice occurring as an essential feature in the worst month's predictions.

Moreover, for the majority of this subset, the one time they emerge as a critical variable for the worst month's predictions, they also occur as a driving feature for the best month's forecasts. This phenomenon involves the FX markets of the quoting currencies JPY and SEK against the base currencies EUR and USD. These markets' Profit and Loss functions lead to both profits and losses. Positive PnLs suggest that these technical indicators act as driving forces in both best and worst-performing months, as most classifications are accurate. Conversely, negative PnLs could also suggest the opposite, in that predicting the specific FX market is challenging, and none of the technical indicators is informative enough to make accurate classifications. Either way, the fact that most of these features occur as drivers of predictions in both the best and worst-performing months instead of solely the worst-performing months suggests that some market-wide characteristics drive these occurrences. Therefore, still supporting the importance of these features in accurate classification.

	Total Occuarances	Best Month Occurances	Worst Month Occurances	USD Occurances	EUR Occurances	Both Best and Worst Occurance
DX	6	4	2	4	2	USD/JPY
CCI	6	5	1	3	3	EUR/JPY
AROON_DOWN	6	5	1	2	4	EUR/JPY
RVGI	6	5	1	3	3	USD/JPY
PERCENT_B	6	5	1	4	2	USD/SEK
ROC	5	4	1	3	2	
ULCERINDEX	5	4	1	4	1	USD/SEK
RVI	5	4	1	4	1	
SO	5	4	1	2	3	
NATR	4	2	2	3	1	USD/JPY
MI	4	2	2	1	3	
MSD	4	4	0	4	0	
PSL	4	2	2	3	1	USD/NZD
DPO	4	3	1	3	1	
TRIX	3	1	2	2	1	
VORTEX_NEG	3	2	1	2	1	
MOM	3	2	1	2	1	USD/NZD
AROONOSC	3	2	1	1	2	
AROON_UP	3	2	1	1	2	
Parabolic_SAR	3	3	0	0	3	
VORTEX_POS	3	2	1	1	2	
Typical_Price	2	1	1	1	1	
CHOP	2	1	1	1	1	
Coppock_Curve	2	2	0	2	0	
RSI	2	2	0	2	0	
DEMA	2	2	0	2	0	
ADX	2	2	0	1	1	
TRIMA	2	2	0	1	1	
MIDPRICE	2	0	2	1	1	
CHANDELIER_SHORT	2	1	1	0	2	
BB_LB	2	1	1	0	2	EUR/SEK
LINREG	1	1	0	0	1	
ACC_MID	1	1	0	1	0	
BB_HB	1	1	0	1	0	
Adj Close	1	1	0	1	0	
CMO	1	1	0	1	0	
High	1	1	0	1	0	
PPO	1	0	1	1	0	
BBWIDTH	1	1	0	1	0	
Low	1	1	0	0	1	
WI	1	1	0	0	1	
ACC_UP	1	1	0	0	1	
DO_UP	1	1	0	0	1	
ATR	1	0	1	0	1	
KAMA	1	0	1	0	1	
APO	1	1	0	0	1	
Open	1	1	0	1	0	
TEMA	1	0	1	0	1	

Figure 9.8: Summary of the times and places in which the essential technical indicators have occured

Something that also stands out for this subset of technical indicators is that they most often occur as drivers in FX markets of USD pairs – suggesting that the importance of these technical indicators predominantly associates with the USD FX pairs. Keep in mind that while they are not predominantly observed, they do occur frequently as driving features in EUR FX pairs as well.

In conclusion, the fact that this subset of technical indicators is most frequently present in the results of Figure 9.7 and predominantly, almost exclusively, for best-performing months provides evidence for the importance of these features in driving accurate classification of the direction of the next day's returns in FX markets. Furthermore, the instances where they do emerge as crucial indicators in the classification during the worst-performing months appear to be attributable to market-wide characteristics.

The technical indicators utilised in this research can be subdivided into Momentum, Trend, and Volatility-based Indicators. While Momentum-based indicators show the strength

of an FX price move, Trend-based indicators provide information about the prevailing direction of the FX market. Furthermore, Volatility-based indicators help measure the uncertainty or stability of an FX pair's price.

Something that stands out is that the technical indicators most frequently classified as important in Figure 9.8 (DX, CCI, $AROON\_DOWN$, $RVGI$, and $PERCENT\_B$) are all Trend or Momentum-based technical indicators. While all measured differently, these indicators gauge the speed and strength of price movements in the underlying FX security and help identify trends and potential trend reversals − suggesting that our interpretation model often recognises these signals, measured in their respective ways, as important in classifying the direction of the next day's return in the various FX markets.

The subset of technical indicators (ROC, $UCLERINDEX$, RVI, and SO) that occur five times in Figure 9.8 cause the first Volatility-based technical indicators to arise in the results. ROC and SO belong to the group of Momentum-based technical indicators, a group of indicators already identified above. Moreover, $UCLERINDEX$ and RVI belong to the group of Volatility-based technical indicators, adding a measurement of the volatility for the underlying FX security. $UCLERINDEX$ measures the downside risk through the percentage drawdown from the most recent high, and RVI focuses on the standard deviation of price changes. These results illustrate that the XGBoost model considers the volatility of the price movements in the FX markets while making its predictions. However, PFI classifies this, on average, as a less critical signal than the speed and strength of price movements and price trends and potential trend reversals to the XGBoost model in making classifications that supplement better trading performance.

Salient Technical Indicators

Some technical indicators that particularly stand out are Moving Standard Deviation (MSD) and Parabolic SAR ($Parabolic\_SAR$). Besides often occurring in the results of Figure 9.7, the element that stands out is where they appear as important drivers of the classifications.

The technical indicator MSD occurs four times in total, all of which are best-month occurrences and are relatively high of positive magnitude. Moreover, all of these occurrences happen in USD-based FX markets − suggesting that this variable is a significant driver in accurately classifying the upward and downward movement of the next day's return in USD-based FX markets.

Furthermore, another technical indicator that particularly stands out is the $Parabolic\_SAR$. This technical indicator occurs three times in the results of Figure 9.7 and is exclusively present in the set of drivers of the best month's classifications. In contrast to the technical indicator MSD, $Parabolic\_SAR$ occurs solely in EUR-based FX markets. Therefore, following the same reasoning, the results suggest that the technical indicator $Parabolic\_SAR$ is a significant variable in driving accurate directional classifications of the next day's return in EUR-based FX markets.

MSD measures the standard deviation of a period's price data. Hence, it belongs to the Volatility-based technical indicators. As mentioned, this technical indicator occurs frequently and solely as an essential feature in the best-performing months of USD-based FX markets. Hence, this result implies that the volatility of the underlying USD-based FX securities seems most important for the XGBoost model in classifying the direction of the next day's return

in these respective markets, resulting in better trading performance.

On the contrary, $Parabolic\_SAR$, highlighting the general direction in which the FX security is moving, belongs to the set of Trend-based technical indicators. Moreover, as discussed, this technical indicator occurs frequently and solely as an essential feature in the best-performing months of EUR-based FX markets. Hence, in contrast to the USD-based FX markets, the trend and potential trend reversals in the EUR-based FX securities seem most important for the XGBoost model in classifying the direction of the next day's return in these respective markets, resulting in better trading performance.

Assessment of Notable FX Market Specifics

The USD/EUR, USD/CHF, EUR/CHF, and EUR/CAD FX markets have in common that the best-performing months are driven by a relatively small subset of important technical indicators. Specifically, the technical indicators Normalised Average True Range ($NATR$) and SO stand out. $NATR$, with a relatively high coefficient of 0.127, is most important in driving the predictions in the best-performing month of the USD/EUR market. By utilising ranges in the different price levels of the FX security, $NATR$ belongs to the Volatility-based technical indicators. Hence, the results suggest that the volatility of the USD/EUR FX market seems to be a critical driver for the XGBoost model in classifying the direction of the next day's return in this respective market, subsequently leading to better trading performance.

Moreover, regarding the USD/CHF market, what stands out is that SO is classified as the only variable important in making predictions for that market's best-performing month's forecasts − suggesting that the momentum, measured as the closing price relative to a range of the FX market's prices, is driving the XGBoost model's classifications and hence trading performance in this respective market.

Moreover, in the EUR/GBP FX market, the technical indicator Vortex Negative ($Vortex\_NEG$) receives the highest PFI coefficient across all FX -markets − namely 0.19 − indicating that this technical indicator, measuring adverse price movements, is a major driver for this market's best-performing month's predictions and thus trading performance.

Something that also stands out concerns the EUR/DKK FX market. No technical indicators seem essential in driving the best-performing month's classifications, but there are noteworthy high coefficients and critical variables during the worst-performing month's predictions. As this study utilises many different markets and timeframes, it can happen that, by coincidence, a particular month performs relatively well in comparison to other months while not having technical indicators drive these classifications. Moreover, the respective PnL could provide further reasoning − as the PnL steadily increases during the year, there do not seem to be significant discrepancies between good and bad-performing months concerning their classification accuracies, providing reasoning for this phenomenon.

Commonalities Among the Important Technical Indicators

As discussed, the technical indicators utilised in this study can be subdivided into Momentum, Trend, and Volatility-based technical indicators. The previous sections discussed the significant technical indicators and, by means of these indicators, examined the underlying FX securities' momentum, trend, and volatility as potential drivers for their respective markets. They all appeared to some extent as drivers for XGBoost's predictions concerning

the direction of the next day's returns. However, whether the FX securities' momentum, trend, volatility or a subset of these was classified by PFI as the primary driver(s) for these predictions and, hence, trading performance, differed per FX market.

Nevertheless, when focusing on the previously identified set of most often occurring indicators in best-performing trading months (DX, CCI, $AROON\_DOWN$, $RVGI$, $PERCENT\_B$, ROC, $UCLERINDEX$, RVI, and SO), with a marginal advantage of one, the Momentum-based technical indicators outnumber the Trend and Volatility-based indicators − suggesting that the speed and strength of FX price movements, measured in various ways, is most often the primary driver of the XGBoost model's classifications in the best-performing month of the next day's returns and thus are driving the better trading performance.

Something that also stands out is the fact that, although the technical indicators $Vortex\_NEG$ and $AROON\_DOWN$ were included at the start of this analysis alongside $Vortex\_POS$ and $AROON\_UP$, only $Vortex\_NEG$, identifying negative price movements, and $AROON\_DOWN$, measuring the downward trend, were classified by the PFI model as important to XGBoost's classifications during several best-performing trading months. Together with the fact that this research identified $UCLERINDEX$, measuring the percentage drawdown from the most recent high, to be one of the most critical technical indicators in XGBoost's classifications in the best-performing months across FX markets suggests that the downward aspect of the Momentum, Trend, and Volatility-based indicators is overall identified as more informative to the XGBoost model's directional classifications of the next day's returns in the best-performing months and, hence, contributes more towards better trading performance in comparison to their upward counterparts.

10. Conclusion

This study explores different interpretation methods and tests their usefulness within Foreign Exchange (FX) markets.

The first analysis regarded the comparison of different types of machine learning models – distinguishing between complex and non-complex models. The complex eXtreme Gradient Boosting (XGBoost) model outperformed all the others based on a balanced accuracy metric. This result allowed this study to add the model-specific tree-based XGBoost feature importance measure to the subsequent analysis.

More specifically, this paper focused on the Supmodular Pick - Locally Interpretable Model-Agnostic (SP-LIME), Permutation Feature Importance (PFI), and XGBoost's feature importance as interpretation methods. In the second analysis, this study introduced a methodology enabling comparing the various types of interpretation methods. This approach used information from the different interpretation models to select five technical indicators to retain for subsequent prediction. Subsequently, based on the balanced accuracy and F1 score metrics resulting from these subsequent predictions, the usefulness of the information provided by the interpretation models and, hence, their performance was evaluated. PFI was ranked as the most informative interpretation method, with the same asymptotic complexity as the other methods.

The final analysis of this study further expanded upon the groundwork laid thus far. It utilised the best-performing classification and interpretation models to perform an interpretability analysis through trading performance. In this manner, this study sought to exemplify the practicality of interpretation methods in the context of the financial domain, particularly within the FX markets.

This study utilised a simple binary trading strategy for this last analysis. The binary trading strategy hinges on a simple principle: we adopt a long position when our model predicts an uptick in the FX rate and opt for a short position when a downtick is projected – terminating the position after each day and assuming one trades with 100 times leverage or can simply trade these bigger positions. The choice to use this trading strategy is rooted in the strategy's simplicity, which establishes a direct connection to the model's predictions while still facilitating the identification of relative differences between months of superior and inferior performance, which is the information this study requires. Hence, a deliberate trade-off is made by prioritising simplicity and clarity over superior overall trading performance.

Profit and Loss functions were created for each FX market, confirming the non-superior trading performance. However, there did exist a marginal advantage for profitable FX markets. The monthly Sharpe Ratios were calculated to identify the best and worst-performing months. January occured most frequently as the best-performing month, which could be explained by the January Effect. Acknowledging that the following discussion involves a degree of speculation, the presence of higher returns in January following the January Effect, along with an increased frequency of upticks in the next day's return within the FX markets for that month, has the potential to simplify the task for our classification model as it utilises information following from Trend and Momentum-based indicators, which could explain the discovered phenomenon.

After applying the PFI model to these respective months, a myriad of intriguing findings surfaced. As could have been expected, the number of essential variables and their positive magnitudes are higher for each best-performing month in comparison to the worst-performing month of a specific FX market − suggesting that the XGBoost model received more informative information in these best-performing months to make accurate predictions, resulting in better trading performance.

The PFI model further identified the Directional Movement Index, Commodity Channel Index, Aroon Down, Relative Vigor Index, and Percent B as the predominant technical indicators, primarily important in best-performing months. Remarkably, all of these features are Trend or Momentum-based technical indicators, gauging the speed and strength of price movements in the underlying FX security and helping to identify trends and potential trend reversals. The technical indicators utilised in this research can be subdivided into Momentum, Trend, and Volatility Indicators. Therefore, the fact that no Volatility-based technical indicators were present suggests the importance of momentum and trend over volatility in classifying the direction of the next day's return in the various FX markets.

The set of technical indicators ranking second in terms of best-month importance occurrences include the Price Rate of Change, Ulcer Index, Relative Volatility Index, and Stochastic Oscillator, introducing the first Volatility-based technical indicators − namely, the Ulcer Index and Relative Volatility Index. These results illustrate that the XGBoost model considers the volatility of the price movements in the FX markets while making its predictions. However, PFI classifies this, on average, as a less critical signal than the speed and strength of price movements and price trends and potential trend reversals to the XGBoost model in making classifications that supplement better trading performance.

Furthermore, the technical indicators Moving Standard Deviation, Parabolic SAR, Normalised Average True Range, Stochastic Oscillator, and Vortex Negative particularly stand out. A measure of volatility, the Moving Standard Deviation, proved to be a significant driver in accurately classifying the upward and downward movement of the next day's return in USD-based FX markets. The Parabolic SAR indicator, highlighting the trend and potential trend reversals of the FX security, displayed similar importance for the EUR-based FX markets. The Normalised Average True Range and Vortex Negative stood out because of the magnitude of their PFI coefficient. With coefficients of 0.127 and 0.19, they demonstrated significant importance in the USD/EUR and EUR/GBP markets, respectively. These coefficients indicate that a volatility measure is of primary importance in driving the predictions in the best-performing month of the USD/EUR market and adverse price movements in the best-performing month of the EUR/GBP market. The Stochastic Oscillator, measuring the closing price relative to a range of the FX market's prices, seemed to be the only driver behind the classifications resulting in the month with the best trading performance in the USD/CHF market.

Moreover, it is noteworthy to mention that while Vortex Negative, Vortex Positive, Aroon Down, and Aroon Up were all included in the analysis, the PFI model classified only the technical indicators identifying negative trends and momentums (Vortex Negative and Aroon Down) as important to some of XGBoost's best-performing classifications − suggesting that the XGBoost model, in making accurate predictions to enhance trading performance, tends to attribute greater informativeness to the downward aspects of trend and momentum compared

to their upward counterparts. Additionally supported by the dominance of the Ucler Index, measuring the downward volatility, in driving many FX market's best month's predictions.

The emergence of significant Trend, Momentum, and Volatility-based technical indicators, each playing distinct roles in varying proportions and combinations, illustrates the complex and diverse nature of the FX markets. However, the results indicate valuable FX market-specific insights that could create trust in financial classification models and, when used appropriately, with knowledge regarding the underlying FX markets, could enhance trading and risk management practices.

This study provided a framework for examining different types of interpretation methods in the context of financial markets, focusing on FX market prediction more specifically. Furthermore, it made an effort to illustrate the usefulness of interpretation methods within the domain of financial markets. Nonetheless, this study recognises its limitations. The comparative analysis between interpretation methods is based on data from 2019, which, together with the complex and constantly evolving nature of FX markets, makes this study's findings dependent on this period. Further research could extend upon this limitation and analyse whether results are consistent across longer and/or differing periods in time − potentially including years during the pandemic, such as 2020. Moreover, this study primarily focused on market-specific technical indicators in its predictive analysis. However, including macroeconomic features, such as interest rates, inflation rates, GDP-growth rates and employment data, or sentiment-based types of variables could improve prediction and, thus, trading performance, potentially providing more insightful interpretation results − exploring these potential improvements is left for future research.

Moreover, this study acknowledges the assumptions and speculative nature inherent in its methodology, such as employing prediction accuracy as an indirect metric for interpretation performance. Nonetheless, it harmonises with the subsequent analysis, illustrating the practical implementation of an interpretation method in the context of FX market prediction and trading. Moreover, this study is well aware of the inherent simplicity of the binary trading strategy. It recognises that the efficacy of employing interpretation methods expands in tandem with the performance of trading methods. Additionally, a more comprehensive understanding of financial intricacies and FX market specifics could have yielded even more insightful conclusions. Nonetheless, it is vital to note that this study's primary objective did not centre on these aspects as they were mainly utilised to illustrate the possibilities associated with interpretation methods, leaving it open for future research to explore in greater depth.

Bibliography

Ahern, Isaac et al. (2019). "NormLime: A new feature importance metric for explaining deep neural networks". In: *arXiv preprint arXiv:1909.04200*.

Ahmed, Salman et al. (2020). "FLF-LSTM: A novel prediction system using Forex Loss Function". In: *Applied Soft Computing* 97, p. 106780.

Alexander, Carol (2001). "Market models". In: *A Guide to Financial Data Analysis* 1.

Altmann, André et al. (2010). "Permutation importance: a corrected feature importance measure". In: *Bioinformatics* 26.10, pp. 1340–1347.

Barton, Neal Andrew et al. (2022). "Predicting the risk of pipe failure using gradient boosted decision trees and weighted risk analysis". In: *npj Clean Water* 5.1, p. 22.

Borovkova, Svetlana and Ioannis Tsiamas (2019). "An ensemble of LSTM neural networks for high-frequency stock market classification". In: *Journal of Forecasting* 38.6, pp. 600–619.

Breiman, Leo (2001a). "Random forests". In: *Machine learning* 45, pp. 5–32.

— (2001b). "Random forests". In: *Machine learning* 45, pp. 5–32.

Carvalho, Diogo V, Eduardo M Pereira, and Jaime S Cardoso (2019). "Machine learning interpretability: A survey on methods and metrics". In: *Electronics* 8.8, p. 832.

Chan, Chun Sik, Huanqi Kong, and Liang Guanqing (May 2022). "A Comparative Study of Faithfulness Metrics for Model Interpretability Methods". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 5029–5038. DOI: 10.18653/v1/2022.acl-long.345. URL: https://aclanthology.org/2022.acl-long.345.

Chen, Tianqi et al. (2015). "Xgboost: extreme gradient boosting". In: *R package version 0.4-2* 1.4, pp. 1–4.

Cortes, Corinna and Vladimir Vapnik (1995). "Support vector machine". In: *Machine learning* 20.3, pp. 273–297.

Cox, David R (1958). "The regression analysis of binary sequences". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 20.2, pp. 215–232.

Doshi-Velez, Finale and Been Kim (2017). "Towards a rigorous science of interpretable machine learning". In: *arXiv preprint arXiv:1702.08608*.

ElShawi, Radwa et al. (2021). "Interpretability in healthcare: A comparative study of local machine learning interpretability techniques". In: *Computational Intelligence* 37.4, pp. 1633–1650.

Fan, Feng-Lei et al. (2021). "On interpretability of artificial neural networks: A survey". In: *IEEE Transactions on Radiation and Plasma Medical Sciences* 5.6, pp. 741–760.

Fisher, Aaron, Cynthia Rudin, and Francesca Dominici (2019a). "All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously." In: *J. Mach. Learn. Res.* 20.177, pp. 1–81.

— (2019b). "All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously." In: *J. Mach. Learn. Res.* 20.177, pp. 1–81.

Fisher Ronald Aylmer, Sir (1936). "The Use of Multiple Measurements in Taxonomic Problems." In: *Annals of Eugenics* 7, pp. 179–188.

Frye, Christopher, Colin Rowat, and Ilya Feige (2020). "Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability". In: *Advances in Neural Information Processing Systems* 33, pp. 1229–1239.

Girardin, Eric and Fatemeh Salimi Namin (2019). "The January effect in the foreign exchange market: Evidence for seasonal equity carry trades". In: *Economic modelling* 81, pp. 422–439.

Gite, Shilpa et al. (2021). "Explainable stock prices prediction from financial news articles using sentiment analysis". In: *PeerJ Computer Science* 7, e340.

Gumelar, Agustinus Bimo et al. (2020). "Boosting the Accuracy of Stock Market Prediction using XGBoost and Long Short-Term Memory". In: *2020 International Seminar on Application for Technology of Information and Communication (iSemantic)*. IEEE, pp. 609–613.

Guo, Runhua, Donglei Fu, and Giuseppe Sollazzo (2022). "An ensemble learning model for asphalt pavement performance prediction based on gradient boosting decision tree". In: *International Journal of Pavement Engineering* 23.10, pp. 3633–3646.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2001). *The Elements of Statistical Learning*. Vol. 22. Springer New York Inc.

Henrique, Bruno Miranda, Vinicius Amorim Sobreiro, and Herbert Kimura (2019). "Literature review: Machine learning techniques applied to financial market prediction". In: *Expert Systems with Applications* 124, pp. 226–251.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long short-term memory". In: *Neural computation* 9.8, pp. 1735–1780.

Islam, SFN, A Sholahuddin, and AS Abdullah (2021). "Extreme gradient boosting (XGBoost) method in making forecasting application and analysis of USD exchange rates against rupiah". In: *Journal of Physics: Conference Series*. Vol. 1722. 1. IOP Publishing, p. 012016.

JingTao, YAO and Chew Lim Tan (2001). "Guidelines for financial forecasting with neural networks". In: *Neural Inf. Process. Shanghai*.

Kumar, Raghavendra, Pardeep Kumar, and Yugal Kumar (2021). "Analysis of Financial Time Series Forecasting using Deep Learning Model". In: *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pp. 877–881. DOI: 10.1109/Confluence51648.2021.9377158.

Kurani, Akshit et al. (2023). "A comprehensive comparative study of artificial neural network (ANN) and support vector machines (SVM) on stock forecasting". In: *Annals of Data Science* 10.1, pp. 183–208.

Laugel, Thibault et al. (2018). "Comparison-based inverse classification for interpretability in machine learning". In: *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations: 17th International Conference, IPMU 2018, Cádiz, Spain, June 11-15, 2018, Proceedings, Part I 17*. Springer, pp. 100–111.

Li, Yunze et al. (2021). "Feature importance recap and stacking models for forex price prediction". In: *arXiv preprint arXiv:2107.14092*.

Linardatos, Pantelis, Vasilis Papastefanopoulos, and Sotiris Kotsiantis (2020). "Explainable ai: A review of machine learning interpretability methods". In: *Entropy* 23.1, p. 18.

Lipton, Zachary C, John Berkowitz, and Charles Elkan (2015). "A critical review of recurrent neural networks for sequence learning". In: *arXiv preprint arXiv:1506.00019*.

Los, C. (2000). "Wavelet Multiresolution Analysis of High-Frequency Asian FX Rates, Summer 1997". In: *International Finance*.

Lundberg, Scott M and Su-In Lee (2017). "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30.

Mahya, Parisa and Johannes Fürnkranz (2023). "An Empirical Comparison of Interpretable Models to Post-Hoc Explanations". In: *AI* 4.2, pp. 426–436.

Marcano-Cedeño, A. et al. (2010). "Feature selection using Sequential Forward Selection and classification applying Artificial Metaplasticity Neural Network". In: *IECON 2010 - 36th Annual Conference on IEEE Industrial Electronics Society*, pp. 2845–2850.

Melvin, Michael and Mark P Taylor (2009). "The crisis in the foreign exchange market". In: *Journal of International Money and finance* 28.8, pp. 1317–1330.

Molnar, Christoph (2020). *Interpretable machine learning*. Lulu. com.

P. Pudil J. Novovičová, J. Kittler (1994). "Floating search methods in feature selection". In: *Pattern Recognition Letters* 15.11, pp. 1119–1125.

Paliari, Iliana, Aikaterini Karanikola, and Sotiris Kotsiantis (2021). "A comparison of the optimized LSTM, XGBOOST and ARIMA in Time Series forecasting". In: *2021 12th International Conference on Information, Intelligence, Systems & Applications (IISA)*, pp. 1–7. DOI: 10.1109/IISA52424.2021.9555520.

Pang, Zhengqi and Ying Kong (2022). "Prediction of Household Carbon Emissions Based on SP-LIME and Ensemble Learning Models". In: *2022 IEEE 5th International Conference on Electronics Technology (ICET)*. IEEE, pp. 650–654.

Peng, Yaohao et al. (2021). "Feature selection and deep neural networks for stock price direction forecasting using technical analysis indicators". In: *Machine Learning with Applications* 5, p. 100060.

Picasso, Andrea et al. (2019). "Technical analysis and sentiment embeddings for market trend prediction". In: *Expert Systems with Applications* 135, pp. 60–70.

Rasheed, Jawad et al. (2020). "Improving stock prediction accuracy using cnn and lstm". In: *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*. IEEE, pp. 1–5.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016a). "" Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144.

— (2016b). "Model-agnostic interpretability of machine learning". In: *arXiv*.

Riff, Annur Syafiqah Abd, Rajendran Parthiban, and Jin Zhe (2022). "Ensemble model that minimizes the misclassification cost for imbalanced class credit data and its explanation using LIME". In: *AIP Conference Proceedings*. Vol. 2465. 1. AIP Publishing LLC, p. 040005.

Ryll, Lukas and Sebastian Seidens (2019). "Evaluating the performance of machine learning algorithms in financial market forecasting: A comprehensive survey". In: *arXiv preprint arXiv:1906.07786*.

Sayavong, Lounnapha (2019). "Research on Stock Price Prediction Method Based on Convolutional Neural Network". In: *2019 International Conference on Virtual Reality and Intelligent Systems (ICVRIS)*.

Schelthoff, Kai et al. (2022). "Feature Selection for Waiting Time Predictions in Semiconductor Wafer Fabs". In: *IEEE Transactions on Semiconductor Manufacturing* 35.3, pp. 546–555.

Schmidt, Philipp and Felix Biessmann (2019). "Quantifying interpretability and trust in machine learning systems". In: *arXiv preprint arXiv:1901.08558*.

Setzu, Mattia et al. (2021). "Glocalx-from local to global explanations of black box ai models". In: *Artificial Intelligence* 294, p. 103457.

Sheta, Alaa F, Sara Elsir M Ahmed, and Hossam Faris (2015). "A comparison between regression, artificial neural networks and support vector machines for predicting stock market index". In: *Soft Computing* 7.8, p. 2.

Stiglic, Gregor et al. (2020a). "Interpretability of machine learning-based prediction models in healthcare". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10.5, e1379.

— (2020b). "Interpretability of machine learning-based prediction models in healthcare". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10.5, e1379.

Strobl, Carolin et al. (2007). "Bias in random forest variable importance measures: Illustrations, sources and a solution". In: *BMC bioinformatics* 8.1, pp. 1–21.

Tsai, Yun-Cheng (2018). "Predict Forex Trend via Convolutional Neural Networks". In: *Journal of Intelligent Systems*.

Varma, Vinay and Machine Learning Engineer (2020). "Embedded methods for feature selection in neural networks". In: *arXiv preprint arXiv:2010.05834*.

Ververidis, Dimitrios and Constantine Kotropoulos (2005). "Sequential forward feature selection with low computational cost". In: *2005 13th European Signal Processing Conference*, pp. 1–4.

Vuong, Pham Hoang et al. (2022). "Stock-price forecasting based on XGBoost and LSTM." In: *Computer Systems Science & Engineering* 40.1.

Wu, Dingming, Xiaolong Wang, and Shaocong Wu (2022). "Jointly modeling transfer learning of industrial chain information and deep learning for stock prediction". In: *Expert Systems with Applications* 191, p. 116257.

Yetis, Yunus, Halid Kaplan, and Mo Jamshidi (2014). "Stock market prediction by using artificial neural network". In: *2014 world automation congress (WAC)*. IEEE, pp. 718–722.

Yıldırım, Deniz Can, Ismail Hakkı Toroslu, and Ugo Fiore (2021). "Forecasting directional movement of Forex data using LSTM with technical and macroeconomic indicators". In: *Financial Innovation* 7, pp. 1–36.

Yu, Yong et al. (2019). "A review of recurrent neural networks: LSTM cells and network architectures". In: *Neural computation* 31.7, pp. 1235–1270.

Zhang, Zhongheng et al. (2018). “Opening the black box of neural networks: methods for interpreting neural network models in clinical applications”. In: *Annals of translational medicine* 6.11.