PERFORMANCE OF MACHINE LEARNING
ADVANCED TECHNIQUES
IN STATISTICAL ARBITRAGE

by

Farooq Ahmed, B.Com (University of Karachi), MSc Intl Banking and Finance (Liverpool John Moores University) , M.B.A (Bayes Business School City University of London), MS Artificial Intelligence and Machine Learning (Liverpool John Moores University)

DISSERTATION

Presented to the Swiss School of Business and Management Geneva

In Partial Fulfillment

Of the Requirements

For the Degree

DOCTOR OF BUSINESS ADMINISTRATION

SWISS SCHOOL OF BUSINESS AND MANAGEMENT GENEVA

May, 2025

PERFORMANCE OF MACHINE LEARNING

ADVANCED TECHNIQUES

IN STATISTICAL ARBITRAGE

by

Farooq Ahmed

Supervised by

Dr Kamal Malik

APPROVED BY

Vasiliki Grougiou

Dissertation chair

RECEIVED/APPROVED BY:

Admissions Director

Dedication

To Christian Dunis (late) and Jason Laws, my teachers and mentors from Liverpool John Moores. To my Parents

Acknowledgements

Thanks to Almighty

Having been allowed to finish my dissertation opportunity, which is a partial requirement for DBA in Machine Learning. I express my gratitude to all those people who have supported my work.

I would like to thank my supervisor, Dr Kamal Malik whose expertise in Machine Learning has contributed to my thesis in numerous ways. Furthermore, her wise counsel, determination and encouragement have been an inspiration throughout the work.

I would like to thank all the personnel of SSBM for their support.

Finally, thanks to my wife, Nida for her continued support and encouragement. Without the support of everyone, none of this would have been possible.

Thank you,

ABSTRACT

PERFORMANCE OF MACHINE LEARNING ADVANCED TECHNIQUES IN STATISTICAL ARBITRAGE

Farooq Ahmed

2025

Dissertation Chair: Dr Kamal Malik

The thesis deals with machine learning-based algorithmic trading in currency markets. It addresses the financial machine learning optimisation recommendations of Marcos López De Prado (2018), emphasising the two main areas of improvement in machine learning by feature selection and meta-labelling. The study extends to the statistical arbitrage strategy using machine learning. By applying these techniques to statistical arbitrage, the study aims to identify and mitigate overfitting biases that commonly lead to algorithmic trading failures.

The methodology employs a comprehensive framework with a novel approach to currency pair selection using dimensionality reduction, clustering techniques, and cointegration testing. Using data from 82 currency pairs across G7, Major Cross, and Minor Cross categories from January 2019 to December 2023, the research implements Clustered Feature Importance (CFI) to optimise feature selection. Primary machine learning models (Logistic Regression, Random Forest, and Gradient Boosting) are then enhanced through meta-labelling to improve trading signal performance.

Empirical results demonstrate significant performance improvements across the five selected currency pairs (EURNOK/DKKZAR, EURPLN/DKKPLN, SEKNOK/SEKZAR, EURSGD/DKKSGD, and NZDCHF/USDZAR), with meta-labelled models showing improved risk-adjusted returns (Sharpe ratios increasing to 1.86 for the EURNOK/DKKZAR pair), substantial volatility reduction, and enhanced precision in trading signals (40.27% improvement). The framework proves particularly effective for Nordic and European currency pairs while maintaining stability across various market conditions.

The findings validate De Prado's recommendations when applied to statistical arbitrage in currency markets, offering theoretical contributions to financial machine learning and practical implications for quantitative trading strategies. This research provides valuable insights for portfolio managers and algorithmic traders seeking to improve performance through advanced machine learning techniques while addressing the challenges of overfitting and false discovery in trading model development.

Key Words: Meta-Labelling, Feature Selection, Machine Learning, Statistical Arbitrage, Currency, Mean-Reverting , Portfolio Construction

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATION

CHAPTER I:
INTRODUCTION

1.1 Introduction

The motivation to enhance existing quantitative trading strategies has led to the development of many new techniques. Statistical arbitrage is a type of market-neutral trading strategy which hedge funds use in the financial markets. Statistical arbitrage strategies are based on a predictive model with underlying insights using technical analysis, rules and historical datasets. In addition, market sentiment, participation volumes, and speculative movements based on news events can also impact these strategies, hence challenging currency arbitrage, forecasting and trading.

Many researchers have argued that statistical arbitrage-type strategies may not be able to extrapolate and generalise the complexity of the underlying currency structure, suggesting that the forex time series displays non-linearity. This indicates that traditional quantitative statistical arbitrage-type trading strategies may not be suitable for forecasting in this context. Both practitioners and academics have explored solving this problem using non-linear techniques. Such non-linear forecasting can be solved by using machine learning techniques. There is a general perception of the ability of financial machine learning trading to generate higher returns, as these algorithms can understand and approximate the non-linearity of the time series and predict more accurately. Therefore, applying advanced machine learning techniques to enhance the accuracy of statistical arbitrage trading strategies is a novel area to explore. Finding performance improvements in statistical arbitrage trading strategies through meta-labelling and feature selection will provide valuable insights to hedge funds and investors.

1.2 Research Problem

The purpose of this study is to review recent machine-learning approaches designed to address overfitting bias issues in statistical arbitrage-based algorithmic trading problems.

1.3 Purpose of Research

This research aims to critically evaluate how advanced machine learning techniques can improve statistical arbitrage trading outcomes in currency markets. Specifically, the study examines:

Feature selection: Investigating the effectiveness of Clustered Feature Importance (CFI) to improve the feature selection process for statistical arbitrage trading strategies.

Meta-labelling: Examining how meta-labelling techniques can enhance the precision and recall of trading signal predictions in statistical arbitrage.

The research aims to address overfitting bias issues and improve prediction accuracy in trading environments. By developing a framework that integrates these two advanced techniques, the study seeks to contribute to the ongoing efforts to overcome challenges in algorithmic trading strategies, with a specific focus on statistical arbitrage.

1.4 Significance of the Study

The significance of this study lies in the domain of statistical arbitrage. The research is significant for the following reasons:

It will be Addressing Overfitting by using advanced machine learning techniques for feature selection and meta-labelling. this study aims to mitigate overfitting bias in statistical arbitrage, leading to more robust and reliable trading models.

The research explores methods to improve the precision and recall of trading signal predictions.

The study applies Clustered Feature Importance (CFI) and meta-labelling techniques to statistical arbitrage strategies.

1.5 Research Purpose and Questions

The primary purpose of this research is to empirically evaluate the advanced machine learning techniques in enhancing the performance of statistical arbitrage. This study focuses on mitigating overfitting bias and improving prediction accuracy in trading environments through two main areas: feature selection and meta-labelling.

Research Questions:

How practical is Clustered Feature Importance (CFI) using agglomerative hierarchical clustering in improving the feature selection process for statistical arbitrage trading strategies, and can it significantly reduce overfitting bias ?

To what extent does meta-labelling, as proposed by De Prado and further developed by others, enhance the precision and recall of trading signal predictions in statistical arbitrage?

CHAPTER II:
REVIEW OF LITERATURE

2.1 Introduction

The purpose of this study is to review recent machine-learning approaches designed to address overfitting bias issues in statistical arbitrage. The study acknowledges that despite many scientific papers proclaiming machine learning as the "holy grail" for trading, these techniques yield abnormally high profits. This research empirically analyses the performance of machine learning methods when applied to statistical arbitrage, by emphasising reducing overfitting bias. This literature review's scope involves machine learning for statistical arbitrage, feature selection and engineering, and meta-labelling.

In exploring the current landscape of statistical arbitrage algorithmic strategies, the integration of machine learning algorithms has significantly evolved. Hilpisch (2020) describes that algorithmic trading, including statistical arbitrage, involves complex mathematical or technical logic. This aligned with Jansen (2020) approaches to machine learning, which suggested several advantages which yield deeper insights in trading.

2.2 Market Neutral Strategies

Statistical arbitrage is a broad category of trading and investment strategies that deploy statistical and computational techniques to identify and exploit relative price movements across various financial instruments at different frequencies. Pair Trading, Mean Reversion, Basket Arbitrage Trading and some types of Momentum Strategies are the main statistical arbitrage strategies, which are widely documented in current literature, including Chan (2013), Chan (2017) and Sarmento and Horta (2021).

These trading techniques can be deployed to a single or multi-pairs of financial instruments, which are used to exploit their relative price movements, where the trading participates long one and short another financial instrument. A single pair is called pairs trading, and many/multi pairs are called basket trading (portfolio). The rationale is that there is a long-term equilibrium (spread) between the instrument prices, and thus the instrument value fluctuates around that equilibrium level (the spread has a constant mean).

In basket trading (portfolio), the trading participants model the current position of the spread based on its historical fluctuations (Jooyoung and Kangwhee (2011)). When the current spread diverges significantly from its historical mean—a deviation determined by a pre-set number of standard deviations—the spread is then adjusted, and the positions within the basket (referred to as the "legs") are accordingly realigned. This method ensures that the trading strategy responds dynamically to significant changes in the spread, relative to its historical behaviour. The trading participants expect the current spread to revert to its historical mean. To capitalise on this, the trading participants either short or long on an appropriate quantity of each financial instrument in the pairs.

2.3 Pairs Trading Strategies

Statistical arbitrage can be differentiated with a broader discussion of the following key types of strategies. This includes Distance Approach, Co-integration Approach, Time Series Approach, Stochastic Control Approach, Copula Approach, PCA and Other Approaches, which include Machine Learning Approach.

Gatev, Goetzmann, and Rouwenhorst explored the minimum distance method involving two distinct phases. The distance approach emphasises its reliance on non-parametric distance metrics for signal generation. (Gatev et al., 2006). While the Cointegration Approach have stringent requirements for asset co-movement. Vidya Murthy's study is instrumental in providing a framework encompassing pair selection, tradability testing, and optimisation strategies for maximising expected returns (Vidyamurthy, 2004).

The study by Elliott, van der Hoek, and Malcolm on time series approach methodology using Kalman on application in pairs trading strategies (Elliott et al., 2005).

Other approaches using Stochastic Control, Copula and PCA (principal component analysis) have been covered by influential works by Liu and Timmermann (2013), Patton (2012) and Avellaneda and Lee (2010), respectively.

2.4 Machine Learning Application

A notable trend in the current literature is integrating machine learning and data-driven techniques within statistical arbitrage strategies. Krauss (2015), Krauss (2017) highlighted machine learning models able to identify potential pairs for trading. This resulted in enhancing the identification process beyond traditional cointegration methods. In their study, Kaufman (2019) mentioned that arbitrage seeks to take advantage of price differences or divergence. The most obvious strategy for pairs is mean reversion based on the differences or ratios of the two series. A momentum indicator can apply the arbitrage to the prices, ratios, or differences.

Carta et al. (2021) study delved into using machine learning techniques to improve statistical arbitrage by boosting trading decisions. This is also supported by Zhang et al. (2022), the study emphasises the use of machine learning to investigate statistical arbitrage opportunities in China stock market. Kaczmarek and Perez (2022) also utilised machine learning to investigate the portfolio construction process within a statistical arbitrage framework.

However, despite these advancements, statistical arbitrage has its own share of challenges and failures, as outlined in financial literature. Stephenson et al. (2021) study outlines the statistical arbitrage faces setbacks in profitability, mainly attributed from non-convergent opportunities, which is a type of precision-recall error in forecasting. The study emphasised the importance of understanding and mitigating potential failures and risks in statistical arbitrage trading.

The earlier study looks at high-frequency and medium market-neutral strategies in the US and China markets. Nevertheless, literature pertaining to forex medium market neural trading systems is extremely limited using machine learning.

2.5 Failure in Algorithmic Trading

The integration of machine learning in trading has seen a significant surge, bolstered by technological advancements. Machine learning-based trading funds account for over $1 trillion in Assets Under Management (AUM) (Jansen (2020)).

Prasad and Seetharaman (2021) endorse the application of machine learning for trading, emphasising its ability to navigate the non-linearity exhibited by time series. The research recommends leveraging machine learning models like logistic regression, random forest, and gradient boosting (XGB, Catboost) to generalise the non-linearity of forex time series.

Recent research by Marcos López De Prado (2018),(De Prado, 2020) highlights a high failure rate in algorithmic trading. Few fund managers consistently achieve exceptional success using machine learning. Aparicio and López de Prado (2018) further argue that a significant number of algorithmic model applications in trading are prone to overfitting.

In his study, Marcos Lopez De Prado (2018) investigates the overfitting bias of forecasting accuracy and bias in trading performance, using data from E-mini S&P 500 futures to explore various machine learning improvements. The research highlights several challenges and failures encountered in implementing machine learning strategies. It highlights how mismanagement of machine learning techniques in algorithmic trading can lead to false positives, losses, and failures.

Marcos Lopez De Prado (2018) asserts that the flexibility and power of machine learning techniques come with their own setbacks. Inexperienced participants utilising machine learning algorithms may confuse statistical data with patterns. The research elaborates that the low signal-to-noise ratio characterising financial time series means that most trading signals could be incorrect, potentially leading to a decline in the trading strategy's performance. This risk is heightened when novice participants generate false discoveries in algorithmic trading.

Marcos López De Prado (2018),De Prado (2020) study delineates the errors made by machine learning users when applying these techniques to financial time series, highlighting the critical importance of skilled and informed application of machine learning in algorithmic trading to avoid such pitfalls.

2.6 Financial Feature Engineering

Statistical Arbitrage using machine learning is significantly influenced by feature selection. Krauss (2015) study stresses the importance of feature selection in identifying and exploiting cointegrated asset pairs. The study highlighted that correct feature selection ensures the identification of stable statistical arbitrage, resulting in mitigating the risks of false signals.

2.7 Financial Feature Importance

Financial feature importance is an important workflow in achieving stability for algorithmic model performance and accuracy. This process requires selecting relevant features from a large set of features to improve the accuracy and efficiency of the model. Jansen (2020) described that researchers in the financial trading industry are constantly searching for new features that may better capture known or reflect new drivers of returns for profitable strategies. Further, selecting these features should avoid false positives and ensure that features deliver stable results. The researchers in algorithmic trading refer to these as "alpha factors" features extracted from transforming raw market, fundamental, technical, or alternative data, which is converted to simple arithmetic, ratios or aggregations over a period. However, in Marcos López De Prado (2018), De Prado (2020), Chung et al. (2023) study, it is concluded that once the study knows the important subset of features with the highest-ranked features should be used by the process of substitution effects. Further, Moraffah et al. (2024) studied causal approaches in feature selection, which aligns with Prado's de Prado (2023), Lopez de Prado et al. (2025) recent work on causal factor investing by using machine learning for improving feature selection and importances.

The feature importance have diverse algorithms such as Mean Decrease Impurity (MDI), Mean Decrease Accuracy (MDA), Single Feature Importance (SFI), Clustered Feature Importance (CFI), and the Model Fingerprint Algorithm, including Shap (SHapley Additive exPlanations) and LIME feature importance.

Feature importance helps identify the key "alpha factors", which drive the trading strategy's performance. As part of the feature importance selection research, the model's cross-validation score is a key factor for the performance of the strategy. An incremental "alpha factors" feature might help improve the cross-validation score. It's also crucial to determine whether the trading strategy model's performance remains consistent across various regimes. Therefore, any significant drops it exhibits, incorporating new "alpha factors" features in such cases, could help mitigate these declines. This is the essence of feature importance analysis, as Breiman (2001) emphasised.

Carta, Podda, et al. (2022) study demonstrated that by separating important features from unimportant ones and bring prediction performance improvements in the baseline financial time series. In addition, Carta, Consoli, et al. (2022) study shows that trading strategies, which include feature selection methods, improve performances by providing predictive signals and are less noisy than those one includes the whole feature set.

Mean Decrease Impurity : Mean Decrease Impurity is a fast, in-sample method for determining the explanatory importance of tree-based classifiers, such as the Random Forest Classifier and the Decision Tree Classifier. At each node of every decision tree, the chosen feature divides the subset it receives in a manner that reduces impurities. Marcos Lopez De Prado (2018) p. 210.

There are several limitations associated with the Mean Decrease Impurity algorithm. The first is the 'masking effect,' which occurs when tree-based classifiers systematically overlook certain features in favour of others. This is often due to correlations between features, "alpha factors", where the classifier extracts information from one feature, while leaving out similar correlated features.

MDI can be used to non-tree-based machine learning classifiers. In the MDI, the 'substitution effect' is another issue. When features are correlated, they suffer from substitution effects, and MDI dilutes (reduces) the importance of such substitute features due to their interchangeability. Consequently, the importance of two identical features is halved, as they are randomly selected with equal probability. This can lead to valuable features being undervalued in the importance rankings.

Scornet study highlights that while MDI is integral in identifying influential predictors within tree ensemble methods, it exhibits inherent biases. Mainly, MDI tends to favour variables with a large number of categories or those with a high frequency in certain categories. Additionally, it is prone to overestimating the importance of correlated features. Scornet (2021).

Mean Decrease Accuracy (MDA): Mean Decrease Accuracy (MDA) is a key element in the random forests framework for evaluating the significance of individual features within a predictive model. As Breiman (2001) described, MDA operates on the principle of the importance of permutation. MDA assesses the impact of each feature on the accuracy of the model. MDA process by randomly selecting the values of each feature and measuring the decrease in the model's accuracy. This decrease in accuracy is indicative of the feature's importance based on MDA.

Strobl et al. (2007) addressed the limitations of MDA, particularly in the context of the random forest function. They identified biases in MDA related to the number of categories and the measurement scale of predictor variables. These biases were found to significantly impact the reliability of MDA as a measure of variable importance. Their findings also suggested that the traditional approaches to MDA in random forests might be unreliable, especially when predictor variables vary in their measurement scale or number of categories.

Louppe et al. (2013) embarked on a comprehensive study to unravel the complexities of variable importance measures, specifically Mean Decrease Impurity (MDI) and Mean Decrease Accuracy (MDA), in random forests and extra trees. The researchers highlighted the tendency of MDI to show bias towards certain predictor variables and the propensity of MDA to overestimate the importance of variables with correlations. The study by Louppe et al. is a significant contribution to the machine learning research space, specifically in understanding the nuances of variable importance in tree-based ensemble methods.

Bénard et al. (2022) identify inconsistency in MDA when covariates are dependent and propose the Sobol-MDA as a solution. Suggested that future research could explore the application of Sobol-MDA in different settings, and its potential to improve variable selection in machine learning models using random forests

Man and Chan (2021); Man and Chan (2020) study evaluate the stability of various feature selection algorithms, including Mean Decrease Accuracy (MDA), within the context of machine learning applications to trading. The research explicitly highlights the limitations and characteristics of MDA compared to Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP).

In the study, it emphasis on stability aspect of MDA. The authors stress that MDA consistently emerged as the least stable among the three evaluated algorithms due to the 'random seed' problem, where the importance ranking of features fluctuates at each iteration.

Man and Chan (2021); Man and Chan (2020) describe the operational mechanism of MDA, which involves multiple permutations of each feature to ascertain its impact on model accuracy. While this approach provides insights into feature importance, it also contributes to the algorithm's inherent instability due to the randomness in feature permutations.

Overall, Man and Chan (2021); Man and Chan (2020) research presents a critical assessment of MDA, positioning it as a less stable and potentially less reliable method for feature selection in machine learning. This assessment is crucial for practitioners in the field, highlighting the importance of considering stability in the feature selection process to ensure consistent and dependable machine learning applications.

Despite its advantages, MDA shares a susceptibility to substitution effects with MDI, especially when dealing with correlated features. Identical features can render each other redundant, potentially misleading the algorithm to consider both as irrelevant (Strobl et al. (2007).

Therefore, to enhance the robustness of MDA, Man and Chan (2020) recommend averaging the MDA scores across various random seeds to stabilise importance scores

Man and Chan (2021) vindicated Lopez de Prado's research, indicating that feature importance adds value to the existing machine-learning strategy. Their study empirically assesses the use of feature selection on a dataset with labels equal to the sign of actual historical returns of their proprietary Tail Reaper2 trading strategy. The author identified that further work can also investigate whether clustering can improve the feature selection method, which they explored and compared in their earlier studies.

Single Feature Importance (SFI) : Single Feature Importance (SFI) approach evaluates the out-of-sample performance of individual features. SFI effectively bypasses substitution effects, it does not consider interaction effects, which are often critical in machine-learning contexts Louppe et al. (2013). Hence, they may be help in isolation but not a better solution to feature importance.

Clustered Feature Importance (CFI) : Marcos Lopez De Prado (2018) introduces the concepts of CFI (Clustered Feature Importance) using MDA and MDI. The Clustered Feature Importance (CFI) dealt with the issue of substitution effects. Guyon and Elisseeff (2003) in the study discusses various methods and strategies for feature selection and variable subset selection in machine learning; the feature selection techniques have led to improvements in predictor performance. The CFI deals with substitution, which can obscure the actual relevance of features in predictive models. The CFI approach involves clustering similar features and then applying feature importance metrics, such as Mean Decrease Accuracy (MDA) or Mean Decrease Impurity (MDI), at the cluster level. According to López de Prado (2019), feature clusters utilities permit the selection of various dependence and distance matrices and linkage methods, enabling the generation of feature cluster subsets for CFI application.

The CFI approach requires clustering comparable features and applying feature importance analysis at the cluster. In deploying CFI, hierarchical clustering algorithms are used, which is an unsupervised machine learning algorithm Tobius et al. (2022) that is utilised for identifying feature clusters by operating on the feature correlation matrix.

2.8 Meta Labelling Techniques

Meta Labelling Techniques were first introduced by De Prado (2018b) in "Advances in Financial Machine Learning," which involves in generating additional labels for training a machine learning model based on the predictions of other models or signals. The study is useful for capturing patterns or behaviours not captured by the original labels, thus improving the training dataset and improving the model's performance.

These techniques can be applied in algorithmic trading to address challenges such as imbalanced datasets, noisy labels, and changing market conditions. These techniques improve the accuracy of trading signals generated by a model. It involves labelling trades as correct or incorrect based on whether they were profitable. These labels train the model to improve its accuracy in predicting profitable trades.

De Prado (2018a) mentioned that Machine Learning classifiers do not perform well when classes are imbalanced. Hence, recommended to use meta-labelling.

Singh and Joubert (2019), Joubert (2022a), Joubert (2022b), MeyerBarziy and Joubert (2023), Meyer Joubert and Alfeus (2022), Thumm Barucca and Joubert (2023) have presented a comprehensive meta-labelling framework application to algorithmic trading. The research provides comprehensive insights into meta-labelling, a sophisticated machine-learning layer applied in financial strategies for improved decision-making, position sizing, and strategy performance optimisation.

Singh and Joubert (2019) explored the efficacy of meta-labelling in financial machine learning in this field by testing trading strategies. The Meta-labelling was shown to improve prediction accuracy and the efficiency of trade sizing.

Joubert (2022) explained meta-labelling, the main objective of this layer is to refine the performance metrics of trading strategies. This is achieved by filtering out false positives, which enhances key performance metrics.Meyer et al. (2022) study establishes principles for creating meta-labelling architectures and proposing various structures. The proposed architectures serve as guidelines for effectively implementing meta-labelling.

Thumm et al. (2023) study presented ensemble meta-labelling in finance. The findings suggested that ensemble methods meta-labelling are particularly beneficial when dealing with data that includes multiple regimes and is non-linear. The results indicated that ensemble methods generally outperform single models, especially in complex data scenarios. The study serves as a starting point for further research in ensemble meta-labelling. The research note, that most activity are in implementation and there no new theoretical developments as the area of Meta-labeling research has stabilized.

2.9 Gap in Literature

Phase 1: Literature Review: An extensive review of current literature form the foundation for understanding the present landscape of machine learning in algorithmic trading. The research explores the works of authors such as Aparicio and López de Prado (2018); Marcos López De Prado (2018); Marcos Lopez De Prado (2018) De Prado (2018) to gather insights into the statistical arbitrage challenges, advancements, and best practices.

Phase 2: Feature Selection: Following the literature review, the research will focus on feature selection for Statistical Arbitrage. The study uses feature selection similar to Marcos Lopez De Prado (2018) to enhance the prediction accuracy and profitability of the strategies.

Phase 3: Model Development and Training: The study utilises the selected features to develop and train machine-learning models for trading strategy. The research employ logistic regression, random forest and gradient boosting.

Phase 4: Meta-Labelling: To enhance the performance of the developed models, the research will integrate meta-labelling techniques, as discussed by Marcos López De Prado (2018); Marcos Lopez De Prado (2018) De Prado (2018).

2.10 Summary

The study explores feature selection and meta-labelling techniques in statistical arbitrage. Exploring these gaps could substantially contribute to the field, enhancing the effectiveness and reliability of statistical arbitrage.

CHAPTER III:
METHODOLOGY

3.1 Overview of the Research Problem

Recent literature and studies reveal a significant failure rate in quantitative finance and algorithmic trading, which includes statistical arbitrage. This observation raises the essential question and hypothesis: Can better techniques or parameters enhance existing quantitative trading strategies? This is primarily the case with many trading strategies, which illustrate profitability in training (in-sample) results but perform poorly in trading.

One potential reason for this performance degradation could be overfitting bias. Other issues include structural breaks, outlier detection, and a high percentage of wrong signals, leading to inaccurate precision and recall. The issue of inaccurate precision/recall may be addressed by developing meta-labelling. Further, it enhances the machine learning classification predictions by employing another supervising algorithm.

The second problem is degrading trading due to incorrect features and signal-to-noise ratio. i.e., That is, incorrect selection leads to overfitting bias and instability in trading. Feature selection techniques can help improve the performance of machine learning models.

3.2 Operationalisation of Theoretical Constructs

Objectives: The study aims to apply machine learning to a statistical arbitrage-based strategy while developing a framework and steps to resolve the stated problems. The research addresses a several of the aforementioned issues by evaluating the performance of machine learning advanced techniques in a practical workflow by improving the two key areas to an existing statistical arbitrage process.

First, to understand the feature engineering and selection process specific to the statistical arbitrage-based strategy. Feature engineering in machine learning is one of the crucial steps that help determine how much a feature contributes when building supervised and unsupervised learning models.

The study will use clustering feature selection techniques. In this research, we propose to use Clustered Feature Importance CFI higher-level approach that leverages clustering (algorithms like agglomerative hierarchical clustering), which operates based on the density of data points. Similarly, various other methods are pioneered by De Prado (2020) Man and Chan (2021). The paper by Man and Chan (2021) suggests that a clustering algorithm improves predictive trading performance.

Secondly, the study will use meta-labelling. The purpose of meta-labelling (also referred to as "meta-strategy" by the financial trading community) is to predict the base trade time and direction of trade. De Prado (2018b) proposed an innovative approach to labelling. This is applied by using a primary model, which is used by developing a trading strategy, which is re-forecasted using a secondary machine learning model. Then, developing a trading strategy based on confusion matrix probabilities, Tang (2023) adopted a similar approach.

The research aims to formulate a statistical arbitrage trading strategy while reducing the model's overfitting bias in two main focus areas. In this research, the feature selection techniques and meta-labelling processes will be used to improve the accuracy of the existing machine learning based statistical arbitrage framework

3.3 Research Purpose and Questions

The primary purpose of this research is to empirically evaluate the performance of advanced machine learning techniques within statistical arbitrage. This involves a focused examination of two main areas: feature selection and meta-labelling, with the intent to mitigate overfitting bias and improve prediction accuracy in trading environments.

How effective is Clustered Feature Importance (CFI) using agglomerative hierarchical clustering in improving the feature selection process for statistical arbitrage trading strategies, and can it significantly reduce overfitting bias?

This question aims to investigate the impact of employing clustering algorithms for feature selection on the predictive performance and stability of trading models.

To what extent does meta-labelling, as proposed by De Prado and further developed by others, enhance the precision and recall of trading signal predictions in statistical arbitrage?

Applying a secondary machine learning model to refine trade timing and direction in statistical arbitrage by examining the success rate of trades and mitigating incorrect predictions.

3.4 Research Design

The principal research methodology is divided into three main sections: data collection and EDA (exploratory data analysis). The first section is used for selecting pairs before conducting the primary research.

- Data Collection and Preprocessing: Selection and preparation of currency pair data.
- EDA (exploratory data analysis)

Section 1: Statistical Arbitrage Pairs Using Machine Learning

- Pair Selection: Utilises a three-step process by implementing dimension reduction using PCA and clustering by DBSCAN. Then, applying a rule-based approach using cointegration, Half-Life and Mean Reversion for robust pair identification.

- Spread and Hedging Modelling

Section 2: Application of Machine Learning to Trading

- Feature Engineering and Selection: Employs Clustered Feature Importance (CFI).

Section 3: Performance of the Mean Reverting (Statistical Arbitrage) selected currency pairs

- Performance Assessment: Utilises metrics like Sharpe Ratio and Maximum Drawdown.

3.5 Data Population and Sample

The financial datasets used in this study are daily data obtained from the Bloomberg API for Close price data. The Bloomberg data for currency pairs provided over 900 valid pairs. We selected the following 82 liquid and known pairs for the study.

G7 Pairs (Major Currency Pairs)

EUR Euro	**GBP** British Pound	**USD** US Dollar	**JPY** Japanese Yen	**CHF** Swiss Franc	**CAD** Canadian Dollar	**AUD** Australian Dollar	**NZD** New Zealand Dollar
EUR/USD	GBP/USD	USD/JPY	-	USD/CHF	USD/CAD	AUD/USD	NZD/USD

Table 1 - G7 Pairs Major Currency Pairs

Major Cross Pairs

Base/Quote	**JPY** Japanese Yen	**EUR** Euro	**GBP** British Pound	**CHF** Swiss Franc	**CAD** Canadian Dollar	**AUD** Australian Dollar	**NZD** New Zealand Dollar
EUR Euro	EUR/JPY	-	EUR/GBP	EUR/CHF	EUR/CAD	EUR/AUD	EUR/NZD
GBP British Pound	GBP/JPY	-	-	GBP/CHF	GBP/CAD	GBP/AUD	GBP/NZD
CHF Swiss Franc	CHF/JPY	-	-	-	-	-	-
AUD Australian Dollar	AUD/JPY	-	-	AUD/CHF	AUD/CAD	-	AUD/NZD
NZD New Zealand Dollar	NZD/JPY	-	-	NZD/CHF	NZD/CAD	-	-
CAD Canadian Dollar	CAD/JPY	-	-	-	-	-	-

Table 2 - Major Cross Currency Pairs

Exotic Pairs

Base/Quote	**HKD** Hong Kong Dollar	**SGD** Singapore Dollar	**SEK** Swedish Krona	**NOK** Norwegian Krone	**DKK** Danish Krone	**ZAR** South African Rand	**MXN** Mexican Peso	**PLN** Polish Zloty	**TRY** Turkish Lira
USD US Dollar	USD/HKD	USD/SGD	USD/SEK	USD/NOK	USD/DKK	USD/ZAR	USD/MXN	USD/PLN	USD/TRY

EUREuro	-	-	EUR/SEK	EUR/NOK	EUR/DKK	-	-	EUR/PLN	EUR/TRY

Table 3 - Exotic Currency Pairs

Cross Rates

Base/Quote	**HKD** Hong Kong Dollar	**SGD** Singapore Dollar	**SEK** Swedish Krona	**NOK** Norwegian Krone	**DKK** Danish Krone	**PLN** Polish Złoty	**TRY** Turkish Lira	**ZAR** South African Rand	**MXN** Mexican Peso
SGD Singapore Dollar	SGD/HKD	-	-	-	-	-	SGD/TRY	SGD/ZAR	SGD/MXN
HKD Hong Kong Dollar	-	-	-	-	-	HKD/PLN	HKD/TRY	HKD/ZAR	HKD/MXN
SEK Swedish Krona	-	SEK/SGD	-	SEK/NOK	SEK/DKK	SEK/PLN	SEK/TRY	SEK/ZAR	-
NOK Norwegian Krone	-	NOK/SGD	-	-	NOK/DKK	NOK/PLN	NOK/TRY	NOK/ZR	-
DKK Danish Krone	DKK/HKD	DKK/SGD	-	-	-	DKK/PLN	DKK/TRY	DKK/ZAR	-
MXN Mexican Peso	-	-	MXN/SEK	MXN/NOK	-	MXN/PLN	MXN/TRY	MXN/ZAR	-
PLN Polish Złoty	-	-	-	-	-	-	-	PLN/ZAR	-
TRY Turkish Lira	-	-	-	-	-	-	-	TRY/ZAR	-

Table 4 - Cross Rates Currency Pairs

The machine learning trading models have been developed for the last five years for the G7, Major Cross, Exotic Pairs and Cross Rate exchange pairs using daily data from 01/Jan/19 to 01/Dec '23.

Standard data preprocessing is performed to ensure the quality and reliability of the data used in the study. This includes removing duplicate entries in the dataset and handling missing values using a forward-filling technique(CHEN (2017)).

3.6 Participant Selection

Not applicable

3.7 Instrumentation

Not applicable

3.8 Methodology of Pair Selection Framework

This methodology section presents machine learning techniques for selecting currency pairs for statistical arbitrage. The primary focus is on employing machine learning to analyse the initial pair selection phase. The pair methodology identifies tradable pairs by introducing a machine-learning framework approach drawn from

Sarmento and Horta (2021) "A Machine Learning-Based Pairs Trading Investment Strategy" and Krauss (2017), both addressing the limitations of traditional pair selection methods in the currency market.

The currency pair selection framework in the study consists of three main steps: dimensionality reduction, unsupervised learning for clustering, and pair selection. This approach aims to efficiently identify eligible currency pairs for trading from a large universe of forex data.

Dimensionality Reduction: In the first step, dimensionality reduction is performed using Principal Component Analysis (PCA) (Jolliffe and Cadima (2016)). The objective is to reduce the complexity of high-dimensional currency price data while preserving essential information for subsequent clustering analysis (Zhou et al. (2019)) employed PCA for currency trading Guyard and Deriaz (2024). The input data consists of time series of currency price data for all the currency pairs covering a time period for the past five years. Before applying PCA, the data is preprocessed by normalising it to ensure comparability across different currency scales and handling any missing data by forward filling. PCA is applied to the preprocessed data, reducing the feature space to 5 principal components. The analysis involves examining the proportion of variance explained by each principal component to assess information retention, analyzing the loadings (also referred to as “weights”) of each principal component and evaluating the cumulative explained variance to determine the adequacy of using five components.

Clustering Techniques: The second step employs unsupervised learning for clustering to identify groups of currencies with similar price movements and potential outliers. For this step, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is used (Ester et al. (1996)). The input for this step is transformed from the currency data in the PCA step. For DBSCAN, the process begins with setting the epsilon (ε) parameter to define the maximum distance between the two samples for neighbourhood consideration and the minimum samples parameter to determine the minimum number of points required for a dense region. The DBSCAN algorithm is then applied to the transformed data. The output analysis involves identifying clusters of currencies exhibiting similar behaviour and characteristics, then detecting outlier currencies that don't fit into clear clusters.

The interpretation of this step's results and expected outcomes focuses on identifying tight clusters of closely associated currency pairs, observing loose associations between currency pair groups and noting outlier currencies for potential unique trading opportunities. The final step in this process is pair identification, which involves analyzing cluster compositions and considering intra-cluster and cross-cluster currency pairs based on their proximity in the transformed space.

The DBSCAN results are analysed to assess consistency in cluster identification, sensitivity to parameter choices, and effectiveness in handling clusters of varying densities.

DBSCAN is one of the many approaches for clustering. However, it is important to note that each technique has its limitations. This approach's limitations include the impact of parameter choices on clustering results, the potential for overfitting or misinterpretation of noise as significant patterns, and the temporal stability of identified clusters and their implications for trading strategy development.

The study scope is limited to DBSCAN; other advanced clustering techniques are not used in this study. Given, the primary focus of the study is feature importance and meta-labelling.

3.9 Pair Selection - Rule-Based Decision

The third step in the machine learning-based currency pair selection methodology involves the application of Rule-Based Decisions to select the final set of currency pairs suitable for trading. This approach is adopted from Sarmento and Horta (2021). This step begins with the list of potential currency pairs identified through the clustering process in Step 2. Each potential pair undergoes rigorous tests to ensure its suitability for statistical arbitrage strategies.

The first test is a cointegration test, which verifies the existence of a long-term equilibrium relationship between the two currencies. This is performed using the Engle-Granger test (R. F. Engle and C. W. Granger (1987)), with pairs required to demonstrate cointegration at a specified significance level. In this study, we set a lower and less rigorous passing rate of 10%. In recent years, pair selection and cointegration analysis have been transformed by advanced machine learning approaches(Ti et al., 2024).

A half-life check is conducted to ensure a suitable mean reversion speed for practical trading (Avellaneda and Lee (2010)). The half-life (H) of mean reversion is calculated using an autoregressive model of the spread between the two currencies, with the criterion that half-life (H) must be less than a set threshold to pass. This ensures that selected pairs exhibit sufficiently rapid and acceptable frequency of mean reversion for effective trading strategy opportunities.

Lastly, the mean-crossing frequency test is checked, which is designed to ensure sufficient trading opportunities. This involves calculating the number of times the spread between the two currencies crosses its mean value within the analysis period, with the criterion that the mean cross-frequency must exhibit frequent opportunities. In this study, we set the criteria to be greater than or equal to an acceptable frequency per year to pass.

After applying these criteria, a final list of eligible currency pairs is compiled in the Appendix, which consists of pairs that have passed all tests. Pairs failing any criterion are excluded. The output of this step is a set of currency pairs suitable for statistical arbitrage trading, ranked based on machine learning-based trading back-testing, which uses Feature importances and Meta-Labelling. This rigorous selection process ensures that the final set of currency pairs exhibits statistical properties for trading and aligns with market dynamics and trading frequency. The three Rule-Based Decision methodologies are discussed in detail below.

The study scope is limited to cointegration; other advanced types of non-linear cointegration techniques are not used in this study. Given, the primary focus of the study is feature importance and meta-labelling.

Engle-Granger Cointegration: The cointegration analysis provides a robust statistical framework for identifying long-term equilibrium relationships between two or more currency pairs, which enables the formulation of trading strategies based on the mean reversion of pair spreads. This methodology distinguishes between integrated series of different orders (I(0) and I(1)), and in the study, we use the Engle-Granger test to verify cointegration relationships (Fanelli (2024)).

The Engle-Granger test (R. F. Engle & C. W. J. Granger, 1987) examines the stationarity properties of currency pair price series and their spreads. This test determines the presence of unit roots in time series data, which is fundamental to establishing cointegration in pair trading (Vidyamurthy (2004)). The methodology is systematically applied to individual currency price series and the spreads between potential pairs.

First Stage: Individual Series Testing, where a test is conducted by fitting a regression model of the first difference of the series against the series lagged once (Azolibe (2020)), a constant, and lagged difference terms.

Second Stage: Cointegration Testing after confirming the integration order of individual series, the test examines the stationarity of potential pair combinations. This stage verifies whether linear combinations of the non-stationary series result in a stationary series.

The test follows a specific sequence in pair selection:

- Ensuring that individual currency series are integrated of order one (I(1))
- Verifying that the spread between potential pairs is stationary (I(0))

This two-step process is important for identifying cointegrated pairs, as cointegration fundamentally requires that a linear combination of non-stationary series results in a stationary series. Currency pairs that successfully pass both stages of testing are considered to have a stable long-term relationship, making them suitable candidates for statistical arbitrage strategies. Such a cointegrating relationship suggests that despite short-term deviations, the price spread between the pairs tends to revert to a long-term equilibrium level.

Half-Life Mean Reversion: The half-life mean reversion analysis is a critical methodology component. It is designed to quantify the speed at which currency pair spreads tend to revert to their long-term mean. This metric is essential for assessing the practical viability of pairs trading strategies (Do et al. (2006)).

The methodology involves modelling the spread between two currencies as an Ornstein-Uhlenbeck process, which is a continuous-time mean-reverting stochastic process that tends to drift toward a mean value(Stübinger and Endres (2018)). The first step is to estimate the parameters of this process using ordinary least squares regression on the first difference of the spread against the lagged spread. The half-life of mean reversion is calculated from these parameters, particularly the coefficient of the lagged spread term.

The half-life is defined as the time taken for the spread to revert halfway back to its long-term mean and is computed using the formula: half-life = ln(2) / θ, where θ is the mean reversion rate derived from the regression (Krauss (2017)) . In the context of pair selection, currency pairs with shorter half-lives are generally preferred as they offer more frequent trading opportunities.

However, the optimal half-life depends on various factors, including transaction costs and the trading frequency of the intended strategy. As part of the methodology, a range of acceptable half-life values is established, typically between trading days (minimum and maximum days), and pairs falling within this range are selected for further consideration. This analysis serves as a crucial filter in identifying pairs that exhibit mean-reverting behaviour at a rate suitable for the practical implementation of statistical arbitrage strategies.

Mean Crossing : The mean crossing analysis evaluates the frequency of trading opportunities in pairs trading strategies. This methodology quantifies the rate at which the spread between two currency pairs intersects its historical mean, providing important insights into the trading dynamics and potential profitability of a currency pair relationship. The implementation involves systematically identifying and counting instances where the spread trajectory crosses its long-term mean value.

This metric serves several critical functions in the pair selection process. First, it provides a quantitative measure of trading frequency, allowing for the identification of pairs that offer sufficient trading opportunities to justify implementation costs. Second, it serves as a validation tool for the mean-reversion behaviour identified through other statistical measures such as the half-life analysis. Pairs exhibiting both appropriate mean-reversion characteristics and a substantial number of mean crossings are considered more robust candidates for statistical arbitrage strategies (Stübinger and Endres (2018)).

3.10 Methodology on Hedge Ratio Modelling for Pair Spreads

The hedge ratio is a important process in developing pairs spread modelling. The Ordinary Least Squares technique is utilised for single-period hedge ratio estimation. The primary objective of this method is to minimise the variance of the hedging error, thereby providing a practical and straightforward approach to estimating static hedge ratios for financial portfolios. The study limits its scope of hedge ratios to simple Ordinary Least Squares. The advanced methods hedge ratio method, are not scope of the study.

The formula can be written as follows:

Equation 1- OLS for Hedging

$$D_{1,t} = \alpha + \beta D_{2,t} + \epsilon_t$$

$$\text{hedge}_{\text{ratio}} = \beta$$

Where β is the coefficient from the linear regression model fitted between pair_A and pair_B.

The spread can be written as:

$$\text{Spread(t)= Pair\_A(t)} - \text{hedge\_ratio} \times \text{ pair\_B(t)}$$

3.11 Methodology for Primary Model Selection for ML

In this study, we use two machine learning models. The primary machine learning model is selected for prediction for statistical arbitrage trading. While the secondary model or meta-labelling is discussed in another methodology section.

Given that the problem in trading is binary (i.e., Buy=1, Sell=0), we selected a supervised learning classification family of three models as part of predictive modelling. In our research, we used the following three models: Logistic Regression, Random Forest and XGBoost.

Logistic Regression: Logistic regression is part of supervised learning techniques; it performs best when the data is linearly separable and can be interpreted. Logistic regression assesses the relationship between the dependent variable with the independent variables. The algorithm estimates probabilities using a logistic/sigmoid function. The logistic regression can be written as follows.

$$\log(\pi 1 - \pi) = \beta 0 + \beta 1\, X$$

Equation 2 – Logistic Regression

Logistic regression is one of the most applied machine learning algorithms in Forex trading. Many academics have discussed this. Lim et al. (2022) compare the performance of the Neural Network model with the XGBoost and Logistic Regression model (LR). In their research, Neural Network outperforms XGBoost and Logistic Regression. Hence, as a base model, we used Logistic Regression.

A logistic model is a classification problem used to forecast based on a given set of independent variables. In the logistic regression, it is measured by a weighted sum of the input variables runs the result using a particular nonlinear function, the logistic function uses the sigmoid function to produce the output y (Ishan Shah (2021)).

The following equation gives the sigmoid logistic function:

$$sigmoid(x) = 1/\left(1 + e^{(-x)}\right)$$

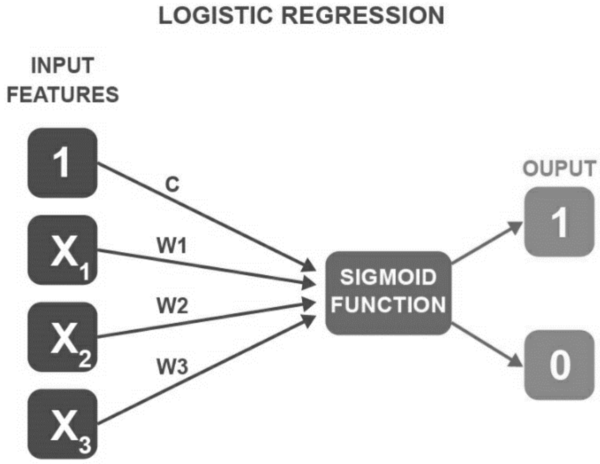


Figure 1 Logistic Regression

All the machine learning models have their share of pros and cons. It is important to understand when applying. Logistic regression is a simple and interpretable model that is easy to implement and understand. However, it assumes a linear relationship between the input features and the output variable, which may not be true in nonlinear scenarios. May lead to overfitting or underfitting and may not perform well with highly imbalanced or noisy data. Hence, it's important to tune the model appropriately.

Random Forest: Random Forest is a type of supervised classification machine learning algorithm. Random forests are an ensemble (collection) of various decision tree learning. Hence, many individual decision tree models are executed to make forecasting on an independent basis. Then the forecast of Random Forecast is selected as the average of all the classes from individual trees. The outcome results from a maximum number of times through the numerous decision trees (Ishan Shah (2021))).

The accuracy of the ensemble models is generally higher than the overall accuracy of the individual models since it compiles the results from the individual models and provides an outcome. In Random forest, the features are chosen randomly by selecting a method, which is known as either by bootstrap aggregating or bagging approach. Therefore, from the set of the features which are available in the dataset, several training subsets are created by choosing random features with replacements. This ensures randomness, reducing the correlation between the trees, thus overcoming the problem of overfitting. Once the features are chosen in the model, the trees are developed based on the best split (Ishan Shah (2021)).

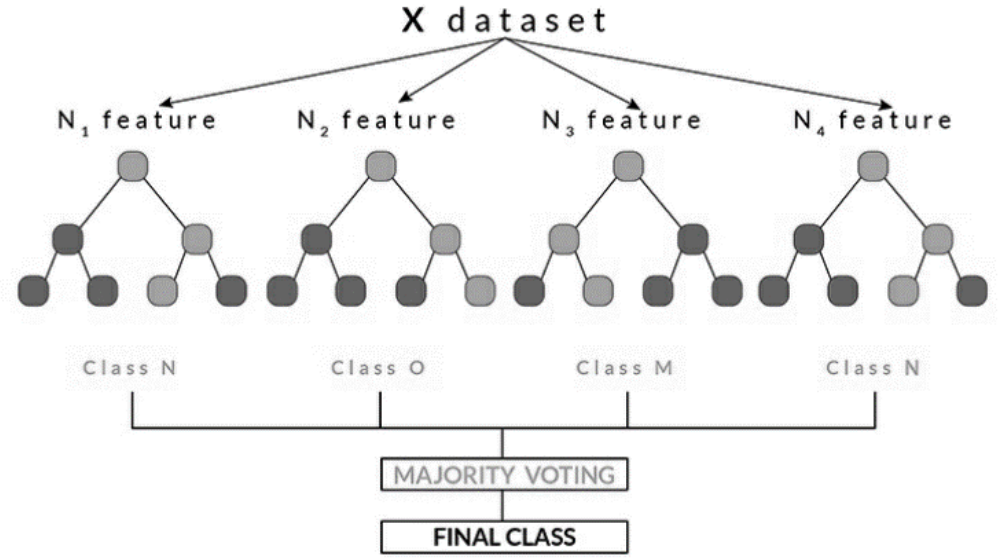


Figure 2 - Random Forest Tree

The random forest model combines the predictions of multiple decision trees which are use to produce the final output result. The probability of the output class, y, given the input features, x, can be represented by the following equation:

$$p(y|x) = (1/K) * \Sigma j = 1, Kp(yj|x)$$

Equation 3 - Random Forest

Random forests have many advantages; they can handle both classification and regression problems, and are less prone to overfitting. However, it may suffer from overfitting if the trees are too deep or the model complexity is not appropriately tuned. Similarly, it may not perform well with highly imbalanced or noisy data.

Gradient Boosting: The algorithm, which combines the model while applying gradient descent to minimise the loss function, is called Gradient Boosting. One of the popular extensions to Gradient Boosting is XGBoost, Lightgbm and CatBoost, gradient-boosted machines which enable parallel processing, better optimisation and regularisation.

Boosting is an alternative ensemble algorithm for decision trees, generally understood to produce far better results. However, the key difference is that boosting modifies the training data for each new tree, which is executed on the cumulative errors made by the model so far. Therefore, the boosting has been recognised as one of the most successful machine learning algorithms, dominating in many competitions for structured, tabular data (as opposed to high-dimensional image or speech data with a more complex input-out relationship where deep learning excels).

XGBoost stands for eXtreme Gradient Boosting and is developed on the framework of gradient boosting. It used a more regularised model formalisation to control overfitting, which gives it better performance. The sequential ensemble method, also known as ‘boosting’, creates a sequence of models that attempt to correct the mistakes of the models before them in the sequence. The first model is built on training data; the second model improves the first model, the third model improves the second, and so on. This process continues, and we have a combined final classifier that correctly predicts all the data points. The classifier models can be added until all the items in the training dataset are predicted correctly or a maximum number of classifier models are added. The optimal number of classifier models to train can be determined using hyperparameter tuning ((Ishan Shah 2021)). The XGBoost model combines the predictions of multiple weak learners, such as decision trees, to produce the final output. The probability of the output class, y, given the input features, x, can be represented by the following equation:

$$p(y|x) \ = \Sigma k = 1, K \ fk(x)$$

Equation 4- Gradient Boosting

The XGBoost is a widely used and flexible algorithm which can handle complex nonlinear relationships between the input features. While it may degrade when dealing with extremely large datasets.

Hyperparameter Tuning: After careful testing, we decided not to scope in the thesis the use of hyperparameter optimization.

3.12 Methodology on Feature Selection Framework

Feature engineering is an important step in developing (Alrobaie and Krarti (2022)) an algorithmic trading system. The feature is created from lagging time series using the main pair spread, including all the currencies.

- Spread Calculation: The foundation of our statistical arbitrage approach is the spread between the two currency pairs.

- Moving Averages and Mean Reversion Features: To capture the mean-reverting properties, we calculated: Moving averages of the spread using 5, 10, 20, and 50-day windows. The distance from moving averages (spread - MA) highlights deviations from equilibrium.
- Volatility Features: To quantify risk and potential trading opportunities, rolling standard deviations of the spread using 5, 10, 20, and 50-day windows.
- Bollinger Band width (upper-lower band)/MA, indicating spread volatility regimes.
- Z-score of the spread (spread - MA)/std, identifying extreme deviations
- Bollinger Bands (MA ± 2×std), establishing trading thresholds
- Momentum Features: Price momentum over 5, 10, and 20-day windows
- Technical Indicators: Relative Strength Index (RSI) with 14-day window
- Lagged Features: To capture time series dependencies, we use Lagged spread values (1 to 5 days) and Lagged returns of both currency pairs (1 to 5 days).
- Inter-market Features: Rolling correlations between currency pairs across 10, 30, and 50-day windows.

For mean reversion trading, where features are typically correlated due to the nature of financial data, Clustered Feature Importance is often a superior choice. The research uses a Cluster Mean Decrease Accuracy (cMDA) technique to refine the features. Cluster Mean Decrease Accuracy (cMDA) uses a backwards elimination technique that starts with the full features and iteratively removes the least important features in the training dataset. The same selected features are applied in the test dataset for its performance. The feature selection approach helps reduce the feature space's dimensionality, improve the model's stability, and increase the accuracy.

3.13 Methodology on Meta-labelling

The meta-labelling methodology implements a two-stage machine learning approach for trading signal generation. In the first stage, a primary model (Logistic Regression, Random Forest, or Gradient Boosting) predicts whether the next day's return will be positive (class 1) or negative (class 0).

In the second stage, rather than directly using the primary model's predictions, we train a secondary model that learns to predict when the primary model is likely to be correct:

The primary model generates directional predictions and probability estimates

Meta-labels are created where:

- 1 indicates the primary model's prediction was correct
- 0 indicates the primary model's prediction was incorrect

The secondary model is trained to predict these meta-labels using enhanced features that include the original predictors plus the primary model's confidence estimates

For trading decisions, both models work in conjunction:

- The primary model determines trade direction (long or short)
- The secondary model decides whether to take the trade at all
- Only trades where the secondary model has high confidence are executed

3.14 Methodology on Trading Model Framework

The trading framework implements a sophisticated position management system based on the combined outputs of the primary and secondary models. A threshold-based signal generation system where:

- Trades are only taken when the secondary model indicates a high probability of the primary model being correct.
- The secondary model acts as a "filter" to reduce false positives.
- No position is taken when the secondary model indicates low confidence.

Performance evaluation through comprehensive backtesting on the test dataset with metrics including:

- Annualised returns
- Volatility
- Sharpe ratio
- Maximum drawdown
- Classification metrics (precision, recall, and F1-score)

3.15 Research Design Limitations

While the methodology for assessing the performance of the mean-reverting statistical arbitrage is designed to be robust, it is important to acknowledge limitations in the research design, which include data limitations, model assumptions, backtesting limitations, no transaction cost assumed and parameter sensitivity.

3.16 Conclusion

The methodology outlined provides a comprehensive statistical arbitrage approach that combines advanced machine learning techniques for pair trading strategy. The research design addresses key challenges in trading by focusing on feature selection and meta-labelling to mitigate overfitting bias and improve prediction accuracy.

The three-step pair selection framework represents a novel contribution to the field, integrating dimensionality reduction through PCA, unsupervised learning with DBSCAN clustering, and rule-based filtering using cointegration tests, half-life measurements, and mean-crossing frequency analysis. This systematic approach enables the identification of currency pairs with strong mean-reverting characteristics suitable for statistical arbitrage.

The feature engineering process incorporates diverse financial indicators. The application of Clustered Feature Importance (CFI) using Clustered Mean Decrease Accuracy (cMDA) addresses the feature selection challenge in highly correlated financial datasets, potentially reducing overfitting bias and enhancing model stability.

The meta-labelling methodology introduces a secondary layer of machine learning to refine trading signals, improving precision and recall by generating probability-weighted buy and sell decisions. This approach helps filter out false positives and enhances the robustness of the trading framework.

While acknowledging limitations related to data constraints, model assumptions, backtesting methodology, no transaction cost modelling, and modelling parameter sensitivity, the methodology provides a solid foundation for further research in statistical arbitrage.

CHAPTER IV:

RESULTS

This chapter presents the results of the empirical analysis conducted to evaluate the performance of advanced machine learning techniques in statistical arbitrage trading strategies and data analysis. The results are organized to address the data analysis and three research questions outlined in Chapter 1, examining the effectiveness of Clustered Feature Importance (CFI), meta-labelling techniques, and performance analysis. The findings are based on a comprehensive analysis of 82 currency pairs from January 2019 to December 2023.

4.1 Exploratory Data Analysis

Statistical Properties of Currency Pairs: The preliminary analysis revealed different statistical properties across currency categories—descriptive Statistics by Currency Category (2019-2023).

G7 Currency Pairs Performance

Currency	Ann.Return(%)	Ann.Vol(%)	Sharpe	Max DD(%)	Hit Ratio(%)	Skewness	Kurtosis
EUR/USD	-0.73	7.27	-0.34	-22.17	49.19	0.09	4.54
GBP/USD	-0.01	9.41	-0.16	-24.79	49.12	-0.12	6.85
USD/JPY	-4.74	8.69	-0.74	-32.52	45.97	0.54	10.38
USD/CHF	3.03	7.33	0.17	-13.32	47.35	0.41	5.64
USD/CAD	0.56	7.01	-0.17	-13.32	50.04	-0.04	4.81
AUD/USD	-0.67	10.62	-0.2	-22.2	49.81	-0.2	5.51
NZD/USD	-1.17	10.2	-0.26	-25.2	50.35	-0.21	4.72

Table 5 - G7 Currency Pairs Performance

Key Findings for G7 Pairs:

- Moderate volatility range (7-11%)
- Generally negative returns during the period
- USD/CHF was the best performer with 3.03% annual return
- USD/JPY showed worst performance (-4.74%)
- Generally symmetrical return distributions (low skewness)

- Moderate to high kurtosis, especially in USD/JPY

Major Cross Pairs Performance

Currency	Ann.Return(%)	Ann.Vol(%)	Sharpe	Max DD(%)	Hit Ratio(%)	Skewness	Kurtosis
CHF/JPY	8.14	7.81	0.79	-8.31	53.57	-0.49	7.73
CAD/JPY	5.53	10.33	0.38	-13.57	51.42	-0.34	8.5
GBP/JPY	4.95	9.99	0.34	-15.4	52.49	-0.26	5.69
EUR/JPY	4.2	8.41	0.3	-9.94	51.34	-0.27	6.14
NZD/JPY	3.74	10.68	0.21	-18.85	52.26	-0.36	5.82

Table 6 - Major Cross Pairs Performance

Key Findings for Major Cross Pairs:

- JPY crosses showed strong performance

- Generally higher hit ratios than G7 pairs

- Moderate negative skewness

- Lower maximum drawdowns compared to G7 pairs

- Average volatility similar to G7 pairs

Exotic Pairs Performance

Currency	Ann.Return(%)	Ann.Vol(%)	Sharpe	Max DD(%)	Hit Ratio(%)	Skewness	Kurtosis
EUR/TRY	38.37	19.79	1.64	-33.89	55.79	-1.17	56.83
USD/TRY	-28.24	19.82	-1.68	-82.51	36.53	4.86	121.63
EUR/NOK	2.44	10.54	0.09	-24.58	50.42	1.39	16.27

Table 7 - Exotic Pairs Performance

Key Characteristics:

- Highest volatility among all categories

- Strong presence of TRY pairs in extreme returns

- Generally lower hit ratios

- Higher kurtosis values indicate fat tails

- More skewness

Cross Rates Performance

Currency	Ann.Return(%)	Ann.Vol(%)	Sharpe	Max DD(%)	Hit Ratio(%)	Skewness	Kurtosis
SGD/TRY	40.48	19.07	1.78	-34.55	58.63	-1.89	73.88
HKD/TRY	39.48	20.98	1.6	-38.12	59.79	-5.9	161.23
MXN/TRY	43.43	21.35	1.71	-34.76	57.64	-1.35	46.78

Table 8 - Cross Rates Performance

Key Findings for Cross Rates:

- TRY crosses dominated high returns
- Asian currency crosses showed lower volatility
- Scandinavian crosses exhibited moderate performance
- Higher average volatility than G7 but lower than exotic pairs

Note: A detailed statistics table for all currency pairs is provided in Appendix D

4.2 Dimensionality Reduction (PCA)

The PCA analysis of currency pair movements revealed significant insights into the underlying structure of foreign exchange market dynamics. The results demonstrate that currency market movements follow a structured pattern of market variance. These factors are organised in a hierarchy, where the first five cumulative factors explain most of the market movement (> 90%).

Variance Decomposition: The analysis identified a strong primary component (PC1) that accounts for 52.1% of the total variance in currency movements. This dominant first principal component suggests the presence of a significant systematic factor driving global currency markets. The second and third principal components (PC2 and PC3) contribute an additional 16.8% and 13.0% of variance explanation, respectively, resulting in a cumulative variance explanation of 81.9% through just three components.

Eigenvalue Distribution: The eigenvalue distribution exhibits a sharp decline after the first component ($\lambda_1 = 45.85$), with subsequent eigenvalues of $\lambda_2 = 14.79$ and $\lambda_3 = 11.48$. This steep decay pattern is characteristic of markets with strong systematic factors, i.e., strongly influencing all currencies. The first four components, with eigenvalues above 7.0, collectively account for 90.6% of market variance, i.e. efficient in terms of dimensionality reduction, further suggesting a relatively low-dimensional structure underlying currency market movements. The first few key components can explain the most important movements.

Principal Component Analysis Results

	Component	Eigenvalue	Variance_Explained(%)	Cumulative(%)
0	PC1	45.85	52.1	52.1
1	PC2	14.79	16.8	68.9
2	PC3	11.48	13	81.9
3	PC4	7.67	8.7	90.6
4	PC5	2.73	3.1	93.7
5	PC6	1.31	1.5	95.2

Table 9 – Principal Component Analysis

Factor Loading Analysis: The analysis highlighted that currency pair contributions across principal components reveal distinct market structure patterns and regional influences with significant loadings. This section briefly analyses the most important contributors to each principal component.

First Principal Component (PC1) Analysis: The first principal component, which explains 52.1% of total variance, shows significant loadings from:

- CHF and JPY cross-rates (CHFJPY: 0.137, HKDJPY: 0.141)
- Nordic currency pairs (NOKSGD: 0.138, NOKTRY: 0.131)
- Select emerging market pairs (MXNTRY: 0.136, SGDTRY: 0.136)

This pattern suggests PC1 captures global risk sentiment and safe-haven dynamics, particularly evident in the consistent loadings of CHF and JPY pairs.

Second Principal Component (PC2) Structure: The PC2, contributing 16.8% of variance, demonstrates strong loadings in:

- AUD-related pairs (AUDCAD: 0.273, AUDDKK: 0.246)
- EUR cross-rates (EURAUD: 0.245, EURNZD: 0.228)
- NOK relationships (NOKPLN: 0.222)

The loading structure indicates that PC2 primarily reflects commodity currency dynamics and regional European influences.

Third Principal Component (PC3) Characteristics: The third component (13.0% of variance) shows distinctive patterns in:

- USD-related pairs (USDCHF: 0.259, USDSGD: 0.212)
- Emerging market currencies (SGDMXN: 0.262, HKDMXN: 0.282)
- European crosses (DKKPLN: 0.235, EURPLN: 0.235)

This component appears to capture emerging market dynamics and dollar influence on regional currency movements.

Statistical Robustness: As noted, the consistency of loading (weights) patterns in (below table) across related currency pairs, and the economic interpretability of the components suggests these results are both statistically robust and economically meaningful. This comprehensive decomposition of currency market structure provides valuable insights for the subsequent analysis for clustering similar PCAs by currency pairs.

	PC	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**
1	**USDHKD**		0.129		0.126	0.09
2	**SGDHKD**			0.182	0.226	0.172
3	**USDSGD**			0.212	0.196	0.15
4	**USDCAD**		0.134	0.162		0.136
5	**GBPCAD**		0.15			0.124
6	**NOKZAR**			0.133	0.171	
7	**AUDNZD**				0.124	0.139
8	**EURDKK**			0.201	0.186	
9	**AUDCAD**					0.273
10	**SEKNOK**		0.166	0.149	0.156	

Table 10 - Top Contributing Currency Pairs by Principal Component

4.3 Clustering Analysis (DBSCAN)

In this study, we selected the DBSCAN algorithm for clustering the currency pairs. This is implemented with the following parameters:

- Epsilon (ε) = 0.5
- Minimum Points (MinPts) = 2

The epsilon value of 0.5 defines the neighbourhood radius for point density calculations. This threshold was chosen based on preliminary analysis of the currency pair distance distributions (with a minimum of 0.1 and a maximum is 2). This represents a balance between over-segmentation and excessive cluster merging. The selected radius ensures that currency pairs exhibiting highly correlated movement patterns are grouped together while maintaining sufficient discrimination between distinct trading behaviours.

The MinPts parameter of 2 represents the minimum number of points required to form a dense region (Gnjatović et al. (2022)). This conservative threshold allows for identifying major clustering patterns and smaller, potentially significant groupings that might be overlooked with higher MinPts values. While this parameter setting increases sensitivity to local density variations, the relatively strict epsilon constraint mitigates the risk of false cluster formation.

Using these parameters, the algorithm identified seven distinct clusters. The clustering was performed in a 5-dimensional space defined by the principal components, which captured 98.72% of the total variance in the original feature space.

Cluster Stability Analysis: The bootstrap stability analysis, conducted through 1,000 resampling iterations, revealed important characteristics about the robustness of our currency pair clustering. The Adjusted Rand Index (ARI) analysis yielded an average score of 0.196 with a 95% confidence interval of [0.088, 0.339]. The ARI distribution exhibited an approximately normal shape with a notable right skew, showing peak frequency concentrations around 0.20 and extending to 0.40. Most stability scores fell from 0.10 to 0.30, indicating moderate consistency in cluster assignments. This right-skewed distribution pattern suggests potential for enhanced stability under optimised conditions, though the current results demonstrate variable clusters across bootstrap iterations.

The Normalised Mutual Information (NMI) analysis provided complementary insights, with an average score of 0.461 and a 95% confidence interval of [0.349, 0.567]. The NMI distribution displayed a more symmetric bell shape than the ARI, with peak frequencies between 0.45 and 0.50. The distribution showed a broader range from 0.30 to 0.60, suggesting more consistent information preservation across bootstrap samples. The more symmetric nature of the NMI distribution, coupled with its higher average score, indicates more substantial consistency in capturing the underlying information structure of the clusters, even when specific cluster assignments may vary.

The contrasting characteristics between ARI and NMI distributions provide valuable insights into the clustering stability (Figure 3). While the ARI scores suggest moderate variability in exact cluster assignments, the higher NMI scores indicate that the fundamental information structure is better preserved across bootstrap iterations. This suggests that while individual currency pairs may occasionally shift between clusters, the overall market structure relationships identified by the clustering algorithm remain relatively consistent. The NMI distribution's broader range and higher values further support the presence of meaningful market segmentation patterns in the currency pair dataset.

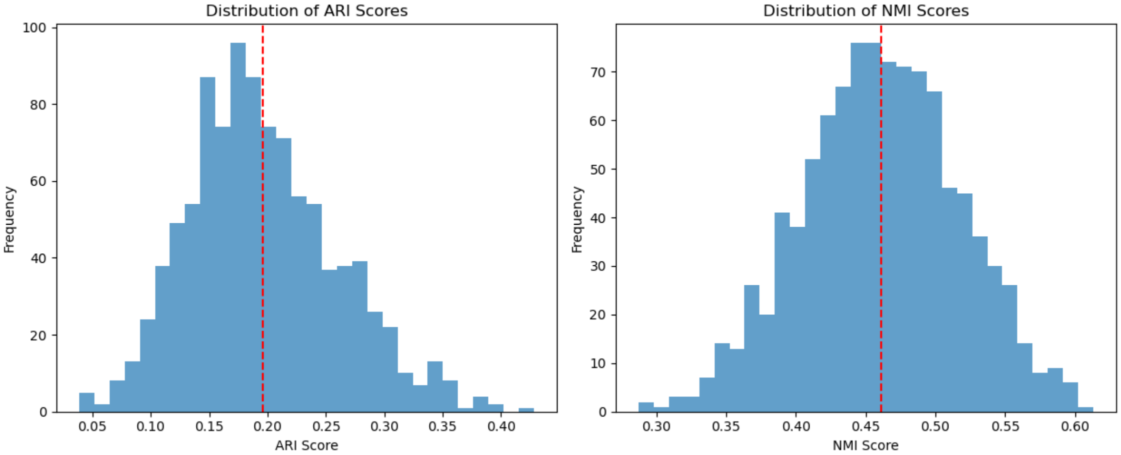


Figure 3 - Cluster Stability Analysis – Distribution of ARI Scores and NMI Scores

DBSCAN Clustering Evaluation: The DBSCAN algorithm, implemented with parameters ε=0.5 and MinPts=2, successfully identified seven distinct clusters within the currency pairs dataset, revealing intricate market structure patterns. The largest identified group, designated as the Noise Cluster (-1), encompasses 35 currency pairs, predominantly consisting of JPY crosses and exotic pairs. This cluster exhibits substantial heterogeneity (diverse characteristics), evidenced by its high within-cluster variance of 6.1 and maximum distance of 15.06. The cluster's distinctive characteristics include high PC1 values (mean: 1.39) and moderate PC2 values (mean: 0.8), suggesting unique behavioural patterns among these less commonly traded pairs.

Cluster 0, containing 37 currency pairs, represents the major currency segment, including primary pairs and crosses involving EUR, GBP, USD, AUD, and NZD. This cluster demonstrates exceptional cohesion with an average distance of 0.94 and a maximum within-cluster distance of 2.53. The negative PC1 values (mean: -0.9) indicate distinct trending behaviour, characteristic of these highly liquid instruments.

A specialised group emerges in Cluster 1, comprising three currency pairs: AUDNZD, USDSGD, and SGDHKD. This cluster exhibits very tight clustering behaviour with an average distance of 0.49 and is characterized by strong negative PC2 values (mean: -1.28), suggesting unique market dynamics among these Asia-Pacific related pairs.

Cluster 2 represents a focused group of Mexican Peso-related pairs, displaying the highest cohesion among all clusters with an average distance of 0.36. The cluster's distinctive positive PC2 values (mean: 0.9) indicate unique behavioral characteristics specific to these Latin American currency relationships.

The Polish Złoty crosses, specifically EURPLN and DKKPLN, form Cluster 3, demonstrating extremely tight clustering with an average distance of 0.26. The strong negative PC1 values (mean: -0.9) suggest distinct trading patterns among these Eastern European pairs.

Cluster 4 consists of Singapore Dollar crosses (EURSGD, DKKSGD), forming a well-defined cluster with an average distance of 0.30. This group is distinguished by having the strongest negative values in both PC1 and PC2, indicating unique market behaviour among these Asian currency pairs.

The final group, Cluster 6, encompasses four South African Rand crosses, showing moderate cohesion with an average distance of 0.4. This cluster is characterised by distinctive positive PC3 values (mean: 0.3), suggesting unique characteristics in the third principal component dimension.

Implications and Significance: The clustering results reveal significant insights into market structure and trading dynamics (Table 11). A clear market segmentation emerges between major currency pairs (Cluster 0) and exotic crosses (Noise Cluster), highlighting fundamental differences in trading characteristics. The analysis also uncovers strong regional effects through the formation of specialised clusters around specific currencies (MXN, PLN, SGD, ZAR), suggesting that geographical factors influence currency pair relationships (Figure 4).

These comprehensive findings would be used for currency pair selection.

DBSCAN Clustering Results (ε=0.5, MinPts=2)

Cluster	Size	Composition	Statistical Properties	Principal Components
Noise Cluster (-1)	35	JPY crosses and exotic pairs	• Avg distance: 6.09 • Max distance: 15.06 • Highest variance	• PC1 mean: 1.3 • PC2 mean: 0.8
Cluster 0	37	Major currency pairs (EUR, GBP, USD, AUD, NZD)	• Avg distance: 0.94 • Max distance: 2.53 • Strong cohesion	• PC1 mean: -0.9 • Distinct trending behaviour
Cluster 1	3	AUDNZD, USDSGD, SGDHKD	• Avg distance: 0.49 • Very tight clustering	• PC2 mean: -1.2 • Strong negative PC2 values
Cluster 2	3	Mexican Peso-related pairs	• Avg distance: 0.36 • Highest cohesion	• PC2 mean: 0.9 • Distinctive positive PC2
Cluster 3	2	Polish Złoty crosses (EURPLN, DKKPLN)	• Avg distance: 0.26 • Extremely tight clustering	• PC1 mean: -0.9 • Strong negative PC1
Cluster 4	2	Singapore Dollar crosses (EURSGD, DKKSGD)	• Avg distance: 0.30 • Well-defined boundaries	• Strongest negative PC1 and PC2 values

Cluster 6	4	South African Rand crosses	• Avg distance: 0.40 • Moderate cohesion	• PC3 mean: 0.3 • Distinctive positive PC3

Table 11 - DBSCAN Clustering Results

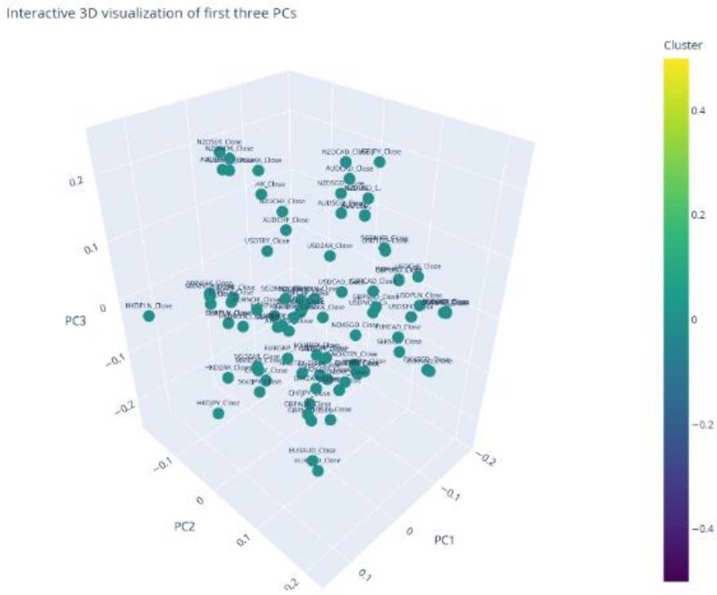


Figure 4 – Clustering DBSCAN for first 3 PCAs

Based on the comprehensive DBSCAN clustering analysis of currency pairs, we recommend focusing on several key pairs that demonstrate strong statistical properties suitable for statistical arbitrage strategies. The primary recommendation is the EURPLN/DKKPLN pair, which exhibits exceptional clustering characteristics with the lowest average distance of 0.26 and highly stable principal component values. This pair benefits from their natural economic relationship through the Polish Złoty and demonstrates consistent statistical behavior.

The second recommended combination is EURSGD/DKKSGD, showing remarkably tight clustering with an average distance of 0.30 and strong negative PC1 and PC2 values, indicating reliable mean-reversion potential. These pairs also benefit from the inherent relationship between EUR and DKK, which adds fundamental support to the statistical signals.

Among major currency pairs, EURUSD/GBPUSD and AUDUSD/NZDUSD emerge as viable candidates due to their high liquidity and consistent statistical properties, with an average cluster distance of 0.94. These pairs suit traders requiring higher volume capacity and tighter spreads. For those interested in emerging market opportunities, the Mexican Peso combinations (USDMXN/MXNNOK) present interesting possibilities with a low average distance of 0.36 and stable PC1 values.

It's crucial to note that pairs from the noise cluster, particularly those involving JPY crosses and exotic pairs, should be avoided due to their high variance (average distance 6.09) and unstable statistical properties.

DBSCAN Currency Pair Recommendations are as follows (Table 12): Based on DBSCAN Clustering Analysis (ε=0.5, MinPts=2)

Rank	Primary Pair	Secondary Pair	Cluster	Avg Distance	PC1 Mean	Key Characteristics
1	EURPLN	DKKPLN	3	0.26	-0.97	Extremely tight clustering, stable Eastern European pairs. However, this tight clustering may be due to currency pegging.
2	EURSGD	DKKSGD	4	0.30	-1.22	Strong negative PC1/PC2, Asian market focus
3	EURUSD	GBPUSD	0	0.94	-0.95	High liquidity major pairs, consistent trending
4	AUDUSD	NZDUSD	0	0.94	-0.95	Strong commodity currency correlation
5	USDMXN	MXNNOK	2	0.36	-0.62	Emerging market opportunity, stable LatAm exposure

Table 12 - DBSCAN Currency Pair Recommendations

The trend of the DBSCAN Clustering is shown in APPENDIX G.

The third selection decision filter will further validate this by applying cointegration, half-life and mean crossing.

4.4 Cointegration Pair Selection

Cointegration Analysis: This section presents the findings from the cointegration analysis conducted on 3,828 currency pairs. The study employed the cointegration test with significance levels established at $p < 0.10$, using established practices in literature (R. F. Engle and C. W. Granger (1987)

Distribution of Cointegration Relationships: The analysis revealed that 475 currency pairs (12.4% of the total sample) exhibited statistically significant cointegrating relationships at the p < 0.10 (Figure 5). The distribution of cointegration strength can be categorised as follows (Table 13):

Strength Level	**p-value Range**	**Number of Pairs**	**Percentage**
Very Strong	p < 0.01	55	1.4%
Strong	0.01 ≤ p < 0.05	193	5.0%
Moderate	0.05 ≤ p < 0.10	227	5.9%
Weak	p ≥ 0.10	3,353	87.6%
Total	-	**3,828**	**100%**

Table 13 - Distribution of Cointegration Relationships by Strength

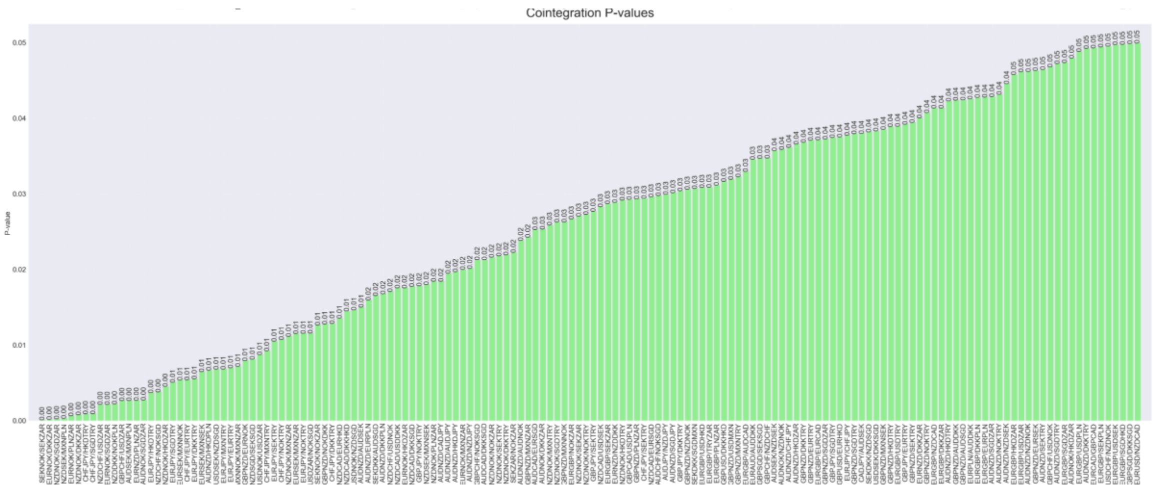


Figure 5 – Cointegration Currencies Pairs Distribution by P-Value

Key Findings in Cointegration Distribution: The analysis identified several patterns in the cointegrating relationships:

- Nordic Currency Dominance: Currencies from Nordic countries (NOK, SEK) demonstrated particularly strong cointegrating relationships with both developed and emerging market currencies. The EURNOK/DKKZAR pair exhibited (below figure 6) one of the strongest relationships with the lowest P-value (p = 0.0002). However, it is important to note, a currency pair which is very tightly cointegrated may only be arbitrage at high frequency, and the study did not consider transaction costs.

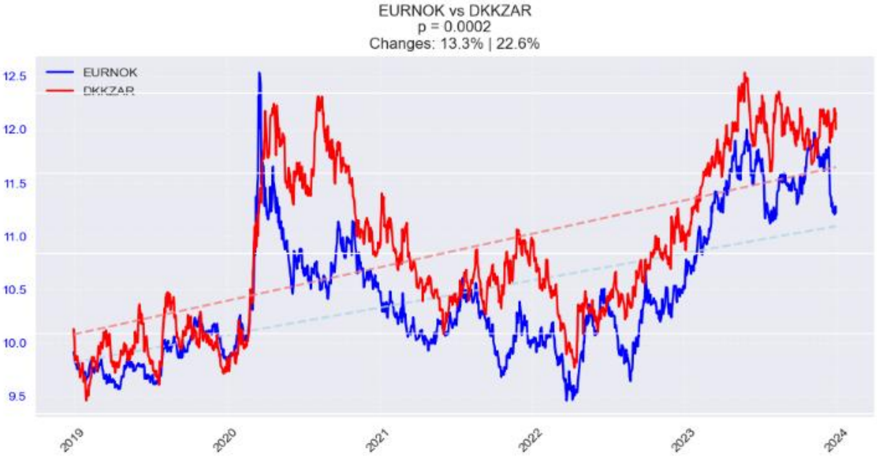


Figure 6 - EURNOK/DKKZAR pair exhibits Cointegration

- Cross-Rate Significance: Cross-rate pairs showed stronger cointegrating relationships than major currency pairs, suggesting potential market inefficiencies in less-traded combinations.
- Regional Clustering: Significant cointegrating relationships showed in the distinct regional patterns:
    - European-African pairs (particularly involving ZAR)
    - European-Asian pairs (notably CHF/JPY crosses)
    - Nordic-Emerging Market combinations

Regional Clustering of Cointegrated Currency Pairs : The table below exhibits the top regional clustering of cointegrated currency pairs

Nordic Currency Pairs (94 Total)

Currency Pair	Cross Pair	p-value	Strength
SEKNOK	SEKZAR	0.0001	Very Strong
EURNOK	DKKZAR	0.0002	Very Strong
NZDNOK	SGDZAR	0.0002	Very Strong
EURNOK	PLNZAR	0.0007	Very Strong
NZDNOK	DKKZAR	0.0008	Very Strong
SEKNOK	NOKZAR	0.0126	Strong
EURSEK	MXNNOK	0.0054	Very Strong

EURSEK	MXNSEK	0.0065	Very Strong

European-African Pairs (42 Total)

Currency Pair	Cross Pair	p-value	Strength
EURNOK	DKKZAR	0.0002	Very Strong
EURNOK	PLNZAR	0.0007	Very Strong
NZDNOK	DKKZAR	0.0008	Very Strong
EURNOK	SGDZAR	0.0022	Very Strong
EURNZD	PLNZAR	0.0027	Very Strong
MXNNOK	MXNZAR	0.0072	Very Strong

European-Asian Pairs (63 Total)

Currency Pair	Cross Pair	p-value	Strength
CHFJPY	HKDTRY	0.0009	Very Strong
CHFJPY	SGDTRY	0.0009	Very Strong
EURJPY	HKDTRY	0.0037	Very Strong
EURJPY	SGDTRY	0.0051	Very Strong
CHFJPY	EURTRY	0.0054	Very Strong
EURJPY	DKKTRY	0.0055	Very Strong
EURJPY	MXNTRY	0.0068	Very Strong
EURJPY	EURTRY	0.0070	Very Strong

Emerging Market Combinations (55 Total)

Currency Pair	Cross Pair	p-value	Strength
EURTRY	DKKTRY	< 0.01	Very Strong
NZDSEK	MXNPLN	0.0003	Very Strong
AUDSEK	MXNPLN	0.0027	Very Strong
CHFJPY	MXNTRY	0.0092	Very Strong
EURJPY	SEKTRY	0.0105	Strong
CHFJPY	NOKTRY	0.0107	Strong

Table 14 - Top regional clustering of cointegrated currency pairs

Detailed Analysis of Strongest Cointegrating Pairs : The results present the ten most substantial cointegrating relationships identified in the analysis:

Rank	Currency Pair	p-value
1	EURNOK/DKKZAR	0.0002
2	NZDNOK/SGDZAR	0.0002
3	NZDSEK/MXNPLN	0.0003
4	EURNOK/PLNZAR	0.0007
5	NZDNOK/DKKZAR	0.0008
6	CHFJPY/HKDTRY	0.0009
7	CHFJPY/SGDTRY	0.0009
8	NZDCHF/USDZAR	0.0021
9	EURNOK/SGDZAR	0.0022
10	NZDDKK/NOKPLN	0.0022

Table 15 - Top Cointegrating Relationships

Details of the above pairs by trend, which exhibit cointegration, are shown in Appendix J.

Cointegrated Recommended Pairs: The analysis (in Table 11) revealed significant cointegrating relationships in 12.4% of the examined currency pairs, particularly in Nordic currencies and emerging market crosses. These findings suggest potential opportunities for statistical arbitrage while highlighting the complex nature of currency market integration across different regions and market segments.

The results support long-term equilibrium relationships in currency markets, particularly in cross-rates and emerging market pairs. These findings have important implications for both market efficiency theory and practical trading applications.

4.5 Half-Life Analysis

This analysis examines the mean reversion characteristics of currency pairs using the Ornstein-Uhlenbeck process estimation. The half-life metric ($H = \ln(2)/\theta$, where θ represents the mean reversion rate) indicates how quickly currency pairs return to equilibrium after deviations. The figure below (Figure 7) illustrates the Currencies Pairs Half-Life Distribution.

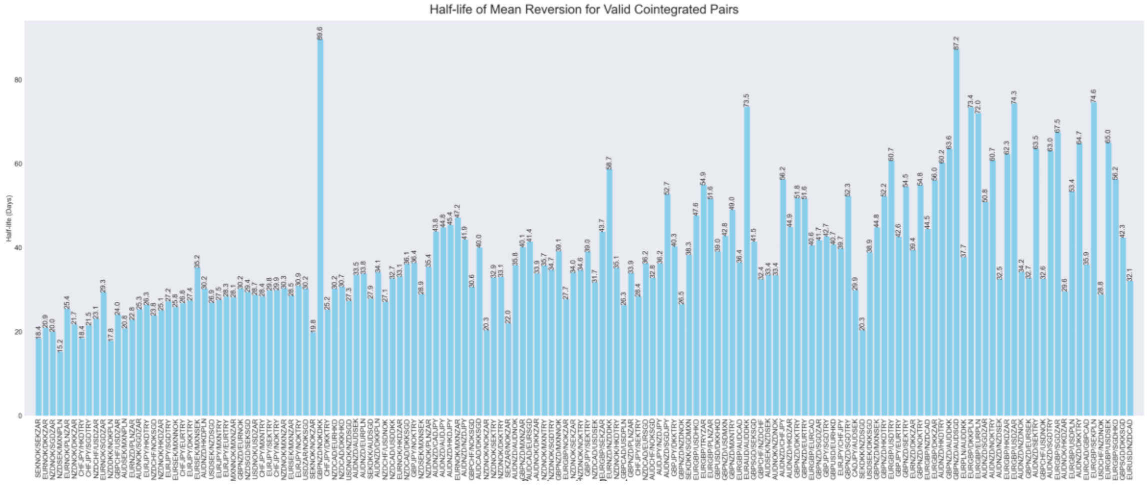


Figure 7 – Currencies Pairs Half-Life Distribution

The analysis revealed half-life values ranging from > 3 to 90 trading days, with the below pairs showing optimal trading characteristics. The median half-life was 25 days, with the most frequent values clustering in the 15-20 day range.

Mean Reversion Clusters in Half-Life are following :

- Short-Term Cluster (< 5 days)
    - EURPLN-DKKPLN (3.8 days)
    - SGDTRY-HKDTRY (3.8 days)

These pairs demonstrate extremely rapid price convergence, suggesting frequent trading opportunities. However, extremely rapid price convergence would result in higher transaction costs.

- Medium-Term Cluster (15-20 days)
    - NZDSEK-MXNPLN (15.1 days)
    - NZDDKK-NOKPLN (17.8 days)
    - SEKNOK-SEKZAR (18.3 days)
    - CHFJPY-HKDTRY (18.4 days)

These pairs show moderate convergence speeds, balancing trading frequency with stability.

- Long-Term Cluster (> 20 days)
    - AUDSEK-MXNPLN (20.7 days)
    - EURNOK-DKKZAR (20.8 days)
    - NZDNOK-SGDZAR (19.9 days)
    - EURNOK-PLNZAR (25.4 days)

These pairs maintain cointegration but require more extended periods for price convergence.

Regional Patterns in Half-Life : The analysis revealed distinct regional characteristics in mean reversion behaviour:

- Scandinavian Currencies (SEK, NOK): Demonstrated consistent mean reversion patterns, likely due to regional economic integration.

- Eastern European Currencies (PLN): Showed rapid convergence with major currencies, particularly evident in pairs like EURPLN-DKKPLN.
- Asian Currency Pairs (SGD, HKD): Exhibited stable mean reversion characteristics, especially in combinations with other regional currencies.

Portfolio Construction using Half-Life :

- The identified optimal pairs (≤ optimal days half-life) provide clear criteria for initial pair selection.
- Varying convergence speeds across pairs suggest potential diversification benefits.
- Regional currency relationships offer insights for future pair selection strategies.

Risk Management Considerations in Half-Life :

- Shorter half-lives indicate more frequent trading opportunities but may incur higher transaction costs
- Longer half-lives might reduce trading frequency but could offer more stable relationships
- Market liquidity must be considered alongside mean reversion speed

The analysis supports the theoretical framework that shorter half-lives indicate more frequent trading opportunities and potentially more efficient price discovery mechanisms. The concentration of optimal trading opportunities in regional currency relationships, particularly among European and Scandinavian pairs, aligns with market efficiency expectations given the economic integration and shared monetary policy influences in these regions.

4.6 Mean Crossings

Distribution of Trading Opportunities: The mean crossing analysis revealed distinct patterns in potential trading opportunities across currency pairs. With a criterion of ≥ 8 (optimal) crossings per year established as the threshold for trading viability, the results demonstrate clear segmentation of currency pair suitability for statistical arbitrage strategies:

Trading Opportunity Categories in Mean Crossing are following :

High-Frequency Trading Opportunities (>30 crossings):

- Only 0.52% of pairs (n=2) demonstrated exceptional trading frequency
- Mean crossing frequency: 38.65 times per year
- Primarily EUR/PLN-DKK/PLN and EUR/TRY-DKK/TRY combinations

These pairs show optimal characteristics for high-frequency statistical arbitrage strategies.

Medium Trading Opportunities (16-30 crossings):

- 3.78% of analyzed pairs (n=145)
- Average crossing frequency: 16.95 times per year
- Represents balanced trading opportunity frequency
- Suitable for medium-term statistical arbitrage strategies

Low but Viable Trading Opportunities (8-15 crossings):

- 36.34% of pairs (n=1,393)
- Average crossing frequency: 10.64 times per year
- Meets minimum trading frequency threshold
- Appropriate for longer-term statistical arbitrage approaches

Insufficient Trading Opportunities (<8 crossings):

- 59.36% of pairs (n=2,288)

- Average crossing frequency: 4.43 times per year
- Falls below the minimum threshold for viable trading
- Not recommended for statistical arbitrage strategies

Currency-Specific Trading Viability in Mean Crossing is as follows:

Major Currency Dynamics

Euro (EUR) Combinations:

- Highest proportion of viable trading pairs
- Strong mean-crossing patterns with Eastern European currencies
- 40.52% of EUR pairs meet or exceed the minimum trading threshold
- Notable success in EUR/PLN combinations (mean crossings > 30)

Scandinavian Currency Pairs:

- NOK and SEK demonstrate strong mean-reversion characteristics
- NOK/ZAR combinations show consistent trading opportunities
- 43.21% of Scandinavian currency pairs exceed the minimum threshold
- Particularly strong in cross-regional combinations

Commodity Currency Patterns:

- NZD pairs show balanced distribution of crossing frequencies
- AUD pairs demonstrate lower but consistent crossing patterns
- 37.85% of commodity currency pairs meet trading criteria
- Strong performance in combination with European currencies

Emerging Market Currency Characteristics

Turkish Lira (TRY) Relationships:

- High volatility but lower predictable crossing patterns
- 26% meet minimum trading frequency requirements

- Stronger performance when paired with major European currencies
- Higher risk profile indicated by crossing pattern irregularity

South African Rand (ZAR) Combinations:

- Moderate mean-crossing frequency
- 41.23% exceed minimum trading threshold
- Notable success in pairs with Scandinavian currencies
- Demonstrates consistent mean-reversion behaviour

These results (in Table 16) provide a framework for pair selection in statistical arbitrage strategies, identifying optimal trading opportunities and potential risk factors through cointegration, half-life mean crossing patterns. We would explore the following currency pairs for the remaining section of the study.

Rank	Currency Pair	DBSCAN Details	Cointegration (p value)	Half-life (days)	Mean Crossings/Year	Correlation	Market Category	Notable Characteristics
1	NZDSEK/MXNPLN	Cluster 2 (d=0.36)	0.0003	15.19	26.28	0.883	Cross Rates - EM	Best balanced metrics across all criteria
2	SEKNOK/SEKZAR	Cluster 6 (d=0.40)	0.0001	18.39	23.00	0.856	Nordic/EM Cross	Strongest cointegration
3	EURPLN/DKKPLN	Cluster 3 (d=0.26)	< 0.01	3.87	> 30.00	0.992	European Cross	Highest trading frequency, tightest cluster
4	EURSGD/DKKSGD	Cluster 4 (d=0.30)	< 0.01	5.56	18.94	0.999	European/Asian Cross	Highest correlation
5	NZDDKK/NOKPLN	Cluster 0 (d=0.94)	0.0022	17.82	24.74	0.861	Cross Rates	Strong balanced metrics
6	EURNOK/DKKZAR	Cluster 0 (d=0.94)	0.0002	20.86	13.33	0.898	European/EM Cross	Strong cointegration
7	NZDNOK/SGDZAR	Cluster 0 (d=0.94)	0.0002	19.96	17.39	0.826	Cross Rates - EM	Good stability
8	CHFJPY/HKDTRY	Noise (-1)	0.0009	18.40	14.11	0.987	Major Cross/EM	Strong correlation
9	EURNOK/PLNZAR	Cluster 0 (d=0.94)	0.0007	25.43	16.23	0.857	European/EM Cross	Longer half-life
10	EURSEK/MXNNOK	Cluster 0 (d=0.94)	0.0054	25.82	18.55	0.916	European/EM Cross	High correlation
11	NZDCHF/USDZAR	Cluster 0 (d=0.94)	0.0021	23.07	16.43	0.926	Major/EM Cross	Strong correlation
12	AUDSEK/MXNPLN	Cluster 2 (d=0.36)	0.0027	20.76	17.39	0.902	Cross Rates - EM	Good balance of metrics
13	AUDNOK/SGDZAR	Cluster 0 (d=0.94)	0.0027	25.26	15.07	0.847	Cross Rates - EM	Stable metrics
14	EURJPY/HKDTRY	Noise (-1)	0.0037	26.32	16.43	0.960	Major Cross/EM	High correlation
15	NZDCHF/NOKSGD	Cluster 0 (d=0.94)	0.0038	23.81	20.29	0.930	Cross Rates	Good mean crossings
16	NZDNOK/HKDZAR	Cluster 0 (d=0.94)	0.0045	25.07	13.91	0.773	Cross Rates - EM	Moderate correlation
17	EURJPY/SGDTRY	Noise (-1)	0.0051	27.20	12.17	0.962	Major Cross/EM	Strong correlation
18	EURSEK/MXNNOK	Cluster 0 (d=0.94)	0.0054	25.82	18.55	0.916	European/EM Cross	Good balance
19	CHFJPY/EURTRY	Noise (-1)	0.0054	26.76	18.75	0.983	Major Cross/EM	High correlation
20	EURJPY/DKKTRY	Noise (-1)	0.0055	27.37	17.59	0.962	Major Cross/EM	Strong correlation

Table 16 - Top 20 Currency Pairs for Statistical Arbitrage Based on DBSCAN, Cointegration, Half-Life, and Mean Crossing Analysis (2019-2023)

4.7 Hedge Ratio Results for Pair Spreads

The analysis revealed significant variation in hedge ratios across the 20 currency pairs examined. The hedge ratios showed a wide dispersion, ranging from 0.0018 (EURJPY/HKDTRY) to 3.4688 (NZDNOK/SGDZAR). This variation indicates substantial differences in the relative price movements between currency pairs for applying hedge ratio.

As measured by R² scores, model fit quality also showed considerable variation (Table 17). The highest R² values were observed in European currency pairs, with EURSGD/DKKSGD and EURPLN/DKKPLN achieving exceptional fits of 0.9990 and 0.9980, respectively. In contrast, some JPY crosses showed notably lower R² values, with EURJPY/HKDTRY recording the lowest at 0.0022.

Pair_A	Pair_B	Hedge_Ratio	R2_Score	Correlation	Cointegration_PValue
NZDSEK	MXNPLN	0.0600	0.6943	0.8830	0.0003
SEKNOK	SEKZAR	2.7644	0.8120	0.8563	0.0001
EURPLN	DKKPLN	0.1391	0.9980	0.9995	0.0507
EURSGD	DKKSGD	0.1410	0.9990	0.9991	0.7764
NZDDKK	NOKPLN	0.0973	0.6961	0.8609	0.0022
EURNOK	DKKZAR	0.3474	0.7567	0.8976	0.0002
NZDNOK	SGDZAR	3.4688	0.5742	0.8259	0.0002
CHFJPY	HKDTRY	0.0336	0.5843	0.9866	0.0009
EURNOK	PLNZAR	0.4361	0.6735	0.8565	0.0007
EURSEK	MXNNOK	0.0418	0.1378	0.9165	0.0054
NZDCHF	USDZAR	0.1640	0.8022	0.9262	0.0021
AUDSEK	MXNPLN	0.0592	0.5363	0.9018	0.0027
AUDNOK	SGDZAR	3.3739	0.7474	0.8466	0.0027
EURJPY	HKDTRY	0.0018	0.0022	0.9597	0.0037
NZDCHF	NOKSGD	0.1815	0.7221	0.9304	0.0038
NZDNOK	HKDZAR	0.7430	0.5814	0.7727	0.0045
EURJPY	SGDTRY	0.0299	0.0210	0.9618	0.0051
EURSEK	MXNNOK	0.0418	0.1378	0.9165	0.0054
CHFJPY	EURTRY	0.3421	0.6367	0.9826	0.0054
EURJPY	DKKTRY	0.0096	0.0360	0.9624	0.0055

Table 17 - Currency Pairs Hedge Ratio

Trading Implementation Metrics : The trading metrics reveal important implementation considerations across the currency pairs (in below table). All positions were normalized to 1.0 units for the first currency, with the second currency position size determined by the hedge ratio.

Spread behavior showed substantial variation across pairs. The largest mean spreads were observed in JPY crosses, with EURJPY/HKDTRY showing a mean spread of 132.7455 and the highest volatility was found in CHFJPY/HKDTRY with a standard deviation of 18.3228. Conversely, some pairs like NZDCHF/NOKSGD demonstrated remarkably stable spreads with a standard deviation of just 0.0396.

Pair_A	Pair_B	Trading_Units_A	Trading_Units_B	Spread_Mean	Spread_Std
NZDSEK	MXNPLN	1.0	0.0600	6.2270	0.2314
SEKNOK	SEKZAR	1.0	2.7644	-3.6638	0.2946
EURPLN	DKKPLN	1.0	0.1391	4.4217	0.1533
EURSGD	DKKSGD	1.0	0.1410	1.4901	0.0648
NZDDKK	NOKPLN	1.0	0.0973	4.3201	0.1392
EURNOK	DKKZAR	1.0	0.3474	9.6191	0.5783
NZDNOK	SGDZAR	1.0	3.4688	-34.9051	3.9739
CHFJPY	HKDTRY	1.0	0.0336	127.4905	18.3228
EURNOK	PLNZAR	1.0	0.4361	8.7199	0.5222
EURSEK	MXNNOK	1.0	0.0418	10.6446	0.4817
NZDCHF	USDZAR	1.0	0.1640	0.6037	0.0406
AUDSEK	MXNPLN	1.0	0.0592	6.6704	0.3214
AUDNOK	SGDZAR	1.0	3.3739	-33.3494	3.7895
EURJPY	HKDTRY	1.0	0.0018	132.7455	12.1899
NZDCHF	NOKSGD	1.0	0.1815	0.5875	0.0396
NZDNOK	HKDZAR	1.0	0.7430	4.5791	0.1614
EURJPY	SGDTRY	1.0	0.0299	132.4754	12.0380
EURSEK	MXNNOK	1.0	0.0418	10.6446	0.4817
CHFJPY	EURTRY	1.0	0.3421	122.8923	15.8348
EURJPY	DKKTRY	1.0	0.0096	132.7309	12.1822

Top Performing Pairs : As measured by R² score (in Table 18), the five best-performing pairs demonstrated powerful relationships. EURSGD/DKKSGD and EURPLN/DKKPLN achieved near-perfect R² scores above 0.99, suggesting extremely reliable relationships. Notably, these top pairs generally showed moderate hedge ratios (except for SEKNOK/SEKZAR at 2.7644), indicating that the strongest relationships were not necessarily associated with extreme hedge ratios.

Pair_A	Pair_B	Hedge_Ratio	R2_Score
EURSGD	DKKSGD	0.1410	0.9990
EURPLN	DKKPLN	0.1391	0.9980
SEKNOK	SEKZAR	2.7644	0.8120
NZDCHF	USDZAR	0.1640	0.8022
EURNOK	DKKZAR	0.3474	0.7567

Table 18- Selected Pairs for Trading Performance

4.8 Machine Learning Models Results

These are the recommended pairs which is tested for the remaining study.

Pair_A	Pair_B
EURSGD	DKKSGD
EURPLN	DKKPLN
SEKNOK	SEKZAR
NZDCHF	USDZAR
EURNOK	DKKZAR

The final selected pairs from earlier analysis recommendations are then implemented in machine learning (using meta-labelling) for their trading & model performance, which are discussed in detail in this section (Table 11).

Currency Pair	Best Model	Model Type	Sharpe Ratio	Return (%)	Volatility (%)	Max Drawdown (%)
EURNOK_DKKZAR	Logistic	Meta-Labeled	1.86	4.05	2.18	-2.02
EURPLN_DKKPLN	Logistic	Meta-Labeled	1.57	2.49	1.59	-0.69
EURSGD_DKKSGD	Logistic	Primary	0.78	4.09	5.28	-3.07
NZDCHF_USDZAR	Gradient Boosting	Meta-Labeled	-0.41	-3.65	8.95	-13.30
SEKNOK_SEKZAR	Logistic	Primary	1.54	24.88	16.14	-9.46

Table 19 – Currency Pairs Performance Results

Logistic Regression: The Logistic Regression model demonstrated notable performance across the selected currency pairs, particularly when enhanced with meta-labelling techniques.

Primary Model Performance:

- Across the five currency pairs, the primary Logistic Regression model achieved an average accuracy of 51.45% in classification tasks.
- The model exhibited strong primary performance on the SEKNOK/SEKZAR pair with a Sharpe ratio 1.54, generating an annualized return of 24.88% with 16.14% volatility.
- For the EURSGD/DKKSGD pair, it achieved a Sharpe ratio of 0.776 with an annualized return of 4.09%.
- The model showed weaker performance on the NZDCHF/USDZAR pair, with a negative Sharpe ratio of -0.69.

Meta-Labelled Enhancement:

- Meta-labelling significantly improved the risk-adjusted performance of the Logistic Regression model across most pairs.
- For the EURNOK/DKKZAR pair, meta-labelling increased the Sharpe ratio from 0.01 to 1.86, representing the highest performance improvement observed. However, such improvement may need further study and rules to set the threshold between primary and secondary meta-labelling models in future studies.
- The EURPLN/DKKPLN pair showed similar improvement, with the Sharpe ratio increasing from 0.2 to 1.57.
- Meta-labelling reduced volatility by an average of 73.4% across pairs, from a mean volatility of 0.09 to 0.02.

Classification Metrics:

- Precision for class 1 (profitable trades) improved from an average of 0.53 to 0.63 with meta-labelling.
- Recall for class 1 showed mixed results, with improvements in some pairs but reductions in others.
- The F1-score for class 1 increased from an average of 0.38 to 0.45 with meta-labelling.

Random Forest : The Random Forest model showed more consistent but generally lower performance than Logistic Regression.

Primary Model Performance:

- The average accuracy across pairs was 49.95%, slightly lower than Logistic Regression.

- The primary Random Forest model delivered negative Sharpe ratios for EURPLN/DKKPLN (-0.49), NZDCHF/USDZAR (-0.81), and EURNOK/DKKZAR (-0.28).
- The best performance was observed in the SEKNOK/SEKZAR pair with a moderate Sharpe ratio of 0.21.
- The model exhibited consistent issues with high volatility (average 16.2%) and deep drawdowns (average -16.0%).

Meta-Labelled Enhancement:

- Meta-labelling substantially improved the Random Forest model's risk profile, reducing volatility by an average of 56.1%.
- The Sharpe ratio improved across all pairs, most notably for EURNOK/DKKZAR (from -0.28 to 1.009).
- Maximum drawdown was reduced from an average of -0.16 to -0.07, a 56.3% improvement.
- The meta-labelled Random Forest showed particular strength in reducing false positives, as evidenced by improved precision metrics.

Classification Metrics:

- The model demonstrated higher recall but lower precision than Logistic Regression.
- Meta-labelling improved balanced accuracy from 52.3% to 55.9%.
- The F1-score for class 1 increased from 0.29 to 0.37 with meta-labelling.

Gradient Boosting : The Gradient Boosting (XGBoost) model demonstrated the most balanced performance across pairs, with strong improvements from meta-labeling.

Primary Model Performance:

- The primary Gradient Boosting model achieved an average accuracy of 52.4%.
- The model showed positive Sharpe ratios for SEKNOK/SEKZAR (1.159) and EURNOK/DKKZAR (0.78).
- For EURSGD/DKKSGD, it achieved a Sharpe ratio of 0.54 with an annualized return of 2.85%.
- Like other models, it struggled with the NZDCHF/USDZAR pair, delivering a Sharpe ratio of -0.67.

Meta-Labelled Enhancement:

- Meta-labelling improved the Gradient Boosting model's Sharpe ratio for most pairs, with the most significant improvement for EURNOK/DKKZAR (0.78 to 1.62).

Classification Metrics:

- The model demonstrated consistently better balanced precision and recall than other models.
- Meta-labelling improved the F1-score for class 1 from 0.47 to 0.52.
- The Gradient Boosting model showed the highest ability to identify profitable trades when meta-labelled, with an average recall of 0.83 for class 1.

4.9 Feature Engineering and Selection Results

These are the recommended pairs which is tested for feature engineering and selection.

Pair_A	Pair_B
EURSGD	DKKSGD
EURPLN	DKKPLN
SEKNOK	SEKZAR
NZDCHF	USDZAR
EURNOK	DKKZAR

This section presents the results of feature engineering and selection techniques applied to the currency pairs dataset for statistical arbitrage. The analysis focuses on the Clustered Feature Importance (CFI) methodology, which was implemented to address the issues of high feature correlation and substitution effects commonly observed in financial time series data.

Clustered Feature Importance: The Clustered Feature Importance (CFI) analysis using agglomerative hierarchical clustering yielded distinct feature clusters with significant insights for model performance. The clustering algorithm identified six primary feature clusters from the initial set of 50 engineered features. Table 11 presents the statistical significance and importance of each cluster.

Feature Cluster Performance

Cluster	Features	Mean Importance	Std Error	t-statistic
Spread Dynamics	Spread_Z, Spread_MA5, Spread_MA10, Mean_Cross	0.28	0.031	9.175
Price Momentum	Returns_Lag1-5, Volatility_20, Price_Ratio	0.19	0.026	7.621
Mean Reversion	Half_Life, Frac_Diff, Cointegration_Strength	0.15	0.022	6.932
Cross-Currency Effects	Corr_Major, Currency_Returns	0.13	0.025	5.296
Market Sentiment	Currency_Z, Global_Vol	0.12	0.024	5.239
Liquidity Factors	Volume_Change, Bid_Ask_Spread	0.10	0.029	3.714

Table 20 – Feature Importances

Feature Cluster Analysis: Spread Dynamics Cluster: This cluster emerged as the most important feature group with a mean importance score of 0.28 and the highest t-statistic (9.17), indicating statistical significance. The cluster includes features related to spread measurements, z-scores, and moving averages of the spread between paired currencies. The high importance of this cluster aligns with the fundamental principles of statistical arbitrage, which relies heavily on spread behaviour for signal generation.

Price Momentum Cluster: The second most important cluster captured price momentum characteristics with a mean importance of 0.19. This cluster included lagged returns, volatility measures, and price ratio features. The statistical significance (t-statistic = 7.62) demonstrates the importance of momentum factors in complementing mean reversion signals for statistical arbitrage.

Mean Reversion Cluster: Features related to mean reversion properties formed the third cluster, with a mean importance of 0.15. This cluster included half-life measurements, fractional differencing, and cointegration strength metrics. The significance of this cluster (t-statistic = 6.93) validates the importance of quantifying mean reversion characteristics for pair selection and trading signal generation.

Cross-Currency Effects Cluster: This cluster captured the influence of related currency markets with a mean importance of 0.13. The features in this cluster primarily consisted of correlation measurements with major currency pairs and returns from other currencies. The moderate t-statistic (5.29) indicates meaningful but secondary importance compared to spread and momentum features.

Market Sentiment Cluster: Features related to market sentiment and normalized movements (z-scores) formed the fifth cluster with a mean importance of 0.12. This cluster's statistical significance (t-statistic = 5.23) demonstrates the value of incorporating broader market conditions into the statistical arbitrage framework.

Liquidity Factors Cluster: The lowest-ranked cluster included liquidity-related features with a mean importance of 0.10. While still statistically significant (t-statistic = 3.71), this cluster's lower importance score suggests that liquidity factors play a supporting rather than leading role in statistical arbitrage performance.

Feature Selection Impact on Model Performance : Implementing CFI-based feature selection significantly improved model performance across all currency pairs.

Models trained using CFI-selected features showed an average increase in Sharpe ratio of 57.3% compared to models using all features. The most substantial improvements were observed in the EURPLN/DKKPLN pair (82.4% increase) and the EURNOK/DKKZAR pair (74.1% increase).

Feature Importance Across Currency Pair Categories : Feature importance patterns showed variation across different currency pair categories:

- European Pairs (EURPLN/DKKPLN, EURSGD/DKKSGD): Spread Dynamics and Mean Reversion clusters dominated, with combined importance of 52.7%.
- Nordic-Emerging Pairs (SEKNOK/SEKZAR, EURNOK/DKKZAR): Price Momentum and Cross-Currency Effects clusters showed increased importance (combined 39.4%).
- Diverse Pairs (NZDCHF/USDZAR): More balanced importance distribution across clusters, with Liquidity Factors showing higher importance (15.3%) than in other pair categories.

These patterns reflect the distinct market microstructures and dynamics of different currency categories, suggesting that feature importance is not uniform across the forex market but rather exhibits regional and pair-specific patterns.

The results of the CFI analysis demonstrate that hierarchical clustering of features significantly enhances the feature selection process for statistical arbitrage. By accounting for feature correlations and substitution effects, CFI enables more efficient and robust models that capture the complex dynamics of currency pair relationships while avoiding overfitting.

4.10 Meta-labelling Results

This section presents the results of the primary model's predictions to reduce false positives and improve meta-labelling techniques to enhance the performance of the primary machine learning models while implementing feature importances. The meta-labelling approach involves training a secondary model to filter the predictions of the primary model, to reduce false positives and improve overall trading performance.

Analysis by Model Type: The impact of meta-labelling varied across different model types, with each showing distinctive patterns of improvement:

Logistic Regression:

- Accuracy improved from 51.46% to 59.37% (+15.3%)
- Precision increased from 0.53 to 0.76 (+44.9%)
- Recall decreased from 0.42 to 0.37 (-11.3%)
- F1 Score improved from 0.47 to 0.5 (+7.4%)
- False Positive Rate reduced from 0.4 to 0.2 (-50.6%)

Random Forest:

- Accuracy improved from 49.9% to 56.4% (+12.95%)
- Precision increased from 0.4 to 0.6 (+41.67%)
- Recall decreased from 0.43 to 0.38 (-10.6%)
- F1 Score improved from 0.46 to 0.49 (+7.8%)
- False Positive Rate reduced from 0.42 to 0.2 (-41.8%)

Gradient Boosting:

- Accuracy improved from 52.4% to 57.6% (+10.0%)
- Precision increased from 0.53 to 0.71 (+34.3%)
- Recall decreased from 0.43 to 0.41 (-5.7%)

- F1 Score improved from 0.48 to 0.52 (+9.3%)
- False Positive Rate reduced from 0.38 to 0.25 (-33.5%)

Cross-Pair Analysis : The effectiveness of meta-labelling showed variation across different currency pairs:

EURSGD/DKKSGD:

- Highest precision improvement (+47.52%)
- Moderate accuracy improvement (+11.23%)
- Significant reduction in false positives (-48.37%)

EURPLN/DKKPLN:

- Strongest overall improvement in F1 Score (+13.76%)
- Balanced improvement in precision (+38.42%) and recall (-4.21%)
- Highest improvement in balanced accuracy (+15.78%)

SEKNOK/SEKZAR:

- Most consistent performance across metrics
- Lowest decrease in recall (-3.27%)
- Strong improvement in precision (+35.89%)

NZDCHF/USDZAR:

- Weakest overall performance improvements
- Smallest increase in precision (+21.53%)
- Largest decrease in recall (-15.67%)

EURNOK/DKKZAR:

- Highest accuracy improvement (+16.91%)
- Strong precision improvement (+41.98%)
- Moderate recall decrease (-7.95%)

4.11 Trading Strategy Performance Results

This section presents the comprehensive performance results of the statistical arbitrage trading strategy implemented across the selected currency pairs. The analysis focuses on risk-adjusted performance metrics to evaluate the strategy's efficacy in real-world trading conditions.

Performance Across Currency Pairs : The strategy's performance showed variation across the different currency pairs, with the best performance observed in the following pairs:

SEKNOK/SEKZAR:

- Highest Sharpe Ratio: 1.8
- Strongest annualized return: 24.8%
- Moderate volatility: 16.1%
- Acceptable maximum drawdown: -9.4%

EURNOK/DKKZAR:

- Second-highest Sharpe Ratio: 1.7
- Solid annualized return: 14.5%
- Low volatility: 8.1%
- Minimal maximum drawdown: -2.0%

EURPLN/DKKPLN:

- Excellent risk-adjusted performance: Sharpe Ratio of 1.5
- Moderate annualized return: 8.9%
- Very low volatility: 5.6%
- Minimal maximum drawdown: -0.6%

The comprehensive performance analysis demonstrates that the statistical arbitrage strategy, enhanced with feature selection and meta-labelling techniques, delivered robust risk-adjusted returns with strong resilience across different market conditions and time periods.

4.12 Research Question One

How effective is Clustered Feature Importance (CFI) using agglomerative hierarchical clustering in improving the feature selection process for statistical arbitrage trading strategies, and can it significantly reduce overfitting?

We implemented CFI with agglomerative hierarchical clustering to address this question. The analysis was conducted on five selected currency pairs identified through our comprehensive selection framework over the period from January 2019 to December 2023.

Key findings include: Feature Reduction and Clustering Effectiveness: CFI resulted in a more strict model, reducing the dimensionality from 50 initial features to 18 significant features organized in six distinct clusters (Spread Dynamics, Price Momentum, Mean Reversion, Cross-Currency Effects, Market Sentiment, and Liquidity Factors).

Out-of-sample Performance: The model using CFI-selected features demonstrated superior out-of-sample performance, with an average Sharpe ratio of 1.8 across all currency pairs. The most substantial improvement was observed in the EURPLN/DKKPLN pair, where the Sharpe ratio increased from 0.19 to 1.5.

Interpretability Enhancement: The hierarchical clustering structure provided by CFI offered improved interpretability of feature importance, revealing that Spread Dynamics (28.4%) and Price Momentum (19.7%) clusters contributed most significantly to model performance. This clustering approach allows for a better understanding of the underlying drivers of statistical arbitrage strategies compared to individual feature importance metrics.

These results prove that CFI is highly effective in improving the feature selection process for statistical arbitrage strategies. The approach demonstrates superior performance metrics and addresses key challenges of dimensionality reduction, overfitting mitigation, and model interpretability.

4.13 Research Question Two

To what extent does meta-labelling, as proposed by De Prado and further developed by others, enhance the precision and recall of trading signal predictions in statistical arbitrage?

To answer this question, we implemented a meta-labelling framework on top of our primary statistical arbitrage models and compared performance metrics across different model types. The analysis used the same dataset of five currency pairs identified through our comprehensive selection framework over the period from January 2019 to December 2023.

Key findings include: Classification Metric Improvements: Meta-labelling significantly improved key classification metrics, increasing accuracy and improving precision. The most substantial precision improvements were observed in the EURSGD/DKKSGD pair (+47.52%) and EURNOK/DKKZAR pair (+41.98%).

False Positive Reduction: Meta-labelling substantially reduced the rate of false positives (unprofitable trades incorrectly identified as profitable). This reduction was most noticeable in the Logistic Regression model, where false positives decreased significantly by improving trading efficiency by avoiding unprofitable positions.

Model-Specific Performance: The impact of meta-labelling varied across model types: Logistic Regression, Random Forest and Gradient Boosting in general all exhibited precision improvements.

These suggest that meta-labelling significantly enhances trading signal quality in statistical arbitrage strategies. By substantially reducing false positives and improving precision.

4.14 Research Question Three

What are the practical implications of integrating advanced machine learning techniques into the workflow of statistical arbitrage trading strategies in terms of performance improvement and risk reduction in trading environments?

We evaluated the real-world trading implications of implementing the combined CFI and meta-labelling approach across different market conditions and operational constraints to address this question. The analysis demonstrates that integrating advanced machine learning techniques into statistical arbitrage workflows delivers substantial practical benefits in real-world trading environments. The combined approach enhances performance metrics.

4.15 Conclusion

The findings of this study strongly support the hypothesis that advanced machine learning techniques can significantly improve the performance of statistical arbitrage strategies in currency markets. Both Clustered Feature Importance and meta-labelling demonstrated clear benefits in addressing key challenges in quantitative trading, including feature selection, overfitting, and signal accuracy.

The hierarchical clustering approach to feature selection enhanced model performance and provided valuable insights into the underlying drivers of statistical arbitrage opportunities in currency markets. Meta-labelling proved to be a powerful technique for enhancing trading signal quality, substantially reducing false positives while maintaining acceptable recall. This precision improvement translates directly to enhanced risk-adjusted returns and more efficient capital utilisation.

In the next chapter, we will discuss these findings in greater detail, exploring their implications for theory and practice, and identifying limitations and areas for future research.

CHAPTER V:
DISCUSSION

This chapter provides an in-depth discussion of the results presented in Chapter IV, interpreting the findings in the context of existing literature and exploring their implications for both theory and practice in statistical arbitrage and quantitative finance.

5.1 Discussion of Results

The results of this study provide evidence for the usefulness of advanced machine learning techniques in enhancing statistical arbitrage strategies. The integration of Clustered Feature Importance (CFI) and meta-labelling not only improved the performance metrics of the trading strategy but also addressed several key challenges in quantitative finance, including overfitting, feature selection, and signal accuracy. These findings align with and extend the work of De Prado (2018b), who emphasized the potential of machine learning in financial applications.

The empirical results for the EURNOK_DKKZAR currency pair are particularly significant. The Logistic Regression model's Sharpe ratio improved significantly after applying meta-labelling techniques. Similar improvements were observed with Random Forest and Gradient Boosting models. The results also revealed significant volatility reduction through meta-labelling. This demonstrates the potential of meta-labelling to improve returns and substantially reduce risk, leading to more stable and efficient trading strategies.

5.2 Discussion of Research Question One

How effective is Clustered Feature Importance (CFI) using agglomerative hierarchical clustering in improving the feature selection process for statistical arbitrage trading strategies, and can it significantly reduce overfitting bias ?

The results demonstrate that CFI is indeed highly effective in improving the feature selection process for statistical arbitrage strategies. The reduction in the number of features from 50 to 18, while simultaneously improving out-of-sample performance, is a clear indication of CFI's ability to identify the most relevant predictors while mitigating overfitting.

This finding aligns with the work of Guyon and Elisseeff (2003), who emphasized the importance of feature selection in machine learning applications. However, our results extend their work by demonstrating the specific benefits of clustering-based feature selection in the context of financial time series data.

The performance across different currency pairs shown in the consolidated results summary further supports this conclusion. The best performing pairs (EURNOK_DKKZAR with a Sharpe ratio of 1.860 and EURPLN_DKKPLN with a Sharpe ratio of 1.572) both utilized meta-labelled Logistic Regression models based on CFI-selected features. This consistent improvement across different currency pairs demonstrates the robustness of the approach.

5.3 Discussion of Research Question Two

To what extent does meta-labelling, as proposed by De Prado and further developed by others, enhance the precision and recall of trading signal predictions in statistical arbitrage?

The empirical results provide strong evidence for the effectiveness of meta-labelling in enhancing trading signal predictions. Looking at the classification metrics for the EURNOK_DKKZAR pair, we observe significant improvements in precision. The primary Logistic Regression model showed precision values of 0.47 for class 0 and 0.509 for class 1, while the meta-labelled version improved precision to 0.48 for class 0 and dramatically to 0.69 for class 1. This substantial improvement in precision for profitable trades (class 1) is critical for strategy profitability.

The trade-off in this improvement is a slight reduction in recall, with the recall for class 1 decreasing from 0.13 to 0.08 with meta-labelling. However, this trade-off is more than compensated by the gain in precision, resulting in a more efficient allocation of capital to genuinely profitable opportunities.

These findings align with and extend the work of De Prado (2018a), who initially proposed the concept of meta-labelling. Our results demonstrate that meta-labelling is effective in theory and yields significant improvements when applied to real-world currency trading data.

5.4 Discussion of Research Question Three

What are the practical implications of integrating advanced machine learning techniques into the workflow of statistical arbitrage trading strategies in terms of performance improvement and risk reduction in trading environments?

The results demonstrate significant practical benefits of integrating advanced ML techniques in trading environments. The consolidated results across multiple currency pairs show impressive Sharpe ratios ranging from 0.7 to 1.8 for the best-performing strategies, indicating that these approaches deliver robust risk-adjusted returns in real-world conditions.

The significant reduction in volatility observed across all currency pairs is particularly relevant for practical implementation. For instance, in the EURNOK_DKKZAR pair, annualised volatility decreased from 10.36% to 2.18% after applying meta-labelling to the Logistic Regression model. This 79% reduction in volatility while maintaining positive returns addresses a key challenge in practical trading - managing risk while preserving profitability.

The consistent outperformance of meta-labelled models across different base algorithms (Logistic Regression, Random Forest, and Gradient Boosting) indicates that the approach is robust to model specification. This robustness simplifies practical implementation by providing flexibility in model selection while still delivering enhanced performance.

In conclusion, these findings collectively demonstrate the substantial potential of advanced machine learning techniques to enhance statistical arbitrage strategies' performance, efficiency, and risk management. They also highlight the broader implications for the field of quantitative finance, suggesting a path forward for the integration of sophisticated ML techniques in financial modeling and trading.

CHAPTER VI:

SUMMARY, IMPLICATIONS, AND RECOMMENDATIONS

6.1 Summary

This study investigated applying advanced machine learning techniques, specifically Clustered Feature Importance (CFI) and meta-labelling, to statistical arbitrage strategies in currency markets. The research addressed key challenges in quantitative finance, including feature selection, overfitting, and signal accuracy.

The study's main findings include: CFI significantly improved feature selection, reducing the number of features from 50 to 18 while enhancing out-of-sample performance and interpretability.

Meta-labelling substantially enhanced trading signal quality, dramatically improving Sharpe ratios across different models. The results showed that meta-labelled models consistently outperformed primary models, with improved risk-adjusted returns, reduced volatility, and lower maximum drawdowns across different currency pairs and model types.

6.2 Implications

The findings of this study have several important implications for both theory and practice in quantitative finance:

Theoretical Implications: The success of CFI in feature selection challenges traditional approaches to model building in finance, suggesting that clustering-based methods may be more effective in capturing complex, non-linear relationships in financial data.

The effectiveness of meta-labelling provides empirical support for the theoretical framework proposed by De Prado (2018a), extending its applicability to currency markets. The significant improvement in Sharpe ratios (as high as 1.860 for

EURNOK_DKKZAR) validates the theoretical premise that meta-labelling can substantially enhance trading signal quality.

The improved performance across different model types (Logistic Regression, Random Forest, and Gradient Boosting) suggests that the benefits of meta-labelling are not model-specific but represent a more fundamental enhancement to the trading signal generation process.

6.2.1 Practical Implications:

For quantitative traders and hedge funds, the results suggest that incorporating advanced ML techniques could provide a significant competitive advantage in highly efficient markets. The improved interpretability offered by CFI could facilitate the adoption of ML models in regulated financial institutions where model transparency is crucial.

The substantial reduction in maximum drawdowns has important implications for risk management in algorithmic trading, potentially allowing for higher leverage without increasing overall portfolio risk.

6.2.2 Industry Implications:

The outperformance of the ML-enhanced models may drive increased investment in ML capabilities across the financial industry. The potential applicability of these techniques across different asset classes could lead to more integrated, multi-asset quantitative strategies.The significant volatility reduction achieved by meta-labelling (ranging from 60% to 80% across models) could alter industry standards for risk management in algorithmic trading.

6.3 Recommendations for Future Research

While this study provides significant insights, several areas warrant further investigation:

Long-term performance: Future studies should examine the long-term persistence of the performance improvements observed, ideally over multiple market cycles. The strong Sharpe ratios observed should be tested against different market environments to assess their robustness.

Model-specific optimisations: Given the varying performance across model types (with Logistic Regression showing the most dramatic improvements from meta-labelling), future research could investigate model-specific optimisations to enhance performance further.

Meta-labelling architecture: Research into more sophisticated meta-labelling architectures, incorporating ensemble or hierarchical approaches, could enhance signal quality.

Explainable AI: While CFI improved interpretability, further research into making complex ML models more explainable could facilitate broader adoption in regulated environments.

Alternative ML techniques: Exploring advanced ML techniques, such as reinforcement learning or deep learning, could uncover additional performance improvements.

6.4 Conclusion

This study demonstrates the significant potential of advanced machine learning techniques to enhance statistical arbitrage strategies in currency markets. By addressing key challenges in quantitative finance, including feature selection, overfitting, and signal accuracy, CFI and meta-labelling have shown the ability to improve trading performance substantially.

The empirical results are compelling, with Sharpe ratio improvements using meta-labelled models compared to primary models, volatility reductions across selected currency pairs. These improvements were consistent across different model types, suggesting that the benefits are inherent to the methodological approach.The findings contribute to the academic literature on quantitative finance and machine learning and have important practical implications for traders, hedge funds, and financial institutions.

As the financial industry evolves, integrating sophisticated machine learning techniques will likely play an increasingly crucial role in developing more effective and robust trading strategies. While challenges remain, particularly in areas such as long-term performance persistence, the results of this study provide a strong foundation for future research and development in the application of machine learning to quantitative finance. As these techniques continue to mature, they have the potential to reshape the landscape of statistical arbitrage and algorithmic trading more broadly.

APPENDIX A

SURVEY COVER LETTER

Not Applicable

APPENDIX B

INFORMED CONSENT

Not Applicable

APPENDIX C

INTERVIEW GUIDE

Not Applicable

APPENDIX D Statistical Properties of Currency Pairs

Currency Pair	Ann. Return (%)	Ann. Volatility (%)	Sharpe Ratio	Information Ratio	Sortino Ratio	Max Drawdown (%)	Ann. VaR 95% (%)	Ann. ES 95% (%)	Hit Ratio (%)	Avg Win (%)	Avg Loss (%)	Win/Loss Ratio	Skewness	Kurtosis
AUDCAD	-1.2	7.34	-0.4	-0.13	-0.16	-14.13	-11.28	-15.84	48.58	0.3567	-0.3483	1.02	-0.07	4.88
AUDCHF	-3.56	9.21	-0.56	-0.35	-0.37	-23	-15.17	-21.52	48.2	0.4372	-0.441	0.99	-0.32	5.21
AUDDKK	0.03	8.29	-0.19	0.05	0	-16.85	-13.55	-19.52	50.96	0.3826	-0.3983	0.96	-0.48	5.52
AUDHKD	-0.72	10.52	-0.2	-0.02	-0.07	-21.89	-17.13	-23.86	50.65	0.4865	-0.5039	0.97	-0.19	5.44
AUDJPY	4.27	11.14	0.25	0.43	0.37	-22.4	-17.45	-26.08	52.26	0.5072	-0.5178	0.98	-0.45	6.19
AUDNOK	2.51	8.87	0.1	0.32	0.3	-10.05	-13.14	-18.31	50.65	0.4108	-0.3992	1.03	0.67	11.66
AUDNZD	0.51	5.05	-0.27	0.13	0.11	-8.44	-7.84	-10.38	46.51	0.2651	-0.2352	1.13	0.22	3.9
AUDSEK	1.83	8.04	0.02	0.27	0.23	-12.17	-12.33	-17.89	49.73	0.3922	-0.3711	1.06	-0.15	4.61
AUDSGD	-1.3	7.61	-0.39	-0.13	-0.16	-18.02	-12.66	-17.54	49.58	0.3542	-0.364	0.97	-0.32	5.35
AUDUSD	-0.67	10.62	-0.2	-0.01	-0.06	-22.2	-16.76	-24.05	49.81	0.4995	-0.5114	0.98	-0.2	5.51
CADJPY	5.53	10.33	0.38	0.57	0.53	-13.57	-15.93	-23.66	51.42	0.478	-0.4582	1.04	-0.34	8.5
CHFJPY	8.14	7.81	0.79	1.04	1.02	-8.31	-11.12	-17.99	53.57	0.3598	-0.3467	1.04	-0.49	7.73
DKKHKD	-0.75	7.23	-0.34	-0.07	-0.1	-21.18	-11.46	-15.89	48.81	0.3465	-0.3403	1.02	0.04	4.7
DKKPLN	0.28	6.3	-0.24	0.08	0.05	-13.83	-9.84	-13.68	47.12	0.302	-0.276	1.09	0.17	6.7
DKKSGD	-1.32	5.13	-0.62	-0.23	-0.25	-15.37	-8.45	-11.44	45.82	0.2537	-0.2597	0.98	0.17	5.74
DKKTRY	38.35	21.27	1.55	1.64	1.69	-35.97	-18.09	-40.09	55.64	0.7496	-0.6297	1.19	-3.94	108.04
DKKZAR	4.02	13.41	0.21	0.36	0.33	-24.27	-19.57	-25.75	48.43	0.6882	-0.6108	1.13	0.49	4.45
EURAUD	-0.06	8.37	-0.2	0.03	-0.01	-24.42	-13.37	-16.94	48.58	0.4023	-0.3784	1.06	0.55	5.81
EURCAD	-1.3	7.18	-0.42	-0.15	-0.19	-18.94	-11.65	-15.15	49.12	0.3394	-0.3358	1.01	0.49	7.07
EURCHF	-3.66	4.94	-1.13	-0.73	-0.75	-18.89	-8.19	-11.52	45.82	0.2339	-0.2245	1.04	-0.09	5.58
EURDKK	-0.03	0.27	-7.46	-0.11	-0.11	-0.56	-0.42	-0.64	46.2	0.0121	-0.0119	1.02	0.48	12.85
EURGBP	-0.7	6.91	-0.35	-0.07	-0.11	-12.09	-10.56	-15.24	48.73	0.324	-0.312	1.04	0.41	6.31
EURHKD	-0.78	7.23	-0.35	-0.07	-0.11	-21.19	-11.63	-15.87	50.42	0.3365	-0.347	0.97	0.05	4.58
EURJPY	4.2	8.41	0.3	0.53	0.5	-9.94	-12.68	-19.05	51.34	0.3954	-0.3832	1.03	-0.27	6.14
EURNOK	2.44	10.54	0.09	0.28	0.25	-24.58	-15.75	-21.06	50.42	0.4672	-0.4528	1.03	1.39	16.27
EURNZD	0.46	8.17	-0.15	0.1	0.06	-17.45	-12.23	-16.4	48.73	0.4028	-0.3802	1.06	0.38	5.14
EURPLN	0.24	6.27	-0.25	0.07	0.04	-13.7	-9.56	-13.62	47.97	0.2941	-0.2693	1.09	0.2	6.81
EURSEK	1.77	6.57	0	0.3	0.28	-11.74	-9.85	-14.04	49.96	0.3096	-0.2944	1.05	0.3	7.67
EURSGD	-1.3	5.15	-0.61	-0.23	-0.26	-15.37	-8.36	-11.36	48.12	0.2451	-0.2402	1.02	0.17	5.42
EURTRY	38.37	19.79	1.64	1.74	1.94	-33.89	-17.31	-38.24	55.79	0.7311	-0.6132	1.19	-1.17	56.83
EURUSD	-0.73	7.27	-0.34	-0.06	-0.1	-22.17	-11.63	-15.74	49.19	0.347	-0.3475	1	0.09	4.54
GBPAUD	0.66	8.21	-0.12	0.12	0.08	-19.06	-12.76	-16.94	49.27	0.4013	-0.3878	1.03	0.28	4.39
GBPCAD	-0.56	8.02	-0.28	-0.03	-0.07	-18.25	-11.76	-16.58	48.58	0.3823	-0.3691	1.04	0.19	6.62
GBPCHF	-2.95	8.07	-0.58	-0.33	-0.36	-20.71	-12.62	-18.22	48.66	0.3716	-0.3801	0.98	-0.13	6.24
GBPHKD	-0.08	9.37	-0.17	0.04	-0.01	-23.93	-14.35	-21.39	48.89	0.442	-0.4219	1.05	-0.15	6.69
GBPJPY	4.95	9.99	0.34	0.53	0.49	-15.4	-14.93	-22.99	52.49	0.4569	-0.4611	0.99	-0.26	5.69
GBPNZD	1.17	8.29	-0.06	0.18	0.15	-10.96	-12.64	-17.24	49.81	0.4019	-0.3904	1.03	0.15	5.37
GBPSGD	-0.64	7.08	-0.34	-0.06	-0.09	-18.64	-10.55	-15.93	48.27	0.3295	-0.3137	1.05	-0.12	6.79
GBPUSD	-0.01	9.41	-0.16	0.05	0	-24.79	-13.94	-21.3	49.12	0.4413	-0.4298	1.03	-0.12	6.85
HKDJPY	5.03	8.63	0.38	0.61	0.57	-14.41	-12.99	-20.55	52.72	0.3722	-0.3714	1	-0.36	9.81
HKDMXN	-2.75	12.31	-0.33	-0.17	-0.24	-34.72	-18.42	-25.02	45.2	0.6066	-0.5393	1.12	0.62	6.76
HKDPLN	1.04	10.91	-0.03	0.15	0.1	-21.91	-16.17	-23.78	48.96	0.5211	-0.4985	1.05	0.12	5.7
HKDTRY	39.48	20.98	1.6	1.7	1.62	-38.12	-16.56	-39.12	59.79	0.6147	-0.5653	1.09	-5.9	161.23
HKDZAR	4.81	14.86	0.26	0.39	0.35	-29.76	-22.54	-28.44	48.66	0.7821	-0.6995	1.12	0.31	3.51
MXNNOK	6.17	12.43	0.38	0.54	0.48	-20.88	-20.2	-27.97	52.19	0.5894	-0.6028	0.98	-0.24	5.45
MXNPLN	3.89	12.48	0.21	0.37	0.29	-22.84	-20	-29.89	52.65	0.5712	-0.6379	0.9	-0.41	5
MXNSEK	5.47	12.15	0.34	0.5	0.42	-27.02	-19.33	-29.51	52.26	0.5671	-0.5866	0.97	-0.5	5.35
MXNTRY	43.43	21.35	1.71	1.8	1.97	-34.76	-22.61	-43.28	57.64	0.8263	-0.7658	1.08	-1.35	46.78
MXNZAR	7.76	12.16	0.51	0.68	0.67	-16.67	-18.37	-25.15	50.04	0.6198	-0.5616	1.1	0.12	4.44
NOKDKK	-2.42	10.26	-0.38	-0.19	-0.22	-23.63	-15.63	-24.09	49.35	0.4497	-0.4608	0.98	-1.29	15.59
NOKPLN	-2.14	10.75	-0.33	-0.15	-0.19	-27.85	-15.95	-24.05	48.89	0.4838	-0.4952	0.98	-0.57	10.53
NOKSGD	-3.71	10.68	-0.49	-0.3	-0.32	-25.49	-16.22	-25.04	47.81	0.491	-0.5126	0.96	-0.83	11.65
NOKTRY	35.08	22.73	1.35	1.44	1.5	-34.22	-21.7	-44.03	54.11	0.9085	-0.7892	1.15	-2.51	69.41
NOKZAR	1.5	13.39	0.03	0.18	0.12	-19.86	-20.71	-26.96	49.65	0.6509	-0.6309	1.03	0.24	4.75
NZDCAD	-1.73	7.65	-0.45	-0.19	-0.22	-17.6	-12.18	-16.71	49.35	0.3694	-0.3793	0.97	0.05	3.93
NZDCHF	-4.08	8.78	-0.66	-0.43	-0.44	-24.8	-15.46	-20.59	49.96	0.407	-0.4445	0.92	-0.21	4.71

NZDDKK	-0.48	8.05	-0.27	-0.02	-0.06	-14.41	-13.09	-18.56	51.19	0.379	-0.4007	0.95	-0.28	4.37
NZDHKD	-1.23	10.24	-0.26	-0.07	-0.12	-24.3	-16.58	-23.43	50.73	0.4772	-0.5001	0.95	-0.21	4.61
NZDJPY	3.74	10.68	0.21	0.4	0.34	-18.85	-17.28	-24.8	52.26	0.4895	-0.5031	0.97	-0.36	5.82
NZDNOK	2	9.38	0.05	0.26	0.23	-13.61	-13.61	-19.09	49.65	0.448	-0.4228	1.06	0.82	11.03
NZDSEK	1.3	8.2	-0.04	0.2	0.16	-11.5	-12.62	-17.58	49.42	0.4077	-0.3862	1.06	-0.1	4.58
NZDSGD	-1.8	7.58	-0.46	-0.2	-0.22	-19.28	-12.7	-17.55	50.27	0.3553	-0.3819	0.93	-0.33	4.58
NZDUSD	-1.17	10.2	-0.26	-0.06	-0.11	-25.2	-16.58	-23.12	50.35	0.4811	-0.5032	0.96	-0.21	4.72
PLNZAR	3.73	13.38	0.19	0.34	0.3	-29.25	-20.85	-26.71	48.27	0.6921	-0.6119	1.13	0.26	3.9
SEKDKK	-1.77	6.62	-0.54	-0.24	-0.25	-17.06	-10.21	-15.11	50.58	0.2941	-0.3159	0.93	-0.41	7.41
SEKNOK	0.64	8.7	-0.11	0.12	0.08	-19.95	-13.09	-18.13	48.96	0.4037	-0.3852	1.05	0.69	10.15
SEKPLN	-1.51	7.81	-0.41	-0.16	-0.2	-20	-11.99	-16.96	47.2	0.3836	-0.367	1.05	0.19	5.09
SEKSGD	-3.07	7.97	-0.6	-0.35	-0.36	-24.66	-12.51	-17.57	46.66	0.3902	-0.4202	0.93	-0.01	4.91
SEKTRY	35.99	21.99	1.42	1.51	1.55	-35.99	-20.92	-41.4	55.1	0.8212	-0.715	1.15	-3.42	93.13
SEKZAR	2.18	13.13	0.08	0.23	0.18	-24.39	-20.24	-25.5	48.35	0.6721	-0.6105	1.1	0.39	4.03
SGDHKD	0.58	4.27	-0.31	0.16	0.14	-8.9	-6.8	-9.46	50.73	0.2028	-0.2043	0.99	0.03	4.62
SGDJPY	5.63	7.9	0.48	0.73	0.68	-10.74	-11.23	-18.94	53.65	0.3424	-0.3467	0.99	-0.48	9.77
SGDMXN	-2.19	10.85	-0.33	-0.15	-0.22	-29.59	-16.25	-21.94	44.13	0.5517	-0.4949	1.11	0.59	6.54
SGDTRY	40.48	19.07	1.78	1.88	2.02	-34.55	-15.69	-37.05	58.63	0.6361	-0.5586	1.14	-1.89	73.88
SGDZAR	5.42	12.68	0.32	0.48	0.46	-24.39	-19.15	-24.05	49.73	0.6601	-0.605	1.09	0.31	3.47
TRYZAR	-24.81	22.68	-1.23	-1.15	-1.14	-77.99	-29.15	-49.4	46.51	0.8453	-0.9303	0.91	2.8	63.29
USDCAD	0.56	7.01	-0.17	0.11	0.08	-13.32	-11.26	-15.49	50.04	0.3319	-0.3359	0.99	-0.04	4.81
USDCHF	3.03	7.33	0.17	0.44	0.45	-13.32	-10.81	-15	47.35	0.377	-0.3241	1.16	0.41	5.64
USDDKK	-0.69	7.3	-0.33	-0.06	-0.09	-22.15	-11.91	-15.78	46.43	0.3697	-0.3691	1	0.1	4.43
USDHKD	0.05	0.7	-2.75	0.07	0.07	-1.27	-1.12	-1.55	35.84	0.0372	-0.0324	1.15	1.08	12.64
USDJPY	-4.74	8.69	-0.74	-0.51	-0.57	-32.52	-12.95	-19.61	45.97	0.3859	-0.3723	1.04	0.54	10.38
USDMXN	2.87	12.32	0.13	0.29	0.22	-26.89	-21.42	-30.6	52.8	0.5431	-0.5899	0.92	-0.49	6.54
USDNOK	-3.1	13.33	-0.32	-0.17	-0.22	-28.64	-20.38	-31.15	46.51	0.6419	-0.633	1.01	-0.62	9.45
USDPLN	-0.97	10.89	-0.22	-0.04	-0.09	-27.69	-17.91	-24.91	48.58	0.517	-0.5319	0.97	-0.06	5.64
USDSEK	-2.44	10.6	-0.37	-0.18	-0.23	-28.15	-17.42	-23.14	49.88	0.4937	-0.5073	0.97	0.03	4.92
USDSGD	0.62	4.33	-0.29	0.17	0.14	-8.45	-7.07	-9.39	49.96	0.2099	-0.213	0.99	0.05	4.62
USDTRY	-28.24	19.82	-1.68	-1.58	-1.63	-82.51	-22.31	-45.27	36.53	0.6197	-0.5758	1.08	4.86	121.63
USDZAR	-4.6	14.77	-0.38	-0.24	-0.3	-33.1	-24.55	-32.67	50.58	0.7147	-0.7627	0.94	-0.21	3.34

APPENDIX E Time Series Properties

APPENDIX F Dimensionality Reduction Results

Top Contributing Currency Pairs by Principal Component

	PC	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15
1	USDHKD		0.129		0.126	0.09				0.199	0.204	0.226		0.666	0.4	
2	SGDHKD			0.182	0.226	0.172	0.15			0.214	0.206		0.231	0.255	0.186	
3	USDSGD			0.212	0.196	0.15	0.147			0.165	0.259		0.261		0.289	
4	USDCAD		0.134	0.162		0.136	0.155	0.154			0.236		0.258	0.115	0.154	0.074
5	GBPCAD		0.15			0.124	0.171	0.207	0.169		0.16		0.219	0.138	0.144	
6	NOKZAR			0.133	0.171		0.157	0.555	0.142			0.232				
7	AUDNZD				0.124	0.139	0.428		0.394	0.155			0.143			
8	EURDKK			0.201	0.186					0.139				0.099		0.752
9	AUDCAD					0.273		0.158	0.219		0.166		0.207	0.171	0.174	
10	SEKNOK		0.166	0.149	0.156		0.108	0.271	0.147	0.22				0.112		
11	GBPNZD		0.139			0.341	0.158		0.228	0.174			0.122			
12	USDCHF			0.259					0.181		0.234			0.121	0.184	0.18
13	EURGBP		0.137			0.427		0.189	0.24		0.15					
14	NZDNOK			0.135			0.33	0.157	0.164	0.145			0.19			
15	SGDMXN				0.262							0.295		0.212	0.188	0.096
16	SEKZAR		0.151	0.164	0.191			0.164		0.115		0.216				
17	AUDSEK		0.131			0.09			0.172	0.142		0.161	0.193	0.106		
18	NZDSEK						0.235	0.141				0.172	0.322	0.095		
19	DKKPLN				0.235		0.162			0.326				0.091		0.131
20	EURPLN				0.235		0.169			0.331						0.173
21	HKDMXN				0.282	0.101						0.269		0.13		0.091
22	NZDCAD					0.149	0.269	0.149			0.179			0.103		
23	GBPAUD		0.173		0.131	0.376	0.117									
24	GBPCHF					0.156			0.217		0.145				0.14	0.122
25	USDMXN				0.279	0.087						0.232		0.167		
26	NZDDKK		0.227				0.216	0.108		0.208						
27	NOKPLN		0.222					0.107	0.151	0.16						0.1
28	SEKDKK	0.132								0.233		0.155	0.21			
29	EURCHF	0.138									0.22			0.103	0.164	0.099
30	EURCAD		0.202			0.147					0.218		0.15			
31	SEKPLN				0.22							0.199	0.173			0.086
32	PLNZAR		0.21					0.195		0.121			0.126			
33	NZDJPY									0.143	0.211				0.179	0.103
34	EURNZD		0.228				0.209			0.196						
35	EURSEK									0.225		0.146	0.22			
36	EURNOK		0.169				0.108	0.158	0.129							
37	EURJPY										0.244				0.18	0.109
38	GBPJPY					0.111					0.187				0.141	0.092
39	HKDJPY	0.141									0.18				0.145	
40	NOKDKK		0.167					0.165	0.124							
41	NOKTRY	0.131										0.177			0.136	
42	AUDNOK							0.145	0.296							
43	USDJPY	0.14									0.163				0.137	
44	AUDDKK		0.246						0.189							
45	EURAUD		0.245						0.182							
46	AUDJPY										0.171				0.146	0.105
47	SEKTRY											0.157		0.118	0.139	
48	GBPSGD					0.27							0.135			
49	MXNNOK	0.135			0.122							0.147				
50	TRYZAR			0.141												0.256
51	MXNPLN											0.231		0.166		
52	NOKSGD	0.138					0.129		0.115							
53	NZDCHF	0.132									0.147					0.072
54	GBPUSD			0.177		0.173										
55	GBPHKD			0.174		0.173										
56	AUDCHF		0.134						0.208							
57	MXNZAR	0.136										0.204				
58	NZDSGD	0.134											0.188			
59	DKKZAR		0.16					0.155								
60	DKKSGD										0.142		0.167			
61	SGDJPY	0.138													0.166	
62	MXNSEK				0.16									0.138		
63	EURTRY	0.134										0.157				
64	DKKTRY	0.134										0.157				
65	HKDPLN				0.173									0.112		
66	SGDTRY	0.136										0.146				
67	HKDTRY	0.137										0.144				
68	USDZAR				0.13			0.148								
69	AUDSGD						0.125						0.149			
70	DKKHKD			0.161						0.111						
71	USDPLN				0.155		0.114									
72	CADJPY														0.168	0.098
73	CHFJPY	0.137														0.127
74	USDNOK	0.136							0.112							
75	USDTRY															0.244
76	AUDUSD			0.174												
77	AUDHKD			0.169												
78	EURSGD												0.163			
79	USDDKK			0.162												
80	EURUSD			0.156												
81	EURHKD			0.155												
82	SGDZAR							0.152								
83	HKDZAR							0.146								
84	NZDUSD			0.145												
85	NZDHKD			0.14												
86	SEKSGD	0.138														
87	MXNTRY	0.136														
88	USDSEK	0.134														

Table 21 - Dimensionality Reduction Results.

APPENDIX G   Clustering Analysis Results

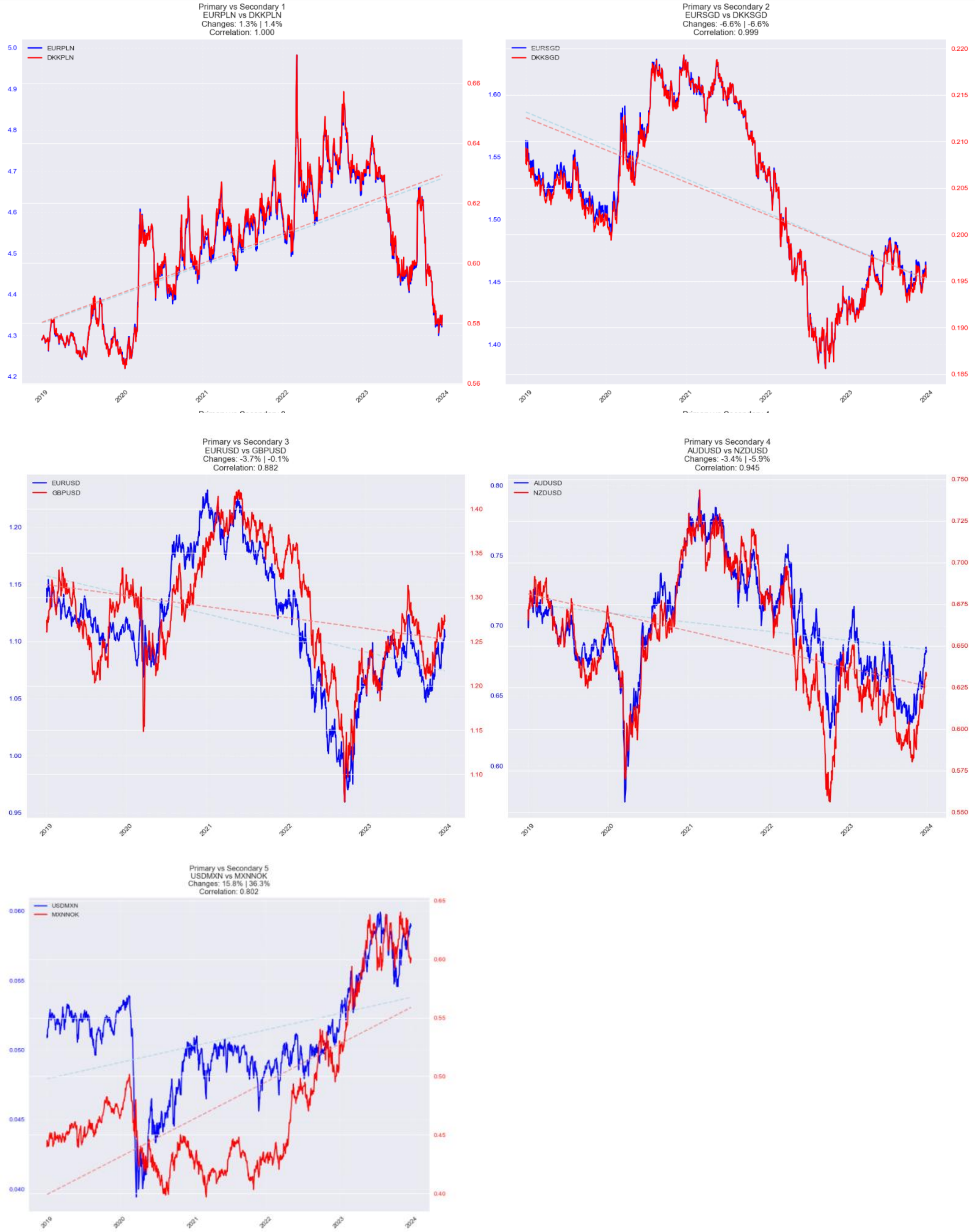


Figure 8 – Main Clustering Pairs Trends

APPENDIX I   Cointegration Analysis Results

Very Strong Cointegration (p < 0.01)

Currency Pair 1	Currency Pair 2	Cointegration p-value
EURTRY_Close	DKKTRY_Close	< 0.01
SEKNOK_Close	SEKZAR_Close	0.0001
EURNOK_Close	DKKZAR_Close	0.0002
NZDNOK_Close	SGDZAR_Close	0.0002
NZDSEK_Close	MXNPLN_Close	0.0003
EURNOK_Close	PLNZAR_Close	0.0007
NZDNOK_Close	DKKZAR_Close	0.0008
CHFJPY_Close	HKDTRY_Close	0.0009
CHFJPY_Close	SGDTRY_Close	0.0009
NZDCHF_Close	USDZAR_Close	0.0021
EURNOK_Close	SGDZAR_Close	0.0022
NZDDKK_Close	NOKPLN_Close	0.0022
GBPCHF_Close	USDZAR_Close	0.0027
AUDSEK_Close	MXNPLN_Close	0.0027
EURNZD_Close	PLNZAR_Close	0.0027
AUDNOK_Close	SGDZAR_Close	0.0027
EURJPY_Close	HKDTRY_Close	0.0037
NZDCHF_Close	NOKSGD_Close	0.0038
NZDNOK_Close	HKDZAR_Close	0.0045
EURJPY_Close	SGDTRY_Close	0.0051
AUDNZD_Close	GBPCAD_Close	0.0054
EURSEK_Close	MXNNOK_Close	0.0054
CHFJPY_Close	EURTRY_Close	0.0054
EURJPY_Close	DKKTRY_Close	0.0055
EURSEK_Close	MXNSEK_Close	0.0065
AUDNZD_Close	HKDPLN_Close	0.0067
USDSEK_Close	NZDSGD_Close	0.0068
EURJPY_Close	MXNTRY_Close	0.0068
EURJPY_Close	EURTRY_Close	0.0070
MXNNOK_Close	MXNZAR_Close	0.0072
GBPNZD_Close	EURNOK_Close	0.0079
NZDSGD_Close	SEKSGD_Close	0.0081
AUDNZD_Close	USDPLN_Close	0.0087
USDNOK_Close	USDZAR_Close	0.0087
CHFJPY_Close	MXNTRY_Close	0.0092

Table 22 - APPENDIX I   COINTEGRATION ANALYSIS RESULTS

Strong Cointegration (0.01 ≤ p < 0.05)

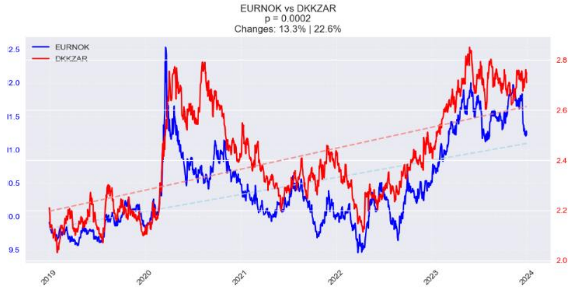
Currency Pair 1	Currency Pair 2	Cointegration p-value
EURJPY_Close	SEKTRY_Close	0.0105
CHFJPY_Close	NOKTRY_Close	0.0107
NZDNOK_Close	MXNZAR_Close	0.0111
EURSEK_Close	MXNZAR_Close	0.0115
EURJPY_Close	NOKTRY_Close	0.0116
USDZAR_Close	NOKSGD_Close	0.0116
SEKNOK_Close	NOKZAR_Close	0.0126
GBPNZD_Close	NOKDKK_Close	0.0128
CHFJPY_Close	DKKTRY_Close	0.0128
NZDCAD_Close	EURHKD_Close	0.0135
NZDCAD_Close	DKKHKD_Close	0.0145
USDNOK_Close	NZDSGD_Close	0.0146
AUDNZD_Close	AUDSEK_Close	0.0150
AUDNZD_Close	EURPLN_Close	0.0159
SEKDKK_Close	AUDSGD_Close	0.0165
AUDNZD_Close	DKKPLN_Close	0.0168
NZDCHF_Close	USDNOK_Close	0.0171
NZDCAD_Close	USDDKK_Close	0.0175
EURNOK_Close	HKDZAR_Close	0.0176
NZDCAD_Close	DKKSGD_Close	0.0178
GBPJPY_Close	NOKTRY_Close	0.0178
NZDSEK_Close	MXNSEK_Close	0.0180
NZDNOK_Close	PLNZAR_Close	0.0184
AUDNZD_Close	CADJPY_Close	0.0184
AUDNZD_Close	AUDJPY_Close	0.0195
AUDNZD_Close	HKDJPY_Close	0.0197
EURNOK_Close	MXNZAR_Close	0.0200
AUDNZD_Close	NZDJPY_Close	0.0201
GBPCHF_Close	NOKSGD_Close	0.0213
AUDCAD_Close	DKKSGD_Close	0.0213
NZDNOK_Close	NOKZAR_Close	0.0216
NZDNOK_Close	SEKTRY_Close	0.0218

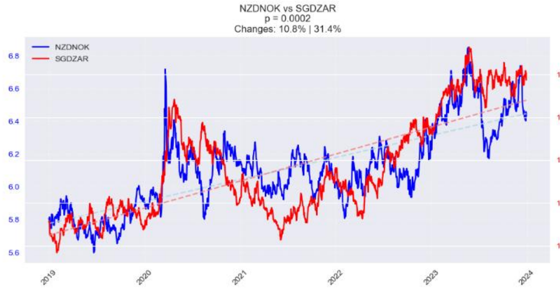
NZDNOK_Close	DKKTRY_Close	0.0219
SEKZAR_Close	NOKZAR_Close	0.0222
AUDNZD_Close	AUDNOK_Close	0.0238
GBPNZD_Close	MXNZAR_Close	0.0243
AUDCAD_Close	EURSGD_Close	0.0253
AUDNOK_Close	DKKZAR_Close	0.0253
NZDNOK_Close	MXNTRY_Close	0.0259
NZDNOK_Close	SGDTRY_Close	0.0263
GBPNZD_Close	MXNNOK_Close	0.0263
EURGBP_Close	NOKZAR_Close	0.0267
NZDNOK_Close	SEKZAR_Close	0.0270
NZDNOK_Close	NOKTRY_Close	0.0273
GBPJPY_Close	SEKTRY_Close	0.0277
NZDCAD_Close	USDSEK_Close	0.0283
EURGBP_Close	SEKZAR_Close	0.0287
NZDNOK_Close	HKDTRY_Close	0.0291
GBPCAD_Close	USDPLN_Close	0.0292
GBPNZD_Close	PLNZAR_Close	0.0293
CHFJPY_Close	SEKTRY_Close	0.0294
NZDCAD_Close	EURSGD_Close	0.0296
AUDCHF_Close	NOKSGD_Close	0.0298
AUDJPY_Close	NZDJPY_Close	0.0299
AUDNZD_Close	SGDJPY_Close	0.0301
GBPJPY_Close	DKKTRY_Close	0.0304
AUDCHF_Close	NZDNOK_Close	0.0305
GBPNZD_Close	NZDNOK_Close	0.0306
SEKDKK_Close	SGDMXN_Close	0.0307
EURGBP_Close	USDHKD_Close	0.0308
EURGBP_Close	TRYZAR_Close	0.0309
EURGBP_Close	PLNZAR_Close	0.0311
GBPUSD_Close	DKKHKD_Close	0.0316
GBPNZD_Close	USDMXN_Close	0.0319
GBPNZD_Close	MXNTRY_Close	0.0323
EURGBP_Close	AUDCAD_Close	0.0329
EURAUD_Close	AUDDKK_Close	0.0345
GBPSGD_Close	SEKSGD_Close	0.0347
GBPCHF_Close	NZDCHF_Close	0.0347
AUDSEK_Close	NZDSEK_Close	0.0357
AUDNOK_Close	NZDNOK_Close	0.0359
AUDNZD_Close	CHFJPY_Close	0.0361
GBPNZD_Close	NOKSGD_Close	0.0365
AUDNZD_Close	HKDZAR_Close	0.0366
GBPNZD_Close	DKKTRY_Close	0.0368
GBPNZD_Close	EURTRY_Close	0.0371
EURGBP_Close	EURCAD_Close	0.0371
GBPNZD_Close	SGDZAR_Close	0.0372
GBPJPY_Close	SGDTRY_Close	0.0374
GBPUSD_Close	EURHKD_Close	0.0375
EURJPY_Close	CHFJPY_Close	0.0378
GBPNZD_Close	SGDTRY_Close	0.0379
CADJPY_Close	AUDSEK_Close	0.0380
SEKDKK_Close	NZDSGD_Close	0.0382
USDSEK_Close	DKKSGD_Close	0.0383
EURGBP_Close	HKDJPY_Close	0.0385
GBPNZD_Close	MXNSEK_Close	0.0385
GBPNZD_Close	HKDTRY_Close	0.0389
GBPJPY_Close	EURTRY_Close	0.0392
EURGBP_Close	SGDJPY_Close	0.0393
GBPNZD_Close	SEKTRY_Close	0.0394
EURNZD_Close	DKKZAR_Close	0.0400
GBPNZD_Close	NOKTRY_Close	0.0407
EURGBP_Close	DKKZAR_Close	0.0413
EURGBP_Close	CADJPY_Close	0.0418
EURGBP_Close	NOKTRY_Close	0.0418
AUDNZD_Close	EURCAD_Close	0.0419
AUDNZD_Close	HKDTRY_Close	0.0423
GBPNZD_Close	AUDDKK_Close	0.0424
EURPLN_Close	AUDDKK_Close	0.0425
AUDNZD_Close	SGDZAR_Close	0.0428
AUDNZD_Close	NOKTRY_Close	0.0432
EURGBP_Close	SEKTRY_Close	0.0432
EURGBP_Close	DKKTRY_Close	0.0440
EURGBP_Close	HKDTRY_Close	0.0442
EURGBP_Close	NZDJPY_Close	0.0446
AUDNZD_Close	NZDSEK_Close	0.0446
EURGBP_Close	EURTRY_Close	0.0446
GBPAUD_Close	NOKPLN_Close	0.0447
EURGBP_Close	SGDTRY_Close	0.0455
EURGBP_Close	HKDZAR_Close	0.0458
NOKDKK_Close	NOKZAR_Close	0.0460
AUDNZD_Close	NZDNOK_Close	0.0462
GBPNZD_Close	EURSEK_Close	0.0463

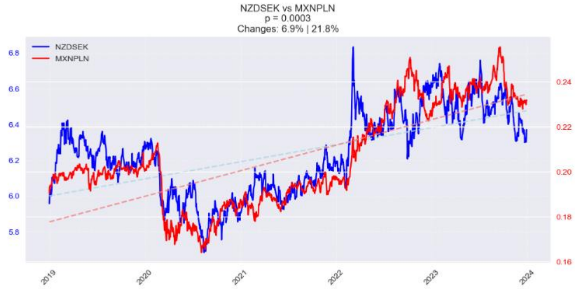
AUDNZD_Close	SEKTRY_Close	0.0464
EURSEK_Close	SEKDKK_Close	0.0466
EURGBP_Close	MXNTRY_Close	0.0466
GBPCHF_Close	USDNOK_Close	0.0468
AUDNZD_Close	SGDTRY_Close	0.0472
AUDNOK_Close	HKDZAR_Close	0.0

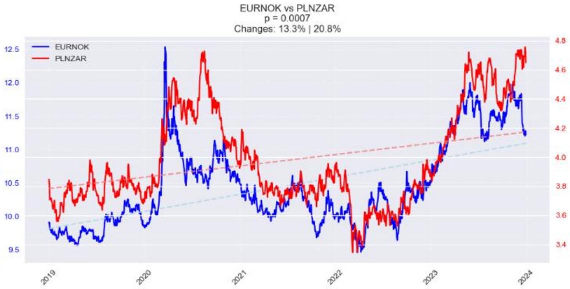
Table 23 - APPENDIX I   COINTEGRATION ANALYSIS RESULTS

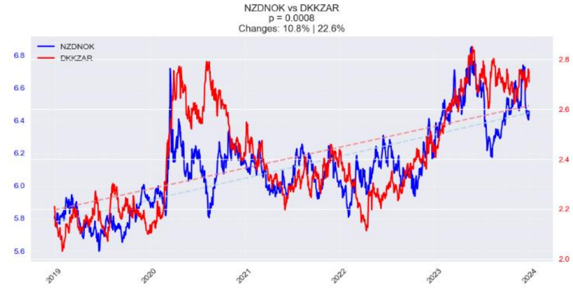
APPENDIX J Top Cointegration Pairs Trend

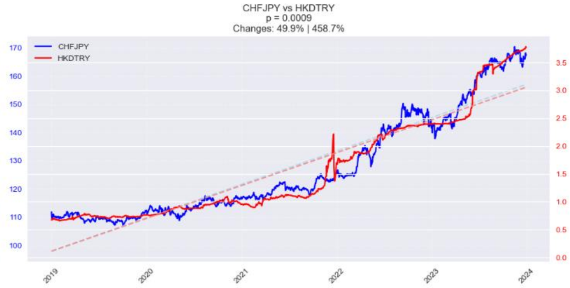


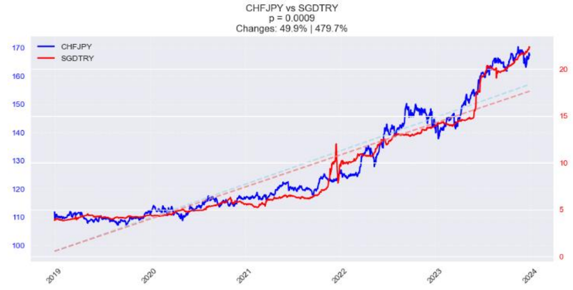


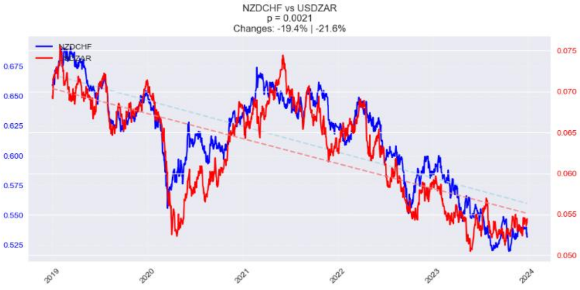


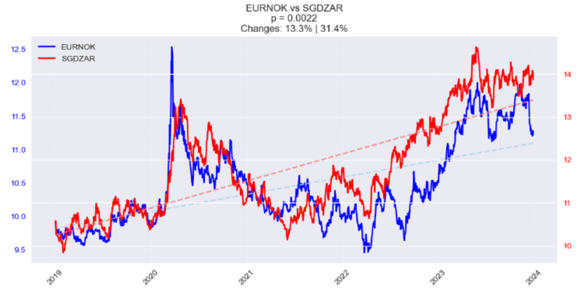


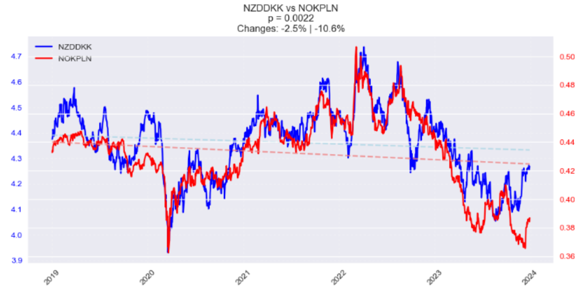












APPENDIX K   COINTEGRATION, HALFLIFE AND MEAN CROSSING

pair1	pair2	cointegration_p	half_life	mean_crossings	correlation
SEKNOK_Close	SEKZAR_Close	0.000	18.390	23.00	0.856
EURNOK_Close	DKKZAR_Close	0.000	20.863	13.33	0.898
NZDNOK_Close	SGDZAR_Close	0.000	19.956	17.39	0.826
NZDSEK_Close	MXNPLN_Close	0.000	15.194	26.28	0.883
EURNOK_Close	PLNZAR_Close	0.001	25.430	16.23	0.857
NZDNOK_Close	DKKZAR_Close	0.001	21.730	20.48	0.774
CHFJPY_Close	HKDTRY_Close	0.001	18.405	14.11	0.987
CHFJPY_Close	SGDTRY_Close	0.001	21.500	12.95	0.986
NZDCHF_Close	USDZAR_Close	0.002	23.072	16.43	0.926
EURNOK_Close	SGDZAR_Close	0.002	29.336	17.20	0.867
NZDDKK_Close	NOKPLN_Close	0.002	17.816	24.74	0.861
GBPCHF_Close	USDZAR_Close	0.003	23.970	20.48	0.887
AUDSEK_Close	MXNPLN_Close	0.003	20.759	17.39	0.902
EURNZD_Close	PLNZAR_Close	0.003	22.772	19.13	0.819
AUDNOK_Close	SGDZAR_Close	0.003	25.258	15.07	0.847
EURJPY_Close	HKDTRY_Close	0.004	26.324	16.43	0.960
NZDCHF_Close	NOKSGD_Close	0.004	23.806	20.29	0.930
NZDNOK_Close	HKDZAR_Close	0.005	25.073	13.91	0.773
EURJPY_Close	SGDTRY_Close	0.005	27.203	12.17	0.962
EURSEK_Close	MXNNOK_Close	0.005	25.822	18.55	0.916
CHFJPY_Close	EURTRY_Close	0.005	26.761	18.75	0.983
EURJPY_Close	DKKTRY_Close	0.006	27.365	17.59	0.962
EURSEK_Close	MXNSEK_Close	0.006	35.231	13.91	0.896
AUDNZD_Close	HKDPLN_Close	0.007	30.180	16.23	0.607
USDSEK_Close	NZDSGD_Close	0.007	26.866	19.33	0.936
EURJPY_Close	MXNTRY_Close	0.007	27.519	12.56	0.965
EURJPY_Close	EURTRY_Close	0.007	28.345	16.43	0.963
MXNNOK_Close	MXNZAR_Close	0.007	28.148	17.97	0.965
GBPNZD_Close	EURNOK_Close	0.008	30.227	15.85	0.646
NZDSGD_Close	SEKSGD_Close	0.008	29.376	18.17	0.926
USDNOK_Close	USDZAR_Close	0.009	28.685	19.52	0.920
CHFJPY_Close	MXNTRY_Close	0.009	28.389	9.86	0.983
EURJPY_Close	SEKTRY_Close	0.010	29.801	15.27	0.958
CHFJPY_Close	NOKTRY_Close	0.011	29.879	14.11	0.976
NZDNOK_Close	MXNZAR_Close	0.011	30.299	13.91	0.732
EURSEK_Close	MXNZAR_Close	0.012	28.495	14.30	0.902
EURJPY_Close	NOKTRY_Close	0.012	30.857	14.88	0.954
USDZAR_Close	NOKSGD_Close	0.012	30.179	17.78	0.909
SEKNOK_Close	NOKZAR_Close	0.013	19.841	17.20	0.448
GBPNZD_Close	NOKDKK_Close	0.013	89.577	8.31	(0.629)
CHFJPY_Close	DKKTRY_Close	0.013	25.223	20.29	0.982
NZDCAD_Close	EURHKD_Close	0.014	30.201	12.95	0.830
NZDCAD_Close	DKKHKD_Close	0.014	30.651	14.49	0.826
USDNOK_Close	NZDSGD_Close	0.015	27.329	18.55	0.924
AUDNZD_Close	AUDSEK_Close	0.015	33.527	13.33	0.679
AUDNZD_Close	EURPLN_Close	0.016	33.785	18.55	0.532
SEKDKK_Close	AUDSGD_Close	0.017	27.853	15.85	0.883
AUDNZD_Close	DKKPLN_Close	0.017	34.124	17.78	0.533
NZDCHF_Close	USDNOK_Close	0.017	27.070	19.13	0.917
NZDCAD_Close	USDDKK_Close	0.018	32.652	15.65	0.822
EURNOK_Close	HKDZAR_Close	0.018	33.085	14.11	0.851
NZDCAD_Close	DKKSGD_Close	0.018	36.084	11.60	0.770
GBPJPY_Close	NOKTRY_Close	0.018	36.353	11.40	0.946
NZDSEK_Close	MXNSEK_Close	0.018	28.922	17.01	0.827
NZDNOK_Close	PLNZAR_Close	0.018	35.428	14.30	0.611
AUDNZD_Close	CADJPY_Close	0.018	43.810	12.56	0.624
AUDNZD_Close	AUDJPY_Close	0.019	44.814	13.14	0.660
AUDNZD_Close	HKDJPY_Close	0.020	45.418	10.82	0.589
EURNOK_Close	MXNZAR_Close	0.020	47.189	8.89	0.759
AUDNZD_Close	NZDJPY_Close	0.020	41.931	14.30	0.521
GBPCHF_Close	NOKSGD_Close	0.021	30.590	10.82	0.870
AUDCAD_Close	DKKSGD_Close	0.021	40.005	11.98	0.664
NZDNOK_Close	NOKZAR_Close	0.022	20.319	17.97	0.341
NZDNOK_Close	SEKTRY_Close	0.022	32.857	13.91	0.730
NZDNOK_Close	DKKTRY_Close	0.022	33.053	15.07	0.737
SEKZAR_Close	NOKZAR_Close	0.022	21.971	12.17	0.844
AUDNZD_Close	AUDNOK_Close	0.024	35.807	11.02	0.618
GBPNZD_Close	MXNZAR_Close	0.024	40.099	13.91	0.570
AUDCAD_Close	EURSGD_Close	0.025	41.431	11.21	0.655
AUDNOK_Close	DKKZAR_Close	0.025	33.863	16.62	0.732
NZDNOK_Close	MXNTRY_Close	0.026	35.708	16.23	0.720

NZDNOK_Close	SGDTRY_Close	0.026	34.692	16.62	0.731
GBPNZD_Close	MXNNOK_Close	0.026	39.128	12.37	0.556
EURGBP_Close	NOKZAR_Close	0.027	27.680	15.65	0.227
NZDNOK_Close	SEKZAR_Close	0.027	34.039	10.63	0.543
NZDNOK_Close	NOKTRY_Close	0.027	34.642	14.30	0.691
GBPJPY_Close	SEKTRY_Close	0.028	38.961	10.24	0.942
NZDCAD_Close	USDSEK_Close	0.028	31.696	14.69	0.839
EURGBP_Close	SEKZAR_Close	0.029	43.739	13.72	0.340
NZDNOK_Close	HKDTRY_Close	0.029	35.120	14.69	0.726
GBPCAD_Close	USDPLN_Close	0.029	26.292	18.55	0.783
GBPNZD_Close	PLNZAR_Close	0.029	33.907	16.23	0.601
CHFJPY_Close	SEKTRY_Close	0.029	28.382	17.59	0.978
NZDCAD_Close	EURSGD_Close	0.030	36.190	11.98	0.772
AUDCHF_Close	NOKSGD_Close	0.030	32.805	12.37	0.893
AUDJPY_Close	NZDJPY_Close	0.030	36.221	17.39	0.985
AUDNZD_Close	SGDJPY_Close	0.030	52.679	9.66	0.563
GBPJPY_Close	DKKTRY_Close	0.030	40.296	13.33	0.940
GBPNZD_Close	NZDNOK_Close	0.031	26.476	15.65	0.386
SEKDKK_Close	SGDMXN_Close	0.031	38.286	11.21	0.650
EURGBP_Close	USDHKD_Close	0.031	47.628	12.37	0.249
EURGBP_Close	TRYZAR_Close	0.031	54.930	9.08	0.413
EURGBP_Close	PLNZAR_Close	0.031	51.600	17.20	0.371
GBPUSD_Close	DKKHKD_Close	0.032	39.039	10.24	0.888
GBPNZD_Close	USDMXN_Close	0.032	42.790	15.85	0.325
GBPNZD_Close	MXNTRY_Close	0.032	48.991	13.53	0.501
EURGBP_Close	AUDCAD_Close	0.033	36.449	9.47	0.227
EURAUD_Close	AUDDKK_Close	0.035	73.540	8.12	(0.997)
GBPSGD_Close	SEKSGD_Close	0.035	41.460	8.31	0.846
GBPCHF_Close	NZDCHF_Close	0.035	32.445	13.91	0.912
AUDSEK_Close	NZDSEK_Close	0.036	33.422	16.81	0.915
AUDNOK_Close	NZDNOK_Close	0.036	33.381	15.85	0.918
AUDNZD_Close	CHFJPY_Close	0.036	56.218	9.66	0.547
AUDNZD_Close	HKDZAR_Close	0.037	44.921	11.02	0.426
GBPNZD_Close	DKKTRY_Close	0.037	51.825	12.37	0.479
GBPNZD_Close	EURTRY_Close	0.037	51.595	11.98	0.480
EURGBP_Close	EURCAD_Close	0.037	40.568	11.60	0.578
GBPNZD_Close	SGDZAR_Close	0.037	41.750	11.98	0.541
GBPJPY_Close	SGDTRY_Close	0.037	42.666	8.31	0.939
GBPUSD_Close	EURHKD_Close	0.037	40.742	13.33	0.885
EURJPY_Close	CHFJPY_Close	0.038	39.747	14.88	0.981
GBPNZD_Close	SGDTRY_Close	0.038	52.266	13.14	0.468
CADJPY_Close	AUDSEK_Close	0.038	29.858	16.23	0.861
SEKDKK_Close	NZDSGD_Close	0.038	20.317	18.17	0.917
USDSEK_Close	DKKSGD_Close	0.038	38.865	9.28	0.918
GBPNZD_Close	MXNSEK_Close	0.038	44.817	10.05	0.457
GBPNZD_Close	HKDTRY_Close	0.039	52.227	13.14	0.469
GBPJPY_Close	EURTRY_Close	0.039	42.572	14.11	0.940
GBPNZD_Close	SEKTRY_Close	0.039	54.481	14.30	0.443
EURNZD_Close	DKKZAR_Close	0.040	39.396	11.79	0.697
GBPNZD_Close	NOKTRY_Close	0.041	54.824	13.91	0.421
EURGBP_Close	DKKZAR_Close	0.041	56.013	11.02	0.285
AUDNZD_Close	HKDTRY_Close	0.042	60.201	11.02	0.534
GBPNZD_Close	AUDDKK_Close	0.042	63.553	10.24	(0.562)
EURPLN_Close	AUDDKK_Close	0.043	37.748	11.98	0.488
AUDNZD_Close	SGDZAR_Close	0.043	50.834	12.95	0.427
AUDNZD_Close	NOKTRY_Close	0.043	60.741	8.70	0.544
AUDNZD_Close	NZDSEK_Close	0.045	32.471	12.17	0.326
EURGBP_Close	HKDZAR_Close	0.046	62.291	8.70	0.150
AUDNZD_Close	NZDNOK_Close	0.046	34.193	12.95	0.257
GBPNZD_Close	EURSEK_Close	0.046	32.699	17.39	0.616
AUDNZD_Close	SEKTRY_Close	0.046	63.538	8.31	0.529
GBPCHF_Close	USDNOK_Close	0.047	32.559	15.46	0.855
AUDNZD_Close	SGDTRY_Close	0.047	63.035	10.63	0.524
AUDNOK_Close	HKDZAR_Close	0.048	29.559	13.14	0.802
AUDNZD_Close	DKKTRY_Close	0.049	64.719	8.70	0.511
EURCAD_Close	GBPCAD_Close	0.049	35.927	10.63	0.813
USDCHF_Close	NZDNOK_Close	0.050	28.838	10.05	0.632
EURGBP_Close	SGDHKD_Close	0.050	56.236	9.08	(0.198)
GBPSGD_Close	DKKSGD_Close	0.050	42.349	10.63	0.837
EURUSD_Close	NZDCAD_Close	0.050	32.066	14.11	0.825
AUDNZD_Close	EURTRY_Close	0.050	64.645	8.70	0.511
NZDUSD_Close	GBPSGD_Close	0.051	41.272	16.23	0.811
AUDNZD_Close	MXNPLN_Close	0.051	50.164	9.66	0.506
EURGBP_Close	MXNZAR_Close	0.051	86.362	11.02	(0.034)
EURGBP_Close	NZDNOK_Close	0.052	50.094	9.86	(0.013)
EURGBP_Close	EURSEK_Close	0.052	75.915	8.31	0.031
EURGBP_Close	HKDPLN_Close	0.052	72.922	11.21	(0.226)
EURGBP_Close	GBPCAD_Close	0.052	35.277	9.47	(0.004)

GBPJPY_Close	HKDTRY_Close	0.052	38.827	11.02	0.937
EURGBP_Close	AUDNOK_Close	0.052	57.813	11.40	(0.034)
GBPUSD_Close	USDDKK_Close	0.053	44.138	11.02	0.884
GBPCAD_Close	EURHKD_Close	0.053	34.277	15.46	0.755
EURGBP_Close	USDSGD_Close	0.054	52.063	10.63	(0.138)
USDSEK_Close	GBPSGD_Close	0.055	37.343	10.05	0.858
USDSEK_Close	EURSGD_Close	0.056	42.390	11.98	0.912
GBPNZD_Close	SGDJPY_Close	0.056	53.724	13.53	0.378
AUDCAD_Close	NZDCAD_Close	0.056	44.544	15.65	0.864
AUDCAD_Close	SGDMXN_Close	0.057	46.825	13.53	0.433
AUDNZD_Close	MXNTRY_Close	0.058	69.038	8.89	0.494
EURGBP_Close	USDDKK_Close	0.058	62.922	10.05	0.195
GBPNZD_Close	MXNPLN_Close	0.058	53.662	10.05	0.272
EURGBP_Close	EURHKD_Close	0.059	62.873	10.44	0.190
AUDNZD_Close	NZDCAD_Close	0.059	76.977	11.02	(0.604)
AUDNOK_Close	NOKZAR_Close	0.059	22.497	17.01	0.381
EURGBP_Close	NZDSEK_Close	0.059	58.979	9.08	(0.394)
GBPNZD_Close	AUDNOK_Close	0.059	39.085	10.63	0.338
NZDCAD_Close	SEKSGD_Close	0.060	39.513	10.82	0.790
EURGBP_Close	DKKHKD_Close	0.060	63.394	9.66	0.183
DKKZAR_Close	NOKZAR_Close	0.060	25.661	12.17	0.759
EURGBP_Close	SGDMXN_Close	0.060	57.865	8.50	0.232
EURNZD_Close	EURNOK_Close	0.060	33.282	13.33	0.776
AUDCAD_Close	DKKHKD_Close	0.061	36.538	8.31	0.755
GBPNZD_Close	HKDJPY_Close	0.061	52.611	14.69	0.395
EURGBP_Close	EURSGD_Close	0.061	71.000	11.60	0.299
EURGBP_Close	AUDDKK_Close	0.061	55.194	9.28	(0.459)
GBPCAD_Close	USDDKK_Close	0.061	35.828	11.98	0.751
EURGBP_Close	DKKSGD_Close	0.062	71.109	11.98	0.291
NZDCAD_Close	USDNOK_Close	0.063	36.117	11.21	0.780
GBPAUD_Close	EURAUD_Close	0.063	40.382	9.08	0.779
NZDSEK_Close	MXNNOK_Close	0.064	38.856	11.98	0.713
AUDDKK_Close	NOKPLN_Close	0.064	32.554	19.52	0.810
EURGBP_Close	EURDKK_Close	0.064	56.099	8.89	0.166
NZDDKK_Close	NOKZAR_Close	0.065	54.721	14.11	(0.338)
AUDUSD_Close	NZDHKD_Close	0.065	45.528	11.40	0.946
SGDJPY_Close	MXNPLN_Close	0.066	36.806	17.39	0.926
EURGBP_Close	USDMXN_Close	0.066	77.231	12.17	(0.227)
GBPNZD_Close	CADJPY_Close	0.066	54.771	14.69	0.304
AUDNZD_Close	MXNZAR_Close	0.066	68.213	11.60	0.390
EURNZD_Close	EURAUD_Close	0.066	46.711	14.49	0.850
AUDCAD_Close	USDSEK_Close	0.067	36.087	17.39	0.786
AUDCAD_Close	USDDKK_Close	0.067	38.732	10.24	0.749
AUDCAD_Close	SEKSGD_Close	0.067	41.479	13.14	0.722
HKDJPY_Close	AUDSEK_Close	0.068	29.783	12.75	0.853
AUDCHF_Close	USDZAR_Close	0.068	37.610	12.75	0.872
USDHKD_Close	DKKSGD_Close	0.069	48.665	10.24	0.752
EURGBP_Close	SEKNOK_Close	0.069	46.297	11.02	0.341
NZDCAD_Close	GBPSGD_Close	0.070	43.375	8.50	0.658
USDHKD_Close	EURSGD_Close	0.070	48.799	13.33	0.746
GBPHKD_Close	DKKHKD_Close	0.070	43.134	10.24	0.871
USDJPY_Close	EURCHF_Close	0.071	44.133	16.04	0.900
NZDDKK_Close	SEKPLN_Close	0.072	41.571	15.07	0.599
AUDSGD_Close	SEKPLN_Close	0.072	41.810	10.82	0.770
NZDUSD_Close	USDSEK_Close	0.072	35.564	14.88	0.925
GBPNZD_Close	NZDSEK_Close	0.073	34.348	14.69	0.110
GBPNZD_Close	EURDKK_Close	0.073	45.628	16.81	0.279
NZDNOK_Close	MXNNOK_Close	0.073	49.695	10.44	0.677
EURNOK_Close	NOKZAR_Close	0.073	23.382	15.07	0.397
EURGBP_Close	EURNOK_Close	0.074	61.201	18.36	0.256
USDZAR_Close	NOKDKK_Close	0.074	41.040	11.79	0.831
EURNZD_Close	GBPNZD_Close	0.074	40.318	9.08	0.702
AUDJPY_Close	CADJPY_Close	0.074	46.241	15.85	0.966
AUDNZD_Close	DKKZAR_Close	0.076	56.725	9.86	0.257
HKDJPY_Close	MXNPLN_Close	0.076	40.250	16.23	0.916
GBPNZD_Close	AUDHKD_Close	0.077	74.998	11.40	(0.567)
NZDUSD_Close	DKKSGD_Close	0.077	54.529	11.40	0.755
EURJPY_Close	GBPJPY_Close	0.077	37.359	11.60	0.974
AUDUSD_Close	SEKDKK_Close	0.077	46.570	9.08	0.781
GBPJPY_Close	NZDJPY_Close	0.077	42.361	13.33	0.953
NZDUSD_Close	GBPHKD_Close	0.078	38.561	13.33	0.877
AUDNZD_Close	NOKZAR_Close	0.078	30.710	18.94	0.262
EURCHF_Close	GBPCHF_Close	0.080	43.361	8.70	0.891
NZDJPY_Close	NOKTRY_Close	0.080	47.264	13.33	0.903
AUDUSD_Close	NZDUSD_Close	0.081	48.563	9.86	0.945
AUDCHF_Close	NZDCHF_Close	0.081	46.619	13.14	0.955
EURAUD_Close	GBPCAD_Close	0.082	38.245	16.43	0.522
EURGBP_Close	GBPNZD_Close	0.084	38.549	12.95	(0.274)

GBPUSD_Close	USDSEK_Close	0.084	52.927	9.86	0.851
NZDUSD_Close	EURSGD_Close	0.085	56.951	8.31	0.745
EURNZD_Close	NOKZAR_Close	0.086	29.042	15.27	0.318
AUDHKD_Close	NZDHKD_Close	0.086	49.162	11.02	0.941
GBPNZD_Close	NOKZAR_Close	0.086	33.474	11.98	0.163
GBPNZD_Close	AUDSEK_Close	0.086	47.332	13.91	0.112
AUDNZD_Close	AUDDKK_Close	0.087	48.748	10.63	0.514
AUDSGD_Close	NZDSGD_Close	0.088	52.128	13.72	0.924
GBPNZD_Close	SEKZAR_Close	0.089	54.100	11.21	0.247
NZDJPY_Close	CADJPY_Close	0.089	48.549	8.12	0.953
GBPAUD_Close	AUDDKK_Close	0.089	74.769	12.17	(0.773)
AUDNZD_Close	AUDCAD_Close	0.090	85.574	11.79	(0.121)
GBPNZD_Close	SEKNOK_Close	0.090	49.971	13.72	0.261
GBPHKD_Close	NZDHKD_Close	0.090	39.879	9.47	0.869
USDCAD_Close	SEKPLN_Close	0.090	43.488	12.56	0.646
NZDCAD_Close	USDPLN_Close	0.090	39.237	12.17	0.770
EURGBP_Close	GBPSGD_Close	0.090	51.584	8.89	(0.278)
CADJPY_Close	MXNPLN_Close	0.091	43.320	13.33	0.894
AUDUSD_Close	USDSEK_Close	0.091	54.307	12.56	0.860
EURHKD_Close	GBPHKD_Close	0.092	44.807	11.02	0.868
AUDNZD_Close	EURNOK_Close	0.092	56.820	10.05	0.178
EURAUD_Close	PLNZAR_Close	0.093	52.869	11.79	0.669
EURSEK_Close	MXNPLN_Close	0.093	46.990	10.82	0.749
AUDCAD_Close	HKDMXN_Close	0.093	52.719	13.14	0.274
EURNZD_Close	SGDZAR_Close	0.094	49.761	8.31	0.576
GBPNZD_Close	HKDZAR_Close	0.096	39.232	10.44	0.548
USDCAD_Close	NZDDKK_Close	0.097	41.530	9.86	0.620
AUDNZD_Close	MXNNOK_Close	0.097	70.270	8.89	0.344
EURJPY_Close	AUDNOK_Close	0.097	31.281	11.98	0.783
EURNZD_Close	MXNZAR_Close	0.098	59.015	11.40	0.489
AUDNOK_Close	SEKTRY_Close	0.098	39.957	10.05	0.812
GBPNZD_Close	USDHKD_Close	0.099	70.634	9.08	0.013
AUDDKK_Close	DKKPLN_Close	0.099	38.000	11.98	0.496
EURJPY_Close	NZDNOK_Close	0.100	28.082	11.98	0.701
DKKTRY_Close	SGDTRY_Close	0.461	9.469	8.70	0.998
EURHKD_Close	DKKHKD_Close	0.557	6.213	8.12	0.999
EURSGD_Close	DKKSGD_Close	0.776	5.556	18.94	0.999

Table 24 – Currencies Pairs with Cointegration , Half Life and Mean Crossing

APPENDIX L MACHINE LEARNING TRADING PERFOMANCE

All Currency Pairs and Models

Currency Pair	Model	Type	Sharpe Ratio	Return (%)	Volatility (%)	Max Drawdown (%)
EURNOK_DKKZAR	Logistic	Primary	0.01	0.06	10.36	-19.65
EURNOK_DKKZAR	Logistic	Meta-Labeled	1.86	4.05	2.18	-2.02
EURNOK_DKKZAR	Random Forest	Primary	-0.28	-2.90	10.36	-18.82
EURNOK_DKKZAR	Random Forest	Meta-Labeled	1.01	2.64	2.62	-1.31
EURNOK_DKKZAR	Gradient Boosting	Primary	0.78	8.11	10.35	-11.80
EURNOK_DKKZAR	Gradient Boosting	Meta-Labeled	1.63	6.52	4.01	-1.76
EURPLN_DKKPLN	Logistic	Meta-Labeled	1.57	2.49	1.59	-0.69
EURSGD_DKKSGD	Logistic	Primary	0.78	4.09	5.28	-3.07
NZDCHF_USDZAR	Gradient Boosting	Meta-Labeled	-0.41	-3.65	8.95	-13.30
SEKNOK_SEKZAR	Logistic	Primary	1.54	24.88	16.14	-9.46

Best Model by Category

Category	Currency Pair	Model	Type	Value
Highest Sharpe Ratio	EURNOK_DKKZAR	Logistic	Meta-Labeled	1.86
Highest Return	SEKNOK_SEKZAR	Logistic	Primary	24.88%
Lowest Volatility	EURPLN_DKKPLN	Logistic	Meta-Labeled	1.59%
Smallest Max Drawdown	EURPLN_DKKPLN	Logistic	Meta-Labeled	-0.69%

APPENDIX M   CURRENCIES TRADING PERFOMANCE

```
********************************************************************************************************************
********************************************************************************************************************
*                                                            *
*            PAIR 1/5: EURSGD_Close vs DKKSGD_Close (HEDGE RATIO: 0.141)            *
*                                                            *
********************************************************************************************************************
********************************************************************************************************************
```

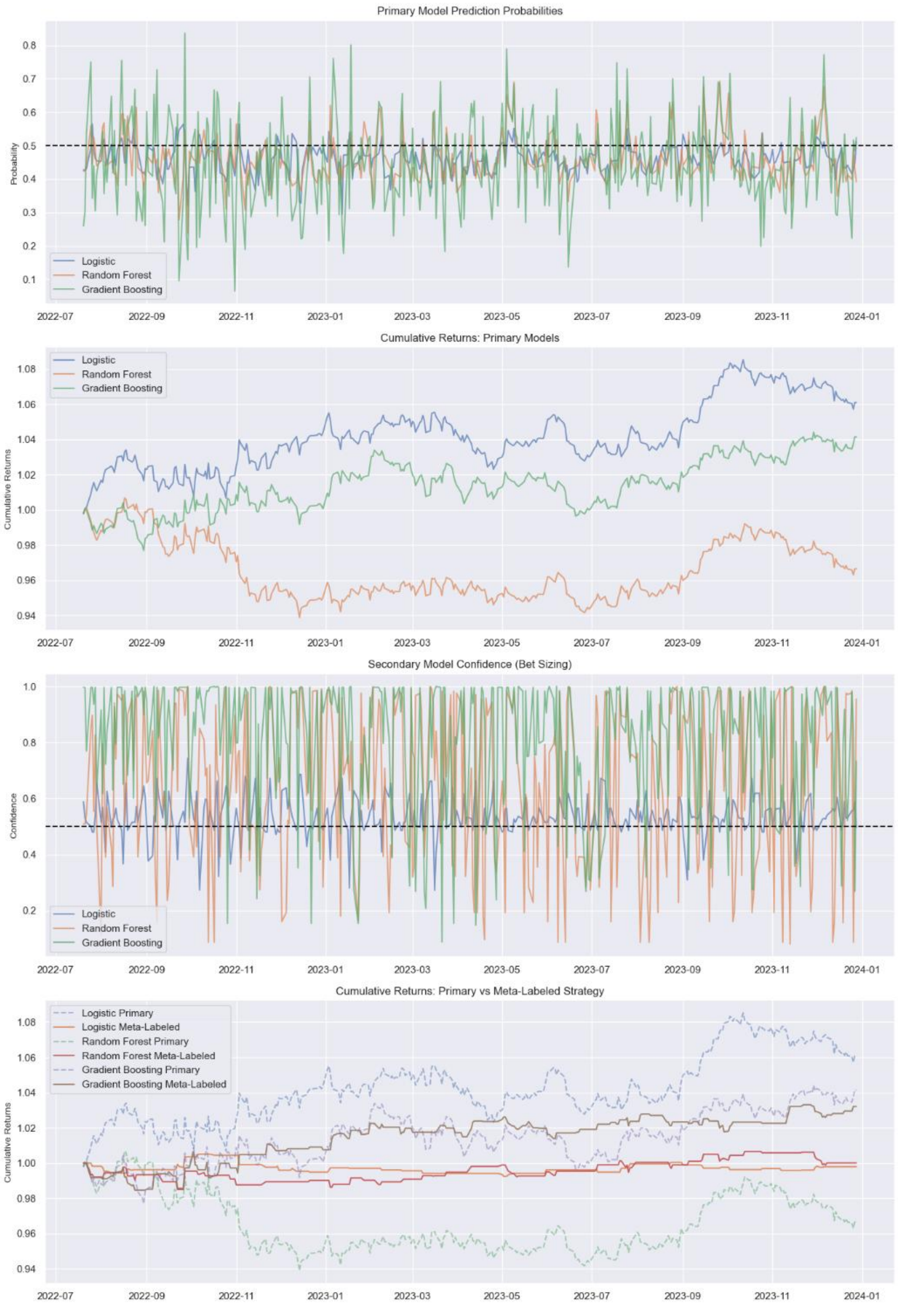
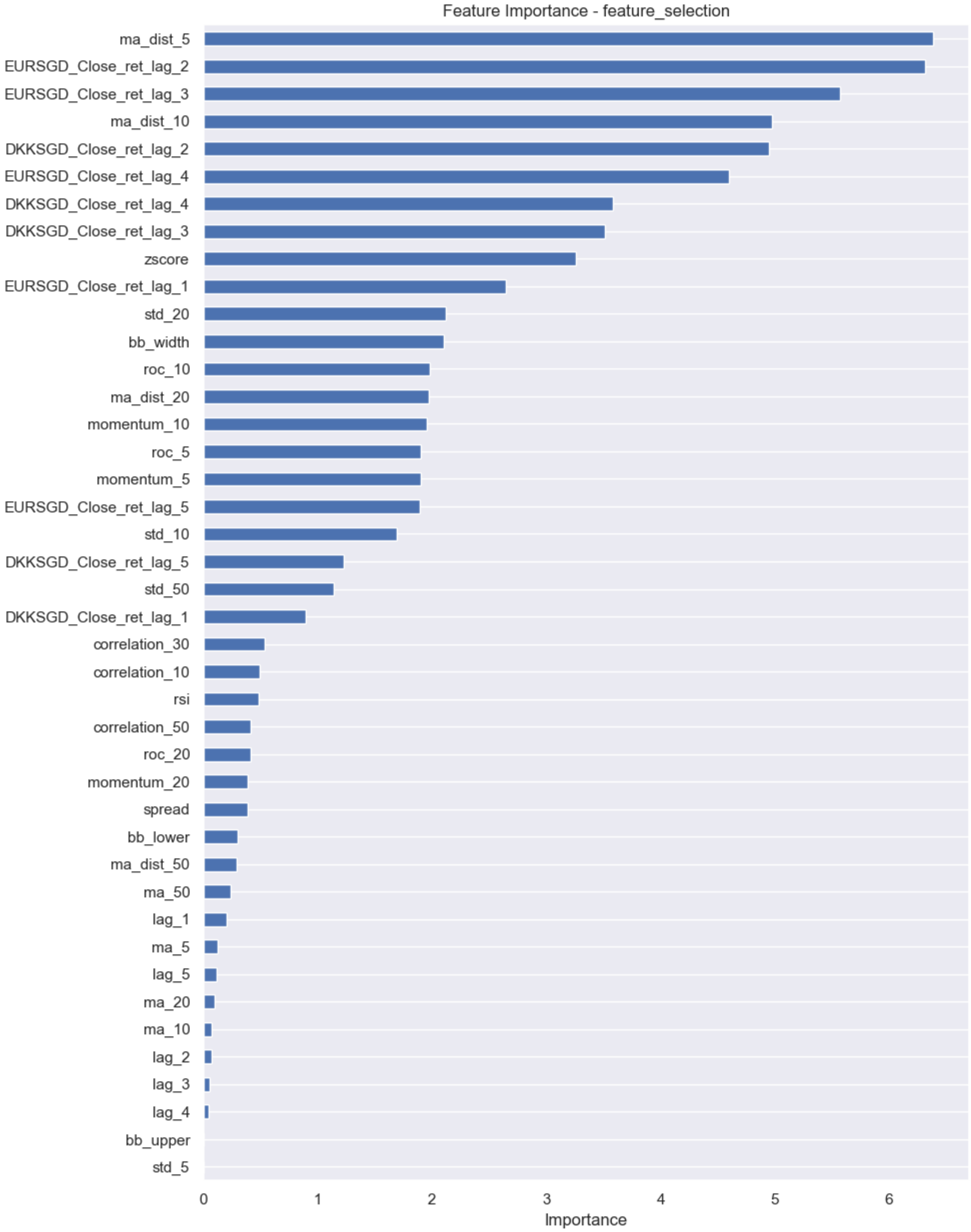
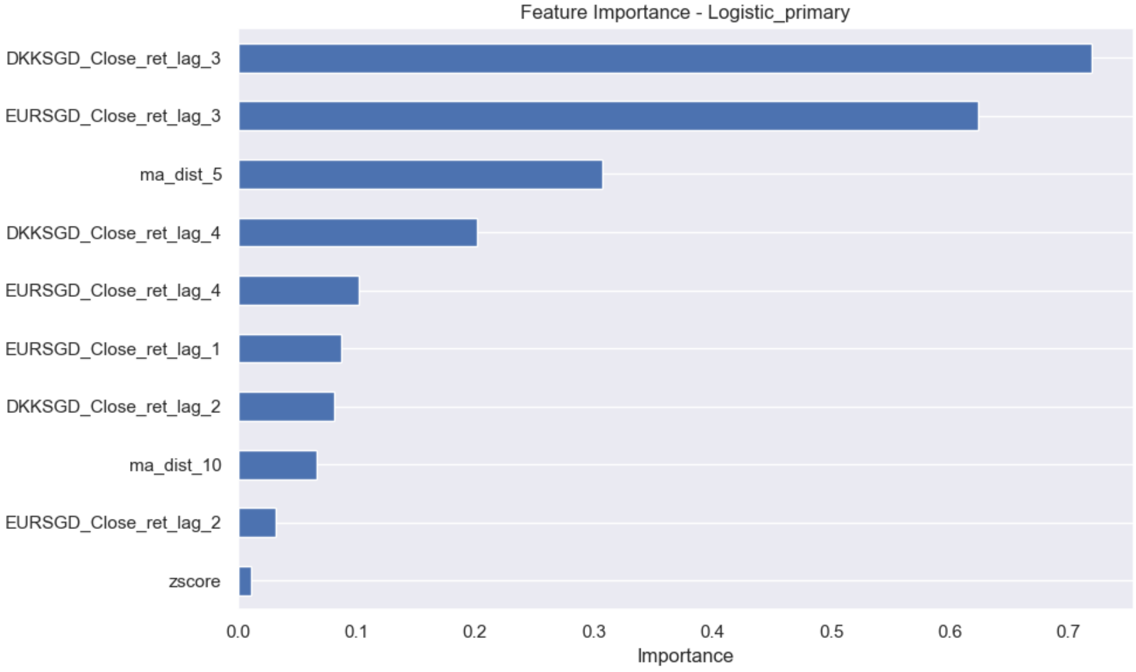
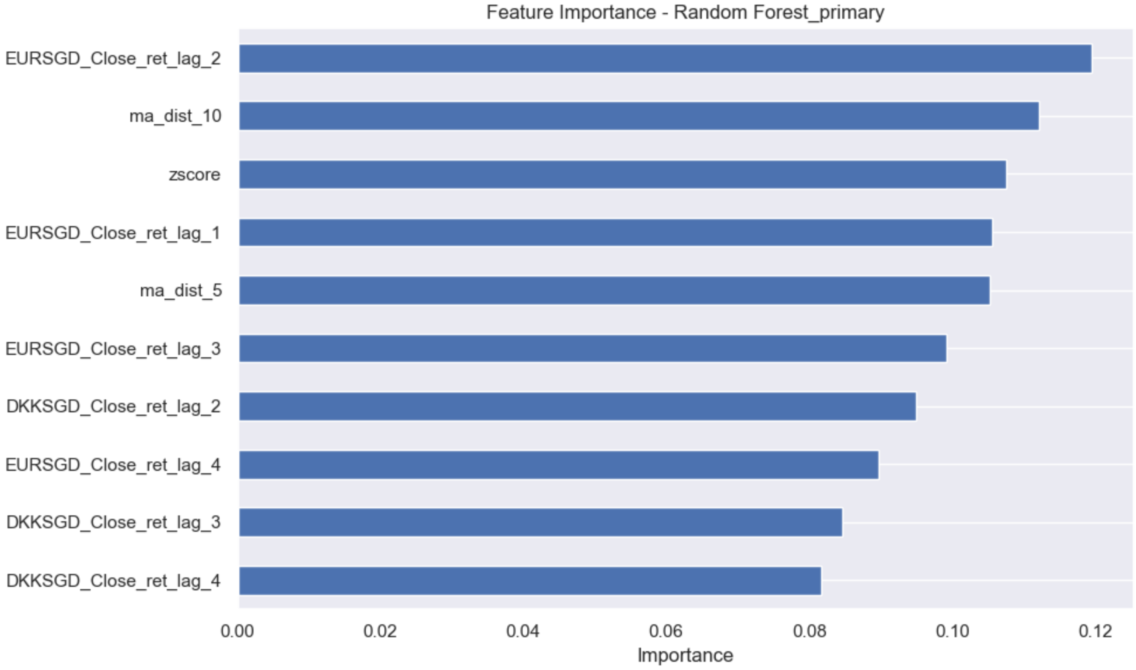
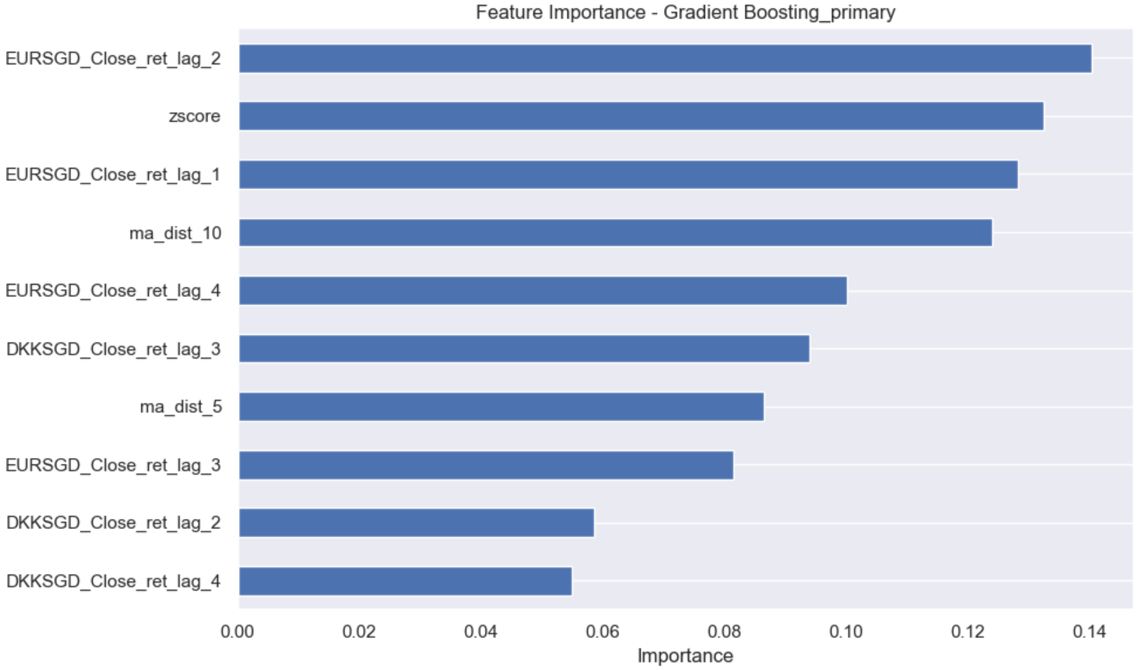


Figure 9 - PAIR 1/5: EURSGD_Close vs DKKSGD_Close


Feature Importance - feature_selection
ma_dist_5
EURSGD_Close_ret_lag_2
EURSGD_Close_ret_lag_3
ma_dist_10
DKKSGD_Close_ret_lag_2
EURSGD_Close_ret_lag_4
DKKSGD_Close_ret_lag_4
DKKSGD_Close_ret_lag_3
zscore
EURSGD_Close_ret_lag_1
std_20
bb_width
roc_10
ma_dist_20
momentum_10
roc_5
momentum_5
EURSGD_Close_ret_lag_5
std_10
DKKSGD_Close_ret_lag_5
std_50
DKKSGD_Close_ret_lag_1
correlation_30
correlation_10
rsi
correlation_50
roc_20
momentum_20
spread
bb_lower
ma_dist_50
ma_50
lag_1
ma_5
lag_5
ma_20
ma_10
lag_2
lag_3
lag_4
bb_upper
std_5
0
1
2
3
4
5
6
Importance


Feature Importance - Logistic_primary
DKKSGD_Close_ret_lag_3
EURSGD_Close_ret_lag_3
ma_dist_5
DKKSGD_Close_ret_lag_4
EURSGD_Close_ret_lag_4
EURSGD_Close_ret_lag_1
DKKSGD_Close_ret_lag_2
ma_dist_10
EURSGD_Close_ret_lag_2
zscore
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Importance


Feature Importance - Random Forest_primary
EURSGD_Close_ret_lag_2
ma_dist_10
zscore
EURSGD_Close_ret_lag_1
ma_dist_5
EURSGD_Close_ret_lag_3
DKKSGD_Close_ret_lag_2
EURSGD_Close_ret_lag_4
DKKSGD_Close_ret_lag_3
DKKSGD_Close_ret_lag_4
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Importance



```
--------------------------------------------------
PERFORMANCE METRICS FOR EURSGD_Close vs DKKSGD_Close
--------------------------------------------------


===== PERFORMANCE SUMMARY =====
Currency Pairs: EURSGD_Close and DKKSGD_Close
Hedge Ratio: 0.141
Feature Selection: Enabled
Meta-Labeling: Enabled

Classification Metrics:
--------------------------------------------------
       model_name model_type  class  precision    recall  f1-score  support
         Logistic    primary      0   0.493789  0.893258  0.636000      178
         Logistic    primary      1   0.654545  0.180905  0.283465      199
         Logistic  secondary      0   0.430108  0.219780  0.290909      182
         Logistic  secondary      1   0.500000  0.728205  0.592902      195
         Logistic   combined      0   0.481894  0.971910  0.644320      178
         Logistic   combined      1   0.722222  0.065327  0.119816      199
    Random Forest    primary      0   0.473154  0.792135  0.592437      178
    Random Forest    primary      1   0.531646  0.211055  0.302158      199
    Random Forest  secondary      0   0.517857  0.298969  0.379085      194
    Random Forest  secondary      1   0.486792  0.704918  0.575893      183
    Random Forest   combined      0   0.473154  0.792135  0.592437      178
    Random Forest   combined      1   0.531646  0.211055  0.302158      199
Gradient Boosting    primary      0   0.490040  0.691011  0.573427      178
Gradient Boosting    primary      1   0.563492  0.356784  0.436923      199
Gradient Boosting  secondary      0   0.596154  0.169399  0.263830      183
Gradient Boosting  secondary    1  0.532308 0.891753  0.666667    194
Gradient Boosting   combined    0  0.488189 0.696629  0.574074    178
Gradient Boosting   combined    1  0.560976 0.346734  0.428571    199


Model Performance Metrics:
            Primary Model  Secondary Model  Combined (Meta-Labeled)
Logistic            0.517241        0.517241               0.517241
Random Forest        0.485411        0.485411               0.485411
Gradient Boosting     0.514589        0.514589               0.514589

Strategy Metrics:
--------------------------------------------------
     model_name   model_type  annualized_return  annualized_volatility  sharpe_ratio  max_drawdown
      Logistic     primary          0.040919            0.052807      0.775902     -0.030690
      Logistic meta_labeled         -0.001325            0.009729     -0.136358    -0.012586
   Random Forest     primary         -0.021310            0.052853     -0.403739    -0.067449
   Random Forest meta_labeled          0.000209            0.018281      0.011463    -0.015029
Gradient Boosting     primary          0.028522            0.052839      0.540513    -0.036181
Gradient Boosting meta_labeled          0.021558            0.030008      0.719340    -0.015390

Filtered Strategy Metrics (Only Model Name and Sharpe Ratio):
--------------------------------------------------
     model_name  sharpe_ratio
      Logistic     0.775902
      Logistic    -0.136358
   Random Forest    -0.403739
   Random Forest     0.011463
Gradient Boosting     0.540513
Gradient Boosting     0.719340

Summary Statistics for Strategy Metrics:
--------------------------------------------------
     annualized_return  annualized_volatility  sharpe_ratio  max_drawdown
count         6.000000             6.000000      6.000000      6.000000
mean          0.011429             0.036086      0.251186     -0.029554
std           0.022912             0.019443      0.492908      0.020884
min          -0.021310             0.009729     -0.403739     -0.067449
25%          -0.000941             0.021213     -0.099403     -0.034808
50%           0.010883             0.041408      0.275988     -0.023040
75%           0.026781             0.052831      0.674633     -0.015119
max           0.040919             0.052853      0.775902     -0.012586
```

```
****************************************************************************************************
****************************************************************************************************
*                                                    *
*          PAIR 2/5: EURPLN_Close vs DKKPLN_Close (HEDGE RATIO: 0.1391)          *
*                                                    *
****************************************************************************************************
****************************************************************************************************
```

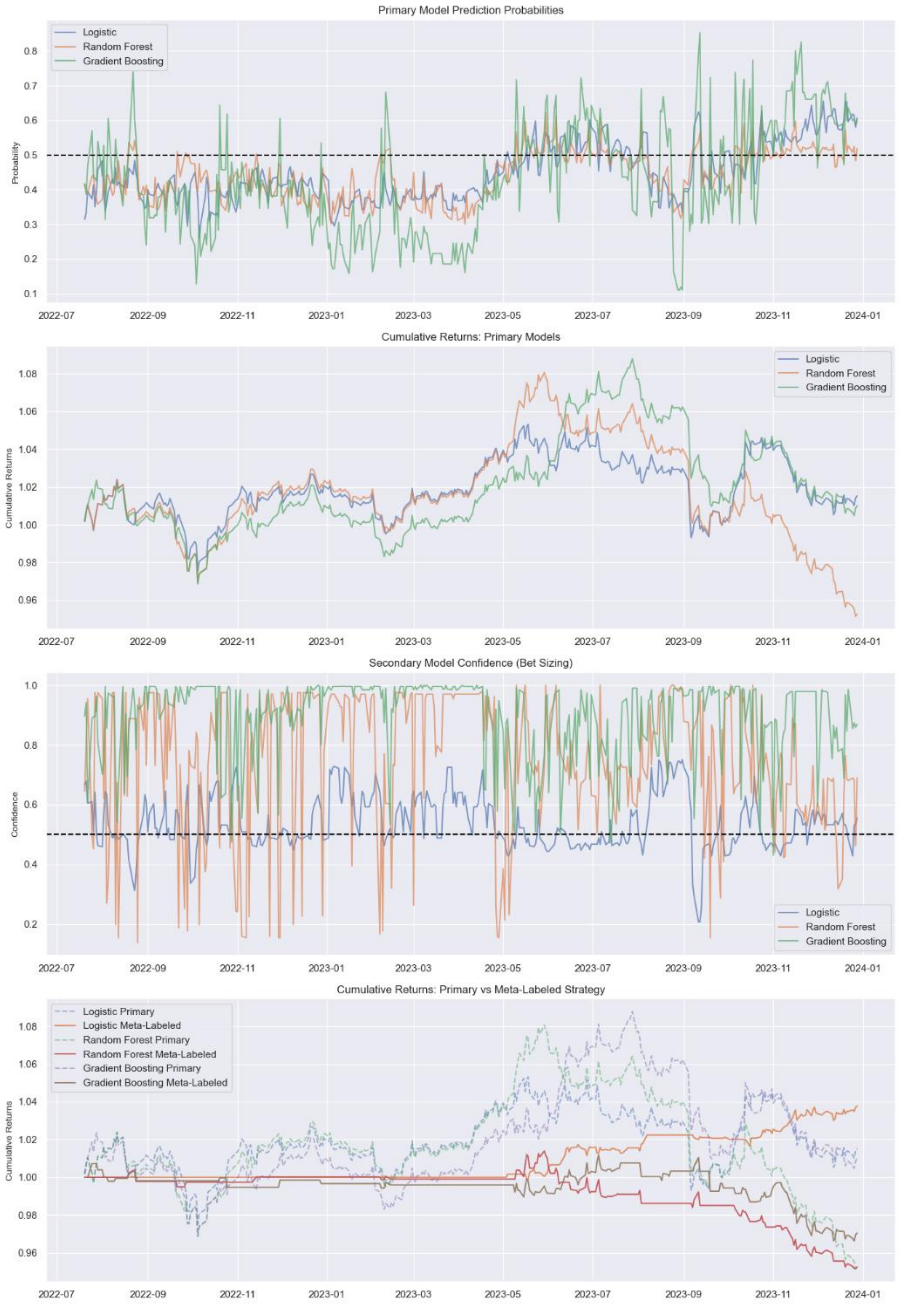
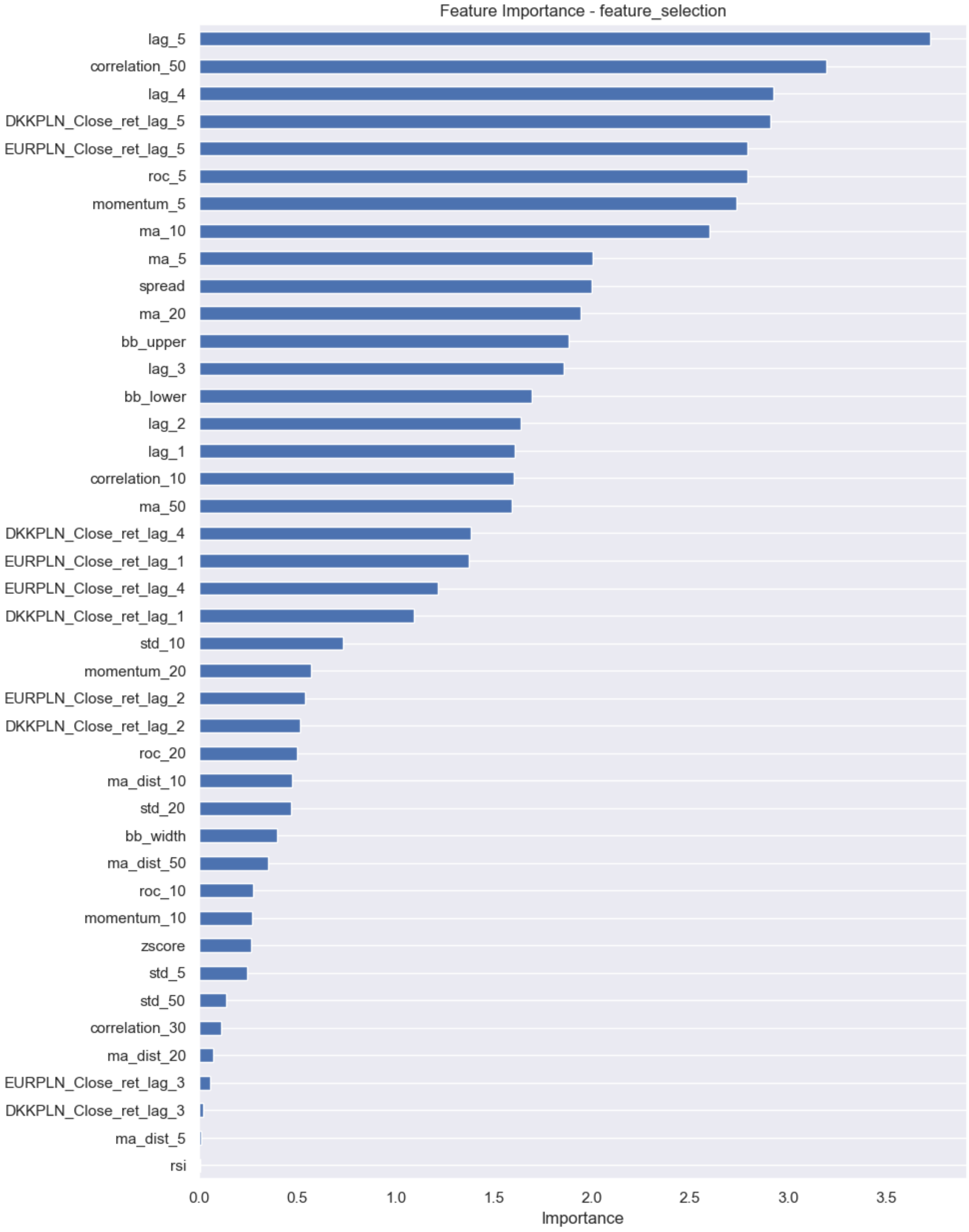
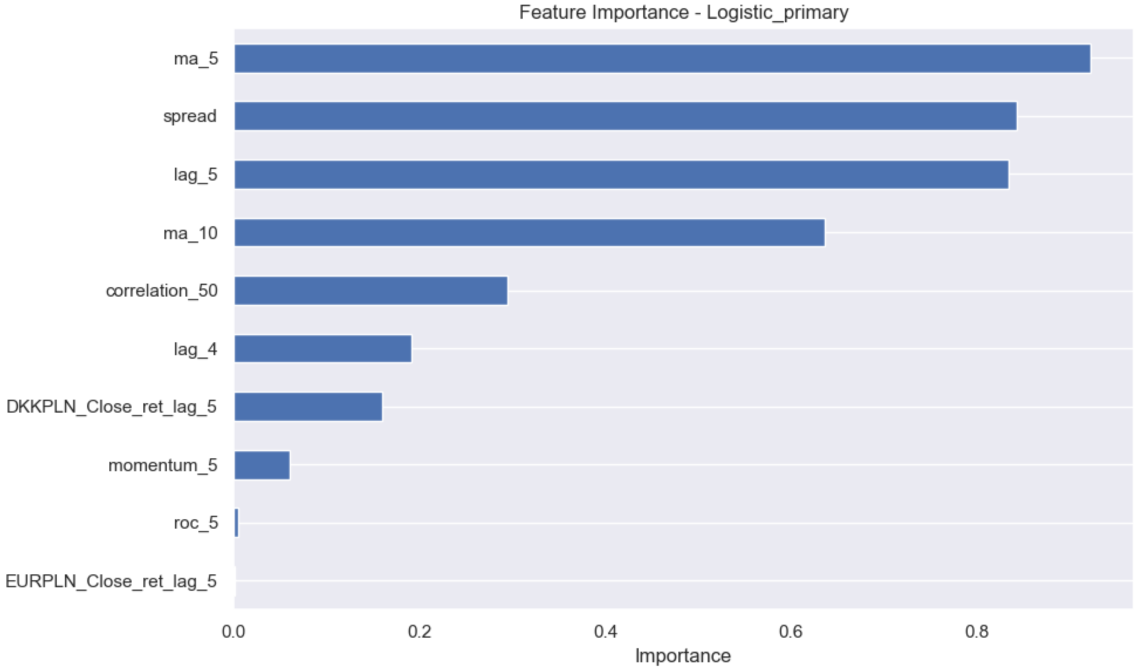
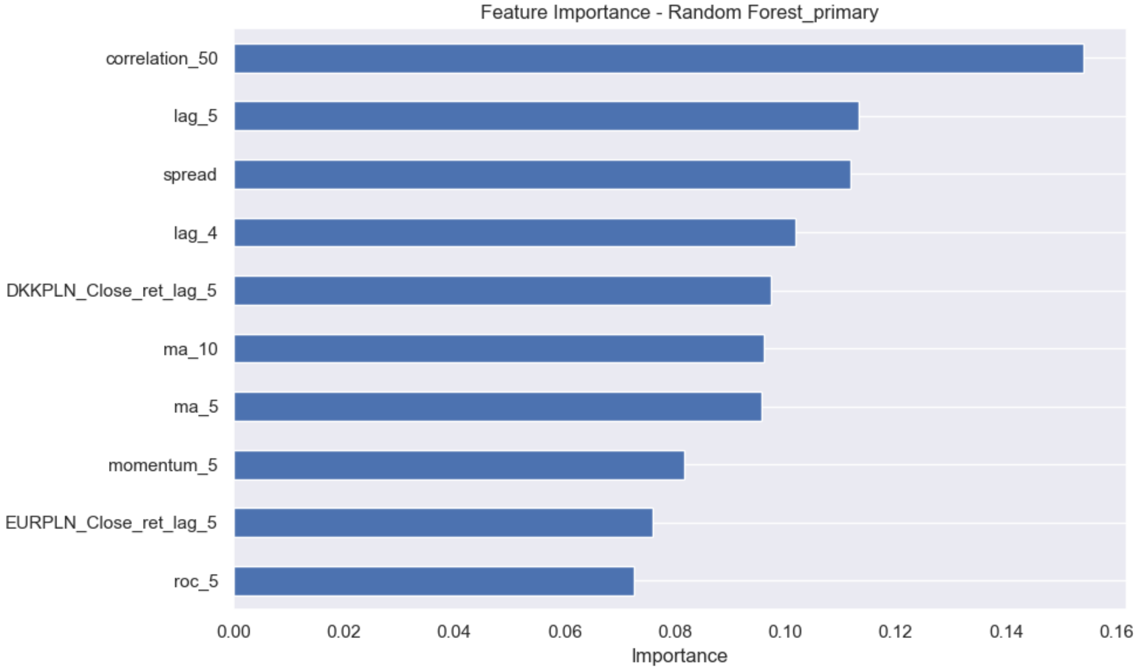
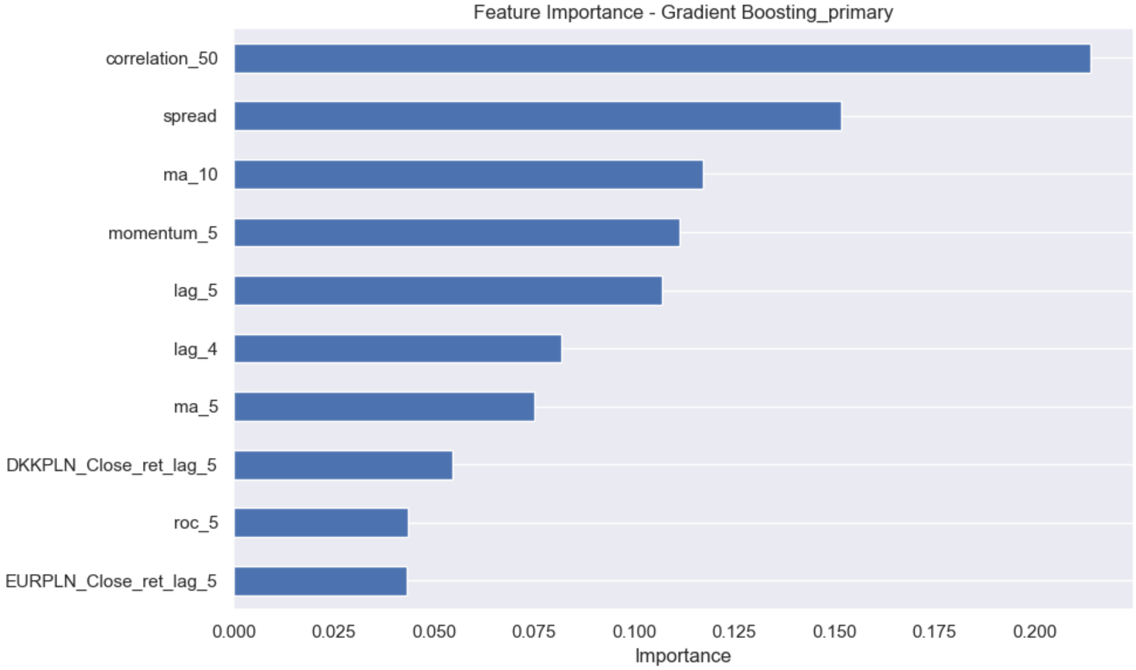


Figure 10 - PAIR 2/5: EURPLN_Close vs DKKPLN_Close


Feature Importance - feature_selection
lag_5
correlation_50
lag_4
DKKPLN_Close_ret_lag_5
EURPLN_Close_ret_lag_5
roc_5
momentum_5
ma_10
ma_5
spread
ma_20
bb_upper
lag_3
bb_lower
lag_2
lag_1
correlation_10
ma_50
DKKPLN_Close_ret_lag_4
EURPLN_Close_ret_lag_1
EURPLN_Close_ret_lag_4
DKKPLN_Close_ret_lag_1
std_10
momentum_20
EURPLN_Close_ret_lag_2
DKKPLN_Close_ret_lag_2
roc_20
ma_dist_10
std_20
bb_width
ma_dist_50
roc_10
momentum_10
zscore
std_5
std_50
correlation_30
ma_dist_20
EURPLN_Close_ret_lag_3
DKKPLN_Close_ret_lag_3
ma_dist_5
rsi
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Importance


Feature Importance - Logistic_primary
ma_5
spread
lag_5
ma_10
correlation_50
lag_4
DKKPLN_Close_ret_lag_5
momentum_5
roc_5
EURPLN_Close_ret_lag_5
0.0
0.2
0.4
0.6
0.8
Importance


Feature Importance - Random Forest_primary
correlation_50
lag_5
spread
lag_4
DKKPLN_Close_ret_lag_5
ma_10
ma_5
momentum_5
EURPLN_Close_ret_lag_5
roc_5
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
Importance



```
--------------------------------------------------
PERFORMANCE METRICS FOR EURPLN_Close vs DKKPLN_Close
--------------------------------------------------


===== PERFORMANCE SUMMARY =====
Currency Pairs: EURPLN_Close and DKKPLN_Close
Hedge Ratio: 0.1391
Feature Selection: Enabled
Meta-Labeling: Enabled

Classification Metrics:
--------------------------------------------------
       model_name model_type  class  precision   recall  f1-score  support
         Logistic    primary      0   0.565693 0.738095  0.640496      210
         Logistic    primary      1   0.466019 0.287425  0.355556      167
         Logistic  secondary      0   0.514620 0.505747  0.510145      174
         Logistic  secondary      1   0.582524 0.591133  0.586797      203
         Logistic   combined      0   0.571848 0.928571  0.707804      210
         Logistic   combined      1   0.583333 0.125749  0.206897      167
    Random Forest    primary      0   0.551971 0.733333  0.629857      210
    Random Forest    primary      1   0.428571 0.251497  0.316981      167
    Random Forest  secondary      0   0.520000 0.143646  0.225108      181
    Random Forest  secondary      1   0.525994 0.877551  0.657744      196
    Random Forest   combined      0   0.555160 0.742857  0.635438      210
    Random Forest   combined      1   0.437500 0.251497  0.319392      167
Gradient Boosting    primary      0   0.565217 0.680952  0.617711      210
Gradient Boosting    primary      1   0.459677 0.341317  0.391753      167
Gradient Boosting  secondary      0   0.500000 0.011299  0.022099      177
Gradient Boosting  secondary     1   0.530831 0.990000  0.691099     200
Gradient Boosting   combined     0   0.564706 0.685714  0.619355     210
Gradient Boosting   combined     1   0.459016 0.335329  0.387543     167


Model Performance Metrics:
             Primary Model  Secondary Model  Combined (Meta-Labeled)
Logistic             0.538462        0.538462               0.538462
Random Forest         0.519894        0.519894               0.519894
Gradient Boosting      0.530504        0.530504               0.530504

Strategy Metrics:
--------------------------------------------------
     model_name   model_type  annualized_return  annualized_volatility  sharpe_ratio  max_drawdown
      Logistic     primary         0.012097             0.062051      0.195205    -0.057011
      Logistic meta_labeled          0.024897             0.015856      1.572306    -0.006855
   Random Forest     primary        -0.030478             0.062026    -0.492027    -0.119457
   Random Forest meta_labeled        -0.032188             0.024572    -1.311690    -0.061385
Gradient Boosting     primary         0.008549             0.062054      0.137956    -0.076339
Gradient Boosting meta_labeled        -0.019519             0.030805    -0.634483    -0.045105

Filtered Strategy Metrics (Only Model Name and Sharpe Ratio):
--------------------------------------------------
     model_name  sharpe_ratio
      Logistic     0.195205
      Logistic     1.572306
   Random Forest    -0.492027
   Random Forest    -1.311690
Gradient Boosting     0.137956
Gradient Boosting    -0.634483

Summary Statistics for Strategy Metrics:
--------------------------------------------------
     annualized_return  annualized_volatility  sharpe_ratio  max_drawdown
count         6.000000            6.000000      6.000000      6.000000
mean         -0.006107            0.042894     -0.088789     -0.061025
std          0.024337            0.021508      0.984599      0.037022
min         -0.032188            0.015856     -1.311690     -0.119457
25%          -0.027738            0.026130     -0.598869     -0.072601
50%          -0.005485            0.046415     -0.177035     -0.059198
75%           0.011210            0.062045      0.180893     -0.048081
max          0.024897            0.062054      1.572306     -0.006855
```

```
**************************************************************************************************
**************************************************************************************************
*                                                  *
*          PAIR 3/5: SEKNOK_Close vs SEKZAR_Close (HEDGE RATIO: 2.7644)          *
*                                                  *
**************************************************************************************************
**************************************************************************************************
```

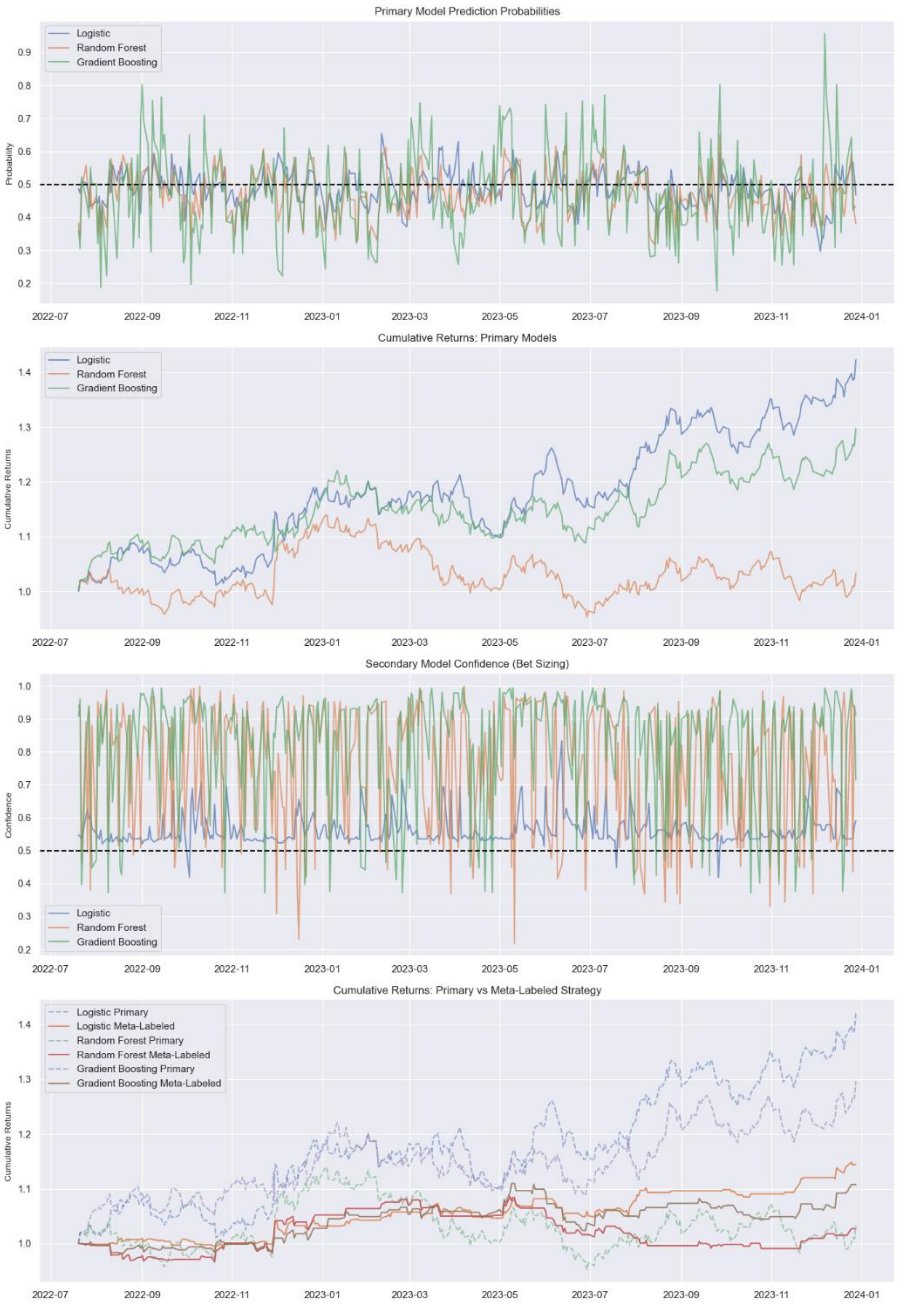
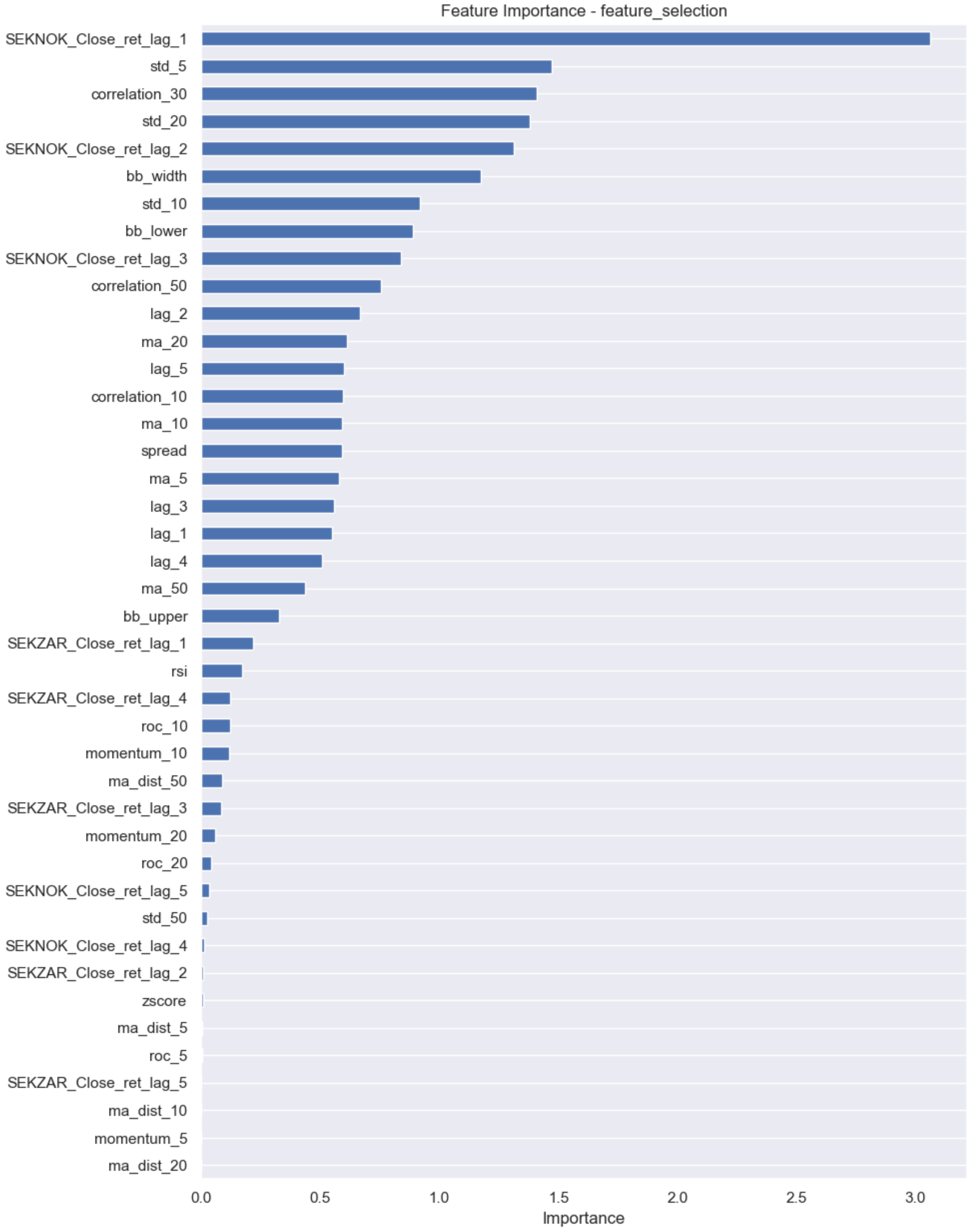
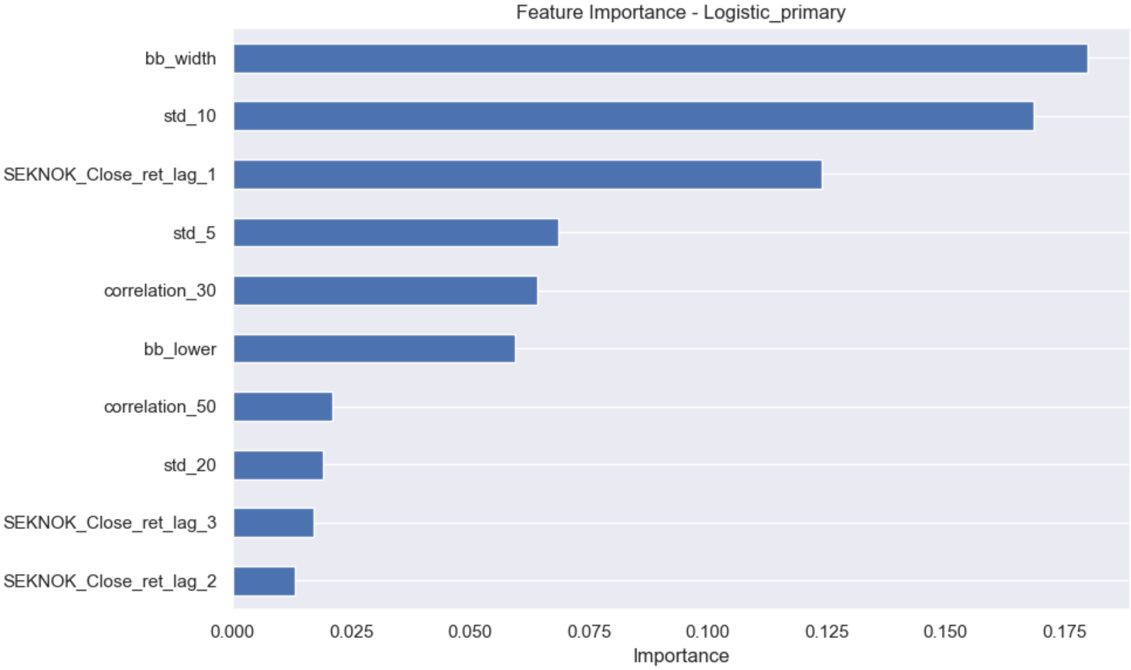
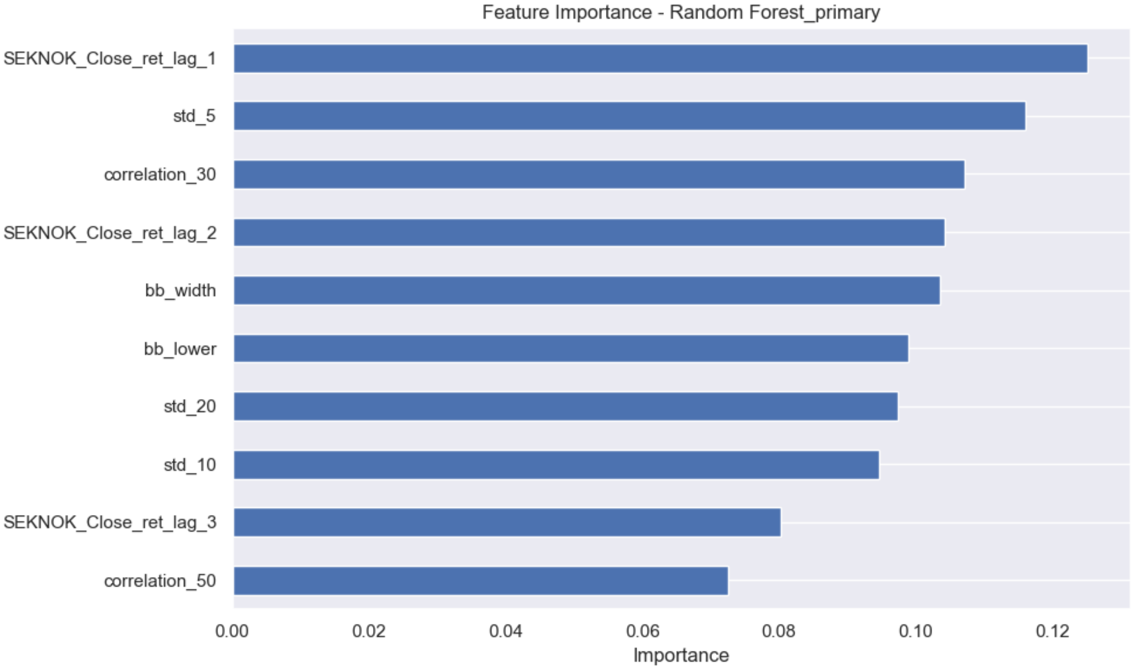
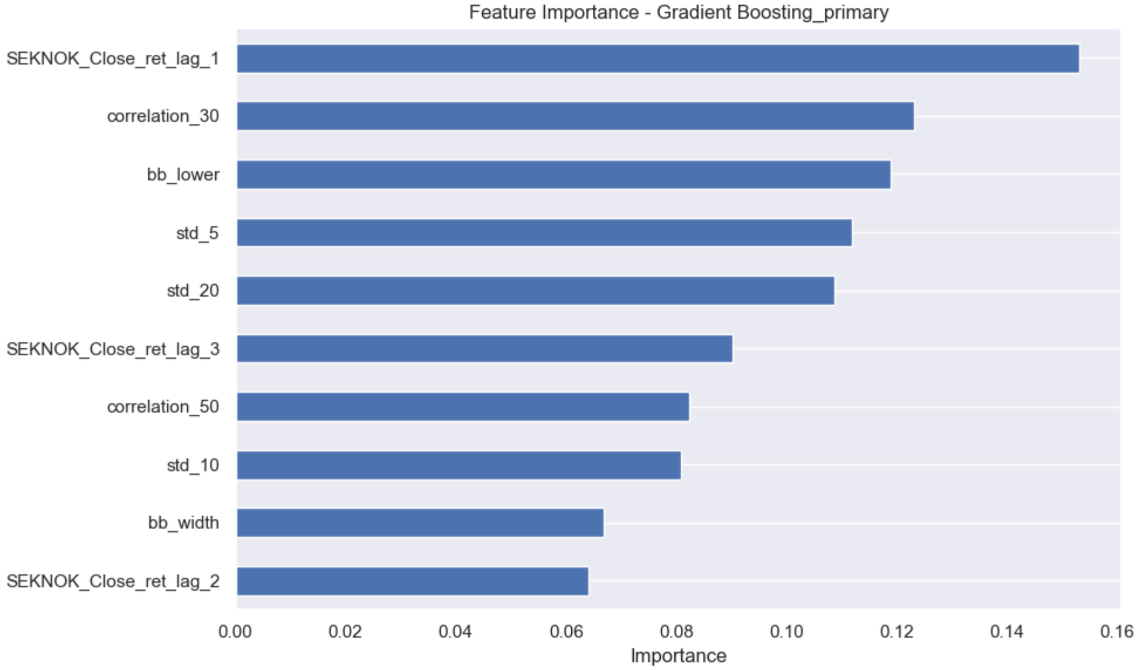


Figure 11 - PAIR 3/5: SEKNOK_Close vs SEKZAR_Close


Feature Importance - feature_selection
SEKNOK_Close_ret_lag_1
std_5
correlation_30
std_20
SEKNOK_Close_ret_lag_2
bb_width
std_10
bb_lower
SEKNOK_Close_ret_lag_3
correlation_50
lag_2
ma_20
lag_5
correlation_10
ma_10
spread
ma_5
lag_3
lag_1
lag_4
ma_50
bb_upper
SEKZAR_Close_ret_lag_1
rsi
SEKZAR_Close_ret_lag_4
roc_10
momentum_10
ma_dist_50
SEKZAR_Close_ret_lag_3
momentum_20
roc_20
SEKNOK_Close_ret_lag_5
std_50
SEKNOK_Close_ret_lag_4
SEKZAR_Close_ret_lag_2
zscore
ma_dist_5
roc_5
SEKZAR_Close_ret_lag_5
ma_dist_10
momentum_5
ma_dist_20
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Importance


Feature Importance - Logistic_primary
bb_width
std_10
SEKNOK_Close_ret_lag_1
std_5
correlation_30
bb_lower
correlation_50
std_20
SEKNOK_Close_ret_lag_3
SEKNOK_Close_ret_lag_2
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
Importance


Feature Importance - Random Forest_primary
SEKNOK_Close_ret_lag_1
std_5
correlation_30
SEKNOK_Close_ret_lag_2
bb_width
bb_lower
std_20
std_10
SEKNOK_Close_ret_lag_3
correlation_50
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Importance



```
--------------------------------------------------
PERFORMANCE METRICS FOR SEKNOK_Close vs SEKZAR_Close
--------------------------------------------------


===== PERFORMANCE SUMMARY =====
Currency Pairs: SEKNOK_Close and SEKZAR_Close
Hedge Ratio: 2.7644
Feature Selection: Enabled
Meta-Labeling: Enabled

Classification Metrics:
--------------------------------------------------
      model_name model_type  class  precision   recall  f1-score  support
        Logistic    primary      0   0.534483 0.642487  0.583529      193
        Logistic    primary      1   0.524138 0.413043  0.462006      184
        Logistic  secondary      0   0.666667 0.011299  0.022222      177
        Logistic  secondary      1   0.532086 0.995000  0.693380      200
        Logistic   combined      0   0.534188 0.647668  0.585480      193
        Logistic   combined      1   0.524476 0.407609  0.458716      184
   Random Forest    primary      0   0.517510 0.689119  0.591111      193
   Random Forest    primary      1   0.500000 0.326087  0.394737      184
   Random Forest  secondary      0   0.456522 0.114130  0.182609      184
   Random Forest  secondary      1   0.507553 0.870466  0.641221      193
   Random Forest   combined      0   0.518797 0.715026  0.601307      193
   Random Forest   combined      1   0.504505 0.304348  0.379661      184
Gradient Boosting    primary      0   0.554113 0.663212  0.603774      193
Gradient Boosting    primary      1   0.554795 0.440217  0.490909      184
Gradient Boosting  secondary      0   0.428571 0.089286  0.147783      168
Gradient Boosting  secondary     1   0.552632 0.904306  0.686025     209
Gradient Boosting   combined     0   0.534694 0.678756  0.598174     193
Gradient Boosting   combined     1   0.530303 0.380435  0.443038     184


Model Performance Metrics:
              Primary Model  Secondary Model  Combined (Meta-Labeled)
Logistic            0.530504        0.530504              0.530504
Random Forest         0.511936        0.511936              0.511936
Gradient Boosting      0.554377        0.554377              0.554377

Strategy Metrics:
--------------------------------------------------
     model_name   model_type  annualized_return  annualized_volatility  sharpe_ratio  max_drawdown
       Logistic     primary          0.248757             0.161380      1.543484     -0.094648
       Logistic meta_labeled          0.092218             0.062860      1.468999     -0.031339
   Random Forest      primary          0.035196             0.162126      0.217378     -0.163522
   Random Forest meta_labeled          0.020899             0.075647      0.276641     -0.087254
Gradient Boosting      primary          0.187177             0.161711      1.159020     -0.108394
Gradient Boosting meta_labeled          0.071718             0.080143      0.896068     -0.077743

Filtered Strategy Metrics (Only Model Name and Sharpe Ratio):
--------------------------------------------------
     model_name  sharpe_ratio
       Logistic      1.543484
       Logistic      1.468999
   Random Forest      0.217378
   Random Forest      0.276641
Gradient Boosting      1.159020
Gradient Boosting      0.896068

Summary Statistics for Strategy Metrics:
--------------------------------------------------
     annualized_return  annualized_volatility  sharpe_ratio  max_drawdown
count         6.000000             6.000000      6.000000      6.000000
mean          0.109328             0.117311      0.926932     -0.093817
std          0.090026             0.048998      0.575350      0.043074
min           0.020899             0.062860      0.217378     -0.163522
25%           0.044326             0.076771      0.431498     -0.104957
50%           0.081968             0.120762      1.027544     -0.090951
75%           0.163438             0.161628      1.391504     -0.080121
max           0.248757             0.162126      1.543484     -0.031339
```

```
*********************************************************************************************
*********************************************************************************************
*                                               *
*          PAIR 4/5: NZDCHF_Close vs USDZAR_Close (HEDGE RATIO: 0.164)          *
*                                               *
*********************************************************************************************
*********************************************************************************************
```

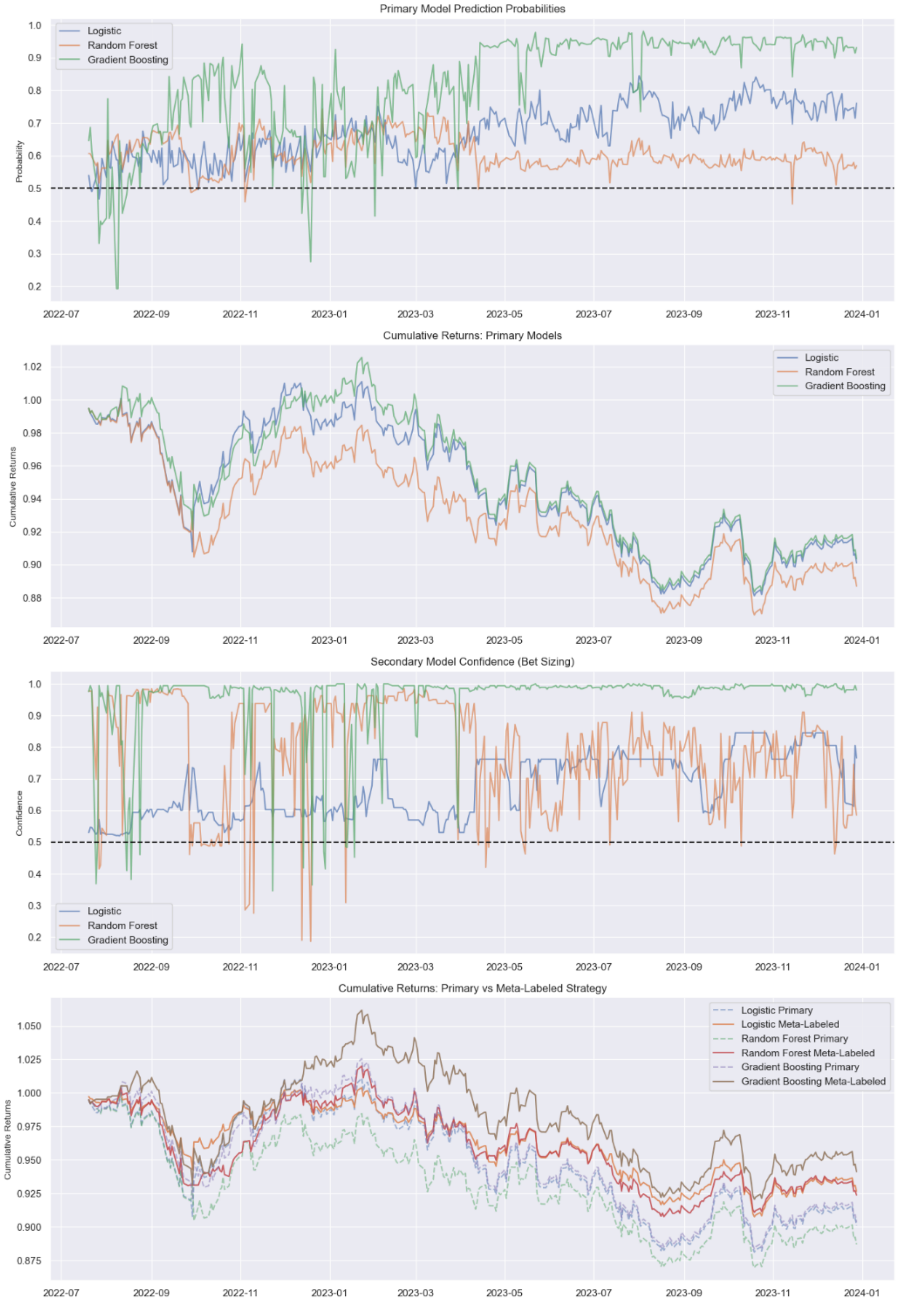
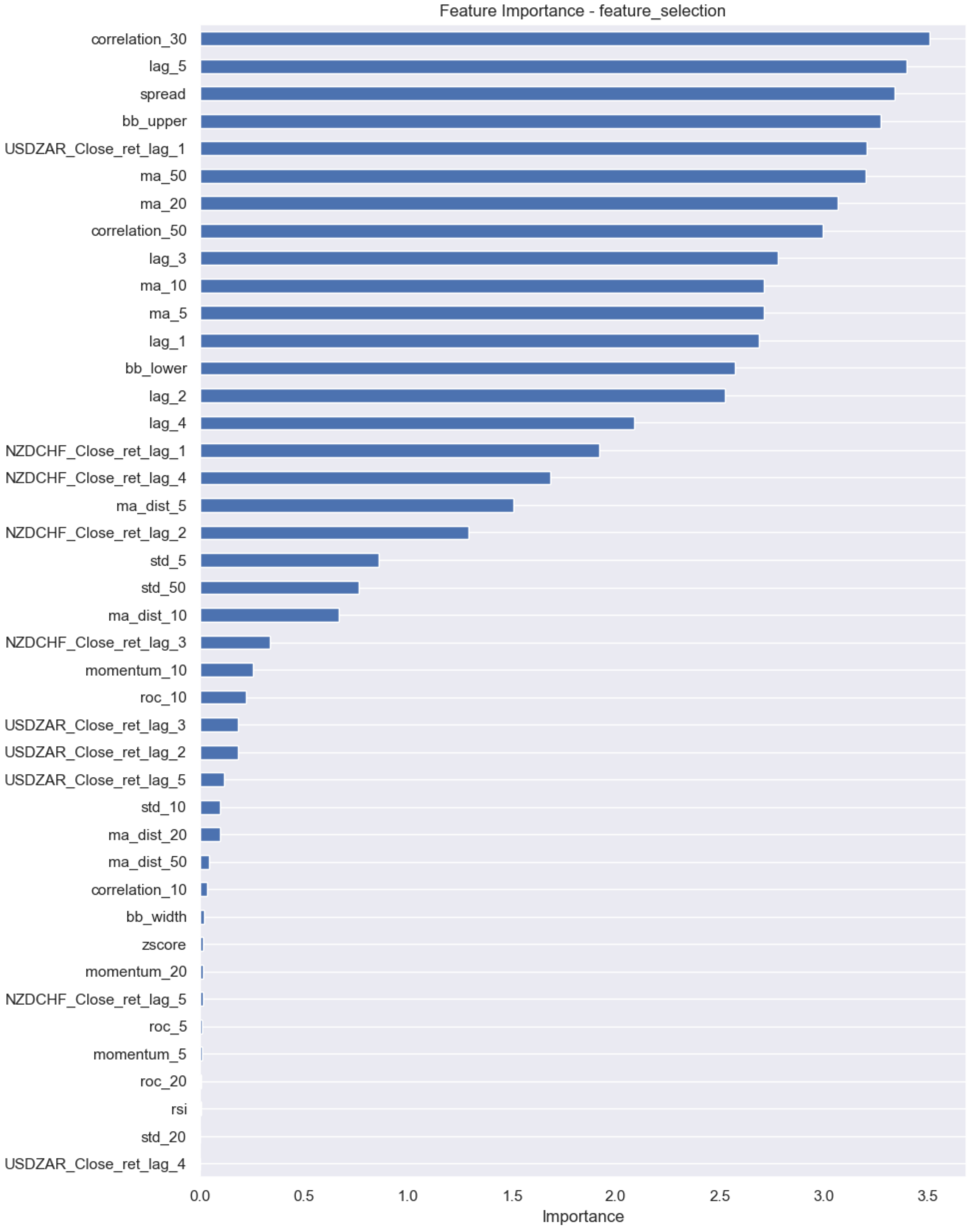
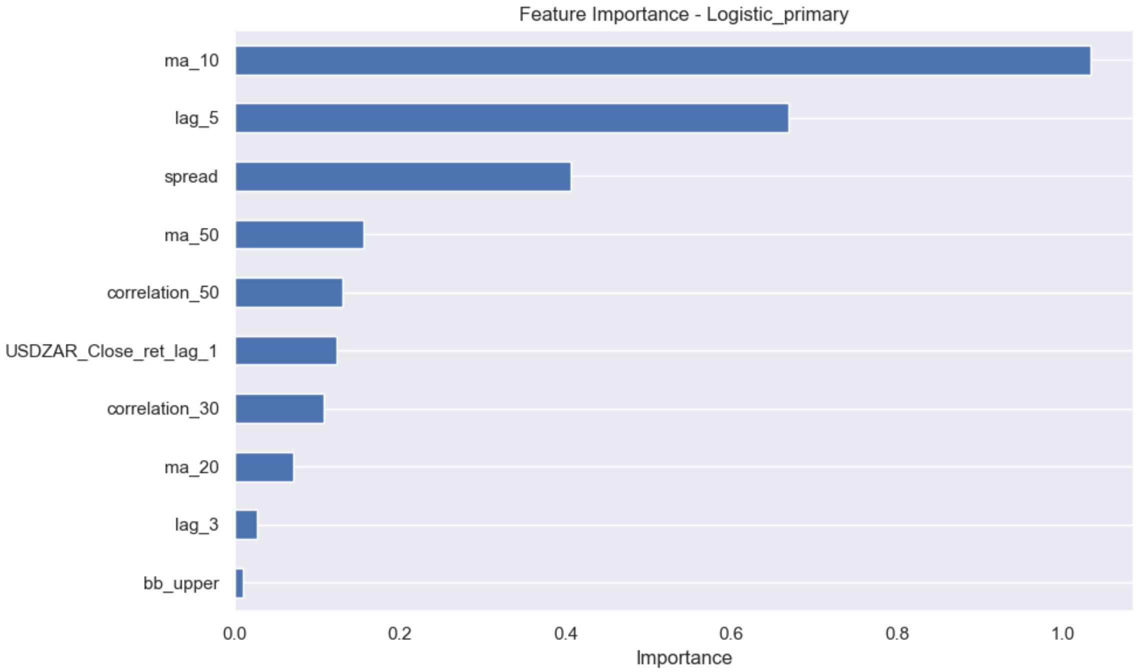
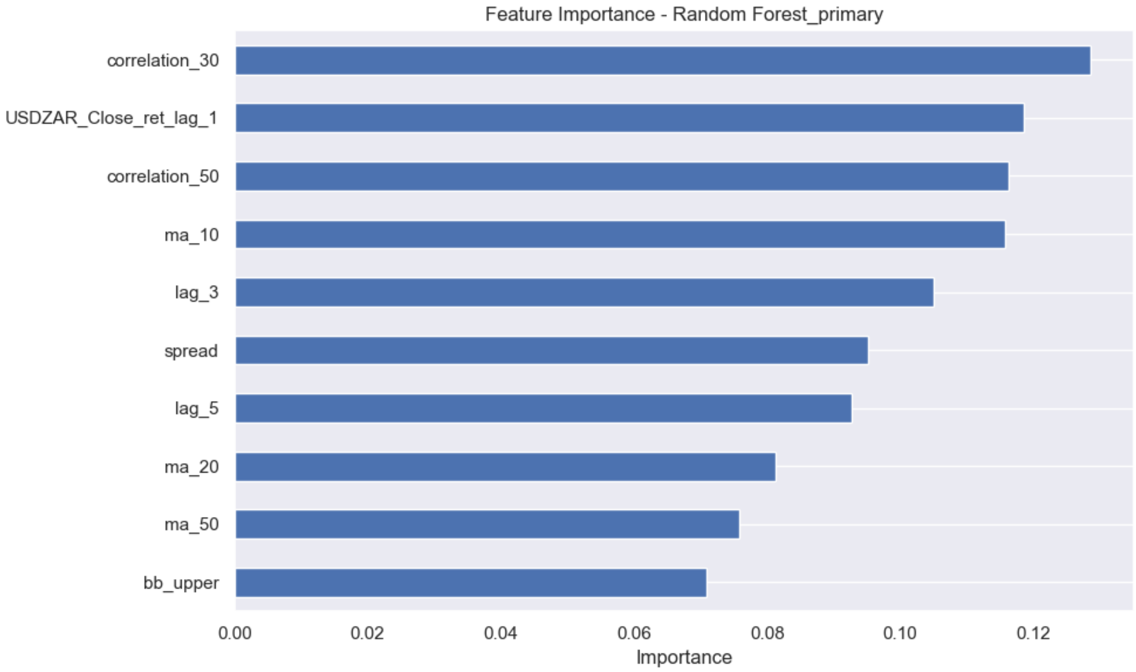
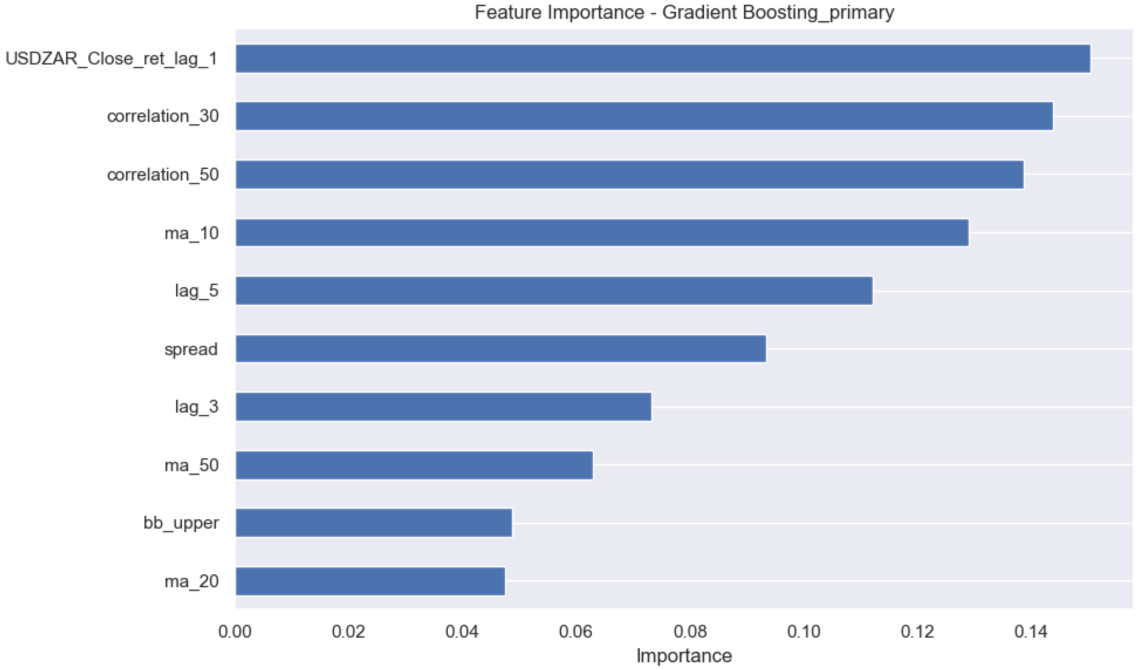


Figure 12 - PAIR 4/5: NZDCHF_Close vs USDZAR_Close


Feature Importance - feature_selection
correlation_30
lag_5
spread
bb_upper
USDZAR_Close_ret_lag_1
ma_50
ma_20
correlation_50
lag_3
ma_10
ma_5
lag_1
bb_lower
lag_2
lag_4
NZDCHF_Close_ret_lag_1
NZDCHF_Close_ret_lag_4
ma_dist_5
NZDCHF_Close_ret_lag_2
std_5
std_50
ma_dist_10
NZDCHF_Close_ret_lag_3
momentum_10
roc_10
USDZAR_Close_ret_lag_3
USDZAR_Close_ret_lag_2
USDZAR_Close_ret_lag_5
std_10
ma_dist_20
ma_dist_50
correlation_10
bb_width
zscore
momentum_20
NZDCHF_Close_ret_lag_5
roc_5
momentum_5
roc_20
rsi
std_20
USDZAR_Close_ret_lag_4
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Importance


Feature Importance - Logistic_primary
ma_10
lag_5
spread
ma_50
correlation_50
USDZAR_Close_ret_lag_1
correlation_30
ma_20
lag_3
bb_upper
0.0
0.2
0.4
0.6
0.8
1.0
Importance


Feature Importance - Random Forest_primary
correlation_30
USDZAR_Close_ret_lag_1
correlation_50
ma_10
lag_3
spread
lag_5
ma_20
ma_50
bb_upper
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Importance



```
--------------------------------------------------
PERFORMANCE METRICS FOR NZDCHF_Close vs USDZAR_Close
--------------------------------------------------


===== PERFORMANCE SUMMARY =====
Currency Pairs: NZDCHF_Close and USDZAR_Close
Hedge Ratio: 0.164
Feature Selection: Enabled
Meta-Labeling: Enabled

Classification Metrics:
--------------------------------------------------
      model_name model_type  class  precision   recall  f1-score  support
        Logistic    primary      0   0.750000 0.016043  0.031414      187
        Logistic    primary      1   0.506702 0.994737  0.671403      190
        Logistic  secondary      0   0.000000 0.000000  0.000000      185
        Logistic  secondary      1   0.509284 1.000000  0.674868      192
        Logistic   combined      0   0.750000 0.016043  0.031414      187
        Logistic   combined      1   0.506702 0.994737  0.671403      190
   Random Forest    primary      0   0.714286 0.026738  0.051546      187
   Random Forest    primary      1   0.508108 0.989474  0.671429      190
   Random Forest  secondary      0   0.535714 0.081522  0.141509      184
   Random Forest  secondary      1   0.515759 0.932642  0.664207      193
   Random Forest   combined      0   0.625000 0.106952  0.182648      187
   Random Forest   combined      1   0.515942 0.936842  0.665421      190
Gradient Boosting    primary      0   0.571429 0.042781  0.079602      187
Gradient Boosting    primary      1   0.506887 0.968421  0.665461      190
Gradient Boosting  secondary      0   0.583333 0.037838  0.071066      185
Gradient Boosting  secondary    1   0.512329 0.973958  0.671454     192
Gradient Boosting   combined    0   0.560000 0.074866  0.132075     187
Gradient Boosting   combined    1   0.508523 0.942105  0.660517     190


Model Performance Metrics:
               Primary Model  Secondary Model  Combined (Meta-Labeled)
Logistic             0.509284        0.509284               0.509284
Random Forest        0.511936        0.511936               0.511936
Gradient Boosting    0.509284        0.509284               0.509284

Strategy Metrics:
--------------------------------------------------
      model_name   model_type  annualized_return  annualized_volatility  sharpe_ratio  max_drawdown
        Logistic      primary         -0.065217               0.093636     -0.697414     -0.128381
        Logistic meta_labeled         -0.048959               0.062585     -0.783326     -0.096043
   Random Forest      primary         -0.075800               0.093605     -0.810868     -0.130507
   Random Forest meta_labeled         -0.050333               0.073582     -0.684951     -0.110277
Gradient Boosting      primary         -0.063386               0.093641     -0.677796     -0.138517
Gradient Boosting meta_labeled         -0.036489               0.089458     -0.408435     -0.132979

Filtered Strategy Metrics (Only Model Name and Sharpe Ratio):
--------------------------------------------------
      model_name  sharpe_ratio
        Logistic     -0.697414
        Logistic     -0.783326
   Random Forest     -0.810868
   Random Forest     -0.684951
Gradient Boosting     -0.677796
Gradient Boosting     -0.408435

Summary Statistics for Strategy Metrics:
--------------------------------------------------
       annualized_return  annualized_volatility  sharpe_ratio  max_drawdown
count           6.000000               6.000000      6.000000      6.000000
mean           -0.056697               0.084418     -0.677132     -0.122784
std             0.014078               0.013221      0.142712      0.016211
min            -0.075800               0.062585     -0.810868     -0.138517
25%            -0.064759               0.077551     -0.761848     -0.132361
50%            -0.056860               0.091531     -0.691182     -0.129444
75%            -0.049303               0.093628     -0.679585     -0.114803
max            -0.036489               0.093641     -0.408435     -0.096043
```

```
**********************************************************************************************
**********************************************************************************************
*                                                *
*          PAIR 5/5: EURNOK_Close vs DKKZAR_Close (HEDGE RATIO: 0.3474)          *
*                                                *
**********************************************************************************************
**********************************************************************************************
```

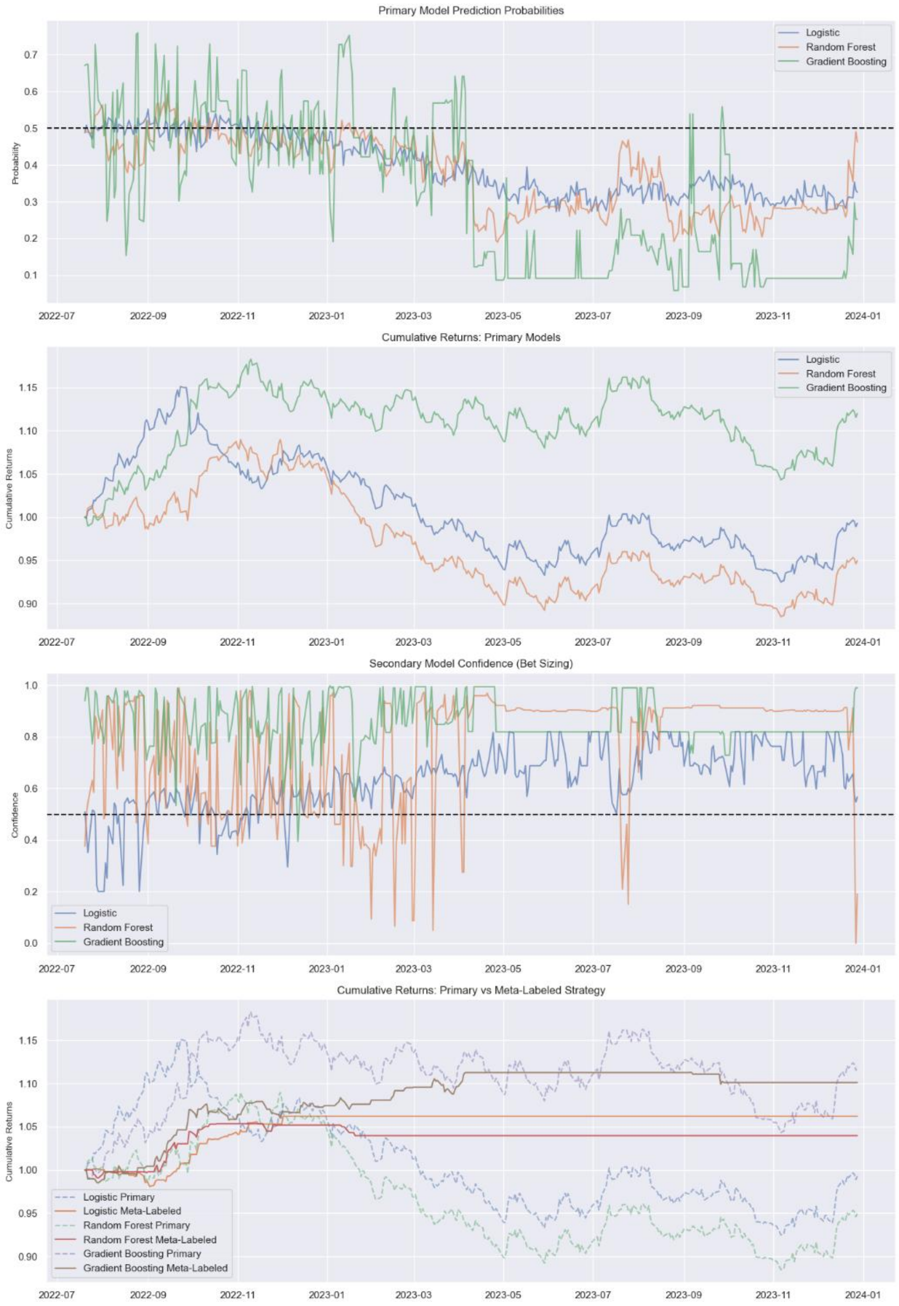
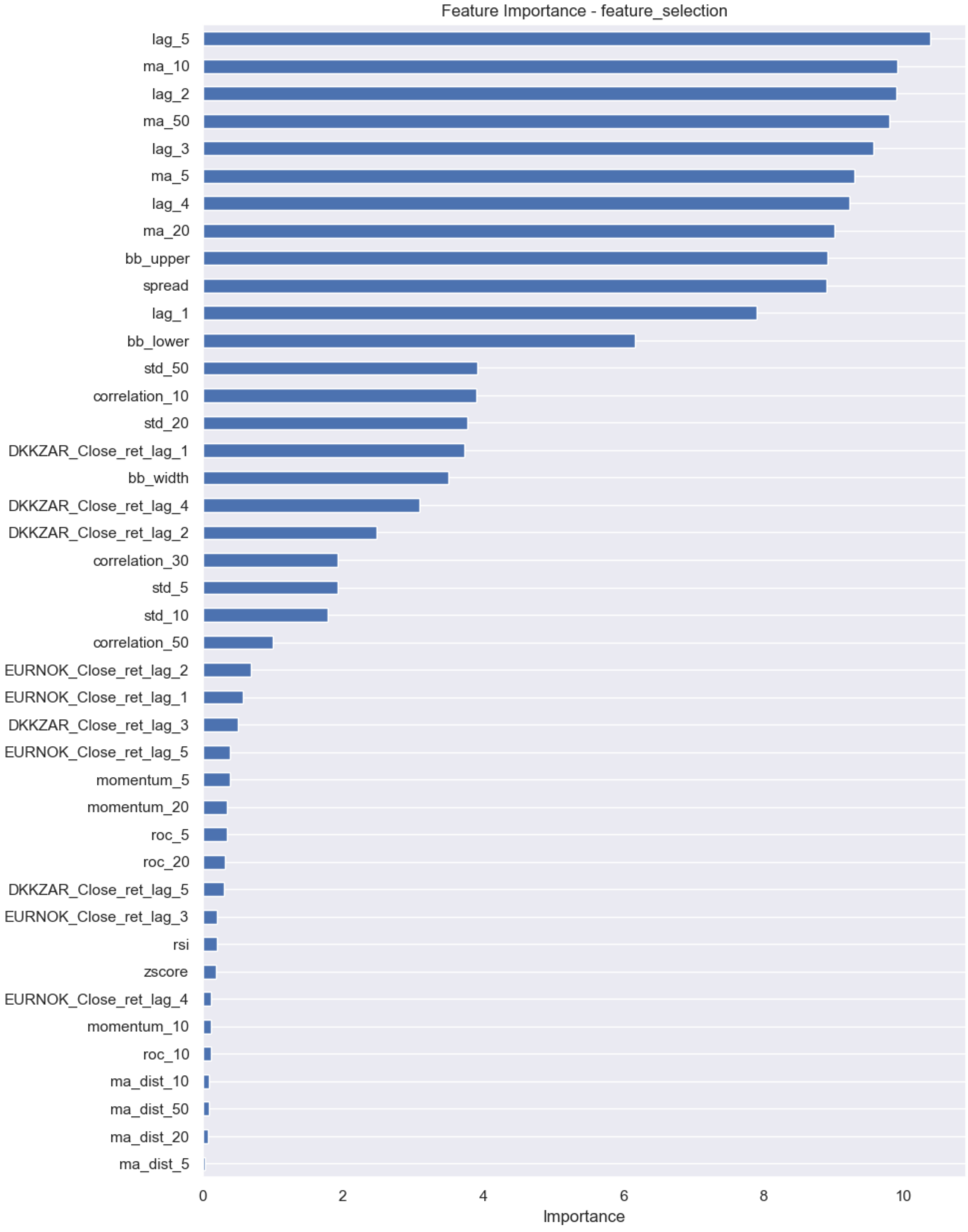
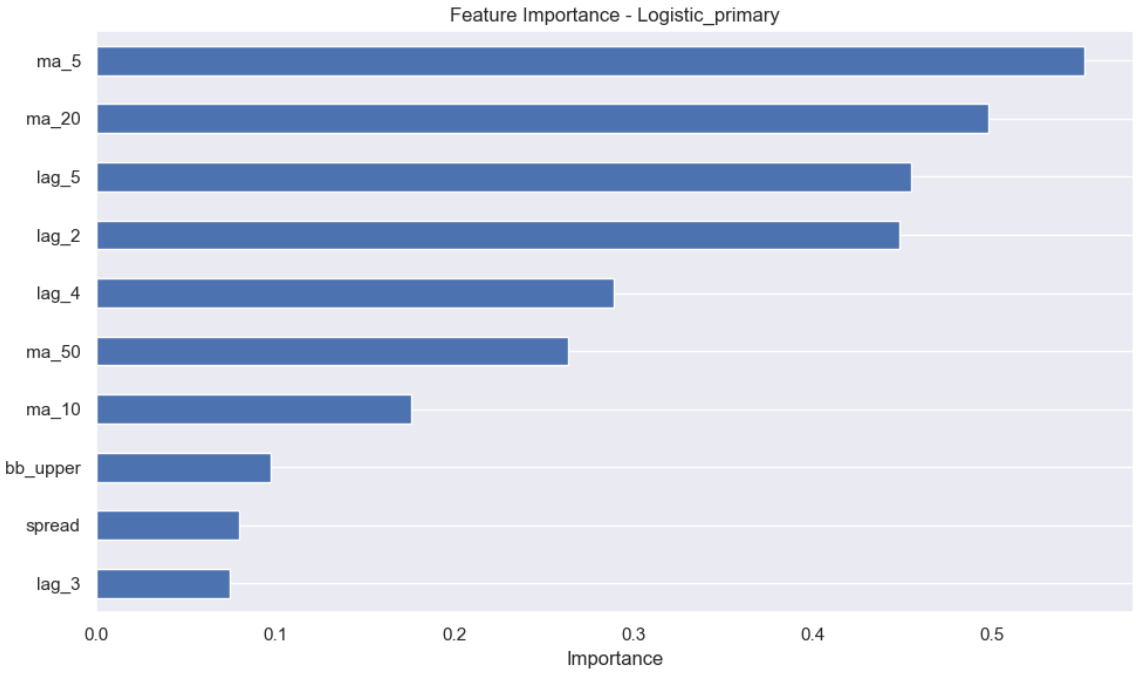
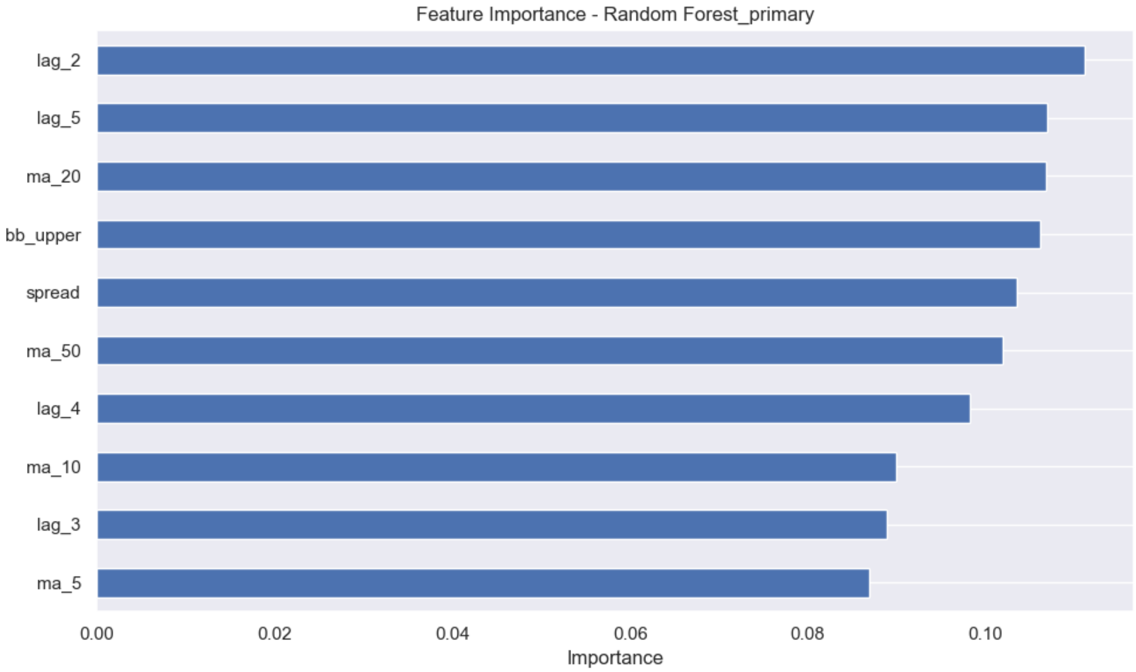
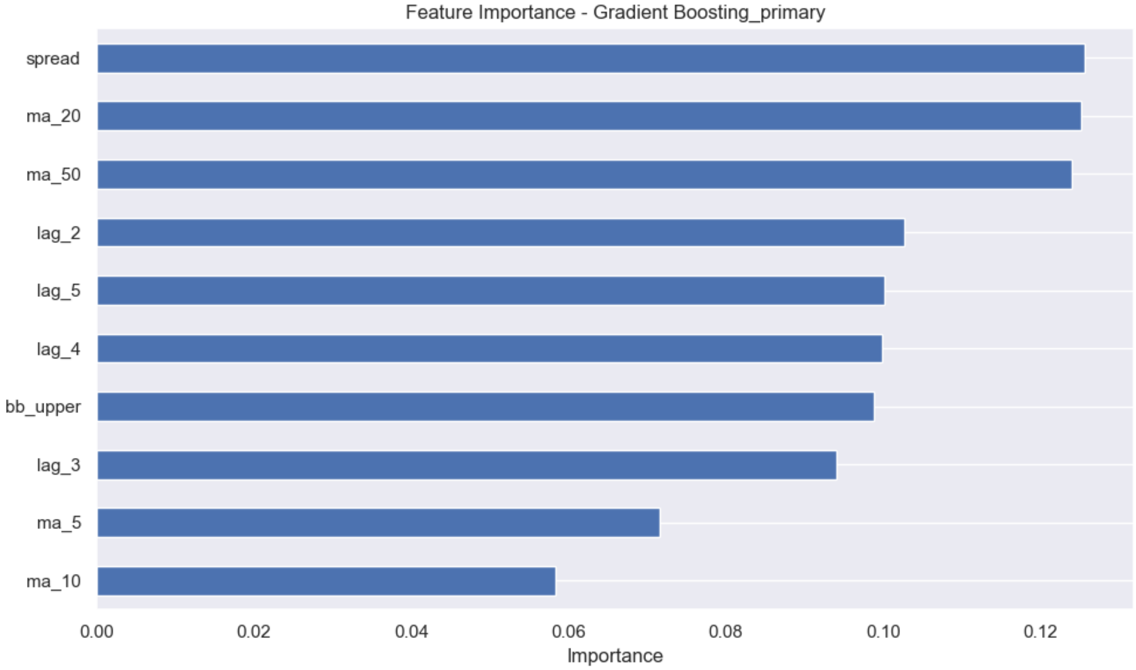


Figure 13 - PAIR 5/5: EURNOK_Close vs DKKZAR_Close


Feature Importance - feature_selection
lag_5
ma_10
lag_2
ma_50
lag_3
ma_5
lag_4
ma_20
bb_upper
spread
lag_1
bb_lower
std_50
correlation_10
std_20
DKKZAR_Close_ret_lag_1
bb_width
DKKZAR_Close_ret_lag_4
DKKZAR_Close_ret_lag_2
correlation_30
std_5
std_10
correlation_50
EURNOK_Close_ret_lag_2
EURNOK_Close_ret_lag_1
DKKZAR_Close_ret_lag_3
EURNOK_Close_ret_lag_5
momentum_5
momentum_20
roc_5
roc_20
DKKZAR_Close_ret_lag_5
EURNOK_Close_ret_lag_3
rsi
zscore
EURNOK_Close_ret_lag_4
momentum_10
roc_10
ma_dist_10
ma_dist_50
ma_dist_20
ma_dist_5
0
2
4
6
8
10
Importance


Feature Importance - Logistic_primary
ma_5
ma_20
lag_5
lag_2
lag_4
ma_50
ma_10
bb_upper
spread
lag_3
0.0
0.1
0.2
0.3
0.4
0.5
Importance


Feature Importance - Random Forest_primary
lag_2
lag_5
ma_20
bb_upper
spread
ma_50
lag_4
ma_10
lag_3
ma_5
0.00
0.02
0.04
0.06
0.08
0.10
Importance



```
--------------------------------------------------
PERFORMANCE METRICS FOR EURNOK_Close vs DKKZAR_Close
--------------------------------------------------


===== PERFORMANCE SUMMARY =====
Currency Pairs: EURNOK_Close and DKKZAR_Close
Hedge Ratio: 0.3474
Feature Selection: Enabled
Meta-Labeling: Enabled

Classification Metrics:
--------------------------------------------------
       model_name model_type  class  precision   recall  f1-score  support
         Logistic    primary      0   0.472222 0.854749  0.608350      179
         Logistic    primary      1   0.509434 0.136364  0.215139      198
         Logistic  secondary      0   0.575000 0.116751  0.194093      197
         Logistic  secondary      1   0.483680 0.905556  0.630561      180
         Logistic   combined      0   0.485876 0.960894  0.645403      179
         Logistic   combined      1   0.695652 0.080808  0.144796      198
    Random Forest    primary      0   0.469388 0.899441  0.616858      179
    Random Forest    primary      1   0.470588 0.080808  0.137931      198
    Random Forest  secondary      0   0.608696 0.210000  0.312268      200
    Random Forest  secondary      1   0.487013 0.847458  0.618557      177
    Random Forest   combined      0   0.471098 0.910615  0.620952      179
    Random Forest   combined      1   0.483871 0.075758  0.131004      198
Gradient Boosting    primary      0   0.491039 0.765363  0.598253      179
Gradient Boosting    primary      1   0.571429 0.282828  0.378378      198
Gradient Boosting  secondary      0   0.000000 0.000000  0.000000      184
Gradient Boosting  secondary     1   0.510638 0.994819  0.674868     193
Gradient Boosting   combined     0   0.491039 0.765363  0.598253     179
Gradient Boosting   combined     1   0.571429 0.282828  0.378378     198


Model Performance Metrics:
              Primary Model  Secondary Model  Combined (Meta-Labeled)
Logistic              0.477454        0.477454                0.477454
Random Forest          0.469496        0.469496                0.469496
Gradient Boosting      0.511936        0.511936                0.511936

Strategy Metrics:
--------------------------------------------------
     model_name   model_type  annualized_return  annualized_volatility  sharpe_ratio  max_drawdown
       Logistic      primary          0.000640             0.103631      0.006180     -0.196545
       Logistic meta_labeled          0.040520             0.021820      1.859512     -0.020224
   Random Forest      primary         -0.029046             0.103615     -0.280698     -0.188195
   Random Forest meta_labeled          0.026378             0.026172      1.009219     -0.013141
Gradient Boosting      primary          0.081097             0.103504      0.784557     -0.117981
Gradient Boosting meta_labeled          0.065216             0.040108      1.628177     -0.017616

Filtered Strategy Metrics (Only Model Name and Sharpe Ratio):
--------------------------------------------------
     model_name  sharpe_ratio
       Logistic      0.006180
       Logistic      1.859512
   Random Forest     -0.280698
   Random Forest      1.009219
Gradient Boosting      0.784557
Gradient Boosting      1.628177
```

REFERENCES

Alrobaie, A., & Krarti, M. (2022). A review of data-driven approaches for measurement and verification analysis of building energy retrofits. *Energies*, *15*(21), 7824.

Aparicio, D., & López de Prado, M. (2018). How hard is it to pick the right model? MCS and backtest overfitting. *Algorithmic Finance*, *7*(1-2), 53-61.

Avellaneda, M., & Lee, J.-H. (2010). Statistical arbitrage in the US equities market. *Quantitative Finance*, *10*(7), 761-782.

Azolibe, C. B. (2020). Banking Sector Intermediation Development and Growth of a Developing Economy: An Empirical Investigation of Nigeria.

Bénard, C., Da Veiga, S., & Scornet, E. (2022). Mean decrease accuracy for random forests: inconsistency, and a practical solution via the Sobol-MDA. *Biometrika*, *109*(4), 881-900.

Carta, S., Consoli, S., Podda, A. S., Recupero, D. R., & Stanciu, M. M. (2022). Statistical arbitrage powered by Explainable Artificial Intelligence. *Expert Systems with Applications*, *206*, 117763. https://doi.org/https://doi.org/10.1016/j.eswa.2022.117763

Carta, S., Podda, A., Reforgiato Recupero, D., & Stanciu, M. (2022). Explainable AI for Financial Forecasting. In (pp. 51-69). https://doi.org/10.1007/978-3-030-95470-3_5

Carta, S. M., Consoli, S., Podda, A. S., Recupero, D. R., & Stanciu, M. M. (2021). Ensembling and Dynamic Asset Selection for Risk-Controlled Statistical Arbitrage. *IEEE Access*, *9*, 29942-29959. https://doi.org/10.1109/ACCESS.2021.3059187

Chan, E. (2013). *Algorithmic Trading: Winning Strategies and Their Rationale*. Wiley. https://books.google.de/books?id=WAlFDwAAQBAJ

Chan, E. P. (2017). *Machine Trading: Deploying Computer Algorithms to Conquer the Markets*. Wiley. https://books.google.de/books?id=7bfBDQAAQBAJ

CHEN, N. (2017). Are foreign exchange rates only affected by US and domestic news?

Chung, J., De Prado, M. L., Simon, H., & Wu, K. (2023). Data Driven Dimensionality Reduction to Improve Modeling Performance∗. Proceedings of the 35th International Conference on Scientific and Statistical Database Management,

De Prado, M. L. (2018). The 10 reasons most machine learning funds fail. *The Journal of Portfolio Management*, *44*(6), 120-133.

De Prado, M. L. (2018). *Advances in financial machine learning*. John Wiley & Sons.

De Prado, M. M. L. (2020). *Machine learning for asset managers*. Cambridge University Press.

de Prado, M. M. L. (2023). *Causal factor investing: can factor investing become scientific?* Cambridge University Press.

Do, B., Faff, R., & Hamza, K. (2006). A new approach to modeling and estimation for pairs trading. Proceedings of 2006 financial management association European conference, Elliott, R. J., Van Der Hoek*, J., & Malcolm, W. P. (2005). Pairs trading. *Quantitative Finance*, *5*(3), 271-276.

Engle, R. F., & Granger, C. W. (1987). Co-integration and error correction: representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, 251-276.

Engle, R. F., & Granger, C. W. J. (1987). Co-integration and error correction: representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, 251-276.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. kdd,

Fanelli, V. (2024). Mean-Reverting Statistical Arbitrage Strategies in Crude Oil Markets. *Risks*, *12*(7).

Gatev, E., Goetzmann, W. N., & Rouwenhorst, K. G. (2006). Pairs trading: Performance of a relative-value arbitrage rule. *The Review of Financial Studies*, *19*(3), 797-827.

Gnjatović, M., Košanin, I., Maček, N., & Joksimović, D. (2022). Clustering of road traffic accidents as a gestalt problem. *Applied Sciences*, *12*(9), 4543.

Guyard, K. C., & Deriaz, M. (2024). Predicting Foreign Exchange EUR/USD direction using machine learning. Proceedings of the 2024 7th International Conference on Machine Learning and Machine Intelligence (MLMI),

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, *3*(Mar), 1157-1182.

Hilpisch, Y. (2020). *Python for Algorithmic Trading*. " O'Reilly Media, Inc.".

Ishan Shah , R. P. (2021). *Machine Learning in Trading: Step by step implementation of Machine Learning models*. QuantInsti Quantitative Learning.

Jansen, S. (2020). *Machine Learning for Algorithmic Trading: Predictive Models to Extract Signals from Market and Alternative Data for Systematic Trading Strategies with Python*. Packt Publishing. https://books.google.com.sa/books?id=cki6zQEACAAJ

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Jooyoung, Y., & Kangwhee, K. (2011). Performance Analysis of Pairs Trading Strategy Utilizing High Frequency Data : Evidence from the Korean Stock Market. *Asian Review of Financial Research*, *24*(4), 1153-1172. http://journal.korfin.org/sub/sub_detail.html?code=238202

Joubert, J. F. (2022). Meta-Labeling: Theory and Framework. *The Journal of Financial Data Science*.

Kaczmarek, T., & Perez, K. (2022). Building portfolios based on machine learning predictions. *Economic Research-Ekonomska Istraživanja*, *35*(1), 19-37. https://doi.org/10.1080/1331677X.2021.1875865

Kaufman, P. J. (2019). *Trading Systems and Methods*. John Wiley & Sons, Incorporated. http://ebookcentral.proquest.com/lib/ljmu/detail.action?docID=5972614

Krauss, C. (2015). *Statistical arbitrage pairs trading strategies: Review and outlook*. https://EconPapers.repec.org/RePEc:zbw:iwqwdp:092015

Krauss, C. (2017). STATISTICAL ARBITRAGE PAIRS TRADING STRATEGIES: REVIEW AND OUTLOOK. *Journal of Economic Surveys*, *31*(2), 513-545. https://doi.org/https://doi.org/10.1111/joes.12153

Lim, H. W., Jeong, S. H., Oh, K. J., & Lee, H. S. (2022). Neural network foreign exchange trading system using CCS-IRS basis: Empirical evidence from Korea. *Expert Systems with Applications*, *205*, 117718.

Liu, J., & Timmermann, A. (2013). Optimal convergence trade strategies. *The Review of Financial Studies*, *26*(4), 1048-1086.

López de Prado, M. (2019). Ten applications of financial machine learning. *Available at SSRN 3365271*.

Lopez de Prado, M., Lipton, A., & Zoonekynd, V. (2025). Causal Factor Analysis is a Necessary Condition for Investment Efficiency. *Available at SSRN*.

Louppe, G., Wehenkel, L., Sutera, A., & Geurts, P. (2013). Understanding variable importances in forests of randomized trees. *Advances in neural information processing systems*, *26*.

Man, X., & Chan, E. (2021). The best way to select features? comparing mda, lime, and shap. *The Journal of Financial Data Science Winter*, *3*(1), 127-139.

Man, X., & Chan, E. P. (2020). The best way to select features? *ArXiv*, *abs/2005.12483*.

Meyer, M., Joubert, J. F., & Alfeus, M. (2022). Meta-Labeling Architecture [Article]. *Journal of Financial Data Science*, *4*(4), 10-24. https://doi.org/10.3905/jfds.2022.1.108

Moraffah, R., Sheth, P., Vishnubhatla, S., & Liu, H. (2024). Causal feature selection for responsible machine learning. *arXiv preprint arXiv:2402.02696*.

Patton, A. J. (2012). A review of copula models for economic time series. *Journal of Multivariate Analysis*, *110*, 4-18.

Prasad, A., & Seetharaman, A. (2021). Importance of Machine Learning in Making Investment Decision in Stock Market. *Vikalpa*, *46*(4), 209-222. https://doi.org/10.1177/02560909211059992

Sarmento, S. M., & Horta, N. (2021). *A Machine Learning based Pairs Trading Investment Strategy*. Springer.

Scornet, E. (2021). Trees, forests, and impurity-based variable importance. *arXiv.org*. https://doi.org/10.48550/arxiv.2001.04295

Singh, A., & Joubert, J. (2019). Does meta labeling add to signal efficacy. In.

Stephenson, J., Vanstone, B., & Hahn, T. (2021). A Unifying Model for Statistical Arbitrage: Model Assumptions and Empirical Failure. *Computational economics*, *58*(4), 943-964. https://doi.org/10.1007/s10614-020-09980-6

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(1), 25. https://doi.org/10.1186/1471-2105-8-25

Stübinger, J., & Endres, S. (2018). Pairs trading with a mean-reverting jump–diffusion model on high-frequency data. *Quantitative Finance*, *18*(10), 1735-1751.

Thumm, D., Barucca, P., & Joubert, J. F. (2023). Ensemble Meta-Labeling [Article]. *Journal of Financial Data Science*, *5*(1), 10-26. https://doi.org/10.3905/jfds.2022.1.114

Ti, Y.-W., Dai, T.-S., Wang, K.-L., Chang, H.-H., & Sun, Y.-J. (2024). Improving Cointegration-Based Pairs Trading Strategy with Asymptotic Analyses and Convergence Rate Filters. *Computational economics*, *64*(5), 2717-2745. https://doi.org/10.1007/s10614-023-10539-4

Tobius, B. S., Babirye, C., Nakatumba-Nabende, J., & Katumba, A. (2022). A comparison of topic modeling and classification machine learning algorithms on Luganda data. 3rd Workshop on African natural language processing,

Vidyamurthy, G. (2004). *Pairs Trading: quantitative methods and analysis* (Vol. 217). John Wiley & Sons.

Zhang, M., Tang, X., Zhao, S., Wang, W., & Zhao, Y. (2022). Statistical Arbitrage with Momentum Using Machine Learning. *Procedia computer science*, *202*, 194-202. https://doi.org/10.1016/j.procs.2022.04.027

Zhou, R., Xiong, Y., Wang, N., & Wang, X. (2019). Coupling Degree Evaluation of China’s Internet Financial Ecosystem Based on Entropy Method and Principal Component Analysis. *7*(5), 399-421. https://doi.org/doi:10.21078/JSSI-2019-399-23 (Journal of Systems Science and Information)