Machine Learning and Technical Analysis for Time-series price forecasting in the financial market an empirical comparison of different approaches

Sub-title of the thesis (leave empty if not required)

Bachelor Thesis

Submitted to
IMC University of Applied Sciences Krems

Bachelor Programme Informatics

by

David Bobek

for the award of academic degree
Bachelor of Science in Engineering (BSc)

under the supervision of
Dipl.-Ing. Deepak Dhungana

Submitted on 27.11.2023

DECLARATION OF HONOUR

I declare on my word of honour that I have written this Bachelor Thesis on my own and that I have not used any sources or resources other than stated and that I have marked those passages and/or ideas that were either verbally or textually extracted from sources. This also applies to drawings, sketches, graphic representations as well as to sources from the internet. The Bachelor Thesis has not been submitted in this or similar form for assessment at any other domestic or foreign post-secondary educational institution and has not been published elsewhere. The present Bachelor Thesis complies with the version submitted electronically.

David Bobek
27.11.2023

ABSTRACT

This study is going to be examining various different approaches to time-series analysis and price forecastin based on Machine Learning and Technical Analysis. I will be focusing on the financial market and will be trying to predict and analyse the performance of various different technical indicators and Machine learning models The purpose of this thesis is to compare the return and performance of my approaches and determine which of them is the most suitable for different scenarios. By carrying out an empirical investigation and analyzing the outcomes, I will evaluate the advantages, disadvantages and discuss potential problems.

Keywords: Machine Learning, Technical Analysis, Time-series price forecasting

ACKNOWLEDGEMENTS

This is an **optional** page. Use your choice of paragraph style for text on this page. Usually, this space is for thanking your supporters and getting emotional about how grateful you are to everyone.

Table of Contents

List of Tables

List of Figures

Chapter 1
INTRODUCTION

In today's fast-paced capitalistically driven society in which everything is based on planning for the future and trying to optimize the present decision, a topic of fore- casting is well suited. In this thesis, I will be combining my passion for Data Science and Machine Learning and will try to predict the market movement based on histor- ically available data. Currently, there are various indicators I would like to explore which machine learning algorithm is suited the best for which type of trend pattern and see how my models can be used in the industry. There are several different problems that I will be facing during this research such as Feature Selection, Over- fitting or Generalization, or potential external factors that could affect the Financial Market.

1.1 Motivation

The goal of this thesis is to explore and help predict market trends and eventually make their money. Trading without prior experience is difficult and complicated. I have decided to pursue this research in order to make this process easier. Most of the people that start trading get too scared and tend to "FOMO" (Fear of missing out) and make incorrect decisions. I would like to there- fore analyze and help people with this issue. By finding and recognizing different patterns I could theoretically help spot trends and elaborate on trend patterns. These models can also be used in the energy sector by predicting electricity or fuel prices which can be crucial for energy companies and consumers. Machine learning models can analyze historical data, weather patterns, and other relevant factors to forecast energy prices accurately. This information helps energy can potentially assist companies with the optimization of production, planning infrastructure investments, and developing pricing strategies. Machine learning models that are based on time-series analysis and forecasting have various different use cases and I see overall a large potential in using

this technology in order to help third parties. Overall, establishing machine learning models that estimate pricing provides useful insights to firms and individuals, allowing them to make informed decisions, improve operations, and plan successfully in dynamic market conditions.

1.2 Research Questions

"What is the comparative performance and predictive accuracy of a machine learning algorithm for price forecasting in financial markets? The aim of this research question is to assess a machine learning algorithm's effectiveness in price forecasting. Through empirical investigation and analysis, I will evaluate the algorithm's advantages, disadvantages, and practical implications in predicting financial market values for specific scenarios.

"How does a machine learning algorithm detect and classify patterns in time-series data? This research question aims to elucidate the methods and strategies employed by machine learning algorithms to identify significant patterns and extract valuable insights from time-series data. This inquiry will delve deeper into the recognition of the initiation and termination of trends within various market trend patterns that may exhibit similarities, potentially posing challenges for the algorithm."

1.3 Research Method

In my research method, I am going to create several Juptyer notebooks that will be used to load process filter analyze and work on the data. I will be using Python as my main programming language and will be using various different technical indicators and machine learning models. During my research I will be trying to find the signals that are going to be indicating the time to buy or sell Each of the notebooks will be focused on a different topic and will be used to display the results of my research. I will be using various different libraries such as Pandas, Numpy, Matplotlib,Plotly, Scikit-learn, Tensorflow, Keras, Backtesting or Bokeh to make interactive visualizations and to make my research more comprehensive.

1.4 State of Art

1. Gathering, analyzing and extracting important information from already existing sources

    - In the first step my main goal is to try to explore the thesis and information from elevant sources that are going to help me understand the complexity of this topic further. In order to stay on the correct path I will be thinking critically and selecting information from trusted and credited sources. My previous knowledge

in financial markets will also be an advantage and will help me broaden the horizon of time-series analysis in financial markets.

2. Data Mining and cleaning : In order to get data of highest quality I will need to access it from trusted and verifiable sources. This data will most likely not be clean and I might need to do manual cleaning and filtering of the data.

    - The second step of my approach requires collecting valuable time-series data of various different markets. I will be trying to mainly focus on the publicly traded stock market of currencies, which will be the most beneficial for me as it is the most traded one and its movement impacts the world the most. I am expecting the collected data to not be specifically ready to use and I will have to get rid of potential issues. For this step libraries such as Pandas and Numpy will be used.

3. Data Science and Analysis: This step will require a lot of Data engineering and digging down into what am I interested in data and which features are going to give us the best result based on our machine learning models and Technical Indicators

    - Third step on my path will be Analyzing the data and performing an umbrella term called 'Data Science' which involves processes like data visualization, data exploration and statistical analysis of the data. This step will help me find different trend patterns in my data and further understand what is required from me to get better results. For this step I will be using libraries such as Matplotlib, Plotly, or Bokeh

4. Picking the Machine Learning models and Technical Indicators. This step is going to require a lot of in-depth knwopledge of the strengths and weaknesses of each Model and Indicator. Some of the models and Indicators I could consider are

    (a) Machine Learning models:

        - ARIMA (Autoregressive Integrated Moving Average) models are commonly used for time series analysis and forecasting. They can handle both stationary and non- stationary time series and capture autocorrelation and offer a simple way to forecast values on a time series.
        - Random Forest: Random Forest is a machine learning model that uses decision trees to make predictions. It is a supervised learning algorithm that uses an ensemble of decision trees to make predictions. Random Forest is one of the most popular machine learning models for classification and could be used for time series forecasting.
        - XGBoost: XGBoost is a machine learning model that also uses decision trees to make predictions. The difference between XGBoost and Random

Forest is that XGBoost uses a gradient boosting algorithm to make predictions meaning that it is a boosting algorithm and not an ensemble algorithm.

- Support Vector Regression (SVR): SVR is a machine learning model that per- forms regression tasks using support vector machines. By including lagged variables as input features, it can be modified for time series forecasting.
- Linear Regression: Linear Regression is the simplest machine learning model on this list, but in certain cases, its shear speed is unrivaled.
- LSTM: Long Short-Term Memory (LSTM) is a type of Recurrent Neural Network (RNN) that can learn long-term dependencies in time series data and is therefore a good candidate for time series forecasting.

(b) Technical Indicators:

- RSI (Relative Strength Index): RSI is a momentum indicator that measures the magnitude of recent price changes to evaluate overbought or oversold conditions in the price of a stock or other asset.
- MACD (Moving Average Convergence Divergence): MACD is a trend-following momentum indicator that shows the relationship between two moving averages of a so called security's price.
- VOLD (Volume Delta): Volume Delta is a volume-based indicator that measures the flow of money in and out of a security.
- Bollinger Bands: Bollinger Bands are a technical analysis tool that measures a security's volatility and provides a relative definition of high and low prices.
- Rolling Windows: Rolling Windows are a type of window function that computes a set of statistics for a fixed window of time and then slides the window across the data by a specified interval in order to calculate the next set of statistics.

5. Training and Testing of Machine learning algorithms: By having high quality data I will be able to train the models in much fewer epochs. In this thesis I will be promoting less higher quality data over a lot of misleading and incorrect data.

- Fourth step involves the actual implementation of the machine learning models based on training it on the train test. I am going to be experimenting with various dif- ferent models and training each model until I can consider its results to be significant enough. In this step I am also going to be implementing the Technical Indicators and combining features and logic of them together in order to potentially get better results. The most crucial part will be the detection of the signals

and updating the financial data accordingly to the signals For this step I will be using libraries such as Scikit-learn, Tensorflow, Keras, or Backtesting

6. Performance Evaluation: In order to accurately evaluate my models I will be need to strategically determinge the best metrics to find the most optimal metric and score the models based on it. I will be using metrics such as: (Return on investment, Buy and hold return, Sharpe ratio, Win Trade, Accuracy, RMSE and many more)

    - The last step of my approach will be evaluating the performance of my Technical Indicators against the series of metrics metrics mentioned before The metrics mentioned below were picked as they represent the most important aspects of the financial market and are the most important ones to consider when evaluating the performance of the models.
        - Return on investment: Return on investment (ROI) is a financial ratio used to calculate the benefit an investor will receive in relation to their investment cost. It is most commonly measured as net income divided by the original capital cost of the investment. The higher the ratio, the greater the benefit earned.
        - Buy and hold return: Buy and hold is a passive investment strategy in which an investor buys at the current market price a financial instrument and holds it for a long period of time, regardless of fluctuations in the market. An investor who uses a buy-and-hold strategy actively selects investments but has no concern for short-term price movements and technical indicators. This metric is used to compare the performance of the models against the Technical Indicators
        - Sharpe ratio: The Sharpe ratio is a measure of risk-adjusted return, which compares an investment's excess return to its standard deviation of returns. The higher the Sharpe ratio, the better the investment's historical risk-adjusted performance.
    - Evaluation of Metrics on the Machine Learning models is going to be different as when using Technical Indicators. While using Machine learning models I will be measuring the accuracy of the model and the RMSE (Root Mean Square Error) of the model and R2 score.
        - Accuracy: Accuracy is the most obvious performance measure and it represents a ratio of correctly predicted observation to the total observations. What accuracy represents in time-series data is the percentage of correct predictions that the model made over time and is therefore a good metric to use when evaluating the performance of the model.
        - RMSE: RMSE is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are;

RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit.

  - R2 score: R2 score is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.

1.5 Structure

This thesis will be composed out of 2 parts.

- Theoretical part will be focused on theory and explanation of my approach and how each of the algorithms works with the trend pattern detection. It will require a lot of research and studying of the topic and than Mathematical explanation of the algorithms and their logics. The result of this part will be a sequential analysis of used Indicators and Machine Learning models. The second part will be a full
- Practical part will be focused on the implementation of the Machine Learning models and Technical Indicators. From data loading through data cleaning and processing to the actual implementation of the models and indicators. The result of this part will be a set of Juptyer notebooks that will be used to display the results of my research and compare the performance of the models and indicators. The technology stack I will use in this project will be explained in the further stages of this research.

Chapter 2
BACKGROUND

In this Chapter, we present theoretical background of related domains and technologies to set up a ground for further discussion

2.1 Time series forecasting

Time series forecasting [**?**] is a statistical technique that predicts future values or trends using past data points collected at regular intervals. It is frequently utilized in many different sectors, including finance. Economic forecasting, weather forecasting, and sales forecasting are all examples of forecasting. The basic premise of time series forecasting is that historical patterns and behaviors in data can provide insights into future patterns. The goal is to find underlying patterns, trends, and seasonality in time series data and utilize that knowledge to create accurate forecasts about future values. In time series forecasting we have several topics we need to look a bit deeper into so we understand this topic more in depth.

- Stationarity Stationarity is assumed for time series data, which means that its statistical features remain constant across time. This comprises a constant mean, constant variance, and autocovariance that is only affected by time lag. Many time series models rely on stationarity to produce accurate forecasts because they require steady statistical features. Stationarity can be spotted by comparing means and RMSE in various different places and if these 2 values do not different in many instances we can claim the data is stationary

- Smoothing Smoothing techniques are used to reduce noise and volatility from time series data, making underlying patterns more visible. Smoothing is frequently accomplished using moving averages and exponential smoothing methods (such as

simple exponential smoothing or double exponential smoothing). By removing noise and volatility from our data we achieve clearer separation in between trends allowing us to make predictions with higher accuracy This step can be compared with removing the outliers, however removing outliers is mostly used to delete the most extreme pieces of data. On the other hand smoothing focuses more on reducing the noises and is more gentle with data deletion.

- Seasonality Analysis Many time series display repeating patterns within a given time period, which is referred to as seasonality. Seasonality can occur on a daily, weekly, monthly, or other basis. Seasonality analysis entails detecting and modeling these periodic patterns distinct from the trend component. For this aim, techniques such as seasonal decomposition of time series (such as classical decomposition or STL decomposition) or Fourier analysis can be used. The topic of seasonality is going to be used very widely in this research as the principle of trend analysis and trend patterns work on recurring patterns and parent patterns (Pattern composed out of multiple patterns)

2.2 Technical analysis

Technical analysis [**?**] is a method of evaluating financial markets that relies on the historical behavior of price. It is examining prior market activity and focuses on identifying patterns, support and rejection levels. Principle of technical analysis is to extract useful information from data and make wise decision based upon it Technical analysis is based on the idea that market behavior is reflecting human psychology and human emotions. Technical analysis takes into account emotions like fear, greed and optimism and it has deducted various different trend patterns that seem to be natural for people. Technical analysis also focused on the study of price charts in order to uncover repeating patterns that might assist anticipate future price changes. These patterns can be as simple as trend lines and support/resistance levels, or as complicated as chart patterns such as head and shoulders or triangles. These patterns are said to provide insights into the market's supply and demand balance as well as the psychological dynamics between buyers and sellers. Technical analysis is going to compose a large part of this thesis and I will be looking deeper into these patterns and trying to use data in which these patterns have been spotted as a traindata for my algorithms

2.2.1 Support and Resistance levels

Support and Resistance levels are an essential part of technical analysis and are a great indicator and insight on the current state of data. The principle of them is to bound already

explored regions in which the trend tends to vary. This bounded zone is capped from the top by a resistance level

Support Levels

Support levels are levels at which demand for a security is strong enough to keep it from falling further are referred to as support levels. As buyers step in and produce enough demand to counteract the selling pressure, they act as a floor or "support" for the price. When the price approaches a support level, it is expected to rebound and rise once more. Support levels can be found by looking for regions where the price has previously reversed its downward trend and begun to rise higher. These levels frequently correspond to prior lows or price consolidation zones on the price chart. Support levels are commonly used by traders to identify potential entry points for purchasing or opening long positions as there is a high probability the trend will reverse and they have a high potential of capitalizing on this

Resistance Levels

Resistance levels, on the other hand, are price levels at which a security's supply is sufficiently robust to prevent it from increasing higher. They operate as a price cap or "resistance" as sellers become more active and outnumber buyers. When the price reaches a resistance level, it is expected to receive selling pressure and possibly reverse its upward trend. Resistance levels are identified by looking for areas where the price has previously encountered selling pressure and reversed its upward trend. These levels often correspond to previous highs or consolidation zones on the price chart. Traders typically use resistance levels to identify potential exit points for selling or taking profits from long positions.

2.3 ARIMA

Autoregressive moving average otherwise known as ARIMA is a Machine learning model used for time-series forecasting. It is used to analyze and predict data points based on their temporal dependencies. It is a combination of three components: autoregression (AR), differencing (I), and moving average (MA).

2.3.1 Autoregression (AR)

Autoregression [**?**] is the process of modeling a variable based on its previous values. The "AR" component in ARIMA represents the relationship between an observation and a set number of lagged observations (i.e., the variable's historical values). The "p" parameter reflects the number of lag observations taken into account by the model. Its mathematical representation is the following:

$AR(p) : y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t$
Where:

- $\Delta y(t)$ represents the differenced value at time t.
- $y(t)$ denotes the original value at time t
- $y(t-1)$ represents the value at the previous time step, $t-1$.

  By applying this differencing formula to each observation in the time series, we can obtain a new series of differenced values. The order of differencing refers to the number of times this process is applied.

2.3.2 Differencing (I)

Differencing is a technique that is incorporated inside of the algorithm ARIMA and its purpose is to flatten out the time series and adjust it in order for it to be stationary. Stationary can be represented as low volatility of mean and variance of a specific amount of time. Differencing is then useful for eliminating any patterns or seasonality. A simple formula for differencing could be represented as:
$\Delta y(t) = y(t) - y(t-1)$
Where:

- $\Delta y(t)$ represents the differenced value at time t.
- $y(t)$ denotes the original value at time t.
- $y(t-l)$ represents the value at the previous time step, $t-l$.

2.3.3 Moving Average MA

The “MA” competent is representing the dependency in between the error term at a certain time point and error time in previous time points. Moving average also uses a parameter named “q”. Its sheer purpose is to represent the number of lagged terms considered in this model. Mathematical formula in order to calculate the moving average for a specific time frame is:
$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q}$

By applying this differencing formula to each observation in the time series, we can obtain a new series of differenced values. The order of differencing refers to the number of times this process is applied.

Chapter 3
Related Work

In this chapter of my thesis I will be focusing on exploring and reviewing already existing academic and commercial knowledge. I will be working closely with academic research and precisely studying it in order to distinguish high quality pieces of information from potential false studies. There is a high potential that some pieces of academic literature might contain incorrect results as the process of data preparation and cleaning could heavily skew the data towards a certain outcome and the party might not be even aware of this mistake. Therefore I am going to be using only studies that seem to have a feasible approach and have not modified the outcome to fit the purpose of their studies.

3.1 Academic Research

Machine learning is a very hot topic in the year 2023 and its combination with the well known stock market is a prevalent combination of the past few years. I have been able to find study such as[**?**] that have focused on in-depth analysis of performance of certain algorithms, however it is not directly what I will be researching . The research of deep learning algorithms used to predict the financial [**?**] were very niche and well performed. I am definitely going to use the already existing findings from the author's essay in order to gain better insight on the current problematic of time series forecasting with deep learning algoriths. The thesis [**?**] has really impacted the way I started looking on this topic as it provides great results and graphs. The challenge it opened is to improve stock trading judgments, PRML, a revolutionary candlestick pattern recognition model based on machine learning methods, is proposed. To begin the pattern recognition schedule, four prominent machine learning methods and 11 different feature types are applied to all potential combinations of daily patterns. To detect the prediction effect at different times, several time windows ranging from one to ten days are used. An investment plan is built around the recognized

candlestick patterns and time span. The PRML model was used on the Chinese market stock for the dates Jan 1, 2000 until Oct 30, 2020. These 20 years were splitted into training and testing data with a ratio 75-25 . This research was using price forecasting with multiple models including PRML. They have concluded that “Empirical results show that the two-day candlestick patterns after filtering have the best prediction effect when forecasting one day ahead” and that applying machine learning methods to two day and three day patterns for one day ahead forecasts can be profitable

Possibly one of the most insightful studies is a thesis named “Machine learning techniques and data for stock market forecasting: A literature review”.[**?**] Its purpose was analyzing multiple machine learning studies and conducting their results. The whole thesis was comparing the performance and quality of the already existing sources. By analyzing a significant amount of studies they have been able to point out which approaches have been performing very well, such as: SVM (Support Vector Machine), or ANN (Artificial neural network) or Fuzzy Theory. These studies were conducted on different markets in order to diversify the machine learning models and also introduce potential differences based on geographical locations of the markets. This study has found out that by using neural networks on the American stock market S&P 500 and as input of indices data and google trends, they were able to achieve over 86.% on a Hit Ratio. These results can be categorized as significant and therefore giving us a good understanding on which models and approaches am I going to focus on in this thesis.

Researchers conducting a study, [**?**] have conducted an in depth research on the long term performance of the already mentioned ARIMA model. They have been focusing on how to properly find the best fit and explained what AICc values are we looking for in order to find the model of best fit and then trained this model on different time series starting from 6 months and up to 23 months. They have conducted testing on 56 different stocks in 6 different industries however information important are the following findings. They have achieved the highest accuracy with the model in the sector Fast Moving Consumer goods of 96% and the standard deviation of 2.03 for the 6 months. The lowest accuracy was in the banking sector with accuracy of 85% and standard deviation also for the 6 months period of 15.7. This standard deviation has almost halved to 8.23 after training the model on 23 months of training data. They have also conducted that the p-values for all possible combinations are high, hence rejecting the null hypothesis is not possible. The null hypothesis will be accepted, which is, the changes in the accuracy for different sizes of training datasets is not significant.

Chapter 4
METHODOLOGY

Data collection and preparation of the data is the start of this research and I have opted to be looking at the publicly traded stock market of currencies. My main interest is going to be the trading pair of EUR/GBP as it is among the most traded pairs in the world and its movement heavily impacts the European Market. The goal of this research is trying to reject the null hypothesis that it is not possible to predict the future price of the stock market and capitalize on it.

4.1 Implementation

My Implementation is going to consist of 3 main parts. Each consisting of multiple steps.

- Data Science
- Technical Indicators
- Machine Learning Models

4.1.1 Data Science

In the section of data analysis I will be focusing on the analysis of the data and trying to gain information about the data via visualizations and statistical analysis. I will be using libraries such as Pandas, Numpy, Matplotlib, Plotly, or Bokeh in order to create visualizations and graphs that will help me understand the data better. I will be looking for patterns and analysing things as: When is the busiest time of the day? Does it have any effect on the price and so on.

- Data Loading: In this step I will be loading the data from the source and will be trying to get the data in the most raw form possible.

- Data Cleaning: In this step I will be trying to find any potential outliers or missing values and will be trying to get rid of them.
- Data Analysis: In this step I will be trying to find any potential patterns and insights that could help me with the next steps.

4.1.2 Technical Indicators

In this section I will be focusing on the implementation of the Technical Indicators and trying to find the best ones that will be used in the next steps. The Technical indicators I am going to use are for example: RSI, MACD, VOLD, Bollinger Bands,Rolling Windows, or Moving Averages and many more.

- Data Preparation: The data used to input into my Technical Indicators will be the data from the previous step and will be used to create new features.
- Implementation of Technical Indicators: In this step I will be implementing the Technical Indicators by at first thoroughly studying them and then extracting their logic and implementing them in Python.
- Signal Detection: In this step I will be capturing the signals that are going to be used in the next steps. These signals will be used to update the financial data and indicate the time to buy or sell.
- Evaluation of singular Indicator: This step will be focused on the evaluation of every single one of the Technical Indicators and visualising their performance.

4.1.3 Machine Learning Models

In this section I will be focusing on the implementation of the Machine Learning models and trying to determine their performance and accuracy. The Machine Learning models I am going to use are for example: ARIMA, Random Forest, XGBoost, Support Vector Regression, Linear Regression, or LSTM.

- Data Preparation: The data used to input into my Machine Learning models will be the data from the previous steps however it wont contain the signals. This data will needed to be prepared in order to be used in the Machine Learning models. As we need to train the models we need to split the data into train and test data. The train data will be used to train the model and the test data will be used to test the model and evaluate its performance.
- Implementation of Machine Learning models: In this step I will be implementing already mentioned Machine Learning models on my time series data. There is a large stigma around the Machine Learning models and their performance on time series

data as this is not a common use case for them. With my thesis I am trying to prove that Machine Learning models can be used for time series forecasting and can be used to predict the future price of the stock market.

- visualization: In this step I will be visualizing the predictions of the Machine Learning models and comparing them with the actual price of the stock market. With this step I wil be able to showcase their performance and give an insight on how well they are performing.
- Evaluation of singular Machine Learning model: This step will be focused on the evaluation of every single one of the Machine Learning models and representing their performance.

4.2 Evaluation metrics

In this section I will be focusing on the evaluation of the performance of the Machine Learning models and Technical Indicators. I will be using metrics such as: (Return on investment, Buy and hold return, Sharpe ratio, Win Trade, Accuracy, RMSE and many more)

- Return on investment: $ROI = \frac{Current\ Value - Original\ Value}{Original\ Value}$
- Buy and hold return: $Buy\ and\ hold\ return = \frac{Current\ Value - Original\ Value}{Original\ Value}$
- Sharpe ratio: $Sharpe\ ratio = \frac{R_p - R_f}{\sigma_p}$ => where R_p is the return of the portfolio, R_f is the risk-free rate of return, and σ_p is the standard deviation of the portfolio's excess return.
- Win Trade: $Win\ Trade = \frac{Number\ of\ winning\ trades}{Total\ number\ of\ trades}$
- Accuracy: $Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$
- RMSE: $RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$
- R2 score: $R2\ score = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$
- MAE: $MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n}$

4.2.1 Libraries and Technologies to be used

I believe it is important to mention the technologies and libraries I am going to be using in this thesis. My approach is not to reinvent the wheel and use already existing data processing and visualising libraries. This will allow me to focus on the main topic of this thesis and not get distracted by the implementation of the libraries. I will be using the following libraries and technologies:

- Python: Python is going to be the main programming language used in this thesis. It is a very popular programming language and is used heavily in the field of Data Science and Machine Learning.

- Jupyter Notebook: Jupyter Notebook is going to be the main IDE used in this thesis. Juptyer Notebook will allow me to create interactive report and visualizations that will be used to showcase the results of my research. As Juptyer Notebook allows me to run the code in blocks and store the results of each block it is going to be very useful for me to showcase the progress in each step of my research.

- Pandas: Pandas is a Python library used for data manipulation and analysis. It is used to manipulate and analyze data in a fast and efficient manner.

- Numpy: Numpy is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices.

- Matplotlib: Matplotlib is also a Python library used for creating static and interactive visualizations. It is a very popular library and is used in many different fields. For my thesis I will be using it to create visualizations of the data mainly in the form of graphs. And static analysis of the data.

- Plotly: Plotly is also a visualising library however it is used to create interactive visualizations. It is a very powerful library and supports a Candlestick charts which is going to be very useful for my thesis as almost all of my data will be in the form of Candlestick charts. (Open, High, Low, Close)

- Bokeh: Bokeh is also a visualising library however it is used to visualising my Backtesting results. By using Bokeh I will be able to track the performance of my indicators at any given momment and see how they are performing.

- Scikit-learn: Scikit-learn is a Machine Learning library for Python. It is among the most popular Machine Learning libraries and I will be heavily using it in my thesis. It contains many different Machine Learning models such as: Random Forest, Support Vector Regression, Linear Regression and many more.

- Backtesting: Backtesting is a Python library used for backtesting trading strategies. It is going to be used to evaluate the performance of my Technical Indicators. In my thesis I will create a class for a Strategy in which I will set TP(Take Profit), SL(Stop Loss) bounderies and the strategy will be able to buy and sell at any given time.

4.2.2 Backtesting

Backtesting is a process of testing a strategy on historical data to see how it would have performed in the past. This concept is often used to evaluate the effectiveness of a trading strategy and is a fairly simple process. How it works step by step:

1. Starting with a set amount of capital, we buy and sell assets according to the strategy we want to test.
2. We record the profits and losses we would have made for each trade.
3. We repeat this process for all trades made by the strategy during the time period we are backtesting.
4. We calculate the total profit or loss made by the strategy by adding up the profits and losses from all trades.
5. We compare the total profit or loss to the amount of capital we started with.

Backtesting as concept is also able to be represented Mathematicaly.

- The formula for calculating the total profit or loss could be represented as:

$$\text{TotalProfitorLoss} = \sum_{i=1}^{n} (Sell\ Price_i - Buy\ Price_i) \tag{4.1}$$

- Therefore the formula for calculating the final capital could be represented as:

$$\text{FinalCapital} = \text{InitialCapital} + \text{TotalProfitorLoss} \tag{4.2}$$

Chapter 5
EXPERIMENTATION

This chapter will be focused on displaying and experimenting with my proposed Technical Indicators and Machine Learning models. Each of the experimentation parts will be focused on a single Technical Indicator or Machine Learning model and will be displaying its performance and accuracy. I will be following the steps outlined in previous sections and will be trying to follow the same structure in order to assure a coherent structure of my research.

5.1 Data Science

5.1.1 Data Loading

The process of Data Loading is starting with finding a trust-worthy source from which I could download the data I am going to be using in my research. As already mentioned I am going to be working on the trading pair of EUR/GBP and a source from which I will be downloading the data is a Swiss Banking group called Dukascopy. Dukascopy have provided me with this historical data in a form of CSV file. The data is in the form of Candlestick charts and contains the following columns: Date, Time, Open, High, Low, Close, Volume. I have been able to get the data in 4 hour intervals which is perfect to determine the trend of the market during daily to weekly intervals.

5.1.2 Data Cleaning

Data cleaning was not particularly difficult as the data was already in a very clean form. As the data provided by Dukascopy was a global data which is traceable backwards it did not contain any missing values. Also in terms of outliers there is no such a concept in the stock market as the price of the market is always changing and there is no such a thing as a price that is too high or too low. The only thing that could be considered as an outlier is an

incorrect value that the data provider could have secretly added to the data. However this is not the case and the data is very clean and almost ready to be used. The only thing that had to be done is renaming the column values as there was an initial issue which caused problems

5.1.3 Data Analysis

I have created a separate Juptyer Notebook ($volume_time_analysis.ipynb$) that contains an analysis of what is the busiest hour during the day and what is the busiest hour during the day and if it somewhow correlates with the price of the stock market. I have been able to find out that the busiest hour during the day and by the chart we can see that the busiest hours are: 11:00 and 15:00. (GMT+2) This is a very interesting finding and also very reasonable as these are the times when London Stock Exchange opens and closes. The London Stock Exchange is the largest stock exchange in Europe and the 3rd largest in the world. Its opening and closing times are 8:00 - 16:30 (GMT+1) and therefore the busiest hours are 11:00 and 15:00 (GMT+2). Its sheer impact on the stock market can easily move the price of the pair (EURGBP). Hedge funds and other large financial institutions are also very active during these hours and therefore it is very reasonable to see this high trading volume during these hours. High volume of trading can change the price however if we have the same amount of buys than sells the price will not be going one way. The market needs to make an "internal agreement" in order to determine the direction it is going to be heading. This theory can also be seen on the actual line chart of how the price moves during the day. In certain cases like $INESERTHEREEEE$ we can see that price is moving in a certain direction and then suddenly it changes its direction and starts moving in the opposite direction. This is the same concept as the "internal agreement" and it is very interesting to see that the price is moving in a certain direction and then suddenly changes its direction. However it could be explained by the opening of the market and the sheer volume of trading that is happening during these hours is able to change the direction of the market.

5.2 Machine Learning models

5.2.1 Neural Network (LSTM)

5.2.2 ARIMA

5.2.3 Random Forest

5.2.4 XGBoost

5.2.5 Support Vector Regression

5.2.6 Linear Regression

5.3 Technical Indicators

5.3.1 RSI

5.3.2 MACD

5.3.3 VOLD & Bollinger Bands

5.3.4 Rolling Windows

5.3.5 Moving Averages

Chapter 6
EXAMPLE CHAPTER

This is only an example of a chapter! Anyways, all thesis should have a problem statement – not necessarily as a separate chapter though. Only after you know the problem, it will be possible for you to evaluate the results of what you did. If you want to see examples of evaluations, have a look at how graph visualizations are evaluated here [1].

6.1 Code and syntax highlighting

You may sometimes want to add code snippets to your thesis. You can do so by using `lstlisting`. Use this with care, as code should not be extensively presented in the thesis. Here is an example.

```
def addition ():
    print("I am adding numbers here!")
    n = float(input("Enter the number: "))
    t = 0 // Total number enter
    ans = 0
    while n != 0:
        ans = ans + n
        t+=1
        n = float(input("Enter another number (0 to end): "))
    return [ans,t]
```

6.2 Labels and References

See chapter 1 for interesting stuff and see a cool logo in Figure 6.1. If you are still not convinced, try adding a footnote[1]. Its easy to add citations, just use a bibtex file to list your references and cite them here like this [3]. If you want to read a cool paper [2], just contact the author of the paper. Haha, that was funny!

6.3 Mathematical Equations and Expressions

Basic equations in LaTeX can be easily "programmed". Fermat's Last Theorem (sometimes called Fermat's conjecture, especially in older texts) states that no three positive integers a, b, and c satisfy the equation

$$a^n + b^n = c^n$$

for any integer value of n greater than 2. The cases $n = 1$ and $n = 1$ have been known since antiquity to have infinitely many solutions. And because its so much fun, here is an integral for you - thank me later!

$$\int_0^1 x^2 + y^2 \, dx$$

Do you want a more complex formula, I have no idea what it means, but it looks pretty.

$$\oint_{i=1}^{n} \sum_{i=1}^{\infty} \frac{1}{n^s} = \prod_p \frac{1}{1 - p^{-s}}$$

6.4 Enumerations and Descriptions

Here is a simple list:

1. The labels consists of sequential numbers.
2. The numbers starts at 1 with every call to the enumerate environment.

Here is another list:

1. The labels consists of sequential numbers.
    - The individual entries are indicated with a black dot, a so-called bullet.
    - The text in the entries may be of any length.

[1] did you like it?

2. The numbers starts at 1 with every call to the enumerate environment.

Maybe such descriptions are also useful. These look neat to me. What do you think? Oh, I forgot, this document is not a tutorial.

Short This is a shorter item label, and some text that talks about it. The text is wrapped into a paragraph, with successive lines indented.

Rather longer label This is a longer item label. As you can see, the text is not started a specified distance in – unlike with other lists – but is spaced a fixed distance from the end of the label.

6.5 Adding images

Adding a simple image is easy. Adding complex images is also easy. What is a complex image anyway?

Figure 6.1: Old IMC Logo

(a) New Logo since 2023

(b) New IMC Logo since 2023

Figure 6.2: Including sub images!

6.6 Colors

1. IMC Blue

```
\textcolor{imcblue}
```

2. IMC Color for Science and Technology

```
\textcolor{imctech}
```

3. IMC Corporate Color

```
\textcolor{imcgray}
```

4. IMC Corporate Color 2

```
\textcolor{imcorange}
```

6.7 Just a poem by Emily Dickinson

I'm nobody! Who are you?
Are you nobody, too?
Then there's a pair of us — don't tell!
They'd banish us, you know.
How dreary to be somebody!
How public, like a frog
To tell your name the livelong day
To an admiring bog!

6.8 Tables

Country List			
Country Name or Area Name	ISO ALPHA 2 Code	ISO ALPHA 3 Code	ISO numeric Code
Afghanistan	AF	AFG	004
Aland Islands	AX	ALA	248
Albania	AL	ALB	008
Algeria	DZ	DZA	012
American Samoa	AS	ASM	016
Andorra	AD	AND	020
Angola	AO	AGO	024

Table 6.1: Example table

BIBLIOGRAPHY

[1] M. Burch, W. Huang, M. Wakefield, H. C. Purchase, D. Weiskopf, and J. Hua, "The state of the art in empirical user evaluation of graph visualizations," *IEEE Access*, vol. 9, pp. 4173–4198, 2021. [Online]. Available: https://doi.org/10.1109/ACCESS.2020.3047616

[2] D. Dhungana, A. Haselböck, and S. Wallner, "Generation of multi-factory production plans: Enabling collaborative lot-size-one production," in *46th Euromicro Conference on Software Engineering and Advanced Applications, SEAA 2020, Portoroz, Slovenia, August 26-28, 2020*. IEEE, 2020, pp. 529–536. [Online]. Available: https://doi.org/10.1109/SEAA51224.2020.00088

[3] B. Huettner, "The elements of technical writing (2nd ed.) - book review," *IEEE Transactions on Professional Communication*, vol. 45, no. 1, pp. 59–60, 2002.

Appendix A

EXAMPLE APPENDIX 1

Appendices should be used for supplemental information that does not form part of the main research. Remember that figures and tables in appendices should not be listed in the List of Figures or List of Tables.

Appendix B

EXAMPLE APPENDIX 2

Appendices should be used for supplemental information that does not form part of the main research. Remember that figures and tables in appendices should not be listed in the List of Figures or List of Tables.