TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Школа Инженерная школа информационных технологий и робототехники
Направление подготовки 09.04.04 Программная инженерия
ООП/ОПОП Технологии больших данных
Отделение школы (НОЦ) Информационных технологий

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА МАГИСТРАНТА

Тема работы
Анализ временных рядов и прогнозирование цен на золото (XAUUSD) с использованием машинного обучения **Predicting the future of time series data of gold prices (XAUUSD) using machine learning**

УДК **004.85:303.446.33:336.743.22**

Обучающийся

Группа	ФИО	Подпись	Дата
8ПМ1И	Халил Марко Эбрахим Тхабет		10.06.2023 г.

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Аксёнов С. В.	к.т.н.		10.06.2023 г.

КОНСУЛЬТАНТЫ ПО РАЗДЕЛАМ:

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОСГН ШБИП	Спицына Л. Ю.	к.ф.н.		

По разделу «Социальная ответственность»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ООД ШБИП	Антоневич О. А.	к.б.н.		

ДОПУСТИТЬ К ЗАЩИТЕ:

Руководитель ООП, должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Губин Е. И.	к.ф.-м.н.		

Томск – 2023 г.

ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОСВОЕНИЯ ООП
по направлению 09.04.04 «Программная инженерия»

Код компетенции	Наименование компетенции
Универсальные компетенции	
УК(У)-1	Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий
УК(У)-2	Способен управлять проектом на всех этапах его жизненного цикла
УК(У)-3	Способен организовывать и руководить работой команды, вырабатывая командную стратегию для достижения поставленной цели
УК(У)-4	Способен применять современные коммуникативные технологии, в том числе на иностранном (-ых) языке (-ах), для академического и профессионального взаимодействия
УК(У)-5	Способен анализировать и учитывать разнообразие культур в процессе межкультурного взаимодействия
УК(У)-6	Способен определять и реализовывать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки
Общепрофессиональные компетенции	
ОПК(У)-1	Способен самостоятельно приобретать, развивать и применять математические, естественно-научные, социально-экономические и профессиональные знания для решения нестандартных задач, в том числе в новой или незнакомой среде и в междисциплинарном контексте
ОПК(У)-2	Способен разрабатывать оригинальные алгоритмы и программные средства, в том числе с использованием современных интеллектуальных технологий, для решения профессиональных задач
ОПК(У)-3	Способен анализировать профессиональную информацию, выделять в ней главное, структурировать, оформлять и представлять в виде аналитических обзоров с обоснованными выводами и рекомендациями
ОПК(У)-4	Способен применять на практике новые научные принципы и методы исследований
ОПК(У)-5	Способен разрабатывать и модернизировать программное и аппаратное обеспечение информационных и автоматизированных систем
ОПК(У)-6	Способен самостоятельно приобретать с помощью информационных технологий и использовать в практической деятельности новые знания и умения, в том числе в новых областях знаний, непосредственно не связанных со сферой деятельности
ОПК(У)-7	Способен применять при решении профессиональных задач методы и средства получения, хранения, переработки и трансляции информации посредством современных компьютерных технологий, в том числе, в глобальных компьютерных сетях
ОПК(У)-8	Способен осуществлять эффективное управление разработкой программных средств и проектов
Профессиональные компетенции	
ПК(У)-1	Способен к созданию вариантов архитектуры программного средства
ПК(У)-2	Способен разрабатывать и администрировать системы управления базам данных
ПК(У)-3	Способен управлять процессами и проектами по созданию (модификации) информационных ресурсов
ПК(У)-4	Способен проектировать и организовывать учебный процесс по образовательным программам с использованием современных образовательных технологий
ПК(У)-5	Способен осуществлять руководство разработкой комплексных проектов на всех стадиях и этапах выполнения работ

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Школа Инженерная школа информационных технологий и робототехники
Направление подготовки 09.04.04 Программная инженерия
ООП/ОПОП Технологии больших данных
Отделение школы (НОЦ) Информационных технологий

УТВЕРЖДАЮ:
Руководитель ООП
_____ _____ Губин Е. И.
(подпись) (дата) (Ф.И.О.)

ЗАДАНИЕ
на выполнение выпускной квалификационной работы

В форме:

магистерской диссертации

(бакалаврской работы, дипломного проекта/работы, магистерской диссертации)

Студенту:

Группа	ФИО
8ПМ1И	Халил Марко Эбрахим Тхабет

Тема работы:

Анализ временных рядов и прогнозирование цен на золото (XAUUSD) с использованием машинного обучения Predicting the future of time series data of gold prices (XAUUSD) using machine learning	
Утверждена приказом директора (дата, номер)	№ 37-58/с от 06.02.2023 г.

Срок сдачи студентом выполненной работы:	10.06.2023 г.

ТЕХНИЧЕСКОЕ ЗАДАНИЕ:

Исходные данные к работе *(наименование объекта исследования или проектирования; производительность или нагрузка; режим работы (непрерывный, периодический, циклический и т. д.); вид сырья или материал изделия; требования к продукту, изделию или процессу; особые требования к особенностям функционирования (эксплуатации) объекта или изделия в плане безопасности эксплуатации, влияния на окружающую среду, энергозатратам; экономический анализ и т. д.).*	The objective of the study is predicting the future prices of gold by study the historical data for 19 years 2004 to present for 1H period. A data pipeline was used to give 100% accuracy data. With the ARIMA model, I obtained a profitable strategy.

Перечень подлежащих исследованию, проектированию и разработке вопросов *(аналитический обзор по литературным источникам с целью выяснения достижений мировой науки техники в рассматриваемой области; постановка задачи исследования, проектирования, конструирования; содержание процедуры исследования, проектирования, конструирования; обсуждение результатов выполненной работы; наименование дополнительных разделов, подлежащих разработке; заключение по работе).*	1. Statistics tests to check time series predictability; 2. Apply power analysis for significant sample size; 3. Overview of data analysis methods; 4. Research of machine learning models; 5. Evaluation of the accuracy of the models; 6. Drive a profitable strategy; 7. Work on the section on financial management, resource efficiency, and resource conservation; 8. Work on the section on social responsibility.
Перечень графического материала *(с точным указанием обязательных чертежей)*	1. Data pipeline structure. 2. Candlestick structure. 2. Dukascopy historical data download interface. 3. Boxplot and histogram for gold returns. 4. pie chart for labels. 4. models' predicted vs actual chart. 5. residual analysis. 5. Ishikawa diagram. 6. Gantt chart.

Консультанты по разделам выпускной квалификационной работы
(с указанием разделов)

Раздел	Консультант
Основная часть	доцент ОИТ ИШИТР, к.т.н., доцент Аксёнов С. В.
Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	доцент ОСГН ШБИП, к.ф.н., доцент Спицына Л. Ю.
Социальная ответственность	доцент ООД ШБИП, к.б.н., доцент Антоневич О. А.

Дата выдачи задания на выполнение выпускной квалификационной работы в соответствии с календарным учебным графиком	1.03.2023 г.

Задание выдал руководитель ВКР:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Аксёнов С. В.	к.т.н., доцент		1.03.2023 г.

Задание принял к исполнению обучающийся:

Группа	ФИО	Подпись	Дата
8ПМ1И	Халил Марко Эбрахим Тхабет		1.03.2023 г.

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Инженерная школа Информационных технологий и робототехники
Направление подготовки (специальность) 09.04.04 Программная инженерия
Уровень образования магистратура
Отделение школы (НОЦ) Информационных технологий
Период выполнения весенний семестр 2022 /2023 учебного года

Форма представления работы:

магистерская диссертация

(бакалаврская работа, дипломный проект/работа, магистерская диссертация)

КАЛЕНДАРНЫЙ РЕЙТИНГ-ПЛАН
выполнения выпускной квалификационной работы

Срок сдачи студентом выполненной работы:	10.06.2023 г.

Дата контроля	Название раздела (модуля) / вид работы (исследования)	Максимальный балл раздела (модуля)
10.06.2023	Основная часть	70
10.06.2023	Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	15
10.06.2023	Социальная ответственность	15

СОСТАВИЛ:
руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Аксёнов С. В.	к.т.н.		

СОГЛАСОВАНО:
руководитель ООП

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Губин Е. И.	к.ф.-м.н.		

Abstract:

The final qualifying work consists of 119 pages, 67 figures, 28 tables, and 21 sources. The goal of this work is to predict the future prices of gold (XAUUSD) using machine learning methods.

Based on statistical analysis and hypothesis tests, I found that my data is predictable and contains some information. However, I need a larger training sample size to obtain significant results. The ARIMA model is the best model for predicting future prices.

I performed hyperparameter tuning and k-fold validation for all my models to validate the results. I collected historical data for the H1 period candlestick from 2004 to the present, which is approximately 19 years of data. After preprocessing the data, my training dataset consisted of 100610 samples, while my testing dataset consisted of 1408 samples.

My research also required testing Transformer models and using data augmentation. However, due to a lack of computational power, I was unable to carry out these tasks.

Definitions, Designations, Abbreviations:

Machine Learning (ML) is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence (AI) based on the idea that systems can learn from data, identify patterns, and make decisions with minimal human intervention.

Artificial Intelligence is intelligence demonstrated by machines, unlike the natural intelligence displayed by humans and animals, which involves consciousness and emotionality. The distinction between the former and the latter categories is often revealed by the acronym chosen.

Time Series. In mathematics, a time series is a sequence of data points indexed (listed or graphed) in time order. Most commonly, a time series is a sequence taken at equally spaced points in time. Therefore, it is a sequence of discrete-time data.

Candlesticks, also known as Japanese candlesticks, are a type of financial chart used to represent the price movements of an asset, such as a stock, commodity, or currency. The chart displays the opening and closing prices, as well as the highest and lowest prices for a given time period, such as a day, week, or month.

Candlestick patterns are a popular way for traders to analyze price movements and make trading decisions. They provide valuable information about market sentiment and can be used in combination with other technical analysis tools to identify potential trading opportunities. Some other popular candlestick patterns include the Doji and the Engulfing pattern.

Technical indicators are mathematical calculations or statistical tools used by traders and analysts to interpret and predict future price movements in financial markets, such as stocks, commodities, or currencies. These indicators are derived from historical price and volume data and are plotted on charts to provide insights into market trends, momentum, volatility, and potential reversal points.

A multivariate time series refers to a type of time series data where multiple variables or observations are recorded and observed at each time step. Unlike univariate time series, which consists of a single variable, a multivariate time series captures the dynamics and dependencies between multiple variables over time.

A univariate time series refers to a type of time series data that consists of a single variable or observation recorded at each time step. It represents a sequence of data points where each point corresponds to a specific time and captures the value of the variable of interest at that particular time.

Covariates, in the context of statistical analysis, are additional variables that are included in a statistical model alongside the main variables of interest. They are often used to control for or account for potential confounding factors that may influence the relationship between the main variables.

Autocorrelation refers to the correlation of a time series with its lagged values. It measures the degree of similarity or relationship between a variable and its past observations at different time lags. Positive autocorrelation indicates a positive relationship between past and current values, while negative autocorrelation indicates an inverse relationship. Autocorrelation is commonly used to identify patterns and dependencies in time series data.

White noise is a type of time series where the values are uncorrelated and have constant variance. In other words, each data point is independent and identically distributed (i.i.d) with a mean of zero and constant standard deviation. White noise is considered a random and unpredictable sequence of values and serves as a baseline for comparing the signal and noise components in a time series.

Stationary refers to a time series that exhibits consistent statistical properties over time. In a stationary series, the mean, variance, and autocovariance structure remain constant across different periods. This implies that the distribution of data does not change over time, and there are no long-term trends or seasonal patterns. Stationary time series are easier to model and forecast as they display stable characteristics.

Random walk theory posits that the future values of a time series are unpredictable and follow a random path. According to this theory, each future value depends solely on the previous value plus a random shock or innovation. In a random walk, the changes in the series are random and not driven by any systematic or predictable patterns. Random walk models are often used to represent and forecast certain types of financial and economic data, such as stock prices, exchange rates, and asset returns.

Joint Probabilities $\boldsymbol{p(x, y)}$ are just the probability of two things occurring at the same time.

Marginal probabilities $\boldsymbol{p(x)}$ are just the probability of one thing occurring.

Table of Contents

1 Theoretical part

1.1 Background and motivation for the study

Gold is a precious metal that has been used as a store of value and a means of exchange for centuries. The price of gold is influenced by a variety of factors, including supply and demand, geopolitical tensions, and economic indicators such as inflation and interest rates. As a result, gold is considered a safe-haven asset that investors often turn to in times of uncertainty.

In recent years, the use of machine learning algorithms for time series prediction has gained popularity in financial forecasting. The ability of these algorithms to analyze vast amounts of data and identify complex patterns makes them a promising tool for predicting the future prices of gold with greater accuracy than traditional forecasting methods.

Despite the potential benefits of machine learning in this area, there is still a need for further research to investigate the performance of different machine learning algorithms for predicting the price of gold. This study aims to address this gap in the literature by using machine learning algorithms to predict the future prices of gold (XAUUSD) with the best possible accuracy.

The gold market is a global market with a wide range of participants, including individual investors, central banks, and financial institutions. The gold market is highly liquid, meaning that there are many buyers and sellers of gold at any given time, and it is open 24 hours a day, five days a week. According to the World Gold Council, the average daily trading value of the gold market is estimated to be around $170 billion. This indicates that the gold market is one of the largest financial markets in the world, and it plays a significant role in the global economy.

Given the size and importance of the gold market, accurate forecasting of gold prices can have significant implications for market participants and policymakers. For example, investors may use price predictions to inform their investment decisions, while central banks may use them to manage their gold reserves and stabilize the financial system. Therefore, the motivation for this study is to provide insights into the performance of different machine learning algorithms for predicting the price of gold. This research is expected to contribute to the development of more accurate forecasting models, which can help investors and financial institutions make better-informed decisions in the gold market. Furthermore, accurate predictions of gold prices can also have significant implications for policymakers and regulators, who need to monitor the stability of the financial system and the potential risks associated with fluctuations in gold prices.

To achieve the objective of predicting the future prices of gold (XAUUSD) with the best possible accuracy, this study will investigate the performance of several machine learning techniques that are commonly used for time series prediction. These techniques include but are not limited to, autoregressive integrated moving averages (ARIMA), artificial neural networks (ANNs), and long short-term memory (LSTM) models.

ARIMA is a statistical model that has been widely used in the field of time series analysis and forecasting. This model is based on the assumption that future values of a time series are a function of its past values and random error terms. ANNs are a type of artificial intelligence algorithm that is effective in handling complex and non-linear relationships in financial data. These models can capture hidden patterns and correlations in data by using multiple layers of interconnected nodes. LSTM models are a type of recurrent neural network that is particularly effective at capturing long-term dependencies and patterns in time series data.

By comparing the performance of these different techniques, this study aims to identify the most accurate method for predicting gold prices as a time series. The results of this study are expected to provide valuable insights for investors, financial institutions, and policymakers, and to contribute to the development of more accurate forecasting models in the field of finance.

1.2 Research problem

Despite the importance of gold as a safe-haven asset and the potential benefits of machine learning algorithms for predicting its prices, there is a lack of comprehensive research that investigates the performance of different machine learning techniques for time series prediction of gold prices. This gap in the literature highlights the need for further research to identify the most accurate method for predicting the future prices of gold.

1.3 Research objectives

The main objective of this study is to predict the future prices of gold (XAUUSD) with the best possible accuracy using machine learning algorithms. To achieve this objective, the following research objectives will be pursued:

1. To review the existing literature on the application of machine learning techniques for time series prediction of financial assets, with a focus on gold prices.
2. To collect and preprocess historical gold price data (XAUUSD) from reliable sources and divide the data into training and testing sets.
3. To implement and evaluate the performance of different machine learning algorithms, including ARIMA, ANN, and LSTM models, in predicting the future prices of gold.
4. To compare the accuracy and efficiency of the different machine learning algorithms and identify the most accurate method for predicting gold prices as a time series.
5. To provide insights and recommendations for investors, financial institutions, and policymakers based on the results of the study, and to identify potential areas for future research in this field.

1.4 Research Questions

To address the research problem and achieve the research objectives, the following research questions will guide this study:

1. Does the historical data of gold prices (XAUUSD) exhibit patterns that can be used to predict future prices using machine learning algorithms?
2. Can the collected features, including historical gold prices, correlated currency pairs, and economic indicators be used to improve the accuracy of gold price predictions?

3. Is the size of the training sample sufficient to train the machine learning algorithms effectively, and how does the size of the training sample affect the accuracy of the predictions?
4. Can the application of machine learning algorithms to predict gold prices result in the development of a profitable trading strategy, and what are the implications of such a strategy for market participants?

The first research question aims to investigate the presence of patterns and trends in the historical data of gold prices that can be used to predict future prices. This question will be answered through the application of various statistics tests.

The second research question will explore the use of additional features, such as economic indicators, and correlated currency pairs to improve the accuracy of gold price predictions. This question will be addressed by applying some statistics tests, and incorporating these features into the machine learning models and evaluating their impact on the accuracy of the predictions.

The third research question will examine the impact of the size of the training sample on the accuracy of the predictions.

1.5 A brief overview of the Methodology of the research

The data used in this research include the historical hourly opening, high, low, and closing prices with the hourly trading volume of XAUUSD, as well as other relevant financial and economic indicators and correlated currency pairs that may influence the price of gold, The data is collected from reliable sources Dukascopy Swiss Banking Group, and pre-processed to ensure its quality and consistency. The data covers a certain period of time, such as the past 22 years, and is divided into training, validation, and testing sets. The training set is used to train the models, the validation set is used to tune the hyperparameters and avoid overfitting, and the testing set is used to evaluate the performance of the models.

1.6 Literature Review

1.6.1 Overview of the gold market and its participants

The gold market is a global market with a wide range of participants, including individual investors, central banks, financial institutions, and governments. It is considered one of the largest and most liquid financial markets in the world. The market operates 24 hours a day, five days a week, and its prices are influenced by various factors, including supply and demand, geopolitical tensions, economic indicators, and investor sentiment.

Individual investors are important participants in the gold market. They can buy gold in various forms, including bars, coins, and exchange-traded funds (ETFs). The demand for gold by individual investors is influenced by various factors, including economic uncertainty, inflation, and currency fluctuations.

Central banks are also significant players in the gold market. They hold gold as part of their foreign exchange reserves, which can help them stabilize their domestic currency and manage their balance of payments. Central banks may also buy and sell gold to manage their reserves and respond to changes in the global economic environment.

Financial institutions, such as investment banks and hedge funds, also participate in the gold market. They may buy and sell gold as part of their investment strategies, including hedging against inflation and currency risk. Financial institutions may also use gold as collateral for loans and other financial transactions.

Governments are another important participant in the gold market. They may hold gold reserves as part of their foreign exchange reserves, which can help them manage their currency and respond to economic shocks. Some governments also produce and sell gold as a source of revenue.

Overall, the gold market is a complex and dynamic system with a wide range of participants and factors influencing its prices. Understanding the behavior of these participants and the factors driving the market is essential for developing accurate forecasting models for predicting future gold prices.

1.6.2 Traditional forecasting methods for gold prices

Traditional forecasting methods for predicting gold prices have been widely used in the financial industry for many years. However, these methods often have limitations, such as difficulty in capturing complex patterns and trends in the data. As a result, there has been an increasing interest in the use of machine learning techniques for time series prediction, which has shown promise in capturing patterns and relationships that may not be apparent with traditional methods.

In this section, we will explore the use of machine learning algorithms for predicting gold prices, including the strengths and limitations of these methods. We will also discuss the different types of machine learning algorithms that have been used for gold price prediction, and provide an overview of their performance compared to traditional methods.

1. Fundamental analysis: This approach involves examining various economic, financial, and geopolitical factors that can affect the supply and demand for gold. For example, changes in interest rates, inflation, and currency values can impact the demand for gold as a safe-haven asset. Fundamental analysts also look at factors such as mining production, central bank buying and selling, and jewelry demand to understand the supply side of the market.
2. Technical analysis: This method involves analyzing charts and using various technical indicators to identify trends and patterns in the gold market. Technical analysts use tools such as moving averages, Bollinger Bands, and relative strength index (RSI) to make predictions about future price movements based on historical patterns.
3. Using indices like the S&P 500 and Dollar index: Some traders use other financial instruments or indices to forecast the direction of gold prices. For example, the S&P 500 index is often used as a proxy for overall market sentiment, and changes in the index can impact the demand for gold. Similarly, the Dollar index, which measures the strength of the US dollar against a basket of other currencies, can impact the price of gold as many investors use gold as a hedge against inflation and currency fluctuations.

4. Trading with economic calendar events: Economic events such as central bank policy announcements, GDP releases, and employment reports can impact the price of gold. Traders who follow these events closely may attempt to make predictions about the direction of gold prices based on the expected impact of the event on the broader economy and financial markets.
5. Trading with the news: Finally, some traders use news events and geopolitical developments to make predictions about gold prices. For example, news of a potential war or conflict can lead to a flight to safety and an increase in demand for gold. Similarly, political instability or changes in government policy can impact the price of gold.

1.6.3 Machine learning algorithms for time series prediction

Machine learning algorithms have emerged as a promising approach for time series prediction in recent years. These algorithms are designed to learn patterns and relationships in data and make predictions based on those patterns. Compared to traditional methods, such as fundamental and technical analysis, machine learning algorithms have the potential to capture more complex and subtle relationships in the data, resulting in more accurate predictions.

One of the most popular machine learning algorithms for time series prediction is the Recurrent Neural Network (RNN) model. RNNs are designed to handle sequential data, making them ideal for time series prediction. A specific type of RNN called the Long Short-Term Memory (LSTM) model, is particularly effective in capturing long-term dependencies in time series data.

Another type of machine learning algorithm that has shown promise in time series prediction is the Support Vector Regression (SVR) model. SVR is a type of supervised learning algorithm that uses support vectors to identify the boundary between different classes or categories. In time series prediction, SVR can be used to identify patterns in the data and make predictions based on those patterns.

Other machine learning algorithms that have been used for time series prediction include Random Forest, Gradient Boosting, and Artificial Neural Networks (ANNs). These algorithms have different strengths and weaknesses and may be more suitable for specific types of data or prediction tasks.

Overall, the use of machine learning algorithms for time series prediction in the gold market has the potential to provide more accurate predictions, leading to more informed investment decisions and better risk management. However, there is still a need for further research to investigate the performance of different algorithms and identify the best approach for predicting gold prices.

1.6.4 Review of previous studies on using machine learning for predicting gold prices

1. "Gold Price Prediction Using Machine Learning Techniques: A Case Study in India" by K. Srinivas and G. K. Patra (2018) — The study found that Random Forest and Gradient Boosting models outperformed other machine learning

models in predicting gold prices in India. The authors also suggested that additional variables related to macroeconomic indicators could improve the accuracy of the models.

2. "Forecasting Gold Price Using Multiple Linear Regression Model and Artificial Neural Network" by M. R. Hasani and M. F. Zarandi (2016) — The authors found that the Artificial Neural Network model performed better than the Multiple Linear Regression model in predicting gold prices. They suggested that incorporating additional variables such as interest rates, exchange rates, and inflation rates could improve the accuracy of the models.
3. "Predicting the Price of Gold Using Machine Learning Techniques" by N. T. Anh and L. T. Trang (2019) — The study compared the performance of three machine learning models (ARIMA, LSTM, and Random Forest) in predicting gold prices. The authors found that the LSTM model outperformed the other two models in terms of accuracy. They also suggested that the inclusion of additional economic indicators could further improve the accuracy of the models.
4. "Gold Price Prediction Using Ensemble Learning Based on Machine Learning Algorithms" by Y. Wang, L. Liu, and W. Lu (2020) — The study proposed an ensemble learning approach that combines multiple machine learning algorithms (Random Forest, Support Vector Regression, and Gradient Boosting) to predict gold prices. The authors found that their proposed approach outperformed individual machine learning models and traditional forecasting methods such as ARIMA and exponential smoothing. They suggested that incorporating additional variables such as political events and gold demand could further improve the accuracy of the models.

2 Methodology

2.1 Data collection and preprocessing procedures

2.1.1 Data Collection

The financial data is represented by something called candlesticks. Candlesticks, also known as Japanese candlesticks, are a type of financial chart used to represent the price movements of an asset, such as a stock, commodity, or currency. The chart displays the opening and closing prices, as well as the highest and lowest prices for a given period, such as a day, week, or month.

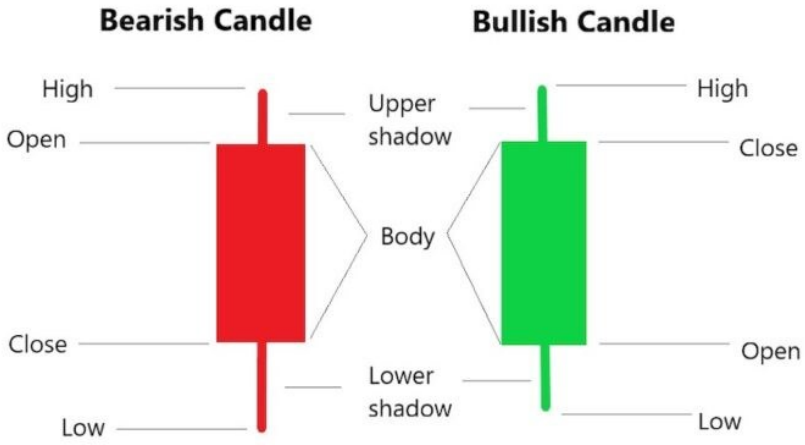

Figure 1 — Candlestick structure

Each candlestick is composed of a rectangle, called the body, and two thin lines protruding from the top and bottom of the body, called the wicks or shadows. The body represents the price range between the opening and closing prices for the period, while the wicks show the highest and lowest prices that occurred during that time.

Dukascopy Swiss Banking Group provides free high-quality Historical data for many financial instruments including many Commodities like Gold (XAUUSD).

The Historical data can be downloaded directly through their website. The interface is simple and you can use it without any instructions needed.

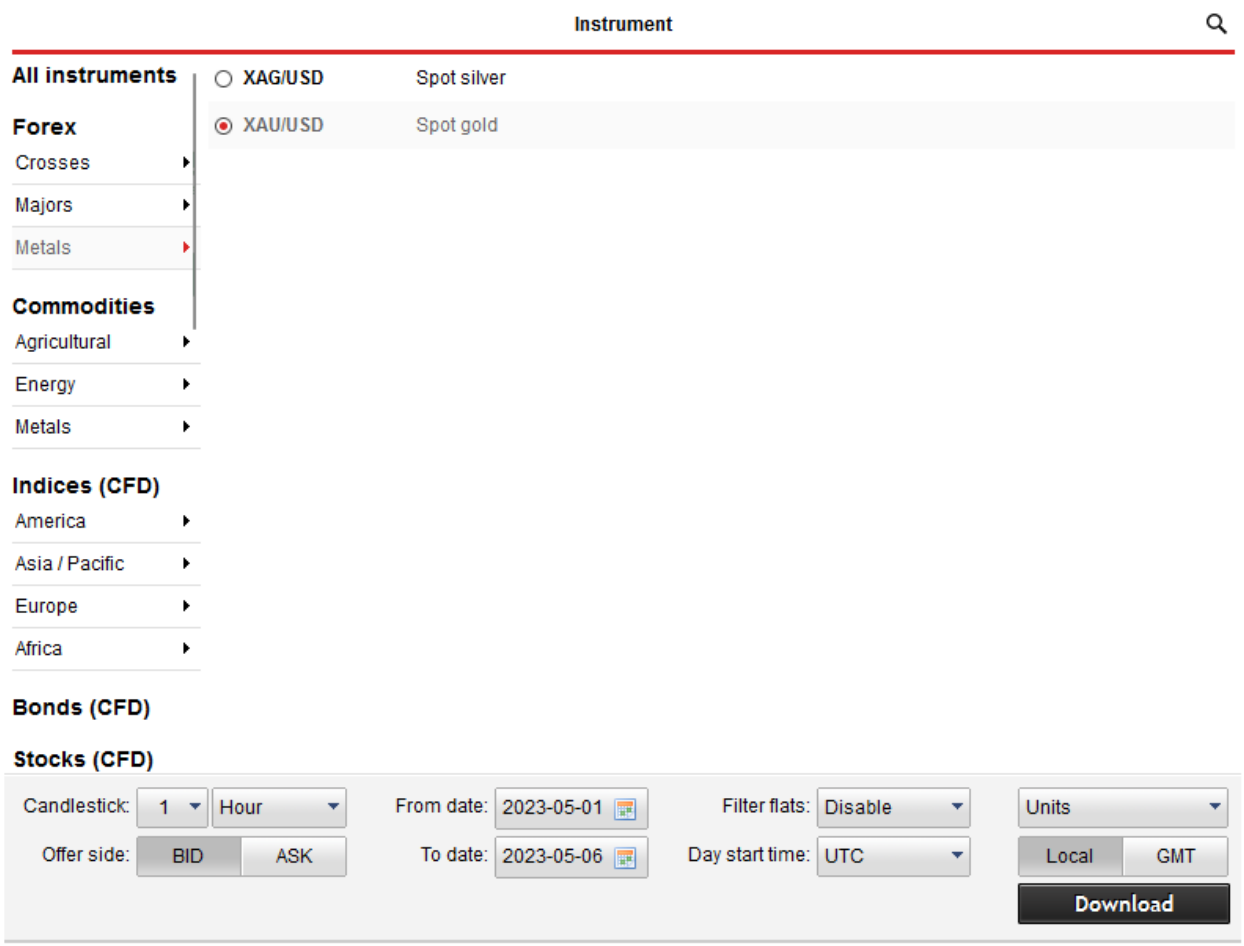

Figure 2 — Dukascopy Historical data download interface

I collected historical data from 2004 till now. about ~19 years of historical data. I collected different timeframes; My main timeframe is 1H and 5M, 15M, 30M, 4H, and 1D for further studies in the future. I followed the download interface to do so.

2.1.2 Data Integrity and solutions

Data integrity refers to maintaining the accuracy, consistency, and reliability of data throughout its lifecycle. This involves ensuring that data is not corrupted, lost, or accessible to unauthorized users. Maintaining data integrity is important for ensuring that data is trustworthy and can be used for decision-making purposes. Measures to ensure data integrity include implementing security protocols, backup and recovery processes, and data validation procedures.

After operating a network analysis for the download packages through the download interface, I conclude that many packages failed to be downloaded through **503** [1] error.

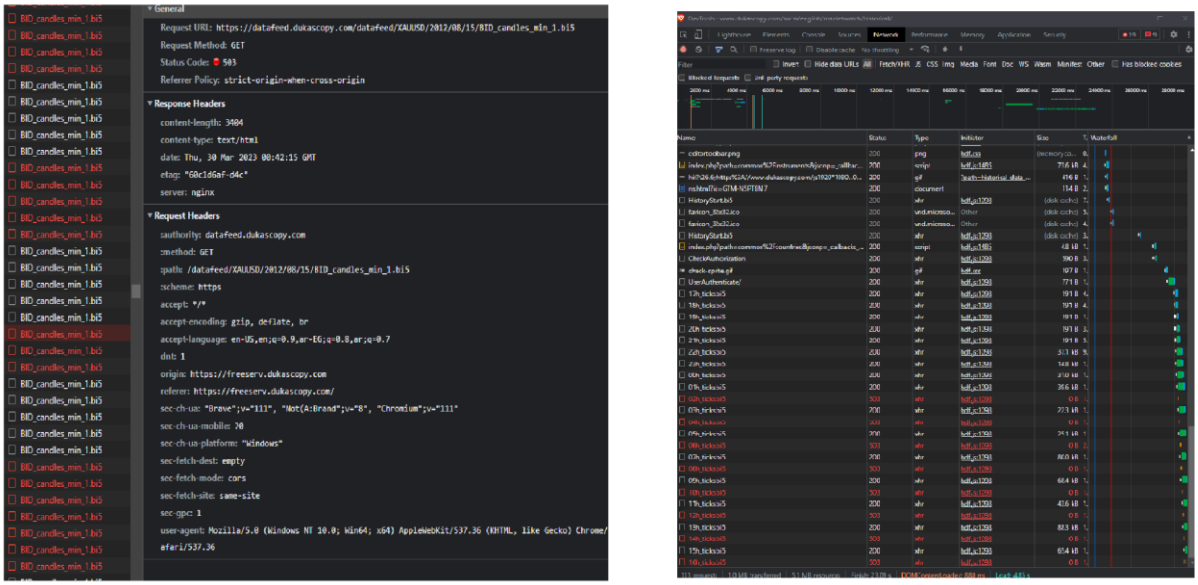

Figure 3 — Network analysis for the download interface

Along with my research for a solution I found, the Node.js library which can download the historical data and iterate n times for the missing packages unit fetching them.

The library is called the Dukascopy-Node library.

2.1.2.1 Dukascopy-Node library

The Dukascopy-Node library is an open-source Node.js library on GitHub that simplifies connecting to the Dukascopy trading platform API, a popular platform for forex trading. The library includes APIs for managing trades, historical data retrieval, and monitoring real-time market data. Users need a Dukascopy trading account and API key, which can be obtained from the Dukascopy website. Once the API key is obtained, users can install the library via npm and start using the APIs. The Dukascopy-Node library is a helpful tool for automating forex trading strategies with the Dukascopy trading platform API, making it easier for users to focus on implementing their strategies.

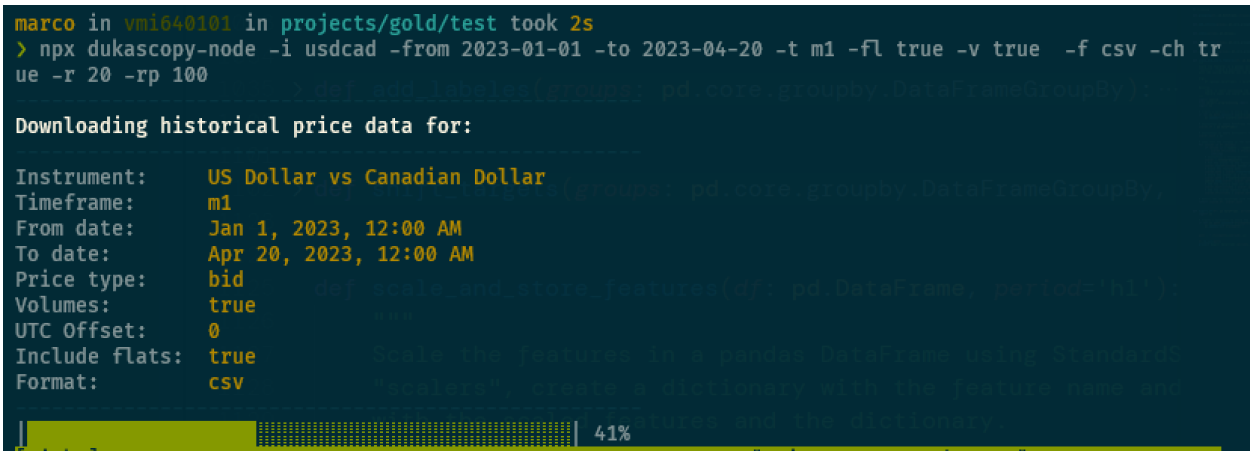

Figure 4 — Dukascopy-Node library CIL prompt

I run this library as a CIL prompt in the bash terminal as follows:

```
npx dukascopy-node -i xauusd -from 1999-01-13 -to 2023-03-26 -t tick -v true  -f csv -ch true -r 10 -rp 500 -d true
```

where *-rp 500* refers to the max iterate number for missing data.

2.1.3 Data Preprocessing

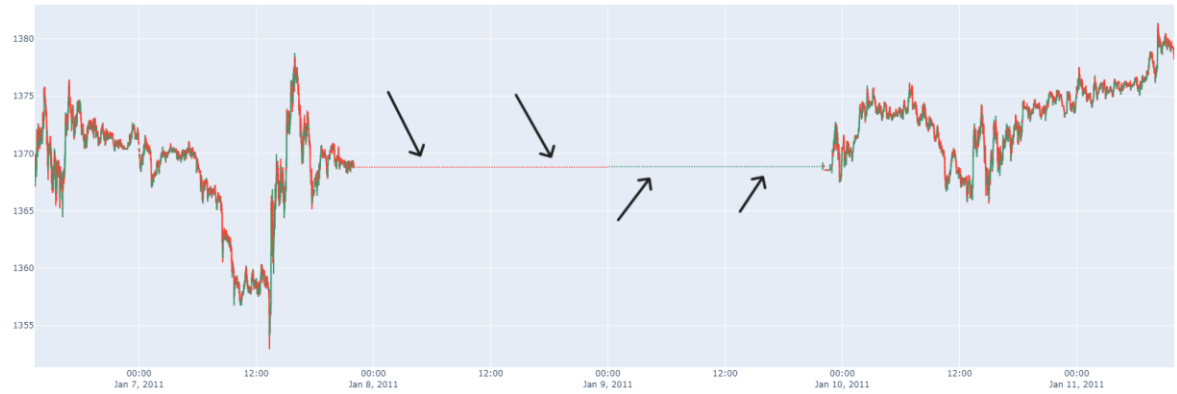

Figure 5 — Example of raw historical data showing the values of the flats.

The Preprocessing data which I download consists of the actual data with the flats, flats mean candlesticks with a volume of 0 for weekends when the market closes, and some missing data. As shown in figure 5

After applying cleaning processes in which I handle missing data, duplicated, and outliers, The final data obtained a quality of 89.5 % with corrupted frequency. The corruption of the frequency will affect every machine model I will create and will cause a huge drop in the model accuracy and my research integrity.

To obtain High-quality data with intact frequency, I developed a data pipeline to achieve my objective.

2.1.4 Data Pipeline

2.1.4.1 Flats Data

Figure 6 — Example of data Flats types

Before Building the data pipeline we need to understand the main concept of the missing data which consists in my dataset which are flats.

Flats data means a candlestick with no volume and there are three types of them:

1. **Weekend Flats** — these candlesticks represent the time at which the market is closed as shown in figure 5. these candlesticks have the same values and there is no trading volume. the forex market is a decentralized market that works 24/7 5 days a week
2. **Consistence Flats** — two or more duplicated candlesticks with no trading volume follow each other's create a series of data of the same value. I interoperated this as a fetal missing data occurred by network 503[1] error. As shown in Figure 6 (blue arrow).
3. **Single Flats** — these are a single value (there is no duplication) has no trading volume. And it represents the actual missing data value for my final data pipeline output and it's less than ≈ 0.4%.

2.1.4.2 Data Pipeline Structure

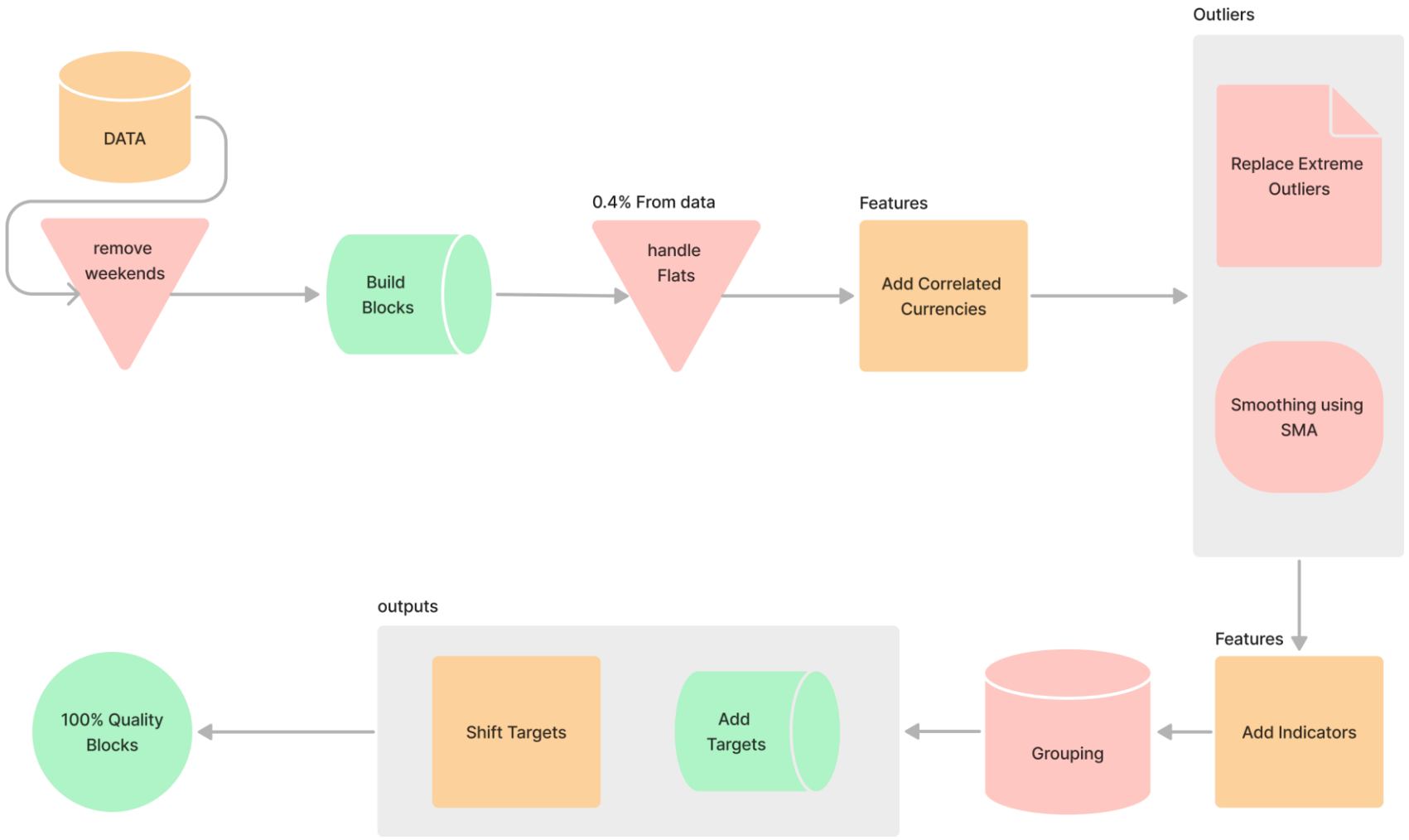

Figure 7 – Data Pipeline Structure

My data pipeline contains different stages as shown in figure 7, and these stages are:

1. Removing weekend flats;
2. Build intact frequented blocks by cutting the time series on consistence Flats and delete these flats;
3. Handle single flats;
4. Adding features 1 — Add correlated currency pairs;
5. Handle Outliers;
6. Adding features 2 — Add Technical indicators;
7. Add a group number for each block;
8. Add Target data (Predicted values) Then shift `-1` to predict the next predicted value by knowing the current predictors' features;
9. Output Pandas Data Frame contains ≈ 100% quality blocks represented by the group column.

All these stages are straight forward and done with coding skills. the processes of handling the Outlier's stage and Add Features stages contain extra details which should be explained.

2.1.5 Handle Outlier

To detect the outliers, I depended on five factors tail, head, body, absolute body, and movement.

$$tail = \min(open, close) - low \quad (1)$$
$$body = \max(open, close) - \min(open, close) \quad (2)$$
$$abs.body = abs(open - close) \quad (3)$$
$$head = high - \max(open, close) \quad (4)$$
$$movement = high - low \quad (5)$$

Where:

- open — is the open price of the candlestick;
- close — is the close price of the candlestick;
- high — is the high price of the candlestick;
- low — is the low price of the candlestick.

By calculating these factors I obtained the following results, head outliers % 6.49, abs. body outliers % 7.16, body outliers % 9.55, tail outliers % 6.96, and movement outliers % 6.48.

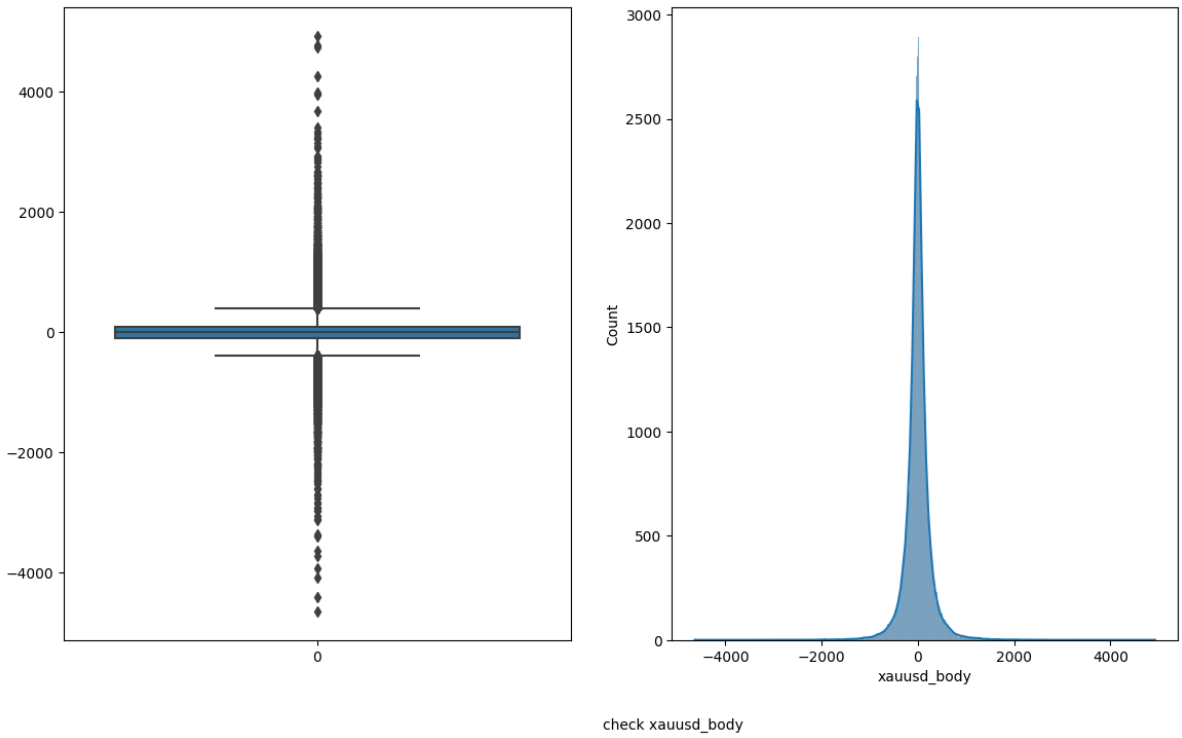

Figure 8 — boxplot and histogram for gold body factor.

2.1.5.1 Outliers Frequency

To understand the natural of the outliers, I needed to make a frequency analysis and this analysis answer a critical question which is Do these outliers are real or do they appear by changing in market characteristics over the years?

I will plot only the outliers and if these outliers cluster in some data range so I can conclude that these are not an outlier but they appear by changing in market characteristics over the years.

Figure 9 — Outlier's clustering

As shown in figure 9, The outliers appear all over the years except ~ three-time ranges (blue arrows) and they didn't create clusters so I can conclude that these are real outliers and proceed with my work.

2.1.5.2 Handle Extreme outliers

We can't just delete outliers from our data because it's time series data and these outliers carry meaningful representation so I tried to come up with another method to handle this problem.

The common equation to get any data outliers is as follows:

$$Q1 = data.quantile(0.25) \quad (6)$$
$$Q3 = data.quantile(0.75) \quad (7)$$
$$IQR = Q3 - Q1 \quad (8)$$
$$filt = out. > (Q1 - whis * IQR) \mid out. < (Q3 + whis * IQR) \quad (9)$$

$whis$ (1.5) value is the magnitude we allow for the values before we consider them outliers. by changing this value to represent the extreme outliers we can normalize these outliers and the process to do so is as follows.

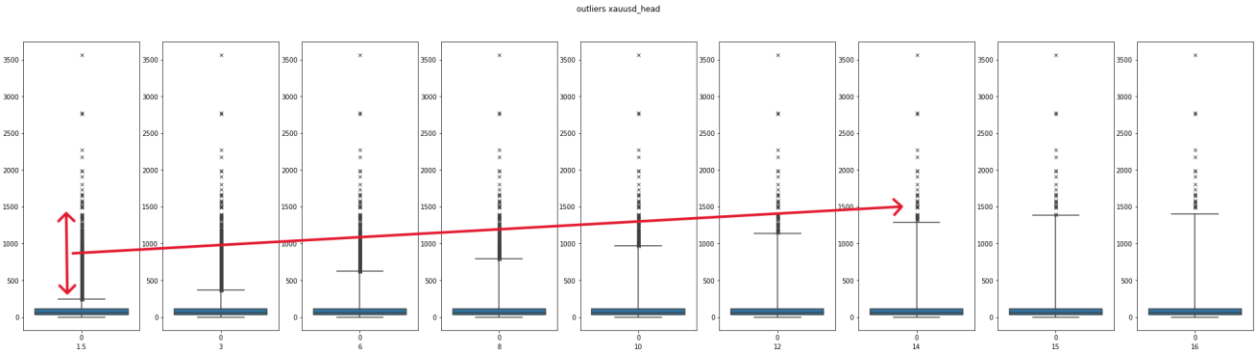

Figure 10 — boxplot with different his values for XAUUSD head values

As shown in figure 10, to handle the extreme value we need to change the whis value in the above equation until to cover the clustered outliers and the outliers became the extreme outliers only. Figure 10 is for the XAUUSD head factor, we need to do the same process to all the five factors. and also, I collect other historical data for correlated currency pairs with gold, AUDUSD, NZDUSD, USDCHF, USDCAD, and EURUSD I will do the same process in them and normalize their extreme outliers. you can see the charts in the appendix.

2.1.5.3 Handle outliers by smoothing

Smoothing is a technique that is often used to reduce noise in a dataset and help identify underlying patterns in the data. One common approach to smoothing is to use a moving average, which involves taking the average of a window of consecutive data points. However, this method can be sensitive to outliers, which can bias the average and distort the underlying patterns in the data.

To address this issue, various methods have been developed to handle outliers in the data when smoothing time series. One such method is the Median Absolute Deviation (MAD), which

involves using the median of the absolute deviations of the data points from the median of the entire dataset. Another method is the Winsorizing technique, which involves setting the extreme values to a certain percentile of the data distribution.

Another popular approach is to use a weighted moving average, where the weights are assigned based on the distance of each data point from the center of the window. This method gives more weight to the data points closer to the center, which reduces the impact of outliers and preserves the underlying patterns in the data.

Overall, smoothing techniques can be useful for identifying underlying trends and patterns in time series data, but care must be taken to handle outliers properly to avoid distorting the results.

I used Exponential Moving Average (EMA) to smooth the open, high, low, and close prices.

Exponential Moving Average is a type of moving average that assigns more weight to recent data points and less weight to older data points. Unlike simple moving averages (SMA), EMA gives greater importance to recent prices.

The formula for EMA is:

$$(EMA)_0 = \left(price_0 * S_{factor}\right) + \left(EMA_1 * (1 - S_{factor})\right) \quad (10)$$

Where:

- $price_0$ – is the $(open \vee high \vee low \vee close)$ prices of the asset today;
- S_{factor} – is the degree of weighting decrease applied to the most recent price, calculated as $\frac{2}{N+1}$, where N is the number of periods;
- EMA_1 – is the EMA value of the previous day.

The smoothing factor in the EMA formula determines the period of the moving average. The shorter the period, the greater the weight given to recent prices, and the more sensitive the EMA is to price changes. Conversely, the longer the period, the less sensitive the EMA is to price changes.

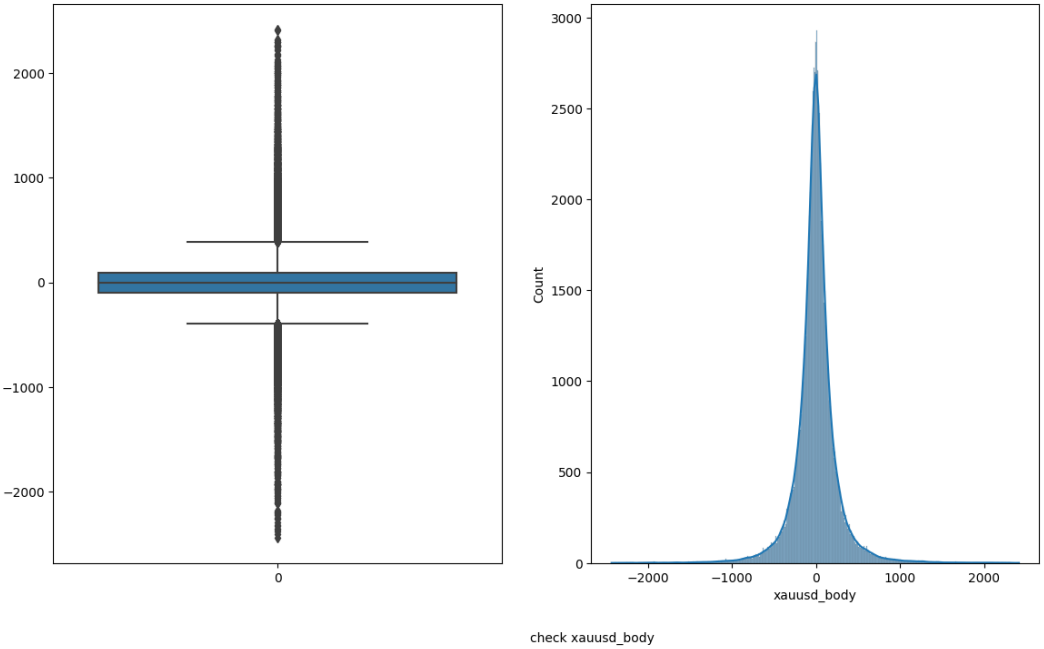

Figure 11 — The body boxplot and histogram after handling the outliers

As shown in figure 11, After deleting the extreme outliers and smoothing the data, the outliers decrease sharply and they cluster with each other's without any extreme values.

2.1.6 Collect Features

The features that I collect consist of some correlated currency pairs with gold, candlesticks patterns, and some technical indicators.

2.1.6.1 Open, High, Low, and Volume

These features came with the dataset. It's important to know this information to n days behind.

2.1.6.2 Trading session (time zones) [8]

When it comes to forex trading, trading sessions play a crucial role in determining the market's behavior. As mentioned earlier, the forex market is open 24 hours a day, five days a week, and this means that the market is always active.

However, the level of activity is not constant throughout the day. The level of activity varies depending on the time of day, the day of the week, and the currency pairs that you are trading.

One of the most significant benefits of understanding the importance of trading sessions is that it can help you identify the best time to trade. Every trading session has unique characteristics that affect the behavior of the market. For example, the European trading session is known to be the busiest trading session since it overlaps with both the Asian and American trading sessions. During this time, the market is highly volatile, which can present many trading opportunities.

Another benefit of understanding trading sessions is that they can help you identify which currency pairs to trade. Some currency pairs are more active during certain trading sessions than others. For example, the AUD/USD pair is more active during the Asian trading session since it involves the Australian and Japanese markets. On the other hand, the GBP/USD pair is more active during the European and American trading sessions since it involves the British and US markets.

In summary, understanding the importance of trading sessions is critical to your success as a forex trader. It can help you identify the best time to trade, which currency pairs to trade, and make more informed trading decisions. The forex market is highly volatile, and knowing when to trade can mean the difference between a profitable trade and a losing one.

The forex market can be broken up into four major trading sessions: the Sydney session, the Tokyo session, the London session, and the New York session.

- Sydney is open from 9:00 pm to 6:00 am UTC
- Tokyo is open from 12:00 am to 9:00 am UTC
- London is open from 7:00 am to 4:00 pm UTC
- New York is open from 1:00 pm to 10:00 pm UTC

I applied to four time zones Sydney, Tokyo, London, and New York. Each time as a categoric feature,

2.1.6.3 Candlestick Patterns:

Candlestick patterns are a popular way for traders to analyze price movements and make trading decisions. They provide valuable information about market sentiment and can be used in combination with other technical analysis tools to identify potential trading opportunities. Some other popular candlestick patterns include the Doji and the Engulfing pattern.

I will use the 7 Candlestick Pattern following a recommendation of a BABYPIPS article [2]. I used Talib [3] library to detect this pattern.

2.1.6.3.1 The Hammer

Figure 12 — The hammer pattern

Is a single candlestick pattern that appears at the bottom of a downtrend. It has a small body, a long lower shadow, and little to no upper shadow. The color of the candlestick is not as important as the position of the shadows. The Hammer indicates that selling pressure has pushed the price down, but buyers have stepped in to push the price back up. This bullish reversal pattern is often used by traders as a signal to buy.

2.1.6.3.2 Bullish and Bearish Engulfing

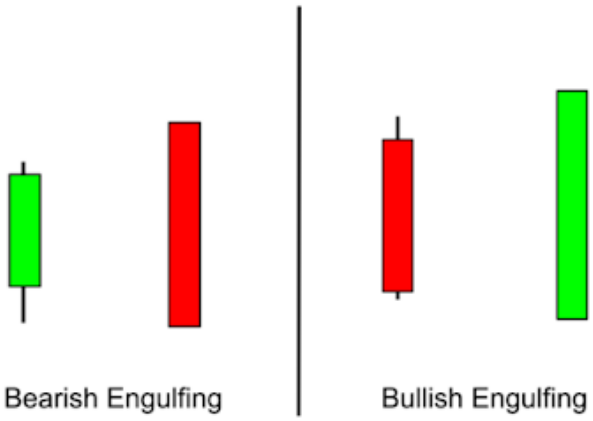

Figure 13 — Bearish and Bullish Engulfing

Bullish and Bearish Engulfing patterns are widely used by traders as a potential signal for trend reversal. However, it is important to note that these patterns should not be used in isolation and should be confirmed with additional technical analysis. Additionally, other factors such as market trends and news events should also be considered before making any trading decisions based on these patterns.

2.1.6.3.3 Shooting Star

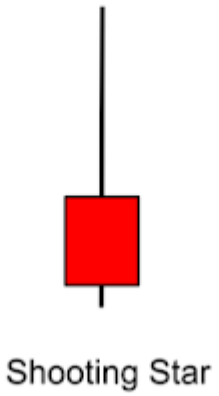

Figure 14 — Shoot Star

The shooting star candlestick is a type of candlestick pattern used in technical analysis. It is formed when a stock's price opens high, then drops throughout the day, but finishes strong with a small body and long upper shadow. This pattern is often used to indicate a potential trend reversal or bearish market sentiment.

2.1.6.3.4 The Doji

Figure 15 — Doji Pattern

The Doji is a candlestick pattern in technical analysis that indicates a potential reversal in price. It is formed when the opening and closing prices of an asset are very close to each other, resulting in a small

or non-existent body and long upper and lower shadows. The Doji can signal indecision in the market and is often used in combination with other technical indicators to confirm a potential trend reversal.

2.1.6.3.5 Doji Star

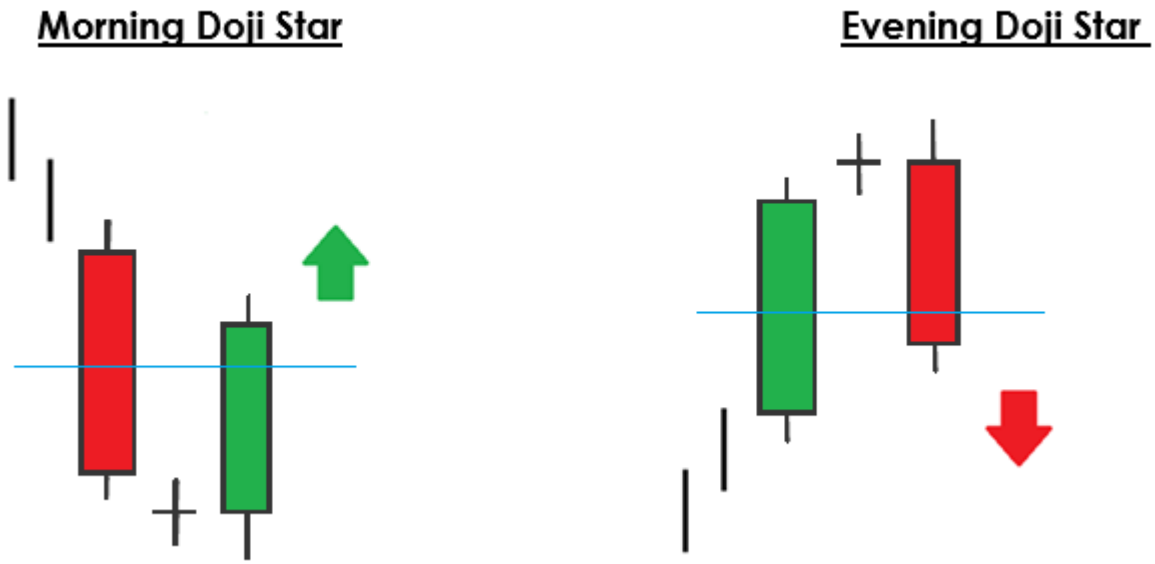

Figure 16 — Doji Star Pattern

A Doji star is a reversal indicator in uptrends and downtrends. It has two types: bullish and bearish. It consists of three candles - the first is long and bullish/bearish, the second is a Doji, and the third is long and bearish/bullish.

2.1.6.3.6 Inside Bar

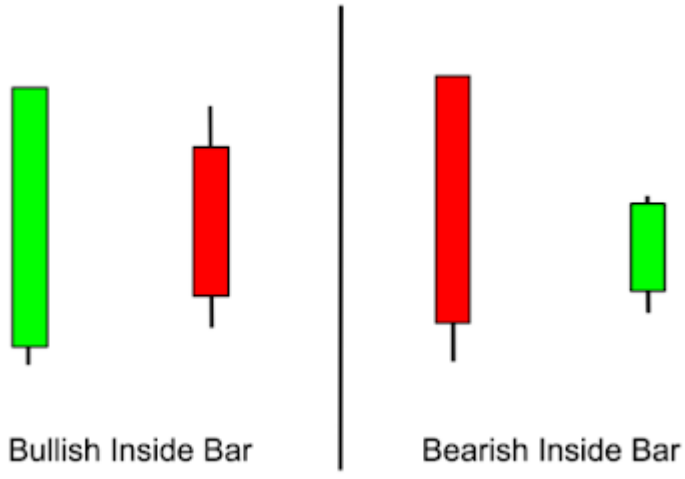

Figure 17 — Bullish and Bearish inside bar

An inside bar candlestick is a two-bar pattern where the second candlestick is completely contained within the range of the prior candlestick. This pattern usually indicates a period of consolidation and can signal a potential breakout in either direction. Traders often use this pattern as a signal to enter or exit trades.

2.1.6.3.7 Morning and Evening Star

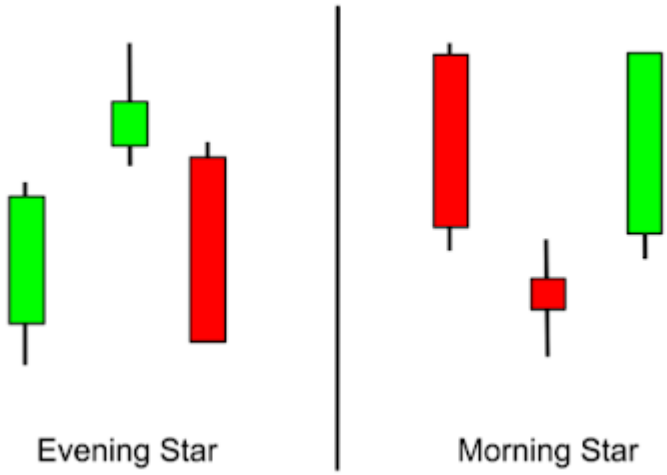

Figure 18 — Evening and Morning Star

The Morning and Evening Star candlestick is a three-candle pattern that is used to signal a potential reversal in the direction of a trend. The pattern consists of a long bearish (or bullish) candlestick, followed by a small candlestick that gaps below (or above) the first one, and then followed by a long bullish (or bearish) candlestick. The small candlestick in the middle is often referred to as a "doji" and represents indecision in the market.

2.1.6.4 Bollinger Bands ® - Volatility indicator

Bollinger Bands ® was developed by technical trader John Bollinger.

The Bollinger Bands indicator is commonly used in technical analysis to determine whether an asset's price is relatively high or low and to identify potential overbought or oversold conditions. When the bands are narrow, it suggests that the market is quiet, while wider bands indicate a more volatile market. The indicator consists of a middle line, which is a simple moving average (SMA) of the asset's price, and upper and lower bands that represent two standard deviations above and below the SMA.

Figure 19 — An Example of Bollinger Bands ® - from Investopedia ®

Formula:

$$BOLU = MA(TP, n) + m * \sigma\,[TP, n] \quad (11)$$
$$BOLD = MA(TP, n) - m * \sigma\,[TP, n] \quad (12)$$

Where:

- BOLU = Upper Bollinger Band;
- BOLD = Lower Bolling Band;
- MA = Moving average;
- TP = (High + Low + Close) / 3;
- n = period;
- m = number of standard deviations (typically 2).

2.1.6.5 Moving Averages

The Moving Average indicator is a widely used technical analysis tool that helps traders identify the trend direction and potential reversals. It uses a set of historical price data and calculates the average price over a specific period of time. This can help smooth out short-term price fluctuations and make it easier to spot trends. There are different types of moving averages, including simple moving averages, exponential moving averages, and weighted moving averages. Each type has its strengths and weaknesses, and traders may choose to use one or a combination of them depending on their trading strategies and goals. Additionally, the Moving Average indicator can be used in conjunction with other technical indicators and chart patterns to confirm trading signals and improve the overall accuracy of trading decisions.

It's a common practice to use a combination of 5-30-62 moving average periods together to catch a trend or breakout. Also, you can use 200 for very long-term trades.

	SMA	EMA
PROS	Displays a smooth chart that eliminates most fake-outs.	Quick Moving and is good at showing recent price swings.
CONS	Slow-moving, which may cause a lag in buying and selling signals	More prone to cause fake outs and give errant signals.

2.1.6.6 Average True Range (ATR)

It was introduced by Welles Wilder in his book, "New Concepts in Technical Trading Systems ".

Average True Range (ATR) is a technical analysis indicator that measures market volatility by analyzing the range of an asset's price movement. The ATR is calculated by taking the average of the True Range (TR), which is the greatest of the following: the difference between the current high and the previous close, the difference between the current low and the previous close, and the difference between the current high and the current low. A higher ATR indicates greater volatility, while a lower ATR indicates less volatility.

Wilder used a period of 7. Other common periods used are 14 and 20.

Formula:

$$ATR = \frac{ATR(n-1) + TR}{n} \tag{13}$$

Where:

- n = number of periods;
- TR = True range.

and TR comes from:

$$TR = Max(|H - L|, |H - C_p|, |L - C_p|) \tag{14}$$

WHERE:

- H - Today's high
- L - Today's Low
- C_p - Yesterday's closing price

2.1.6.7 Standard Deviation

The standard deviation (SD) indicator is a statistical measure that helps forex traders gauge the amount of market volatility. It shows how much prices have deviated from their average value over a period of time. Traders can use the SD indicator to identify potential trading opportunities, such as breakouts or reversals, and to set stop-loss and take-profit levels.

2.1.6.8 Relative Strength Index (RSI)

Relative Strength Index, or RSI, is a popular indicator developed by a technical analyst named J. Welles Wilder [5].

The Relative Strength Index (RSI) is a technical indicator used to measure the strength or weakness of a currency pair by comparing its up movements versus its down movements over a given time period. RSI is commonly used by traders to determine whether a currency is overbought or oversold. When the RSI is above 70, the currency is considered overbought and may be due for a correction. Conversely, when the RSI is below 30, the currency is considered oversold and may be due for a rebound.

Figure 20 — Example of RSI Indicator with a description of overbought and oversold zones.

Formula:

$$RSI = 100 - \frac{100}{1 - RS} \tag{15}$$

$$RS = \frac{Average\ up\ closes}{Average\ down\ closes} \tag{16}$$

Where:

- RSI = Relative strength index

2.1.6.9 Stochastic

It was developed by George C. Lane [6] in the late 1950s. He believed that momentum changes before price so he created the Stochastic Oscillator to follow the "speed" or momentum of price.

The Stochastic indicator is a popular technical analysis tool used by traders to identify potential trend reversals and overbought or oversold conditions in the market. It was developed by George Lane in the 1950s and is widely used by traders to this day.

The Stochastic oscillator is based on the idea that when an asset is trending up, its closing price tends to be near the high of the trading range. Conversely, when an asset is trending down, its closing price is often near the low of the trading range. The Stochastic indicator measures the relationship between the closing price and the high-low range over a set period of time, typically 14 days.

Figure 21 — Example of Stochastic Indicator

The Stochastic oscillator consists of two lines: the $\%K$ line and the $\%D$ line. The $\%K$ line is the main line and is calculated as follows:

$$\%K = 100 * \frac{(current\ close - lowest\ low)}{(highest\ high - lowest\ low)} \quad (17)$$

The line is a moving average of the $\%K$ line and is calculated as follows:

$$\%D = 100 * (sum\ of\ prior\ n\ \%K\ values / n) \quad (18)$$

Traders use the Stochastic indicator to identify overbought and oversold conditions. When the %K line crosses above the %D line, it is considered a bullish signal, indicating that the security may be oversold and that a potential uptrend may be coming. Conversely, when the line crosses below the %D line, it is considered a bearish signal, indicating that the security may be overbought and that a potential downtrend may be coming.

2.1.6.10 Average Directional Index (ADX)

The average directional movement index (ADX) was developed in 1978 by J. Welles Wilder [7]

It is used to measure the strength of a trend. It is a non-directional oscillator, meaning it does not determine whether a trend is bullish or bearish. Rather, it is used to determine the strength of a trend, regardless of its direction. ADX values range from 0 to 100, with readings below 20 indicating a weak trend and readings above 50 indicating a strong trend.

ADX is calculated using a formula that takes into account the difference between two directional indicators: the positive directional indicator (+DI) and the negative directional indicator (-DI). The ADX line itself represents the average of these two directional indicators, and it is typically plotted alongside the +DI and -DI lines.

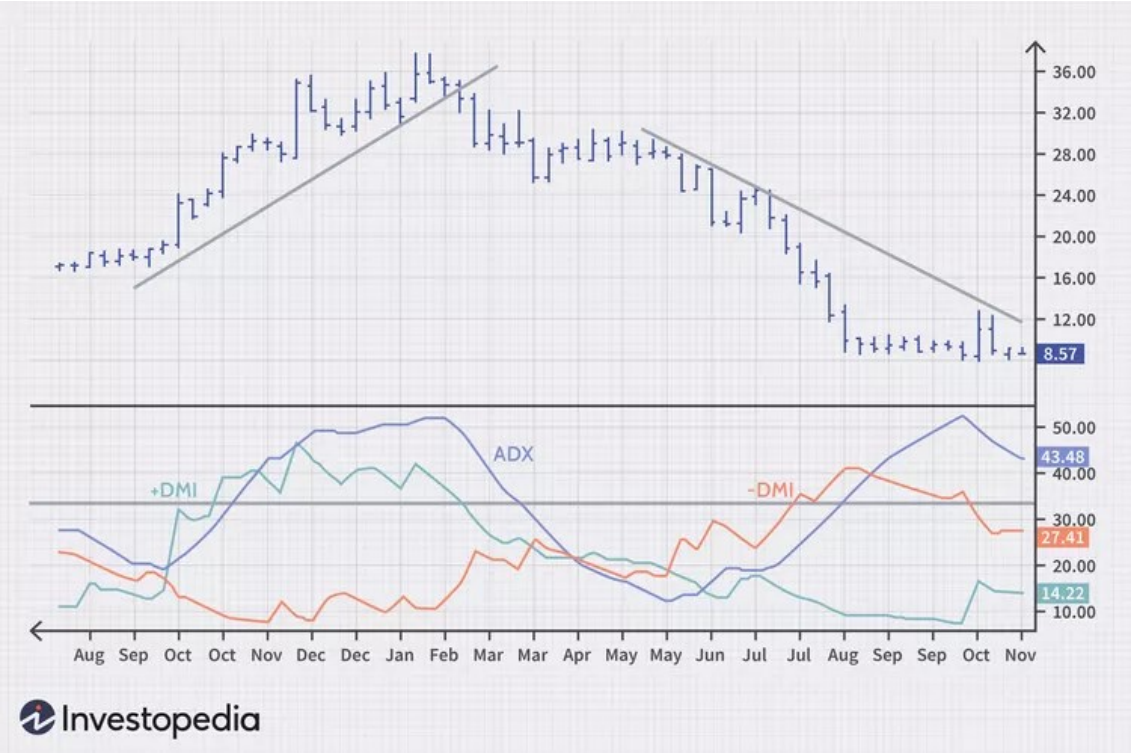

Figure 22 — Example of ADX Indicator

Traders use ADX to identify whether the market is ranging or starting a new trend. When the ADX line is below 20, it indicates that the market is in a ranging phase, with no clear trend in either direction. When the ADX line is above 50, it indicates a strong trend, which can be either bullish or bearish, depending on the direction of the trend.

Overall, ADX can be a useful tool for traders looking to identify the strength of a trend, as well as potential entry and exit points. However, like all technical analysis indicators, it should be used in conjunction with other tools and analysis techniques to make informed trading decisions.

2.1.6.11 Moving Average Convergence Divergence (MACD)

MACD stands for Moving Average Convergence Divergence. It is a popular technical indicator used in financial analysis to identify changes in trends and momentum. The MACD is calculated by subtracting the 26-period exponential moving average (EMA) from the 12-period EMA, with a 9-period EMA signal line used to identify potential buy or sell signals.

When the MACD line (12-period EMA - 26-period EMA) crosses above the signal line (9-period EMA), it is considered a bullish signal indicating a potential buy opportunity. Conversely, when the MACD line crosses below the signal line, it is considered a bearish signal indicating a potential sell opportunity.

The MACD is also used to identify the divergence between the MACD line and the price of the asset being analyzed. Divergence occurs when the price of the asset is moving in one direction, while the MACD line is moving in the opposite direction. This can be a signal of a potential trend reversal.

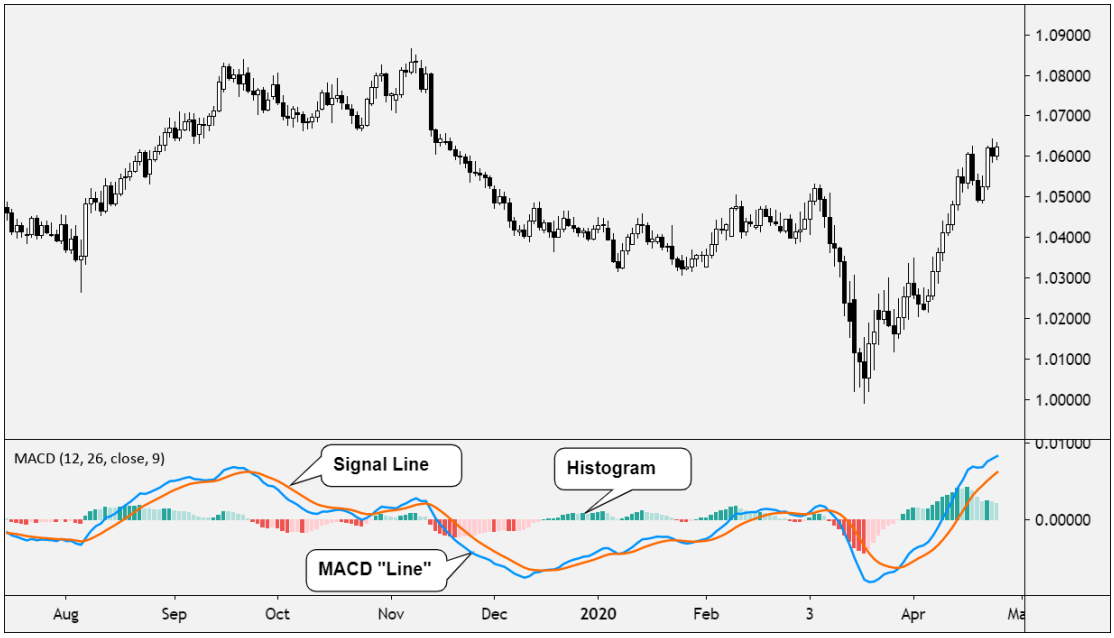

Figure 23 — Example of MACD Indicator

It is important to note that the MACD is just one tool among many used in technical analysis, and should not be relied upon solely to make investment decisions. It is best used in combination with other indicators and analysis techniques to get a more comprehensive understanding of market trends and potential opportunities.

In summary, the MACD is a widely used technical indicator that can help traders and investors identify changes in trend and momentum, as well as potential buy and sell signals. By using the MACD in conjunction with other analysis tools, investors can make more informed decisions and potentially improve their overall investment performance.

2.1.6.12 Chaikin Money Flow (CMF)

Chaikin Money Flow (CMF) is a technical indicator developed by Marc Chaikin that calculates the volume-weighted average of accumulation and distribution over a specified period, typically 21 days. The key principle behind the Chaikin Money Flow is that the proximity of the closing price to the high indicates a greater accumulation of stock, while the proximity of the closing price to the low indicates a greater distribution of stock.

To calculate the Chaikin Money Flow, the user must first determine the Money Flow Multiplier, which is the relationship between the current closing price and the midpoint of the high and low prices. The Money Flow Volume is then calculated by multiplying the Money Flow Multiplier by the volume for the period. The CMF is then calculated by summing the Money Flow Volume for each period and dividing the result by the total volume for the period.

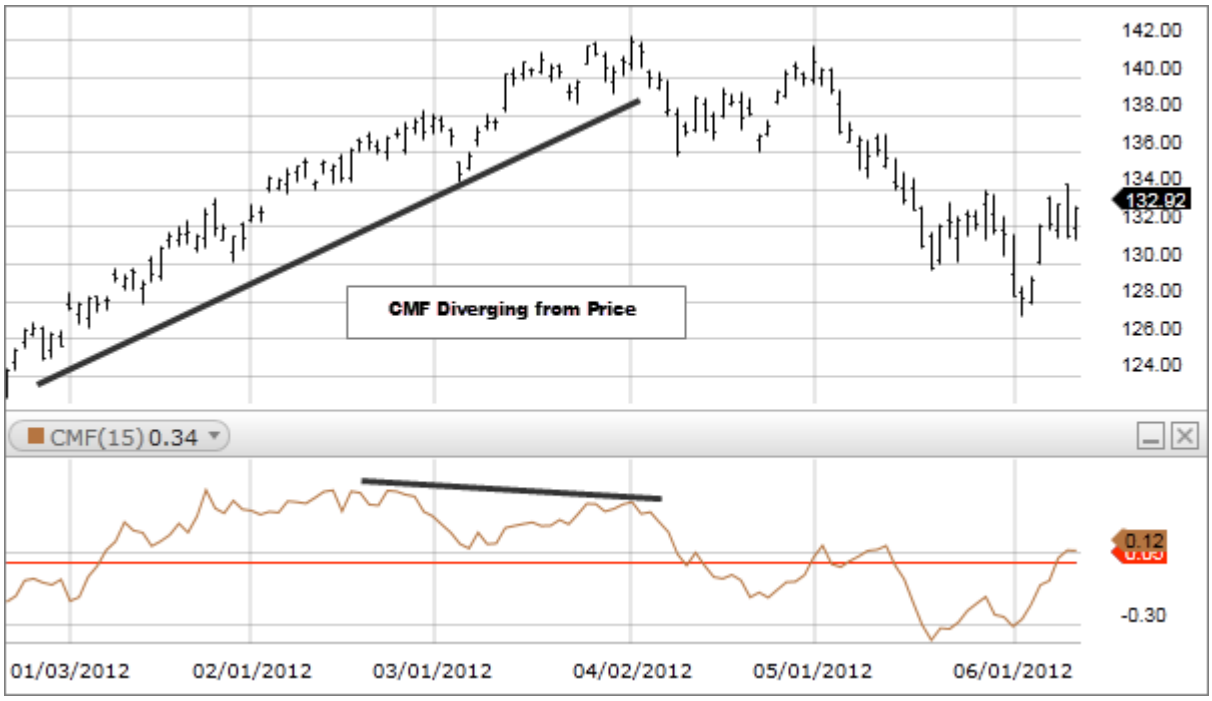

Figure 24 — Example of CMF

When the price action consistently closes above the midpoint of the bar on increasing volume, the Chaikin Money Flow will be positive, indicating a bullish trend. Conversely, when the price action consistently closes below the midpoint of the bar on increasing volume, the Chaikin Money Flow will be a negative value, indicating a bearish trend. Traders can use the Chaikin Money Flow to confirm trends, identify potential reversals, and generate trading signals.

$$\mathrm{CMF} = \sum_{n}^{1} \frac{(\mathrm{C} - \mathrm{L}) - (\mathrm{H} - \mathrm{C})}{(\mathrm{H} - \mathrm{L})} \times \mathrm{Vol} \qquad (19)$$

Where:

- n = number of periods, typical 21
- H = High
- L = low
- C = close
- Vol = volume

2.1.6.13 On Balance Volume (OBV) - Volume —

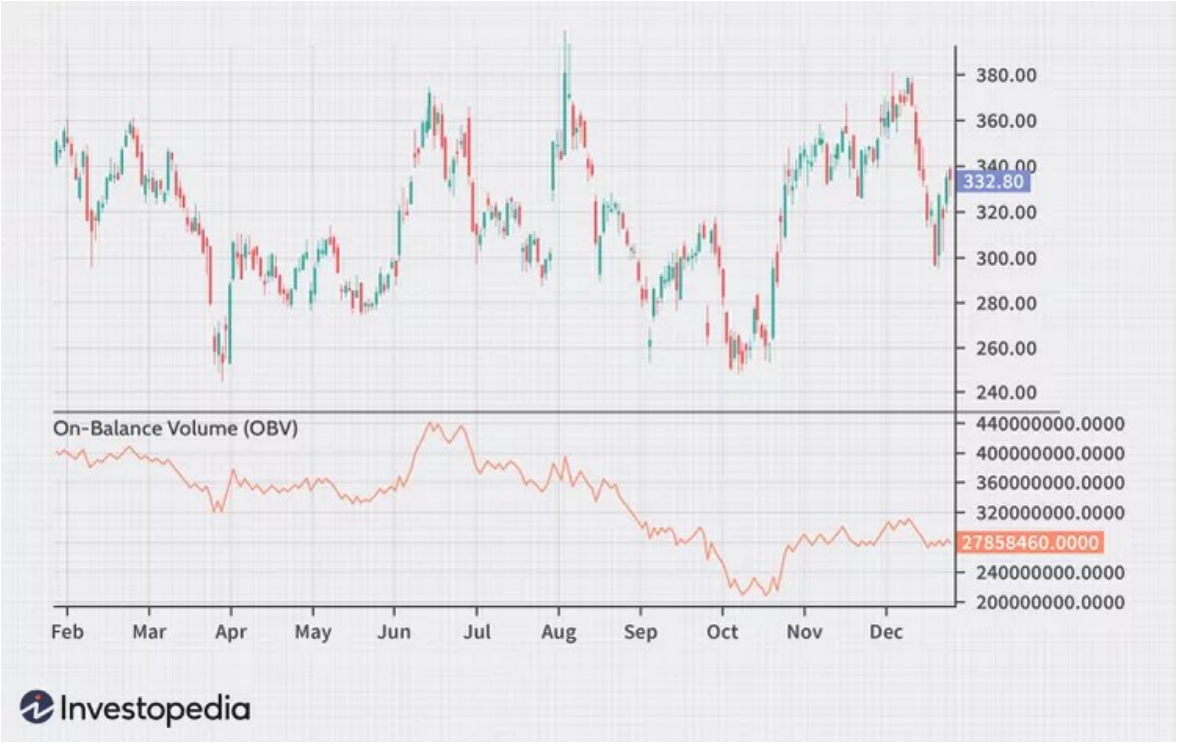

Figure 25 — Example of OBV Indicator

Balance Volume (OBV) is a technical analysis indicator that uses volume flow to predict changes in stock price. OBV measures buying and selling pressure as a cumulative indicator, adding volume on up days and subtracting it on down days. The theory behind OBV is that changes in volume precede price movement, so tracking volume can help predict changes in price trends.

OBV was developed by Joe Granville in the 1960s and has been widely used ever since. Traders use OBV to confirm price trends, spot divergences, and identify potential breakouts. When the OBV line is trending up, it suggests that buying pressure is increasing and the uptrend is likely to continue. Conversely, when the OBV line is trending down, it suggests that selling pressure is increasing and the downtrend is likely to continue.

However, it's important to note that OBV is not a standalone indicator and should be used in conjunction with other technical analysis tools for more accurate predictions.

2.1.6.14 Currency correlated with XAUUSD

2.1.6.14.1 AUD/USD

Australia is the third biggest gold producer in the world, sailing out about $5 billion worth a year.

2.1.6.14.2 NZD/USD

New Zealand (rank 25) is also a large producer of gold.

2.1.6.14.3 USD/CHF

Over 25% of Switzerland’s reserves are backed by gold. As gold prices go up, the pair moves down (CHF is bought).

2.1.6.14.4 USD/CAD

Canada is the 5th largest producer of gold in the world. As the gold price goes up, the pair tends to move down (CAD is bought).

2.1.6.14.5 EUR/USD

Since both gold and the euro are considered “anti-dollars,” if the price of gold goes up, EUR/USD may go up as well.

2.2 Statistics analysis

Conducting statistical analysis is a crucial part of any research project, including a master's thesis. Statistical analysis helps to make sense of the data collected and determine whether the hypotheses or research questions are supported by the data. Statistical analysis provides a way to organize, summarize, and interpret the data, allowing researchers to draw conclusions and make informed decisions.

Statistical analysis can also help to identify patterns, trends, and relationships within the data, providing insights that may not be apparent from simple descriptive statistics. By using statistical techniques such as correlation, regression analysis, and hypothesis testing, researchers can assess the strength and significance of relationships between variables and conclude the population from the sample data.

In summary, conducting statistical analysis is essential for any research project, as it helps to make sense of the data, test hypotheses, and draw meaningful conclusions.

My dataset contains 160 Features with 6 Target (predicted) values shifted -1 (the previous candlestick features used to predict the current candlestick targets).

Targets:

- xauusd_close — close prices without handle outliers
- xauusd_close_smooth — close prices without outliers
- return — return in points for close prices before clean outliers
- return_smooth — return in points for close prices after clean outliers
- labels — classification labels depend on the candlestick return
- labels_smooth — classification labels depend on the candlestick return_smooth

Labels:

For previous experiments, I found that binary classification didn’t get any acceptable results so I decided I will break the binary classification problem into a multi-classification problem.

I deepened on returns values to classify my data into 6 labels up-strong, up, up-weak, doji, down-weak, down, and down-strong as shown in figure 26

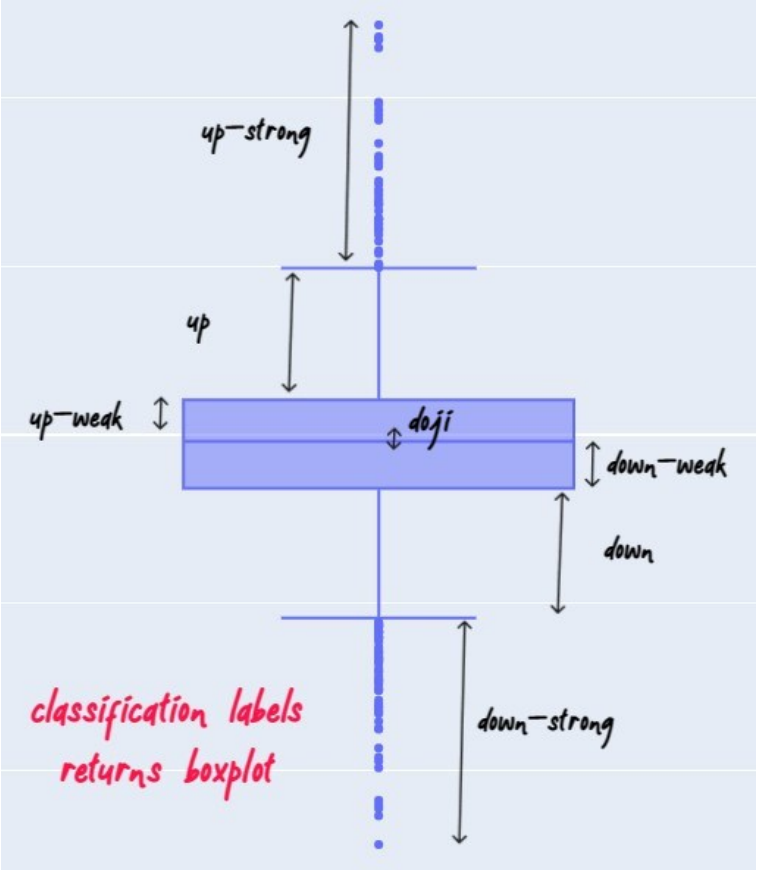

Figure 26 – classification labels for returns

2.2.1 Relationship between values before & after smoothing outliers

Returns:

I compare the predicted sign for each candlestick and I found the percentage of different candlestick signs between return & return_smooth was 35.31 %. and they don't have any linear regression (correlated) feature between each other as shown in Figure 27

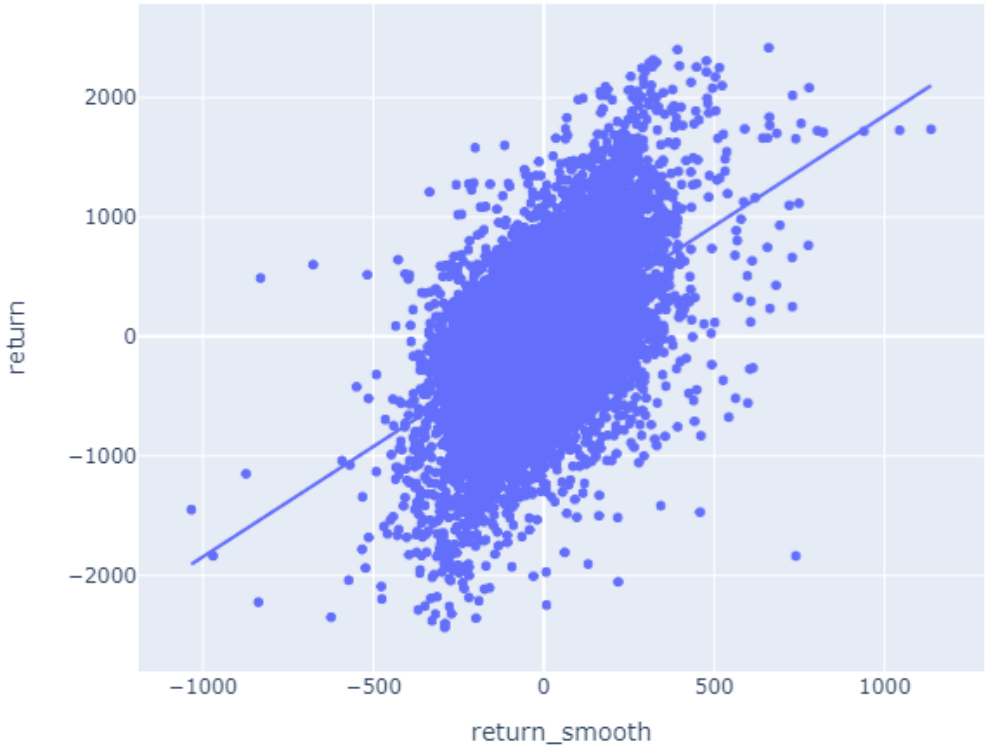

Figure 27 — Linear regression relationship between return & return_smooth

Close prices:

As shown in figure 29, they are a strongly correlated relationship between the two predicted values, so I include that I can use close_smooth values to predict close prices.

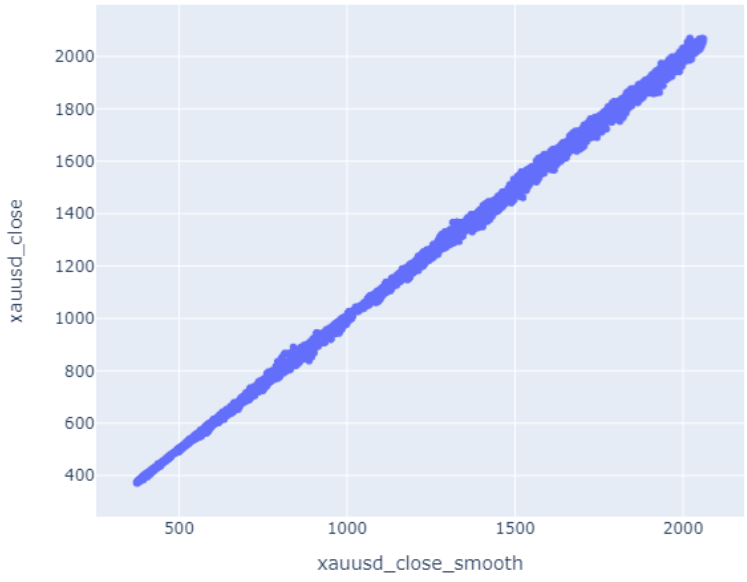

Figure 28 – Linear regression relationship between close & close_smooth prices

Labels:

As shown, for the two classifications I had an imbalance problem which I have to it under consideration when applying classifier models. Also, the class label size got effected my smoothing process.

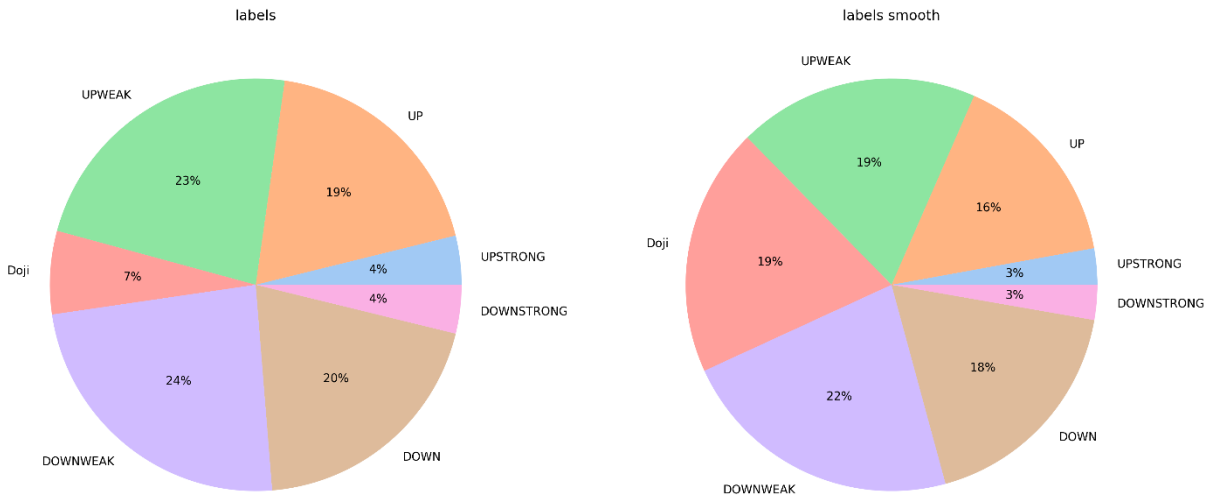

Figure 29 – Labels & Labels smooth Pie chart

Conclusion:

By all the above; I can use smooth values with their outputs to train my models but I can't use them to predict the non-smooth values. also, I can use them as a data augmentation method, if I needed more data to train.

2.2.2 Returns Statistics analysis:

Some statistics properties for return_smooth. I will use smooth to train all my models then I will test with a non-smooth value. so, I will focus my work to find the properties of the smooth parameters.

Table 1 – statistical analysis of returns

Mean:	10. -12.85
Median:	-8.43
Max value:	487.21
Min value:	-488.09
Std:	110.69
Q1:	-64.04
Q3:	40.84
IQR:	104.88
Limit up:	198.16
Limit down:	-221.37

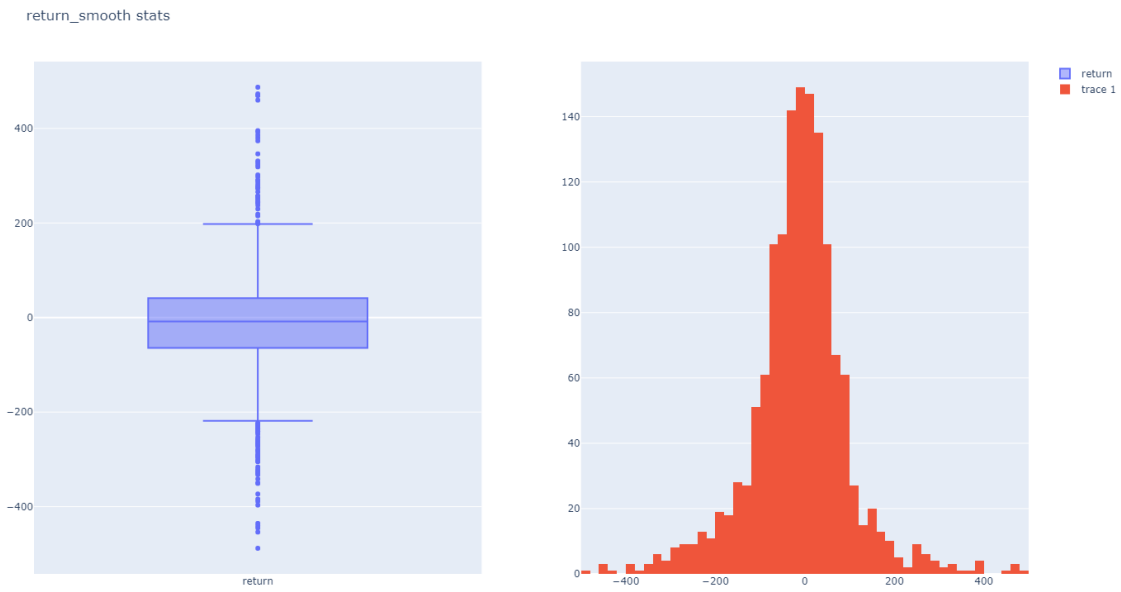

Figure 30 – Boxplot & Histogram for returns

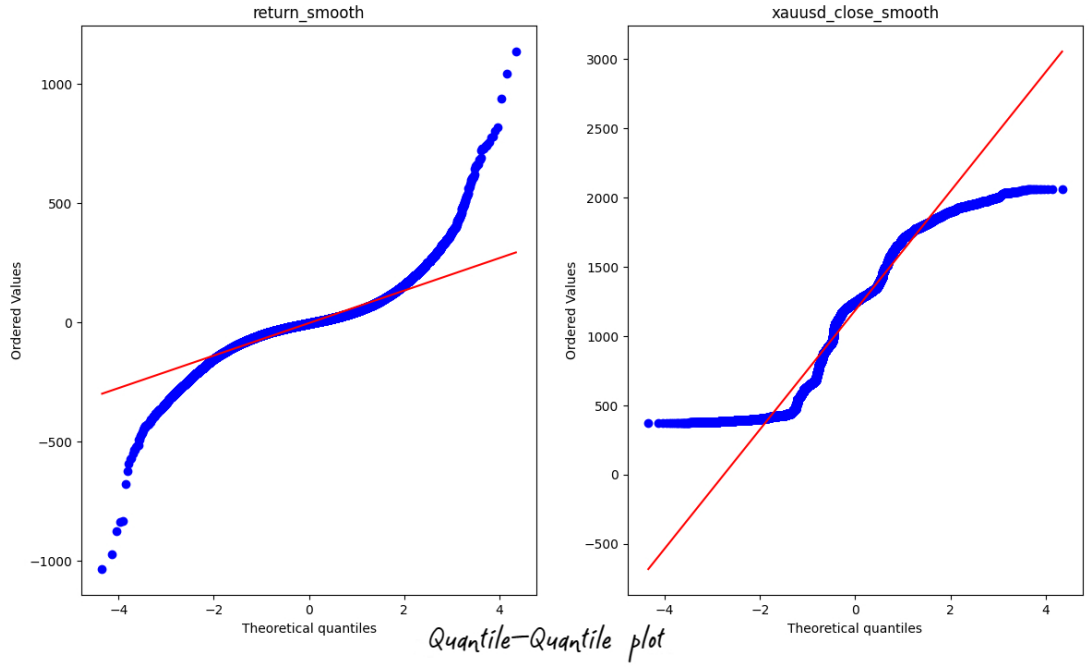

Figure 31 – Quantile-Quantile plot to check normality

2.2.3 Check If in the long run the population mean will be as the sample mean (returns, close prices)

Hypothesis Test:

Applying a t-test for one sample.

Hypothesizes:

- H_0 - In the long run, the mean population will be the sample mean.
- H_1 - In the long run, the mean population will not be as the sample mean.

Results:

- We fail to reject the null hypothesis. The population mean is equal to the sample mean of returns. with p-value: 1.0
- We fail to reject the null hypothesis. The population means is equal to the sample mean of close prices. with p-value: 1.0

2.2.4 Check if my time series data is predictable:

How much information is contained in your past data that can be used to forecast future values?

Before applying any forecasting method to a time series, it is important to check if the time series is predictable. Time series predictability is a metric that quantifies the highest possible prediction accuracy for a given time series. It can be used to evaluate the performance of forecasting algorithms and to understand the underlying patterns and regularities in the data. To check if your time series is predictable, you need to calculate its predictability score using methods such as entropy-based measures, mutual information, Kaboudan metric, or contextual models. These methods can help you estimate how much information is contained in your past data that can be used to forecast future values. A higher predictability score means a higher potential for accurate forecasting. However, predictability also depends on factors such as data quality, resolution, stationarity, and topological constraints.

2.2.4.1 Non-linear approaches:

2.2.4.1.1 Entropy-based measures:

Entropy measures for time series are techniques used to quantify the amount of regularity, variability, or randomness of fluctuations over time series data. Entropy, as it relates to information theory and dynamical systems theory, can be estimated in many ways, with different advantages and limitations.

Some examples of entropy measures for time series are:

- **Shannon entropy:** The basic measure of entropy that evaluates the average amount of information contained in a time series. It is based on the probability distribution of the values in the time series[1].
- **Conditional entropy:** The measure of entropy that evaluates the average amount of information needed to describe a time series given another time series. It is based on the joint probability distribution of a two-time series[1].
- **Permutation entropy:** The measure of entropy that investigates the permutation pattern in time series. It is based on the ordinal relation among consecutive values in the time series.
- **Approximate entropy:** The measure of entropy that evaluates the regularity or complexity of a time series. It is based on the logarithm of conditional probability that two sequences similar form points remain similar at the next point.
- **Sample entropy:** The measure of entropy that evaluates the regularity or complexity of a time series. It is similar to approximate entropy but does not include self-matching cases. It is more consistent and less dependent on data length and noise.

These are some of the common entropy measures for time series, but there are also other methods such as spectral entropy, multiscale entropy, fuzzy entropy, and so on. Entropy measures can be used for various applications such as biomedical engineering, finance, physiology, human factors engineering, and climate sciences.

Approximate entropy:

Approximate Entropy (ApEn) is a statistical technique used to measure the complexity or irregularity of time series data. It was first introduced by Pincus in 1991 and has since been used in various fields, including biology, finance, and engineering. ApEn measures the likelihood that similar patterns or sequences of data points will not repeat themselves within a defined tolerance level. A low ApEn value indicates a high level of regularity and predictability, while a high ApEn value indicates a higher level of complexity and unpredictability. ApEn is commonly used as a diagnostic tool to identify patterns or anomalies in time series data and to assess the predictability of a system.

Approximate entropy (ApEn) equation [13]:

$$\phi^m = \frac{1}{N-m+1} \ln\left(\frac{\sum d(x_i X_j) \le r}{N-m+1}\right) \tag{20}$$

$$ApEn(m,r) = \phi^m - \phi^{\{m+1\}} \tag{21}$$

$$2N(\ln s - ApEn(m,r)) \approx x\text{^}2\ (s^{m+i} - s^m \tag{22}$$

Table 2 – result for approximate entropy (ApEn) for returns degree of 4 and close prices degree of 2

M – return	4
M – closes	2
Return p-value – return	2.74e-41
Return p-value - closes	0.0
ApEn – return	0.66
ApEn - closes	0.08

After obtaining the ApEn value and applying chi-square analysis for the distances between each value and its m different level the results conclude that returns and closes values have a low entropy which indicates that they are some information patterns in my datasets that I can train a machine learning model to detect them.

2.2.4.1.2 Hurst Exponent Test:

The Hurst Exponent Test is a statistical method that is commonly used in time series analysis to determine the long-term memory of a given time series data. The Hurst exponent value ranges from 0 to 1, where a value of 0.5 indicates that the time series data has a random walk, while a value greater than 0.5 indicates that the data exhibits persistence or long-term memory.

The test is important in predicting the behavior of a time series, as it can determine whether a trend is likely to continue or whether it is likely to revert to the mean. A Hurst exponent value greater than 0.5 indicates that the time series is trending, while a value less than 0.5 suggests that the data is mean-reverting.

The Hurst Exponent Test is widely used in finance to study asset prices and to identify trading opportunities. By understanding the long-term memory of a time series, traders can better anticipate future trends and make more informed investment decisions.

Applying hurst exponent test to a time series to its n times back to itself, we got the following results.

Table 3 – the result of the Hurst Exponent Test:

Return	0.004
Closes	0.492
Return smooth	0.227
Closes smooth	0.712

Conclusion:

By examining these results we can expect that the price movement tends to be close to its local mean. The close prices don’t save any information and tend to move like random noise. return smooth and close smooth are shown as good signs of a tendency for saving memory.

2.2.4.1.3 Mutual information

Mutual information is a measure of the amount of information that two variables share. It quantifies the amount of information obtained about one variable by observing the other variable. Mutual information measures the extent to which the knowledge of one variable reduces the uncertainty of the other variable.

Mean-reversion assumes that properties such as stock returns and volatility will revert to their long-term average over time. Mathematically, such a time series is referred to as an Ornstein-Uhlenbeck process. In such strategies, investors try to make money by assuming that after some extreme events (either positive or negative), the stock price will revert to the long-term pattern.

$$mutual\ information = \sum_{x \in X} \sum_{y \in Y} p(x,y)\ log\left[\frac{p(x,y)}{p(x)p(y)}\right] \quad (23)$$

Where:

- $p(x,y)$ – Joint Probabilities
- $p(x), p(y)$ – Marginal Probabilities

Applying mutual information to Window 1, I obtained the results in Table 4.

Table 4 – results of mutual information

Mutual information returns	0.05
Mutual information returns smooth	0.81
Mutual information closes	5.03
Mutual information closes smooth	6.31

Conclusion:

As windows of 1. Close, close smooth prices, and returns smooth tend to save information. As for returns, there is no temptation to save information. If returns and returns smooth had more relationship, I would predict returns by knowing the returns smooth prediction but the candlestick correlated sign between returns, and returns smooth is 35.31 %.

2.2.4.2 Linear approaches (autocorrelation and seasonality)

In this part, I will check if my time series (returns, close prices) have any autocorrelation and the market move in a repetitive pattern or if it's just a Random walk

Before we began here are summaries of some important concepts:

Stationary time series:

A stationary time series is one whose statistical properties such as the mean, variance, and autocorrelation are all constant over time. Hence, a non-stationary series is one whose statistical properties change over time.

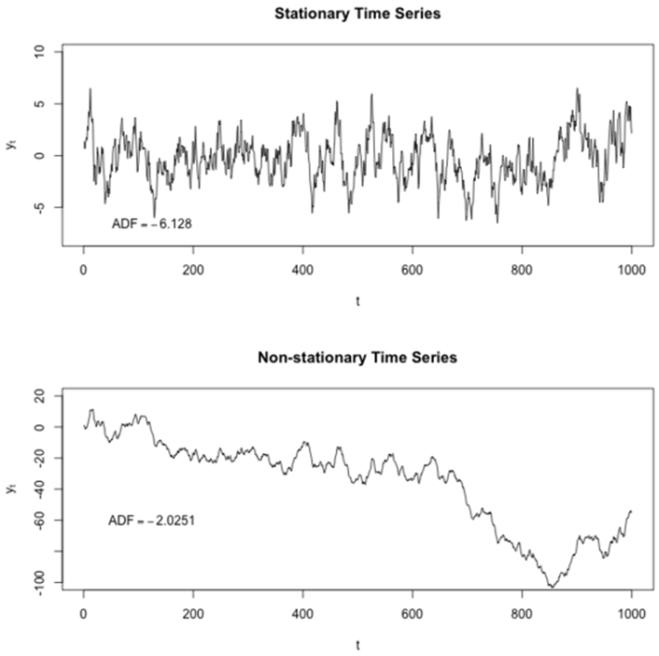

Figure 32 – Stationary vs non-stationary time series example

Autocorrelation:

Autocorrelation is a mathematical representation of the degree of similarity between a given time series and a lagged version of itself over successive time intervals.

Autocorrelation can be checked in Python by ACF and PACF Tests. As shown in this figure 33.

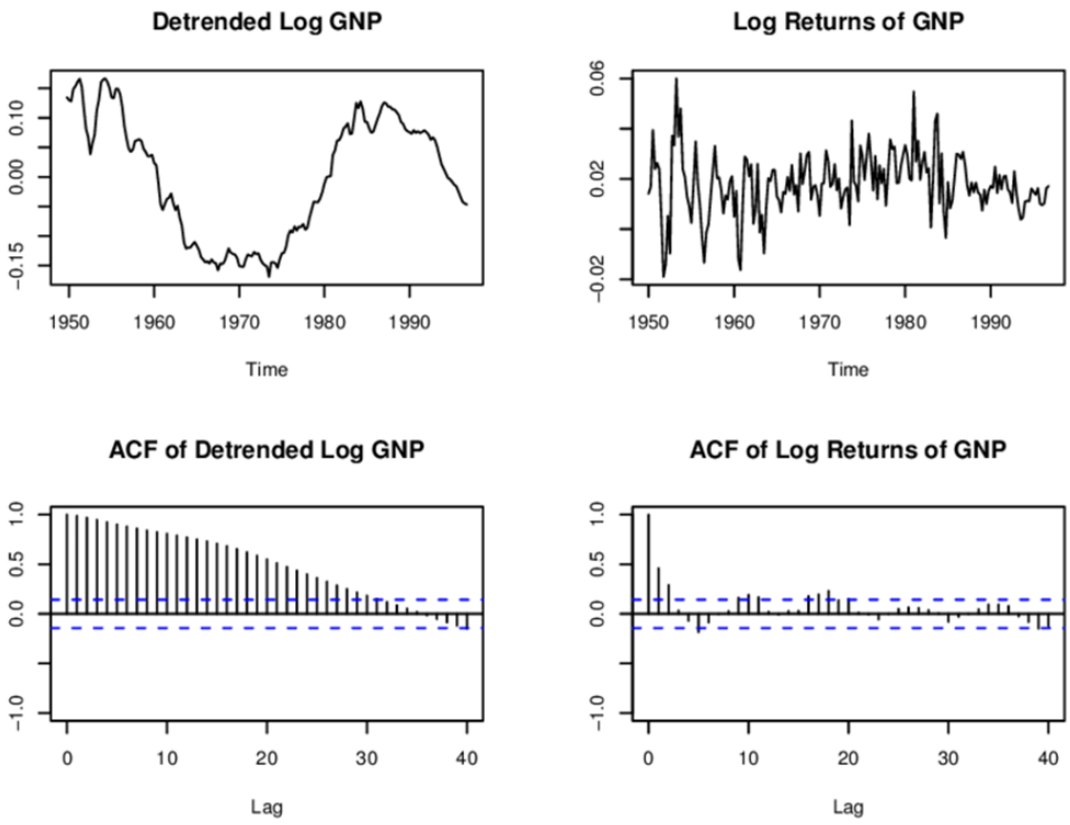

Figure 33 – ACF Test Example on Time series data

White noise:

The white noise is a stationary time series or a stationary random process with zero autocorrelation. In other words, in white noise $N(t)$ any pair of values $N(t\_1)$ and taken at different moments t_1 and t_2 of time is not correlated.

The white noise has true randomness and can't be predicted.

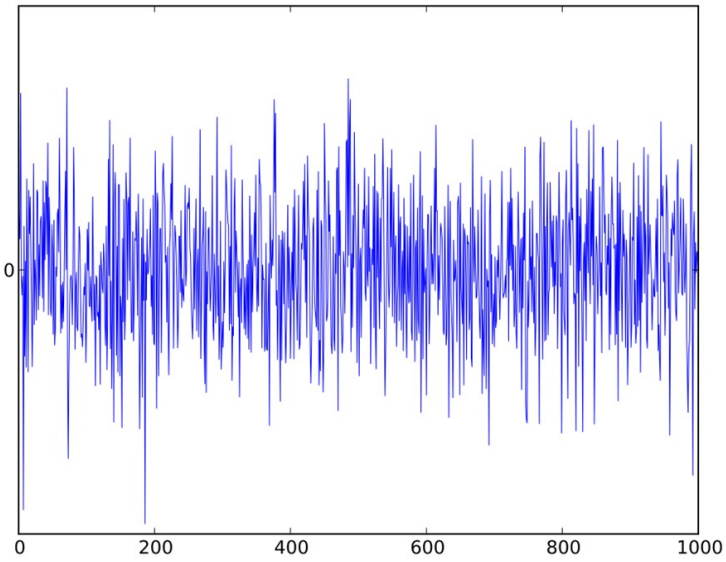

Figure 34 – white noise data example

Random walk theory:

Random walk theory suggests that changes in stock prices have the same distribution and are independent of each other. Therefore, it assumes the past movement or trend of a stock price or market cannot be used to predict its future movement.

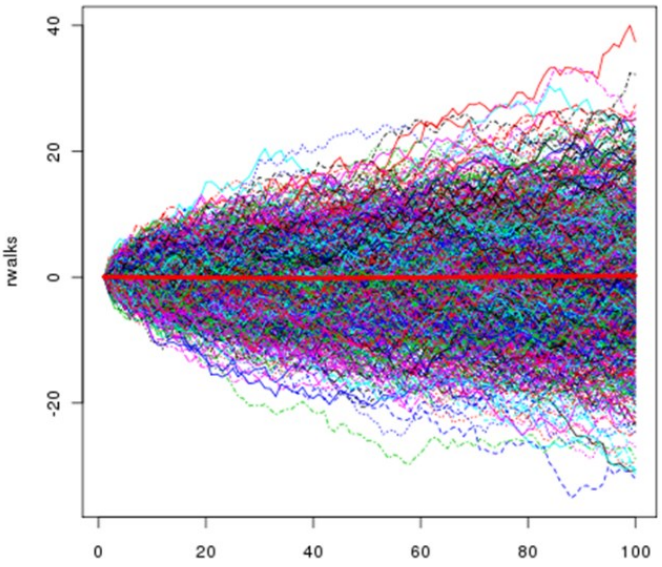

Figure 35 – Random walk example

$$P_t = P_{t-1} + \varepsilon_t \quad (24)$$

Where:

- $\varepsilon_t \approx WN(\mu, \sigma^2)$— ε_t is a white noise.

2.2.4.2.1 Seasonality analysis:

Time series seasonality analysis is a technique used to identify and quantify patterns that repeat in a data set at regular intervals, such as daily, weekly, or monthly. The analysis helps to determine if time series data has a seasonal component, which can affect the accuracy of forecasting models. It is important to perform this step in a thesis as it provides insights into the underlying factors that drive the data and can help to identify the best forecasting methods to use. Seasonality analysis can be done using different techniques such as the seasonal sub-series plot, seasonal pattern decomposition, and spectral analysis. By understanding the seasonal patterns of time series data, it is possible to develop more accurate forecasting models that take into account periodic trends.

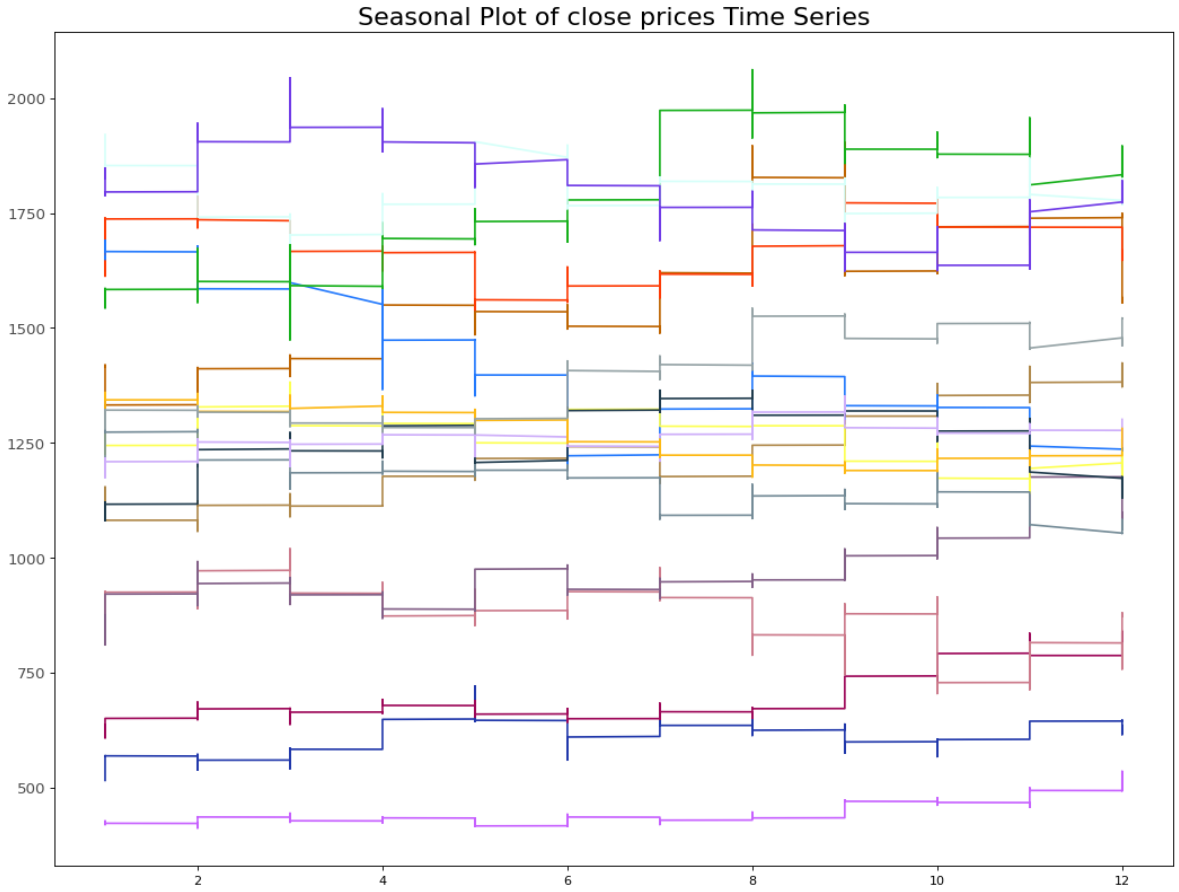

Figure 36 — Seasonal plot for close prices

Additive Time series analysis —

Additive time series analysis is a method used to decompose a time series into its underlying components, including trend, seasonality, and irregularity. The equation used in additive time series analysis is as follows:

$$Y_t = T_t + S_t + I_t \quad (25)$$

Where:

- Y_t — The observed value at time t;
- T_t — The trend component at time t;
- S_t — The seasonal component at time t;
- I_t — The irregular component at time t.

The importance of additive time series analysis for a thesis lies in its ability to separate the different components of a time series, which can help in understanding and interpreting the data. By identifying the trend and seasonal patterns in the data, one can make more accurate predictions and develop better forecasting models. Additionally, the irregular component can provide insights into the random fluctuations and unexpected events that may impact the data.

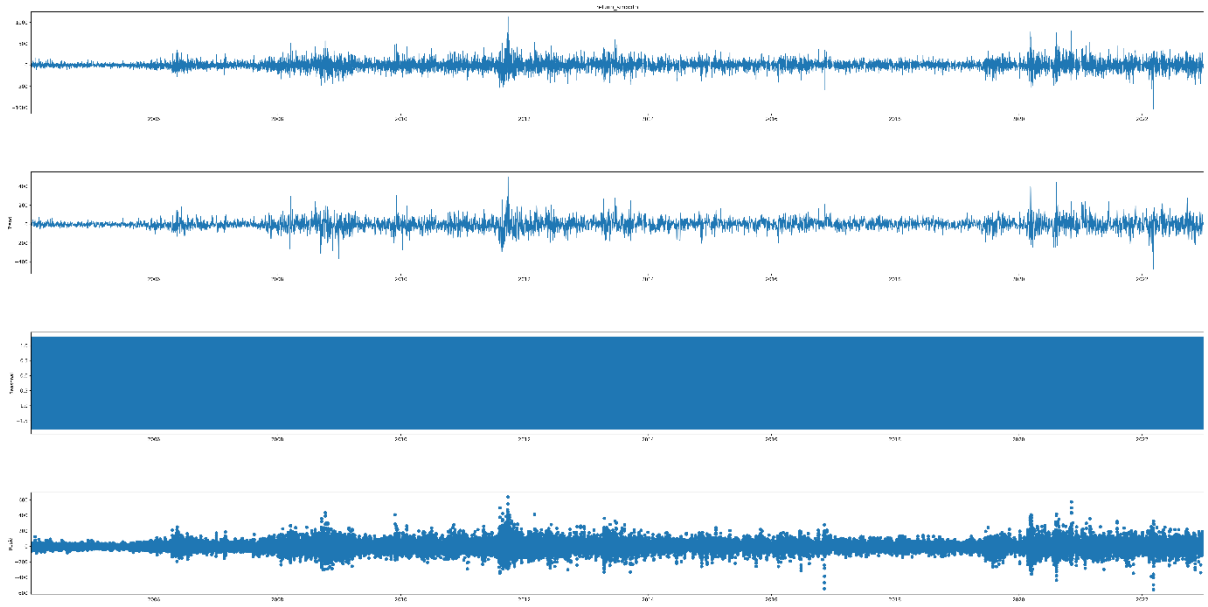

Figure 37 – Seasonality Test for close prices

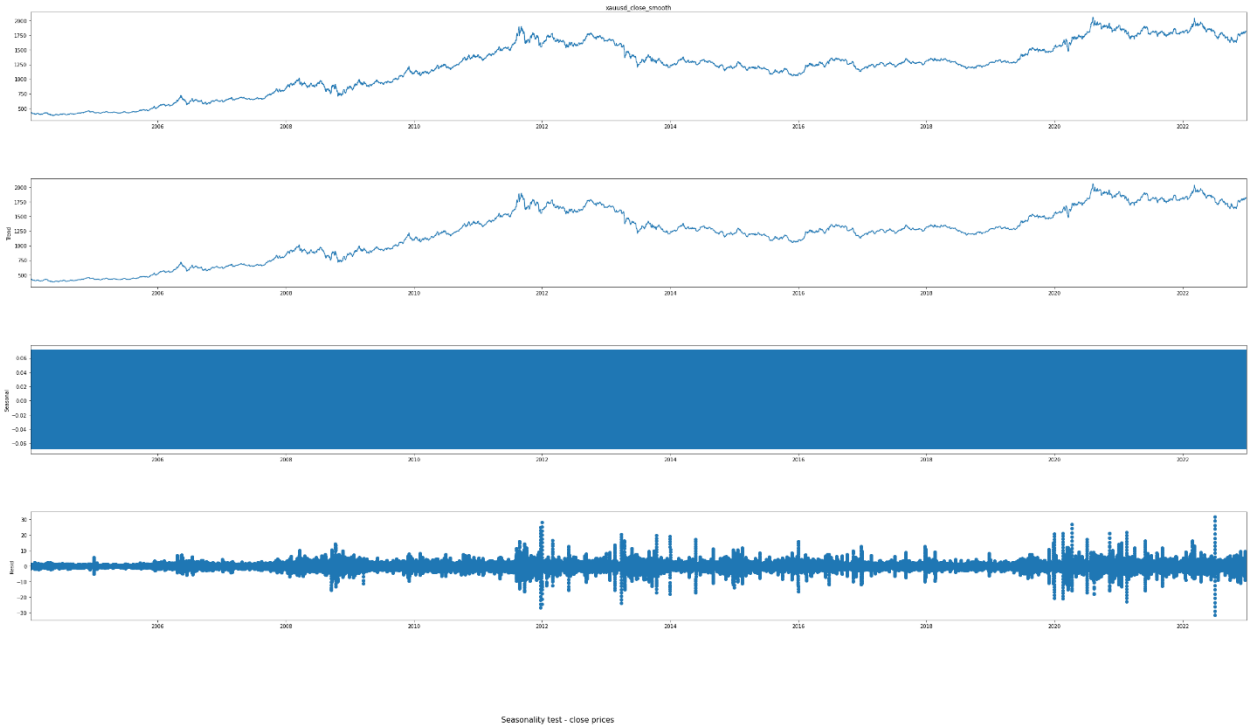

Figure 38 – seasonality test for close prices smooth

Conclusion:

From Figures 37, 38, and 39, we can see that there is no acceptable seasonality I can depend on.

2.2.4.2.2 Autocorrelation Analysis:

Autocorrelation analysis is a statistical method that examines the correlation between a time series and its past values at different lags. It is a measure of the similarity between observations as a function of the time lag between them. Autocorrelation analysis is an essential tool for time series analysis as it helps to identify patterns and relationships in the data.

The autocorrelation function (ACF) plot is a common tool used to visualize the autocorrelation analysis. The plot shows the correlation between the time series and its past values at different lags. A significant autocorrelation at a particular lag indicates that the time series is dependent on its past values at that lag.

Autocorrelation analysis is important for time series forecasting as it helps to identify the appropriate lag for a time series model. By analyzing the ACF plot, we can determine the lag value that has the highest correlation with the time series. This value can be used as the lag parameter in the autoregressive integrated moving average (ARIMA) model or other time series models.

$$\mathrm{ACF(k)} = \frac{\gamma_k}{\gamma_0} \tag{26}$$

Where:

- γ_k - The autocovariance at lag k;
- γ_0 - The variance of the time series.

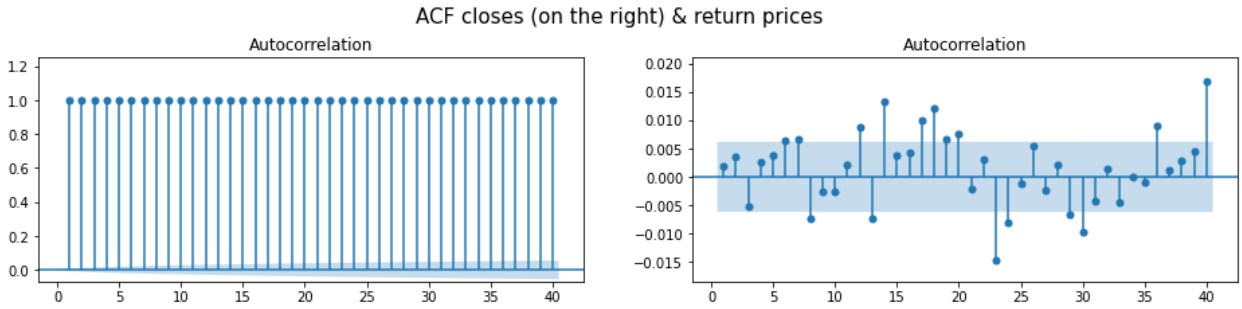

Figure 39 — ACF plots for returns & closes prices

The partial autocorrelation function (PACF) measures the correlation between a time series and its lagged values while controlling for the effect of other lags in between. It is a useful tool for identifying the order of an autoregressive (AR) model. The equation for PACF can be expressed as:

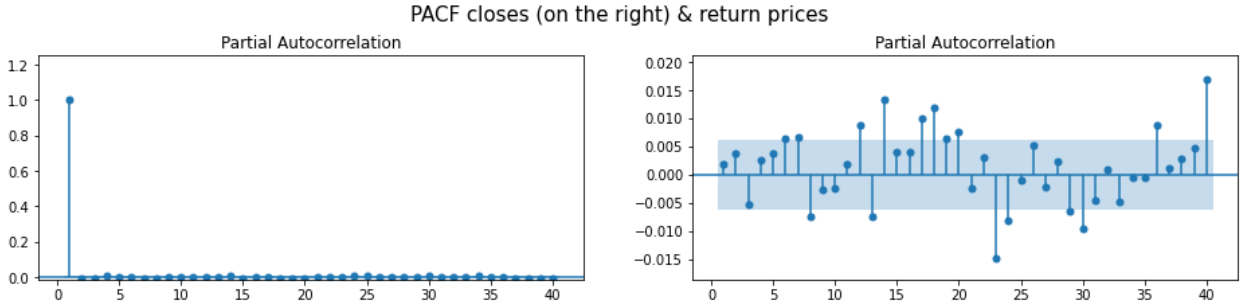

Figure 40 – PACF plots for returns & close prices

$$\alpha_k = \frac{cov(y_t, y_{\{t-k\}} - \hat{y}_{\{t-k\}})}{var(y_t)} \tag{27}$$

Where:

- y_t - The time series at time t;
- $y_{\{t-k\}}$ - The time series at a time t_k;
- $\hat{y}_{\{t-k\}}$ - The predicted value of $y_{\{t-k\}}$ from the AR model;
- cov - The covariance function;
- var - The variance functions.

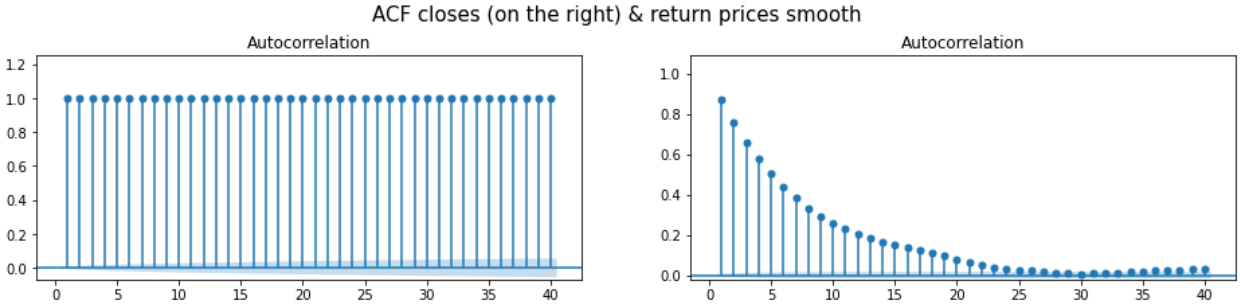

Figure 41 — ACF plots for returns & closes prices

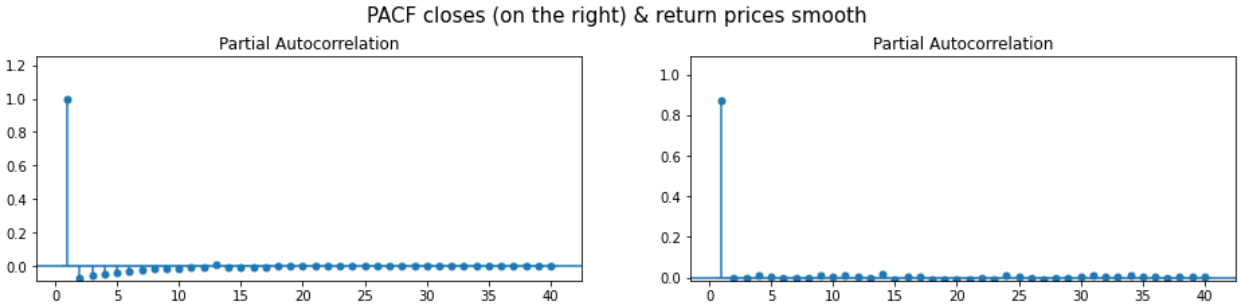

Figure 42 – PACF plots for returns & close prices

Conclusion —

From Figures 39, 40, 41, and 42, we can conclude that returns and close prices have an autocorrelation feature in them and therefore they are not a white noise time series and we can extract some patterns for future predicting.

2.2.4.2.3 Lag Plots:

Lag plots are a graphical method used to test the presence of autocorrelation in a time series. In a lag plot, each observation in the time series is plotted against its lagged values, which are usually the values at a particular lag time. A diagonal line in the plot represents the absence of autocorrelation. If the observations are clustered around the diagonal line, it indicates positive autocorrelation, while a scatter of points indicates no correlation. A vertical pattern in the plot suggests seasonality, while a horizontal pattern suggests a trend.

The lag plot method is a simple and intuitive way to assess autocorrelation and identify the presence of patterns in the data. It can be a useful exploratory tool for checking whether a time series has any autocorrelation before further analysis is conducted. The formula for calculating lag plots is simply plotting each observation against its lagged value at a specific lag time.

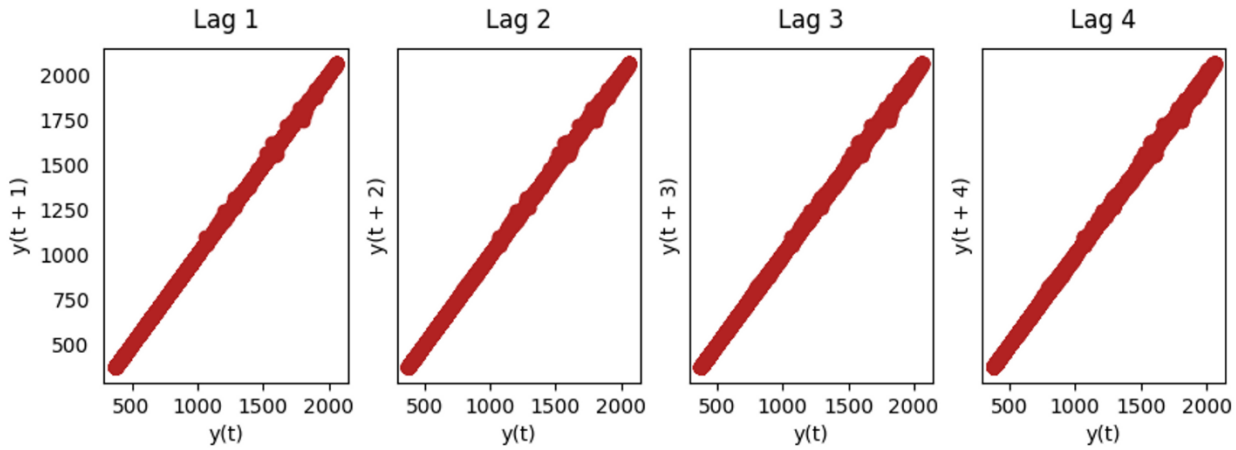

Figure 43 — Lag plots for close prices smooth

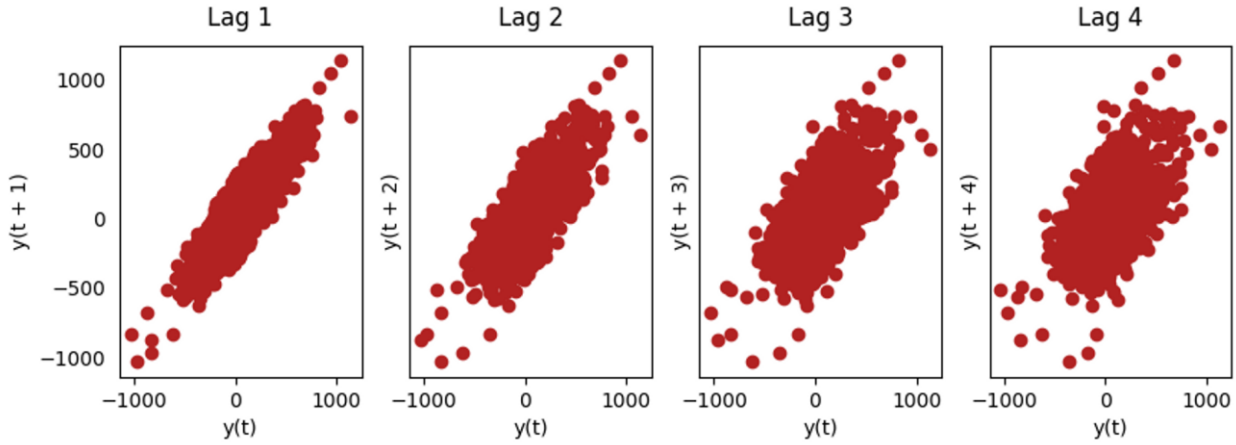

Figure 44 — Lag plots for returns smooth

As shown in Figures 43, and 42, there is a linear regression relationship for close prices and their lags but there is not one for returns values.

2.2.4.2.4 Granger Causality test:

The Granger causality test is a statistical hypothesis test that is used to determine whether one-time series is useful in forecasting another. It is based on the idea that if a time series (A) Granger causes another time series (B), then past values of A should help improve predictions of B. The null hypothesis of the test is that past values of A do not provide any additional information in predicting B beyond what can already be predicted from B's past values.

The Granger causality test can be written in equation form:

- H0: Y(t) does not Granger-cause X(t)
- H1: Y(t) Granger-causes X(t)

where X(t) and Y(t) are two-time series, and the null hypothesis (H0) is rejected if the p-value is less than a specified significance level.

Applying then on returns and different n lags of itself. I obtained these results:

Table 5 – results for Granger Causality

Number of lags	p-value
1	0.7213
2	0.9093
3	0.0464
4	0.0491

As shown in Table 5, the p-value in lag 4 is less than 0.05 so we can reject the null hypothesis and conclude that the returns values have granger causality to its copy of lag 4 and this lag number is the same number I obtained from entropy-based measures in Table 2.

2.2.5 Calculate significant training size:

I will be using power analysis to calculate significant training size.

Power analysis:

Power analysis is a statistical technique used to determine the minimum sample size needed for a study to detect a statistically significant effect, given a certain level of power. Power is the **probability of rejecting the null hypothesis** when the alternative hypothesis is true. The higher the power, the greater the likelihood of detecting a true effect.

The sample size formula for power analysis depends on various factors such as the desired effect size, the level of significance, power, and the number of predictors in the model. In general, the formula for calculating the sample size for a two-sample t-test is:

$$n = \frac{\left(\frac{Z_\alpha}{2} + xZ_\beta\right)^2}{ES^2} \tag{28}$$

Where:

- n - sample size
- $\frac{Z_{\{\alpha\}}}{2}$ - The critical value for the chosen level of significance (e.g., 1.96 for α = 0.05)
- $Z_{\{\beta\}}$ -The critical value for the chosen power (e.g., 0.84 for 80% power)
- ES^2 - The desired effect size (difference between the means of the two groups)

It's important to note that the sample size formula can vary depending on the type of statistical test being performed and the specific assumptions of the analysis. Additionally, power analysis is a complex process that requires careful consideration of many factors, and it's often best to consult with a statistician or use specialized software to perform power analysis.

To apply power analysis to a machine learning model to know which training size requires to obtain significant results, I need to calculate the effect size for two predicted distributions from the same model after changing it a little bit. So, it's like how much effect in predicted values distribution If I change the model a little bit.

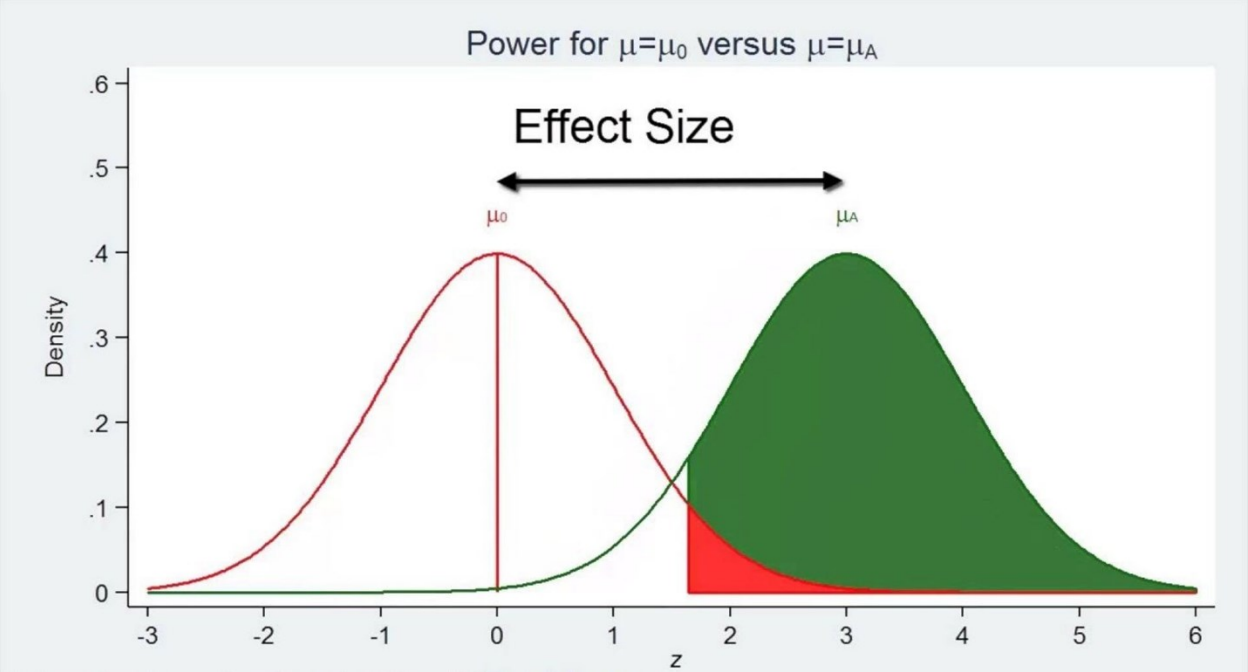

Figure 46 — Effective size example for calculate sample size - power analysis

Some approaches that can be used to estimate the effect size for neural network models:

1. **Difference in means:** If you are interested in comparing the performance of two or more neural network models on a given task, you could estimate the effect size using the difference in means between the performance metrics of the models. For example, you could calculate the difference in mean accuracy, mean F1 score, or mean area under the receiver operating characteristic curve (ROC AUC) between the models.

2. **Cohen's d:** Cohen's d is a widely used effect size measure in statistical analyses. It is calculated as the difference in means between two groups divided by the pooled

standard deviation. For neural network models, Cohen's d could be calculated by subtracting the mean performance metric of one model from the mean performance metric of the other model and dividing it by the pooled standard deviation of the performance metric.

3. **Percentage improvement:** If you are interested in comparing the performance of a neural network model before and after a particular modification, such as changing the architecture, adjusting hyperparameters, or adding regularization, you could estimate the effect size using the percentage improvement in performance. For example, you could calculate the percentage improvement in accuracy, F1 score, or ROC AUC after modifying.

4. **Correlation coefficient:** If you are interested in the relationship between the size of the training set and the performance of a neural network model, you could estimate the effect size using a correlation coefficient, such as Pearson's r or Spearman's rho.

Cohen's d —

To calculate Cohen's d to estimate the effect size from available data for a neural network model, you can follow these steps:

1. Calculate the mean performance metric of two groups of the neural network model that you want to compare. For example, you could calculate the mean accuracy of two models on a binary classification task.
2. Calculate the pooled standard deviation of the performance metric for the two groups. The pooled standard deviation is a weighted average of the standard deviations of the two groups. You can calculate it using the formula:

$$pooled\ SD = \sqrt{(n_1 - 1)s_1^2} + \frac{(n_2 + 1)\ s_2^2}{(n_1 + n_2 - 2)} \tag{29}$$

Where:

- -n1 and n2 are the sample sizes of the two groups
- SD1 and SD2 are the standard deviations of the performance metric for the two groups.

3. Calculate the difference in means between the two groups of the performance metric. For example, you could subtract the mean accuracy of one model from the mean accuracy of the other model.
4. Calculate Cohen's d using the formula:

$$\text{Cohen's d} = \frac{\mu_1 - \mu_2}{pooled\ SD} \tag{30}$$

where:

- μ_1 and μ_2 are the means of the performance metric for the two groups.

After applying power analysis for an LSTM and a linear regression model, the results were that I need 160 194 training sample size for the LSTM model to reach a significant result which I had only 78 924 sample size. For the linear regression model, the results were that I need $\approx 2^{30}$ sample size to reach a significant result which I had 100 601 sample sizes.

2.3 Machine learning algorithms used in the research:

2.3.1 Multiple Linear regression:

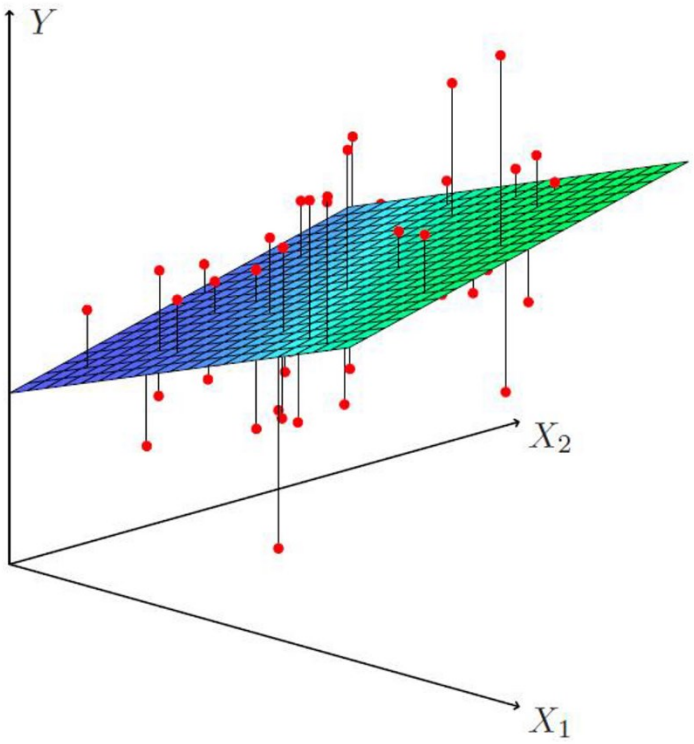

Figure 46 – Multiple linear regression for 3d example

The multiple linear regression model is used when there are two or more independent variables that can be used to explain the variation in the dependent variable. For example, we may want to understand how the price of a commodity is related to its demand, supply, and inflation rate. In this case, price is the dependent variable, and demand, supply, and inflation rate are the independent variables [17].

The multiple linear regression model can be represented mathematically as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_n x_n + \epsilon \quad (31)$$

Where:

- y —Dependent variable
- $x_1, x_2, x_3, \ldots, x_n$ — Independent variables
- β_0 — Intercept
- $\beta_1, \beta_2, \beta_3, \ldots, \beta_n$ — Slope coefficients

- ϵ — Error term

The intercept (β0) is the value of the dependent variable when all the independent variables are zero. The slope coefficients (β1, β2, β3, …, βn) represent the change in the dependent variable when the corresponding independent variable changes by one unit. The error term (ε) is the difference between the predicted value and the actual value of the dependent variable.

Assumptions of Linear Regression

Before using the linear regression model, it is important to ensure that certain assumptions are met. These assumptions include:

1. **Linearity:** There should be a linear relationship between the dependent variable and the independent variable(s).

2. **Homoscedasticity:** The variance of the error term should be constant across all values of the independent variable(s).

3. **Independence:** The error terms should be independent of each other.

4. **Normality:** The error terms should be normally distributed.

If these assumptions are not met, the results of the linear regression model may not be reliable.

Interpreting the Results of Linear Regression

Once the linear regression model is fitted to the data, we can interpret the results by examining the coefficients and the goodness of fit measures. The coefficients provide information about the relationship between the dependent variable and the independent variable(s). For example, a positive coefficient indicates a positive relationship, while a negative coefficient indicates a negative relationship.

The goodness of fit measures, such as R-squared, provide information about how well the model fits the data. R-squared ranges from 0 to 1, with higher values indicating a better fit. However, it is important to note that R-squared should not be used in isolation to assess the

2.3.2 Lasso & Ridge regression:

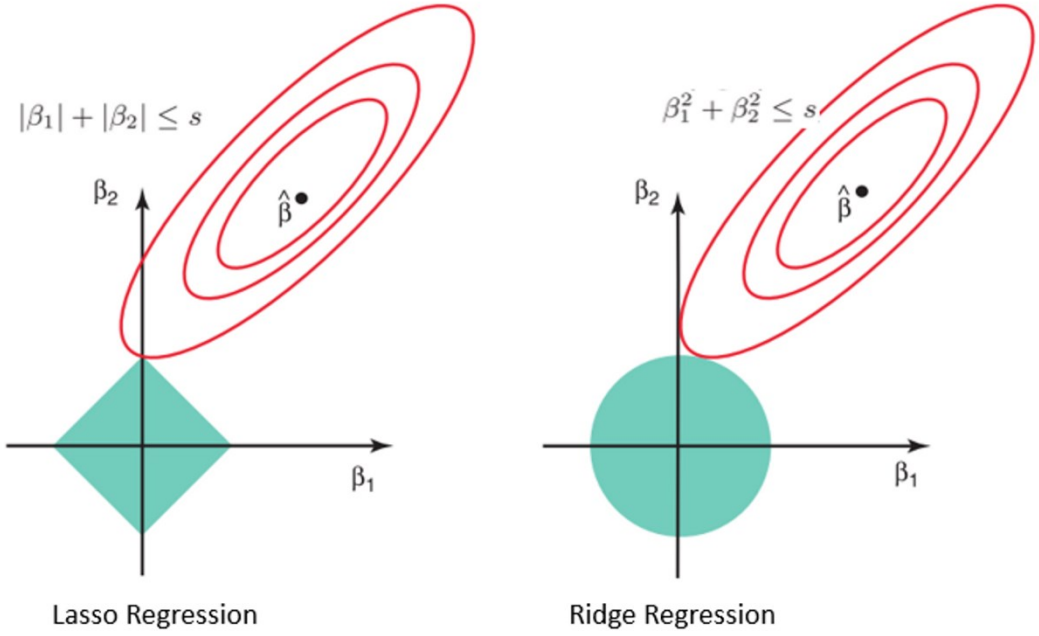

Figure 47 — Lasso & Ridge Regression which applies penalty of coefficients

Ridge regression is a regularization technique used to prevent overfitting in linear regression models. It was introduced by Hoerl and Kennard in 1970 and is also known as Tikhonov regularization [16].

In linear regression, the goal is to minimize the sum of squared errors between the predicted values and the actual values. However, if the number of features (also known as predictors or independent variables) is high compared to the number of observations in the dataset, there is a risk of overfitting. This means that the model will perform well on the training data, but will not generalize well to new, unseen data.

Ridge regression addresses this problem by adding a penalty term to the sum of squared errors. This penalty term is proportional to the square of the magnitude of the coefficients of the features. By adding this penalty term, the model is encouraged to select smaller coefficients for the features, thus reducing the complexity of the model and preventing overfitting.

The ridge regression model is represented by the following equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_n x_n + \epsilon \quad (32)$$

where y is the dependent variable, x1, x2, …, xn are the independent variables, β0, β1, β2, …, βn are the coefficients, and ε is the error term.

The objective of ridge regression is to minimize the following cost function:

$$Cost = RSS + \lambda \Sigma \beta i2 \quad (33)$$

where RSS is the residual sum of squares (i.e., the sum of squared errors), Σβi2 is the sum of the squares of the coefficients, and λ is the regularization parameter. The value of λ is a hyperparameter that needs to be tuned using cross-validation.

The effect of the penalty term is to shrink the coefficients towards zero, but not to exactly zero. This means that all the features are still included in the model, but some of them have smaller coefficients than others. The magnitude of the penalty term determines the amount of shrinkage, with larger values of λ leading to more shrinkage.

Ridge regression is a powerful technique for preventing overfitting in linear regression models, especially when dealing with datasets that have a large number of features. It is widely used in many fields, including finance, healthcare, and engineering, and has been shown to improve the predictive performance of linear regression models.

2.3.3 Logistic regression:

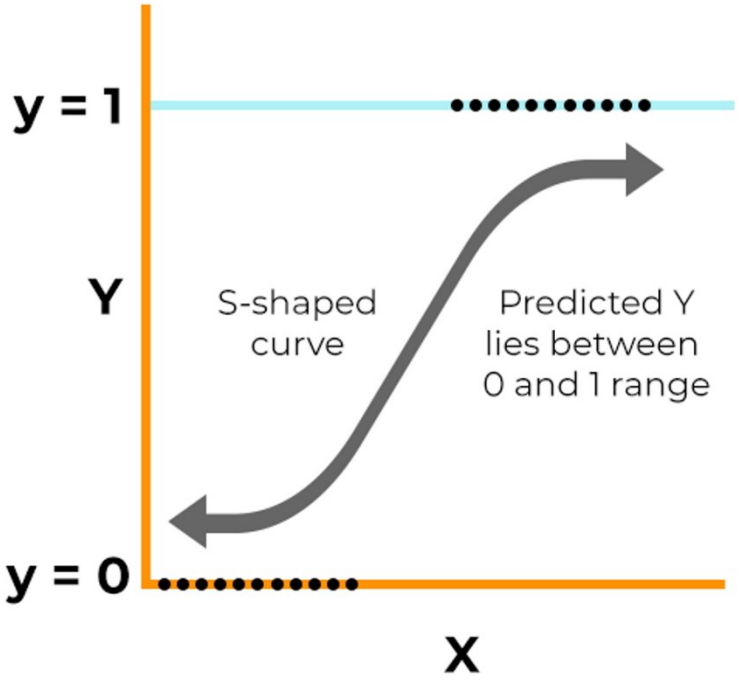

Figure 48 – logistic regression classifier

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). The logistic regression model provides a way to model this binary outcome as a function of the independent variables.

The logistic regression model is based on the logistic function, which is an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits. This is useful because the logistic function can be used to model the probability of an event occurring as a function of other factors.

The equation for the logistic function is as follows:

$$p(x) = \frac{1}{1 + e^{-z}} \tag{34}$$

where p(x) is the predicted probability of the dependent variable, e is the base of the natural logarithm, and z is the linear combination of the independent variables.

Types of Logistic Regression:

1. Binary Logistic Regression: Binary logistic regression is used when the dependent variable is binary (i.e., has only two possible outcomes). The independent variables can be continuous or categorical. For example, binary logistic regression can be used to predict whether a person is likely to buy a product or not based on demographic information.
2. Multinomial Logistic Regression: Multinomial logistic regression is used when the dependent variable has more than two categories. For example, multinomial logistic regression can be used to predict which type of product a person is likely to buy based on demographic information.
3. Ordinal Logistic Regression: Ordinal logistic regression is used when the dependent variable is ordered (i.e., has a natural ordering of the categories). For example, ordinal logistic regression can be used to predict the level of education a person is likely to have based on demographic information.

Advantages of Logistic Regression:

1. Logistic regression is easy to implement and interpret, making it a popular choice for many applications.
2. It can handle both continuous and categorical independent variables.
3. It provides the probability of the outcome, which can be useful in decision-making.
4. It can handle non-linear relationships between the independent variables and the dependent variable.

Disadvantages of Logistic Regression:

1. It assumes a linear relationship between the independent variables and the log odds of the dependent variable, which may not always be true.
2. It may not perform well when there are a large number of independent variables.
3. It may be sensitive to outliers and multicollinearity.
4. It assumes that the observations are independent of each other, which may not be true in some cases.

2.3.4 Random Forest:

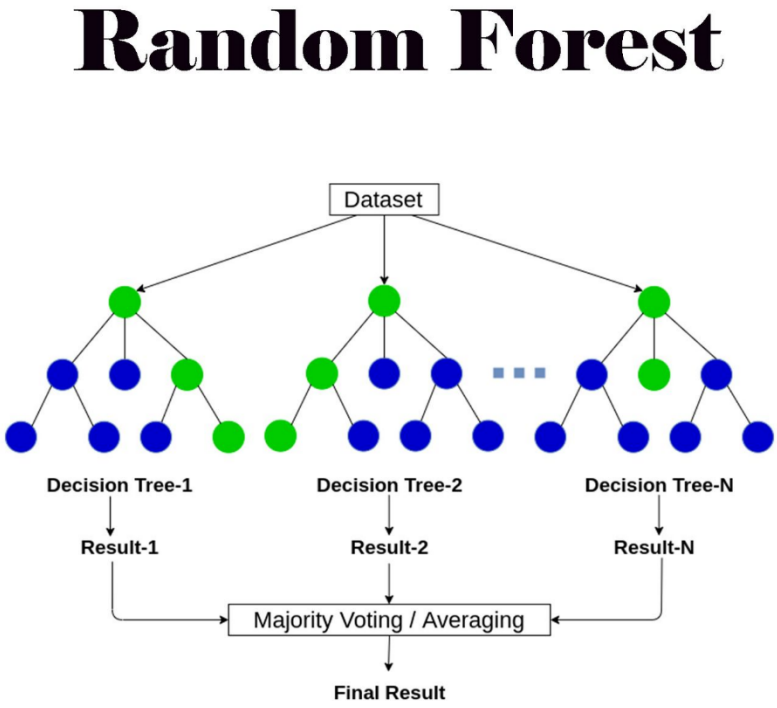

Figure 49 – Random Forest classifier model structure

Random Forest is a popular machine learning algorithm that is used for both classification and regression problems. It is a type of ensemble learning method that uses multiple decision trees to make predictions. Random Forest is a powerful algorithm that can handle complex and high-dimensional datasets, and it is known for its high accuracy and ability to handle missing values and outliers.

Random Forest works by creating a set of decision trees, where each tree is trained on a random subset of the data and a random subset of the features. The trees are built independently of each other, and each tree makes a prediction based on its own set of rules. The final prediction is made by combining the predictions of all the trees in the forest.

One of the advantages of Random Forest is that it can handle missing values and outliers, which can be problematic for other machine learning algorithms. When building the trees, Random Forest uses a technique called bagging, which randomly samples the data and the features to reduce the effect of outliers and noise. This makes Random Forest more robust and less likely to overfit the data.

Random Forest also has a built-in feature selection method, which helps to identify the most important features for the prediction. This is done by measuring the decrease in impurity when a feature is used to split the data. The features that result in the highest decrease in impurity are considered to be the most important.

The main hyperparameters of the Random Forest algorithm are the number of trees in the forest and the maximum depth of each tree. The number of trees should be large enough to capture the complexity of the data, but not so large that it leads to overfitting. The maximum depth of each

tree controls the level of detail in the decision rules and can be adjusted to balance between overfitting and underfitting.

2.3.5 ARIMA:

ARIMA = Auto Regressive Integrated Moving Average

ARIMA (Autoregressive Integrated Moving Average) is a commonly used time series forecasting model. It consists of three components: the autoregressive (AR) component, the integrated (I) component, and the moving average (MA) component. The AR component models the dependency of the current value on past values, the MA component models the dependency of the current value on past forecast errors, and the I component deals with non-stationary data by taking the difference between consecutive values.

AR — Autoregression —

A model that uses the dependent relationship between an observation and some number of lagged observations.

Integrated —

The use of differencing of raw observations (e.g., subtracting an observation from an observation at the previous time step) to make the time series stationary.

Moving Average —

A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations.

$$ARIMA(p, d, q) \tag{35}$$

Where:

- p — The number of lag observations included in the model, so-called the lag order.
- d — The number of times that the raw observations are differenced, also called the degree of differencing.
- q — The size of the moving average window, also called the order of moving average.

$$Y_t = c + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \theta_1 e_{t-1} + \cdots + \theta_1 e_{t-q} + e_t \tag{36}$$

Where:

- Y_t is the value of the time series at time t
- c is a constant term
- ϕ_1 to ϕ_p are the autoregressive coefficients
- $Y_{\{t-1\}}$ to $Y_{\{t-p\}}$ are the past values of the time series
- θ_1 to θ_q are the moving average coefficients
- $e_{\{t-1\}}$ to $e_{\{t-q\}}$ are the past forecast errors
- e_t is the error term or noise at time t

The ARIMA model can be used for both time series forecasting and modeling. It is widely used in finance, economics, and other fields to predict future values based on historical data. The model parameters can be estimated using various techniques, including maximum likelihood estimation, least squares estimation, and Bayesian methods.

2.3.6 LSTM:

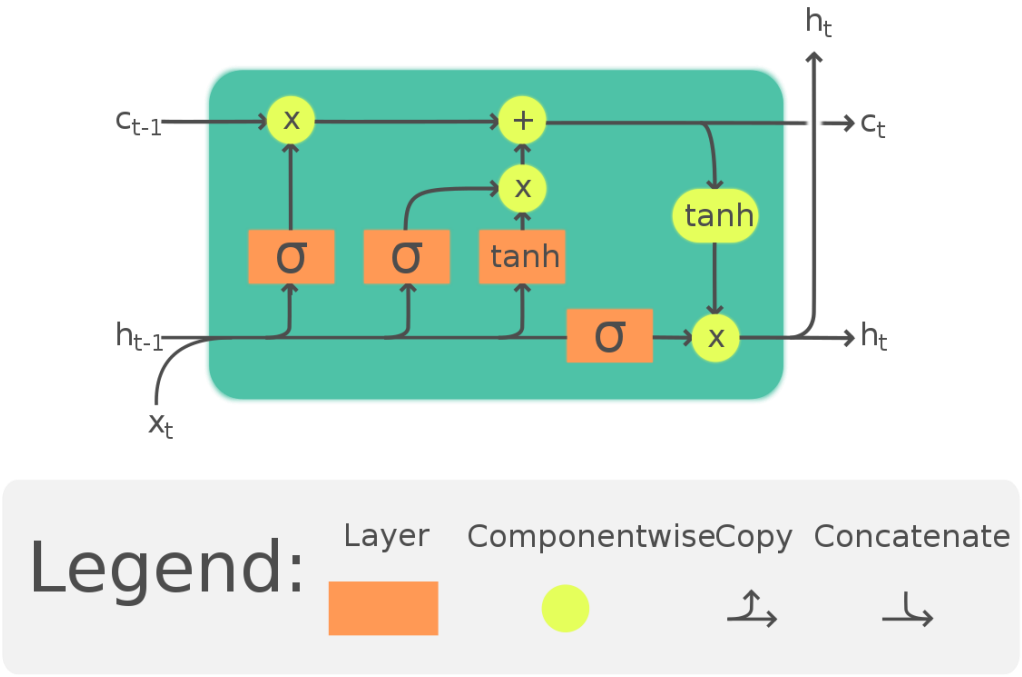

Figure 50 – LSTM unit architecture

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture that is capable of learning long-term dependencies in sequential data, such as time series. LSTMs are especially useful when the data contains long-term temporal dependencies that are difficult to capture with traditional feedforward neural networks.

The key to LSTM's success lies in its ability to selectively remember and forget information. It does this through the use of special memory units, called cells, that are connected by gates. The gates control the flow of information into and out of the cells, allowing the LSTM to selectively remember or forget information as needed.

The LSTM architecture consists of four main components: the input gate, the forget gate, the output gate, and the cell state. The input gate controls which information is allowed into the cell, the forget gate controls which information is discarded from the cell and the output gate controls which information is output from the cell. The cell state is responsible for storing and updating the information in the cell.

The LSTM model is trained using a variant of backpropagation called backpropagation through time (BPTT), which involves unrolling the network over time and applying the standard backpropagation algorithm to each time step. During training, the LSTM learns to adjust the weights and biases of its neurons to minimize a loss function, typically the mean squared error (MSE) or the binary cross-entropy (BCE) loss, depending on the problem being solved.

LSTM models have been successfully applied to a wide range of problems, including speech recognition, natural language processing, image captioning, and time series prediction. In particular, they have been shown to outperform traditional time series forecastings methods such as ARIMA and exponential smoothing on many datasets.

In summary, LSTM is a powerful deep-learning model that is capable of learning long-term dependencies in sequential data, making it well-suited for time series prediction tasks. Its ability to selectively remember and forget information, combined with its ability to handle variable-length input sequences, make it a popular choice for a wide range of applications.

2.4 Evaluation Metrics:

2.4.1 Mean Squared Error (MSE):

MSE measures the average of the squared differences between the predicted and actual values. It is commonly used for regression problems.

$$MSE = \frac{1}{n} * \sum \left(y_{actual} - y_{pred}\right)^2 \quad (37)$$

2.4.2 Root Mean Squared Error (RMSE):

RMSE is the square root of the mean squared error. It gives an idea of how much the predictions deviate from the actual values.

$$RMSE = \sqrt{MSE} \quad (38)$$

2.4.3 R-squared:

R-squared measures the proportion of variance in the target variable that is explained by the independent variables. It ranges from 0 to 1, with 1 indicating a perfect fit.

$$R^2 = 1 - (SS_{\boldsymbol{res}} - SS_{\boldsymbol{tot}}) \quad (39)$$

Where:

- SS_{res} - the sum of the squared residuals
- SS_{tot} - the total sum of squares.

2.4.4 MAPE (Mean Absolute Percentage Error):

MAPE is another widely used metric for evaluating the accuracy of a regression model, especially when dealing with time-series data. It measures the average percentage difference between the predicted and actual values. It is calculated as:

$$MAPE = \frac{1}{n} \times \sum \frac{|actual - predicted|}{actual} \times 100 \quad (40)$$

where n is the number of observations, actual is the actual value, and predicted is the predicted value.

The MAPE value ranges from 0% to infinity. A lower MAPE value indicates better model performance as shown in table 3. However, MAPE has some limitations, such as it treats over-predictions and under-predictions equally and can be sensitive to extreme values in the data.

Table 6 — MAPE score interpretation [18]

MAPE score	Interpretation of score
11. $> 50\%$	Poor
$20\% - 50\%$	Relatively good
$10\% - 20\%$	Good
$< 10\%$	Great

2.4.5 Accuracy:

Accuracy tells us how often we can expect our machine learning model will correctly predict an outcome out of the total number of times it made predictions.

$$accuracy\ score = \frac{TP - TN}{TP + FN + TN + FP} \tag{41}$$

2.4.6 Precision:

The model precision score measures the proportion of positively predicted labels that are actually correct.

$$precision\ score = \frac{TP}{FP + TP} \tag{42}$$

For multiclass classification: we use OvA one-vs.-all.

$$PRE_{macro} = \frac{PRE_1 + \cdots + PRE_K}{K} \tag{43}$$

2.4.7 Recall:

The model recall score represents the model's ability to correctly predict the positives out of actual positives.

$$recall\ score = \frac{TP}{FN + TP} \tag{44}$$

OvA multiclass recall:

$$REC_{macro} = \frac{PRC_1 + \cdots + REC_k}{k} \tag{45}$$

2.4.8 F1-score:

A harmonic mean for recall, precision score.

$$F1\ score = \frac{2 \times precision\ score \times recall\ score}{precision\ score + recall\ score} \quad (46)$$

$$F1\ score = \frac{TP}{TP + \frac{1}{2}\ (FP + FN)} \quad (47)$$

2.4.9 Simple strategy profit in dollars:

Add the daily return if the predicted and the actual returns have the same signs, subtract otherwise.

$$f(x, y) = \begin{cases} x, & if\ \ xy > 0 \\ -x, & otherwise \end{cases} \quad (48)$$

Where:

- x is the actual return
- y is the predicted return
- xy represents the product of actual and predicted returns

2.4.10 Pearson correlation:

Pearson correlation is a statistical method that measures the strength of the linear relationship between two continuous variables. It measures how much two variables are related to each other, and the direction of the relationship (positive or negative). The correlation coefficient, denoted as "r", ranges from -1 to 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation.

$$r = \frac{n \sum xy - \sum x \times \sum y}{\sqrt{n \sum x^2 - (\sum x)^2} \times (n \sum y^2 - (\sum y)^2)} \quad (49)$$

Where:

- r is the Pearson correlation coefficient
- n is the number of observations
- Σxy is the sum of the products of the paired deviations of x and y from their respective means
- Σx and Σy are the sums of the deviations of x and y from their respective means
- Σx^2 and Σy^2 are the sums of the squares of the deviations of x and y from their respective means

Pearson correlation is commonly used in data analysis to identify the relationship between variables and to identify potential patterns or trends in the data. It can help in making predictions, identifying outliers, and selecting relevant features in machine learning models.

2.5 Feature Selection and Extraction Techniques:

2.5.1 Multicollinearity

Multicollinearity is a phenomenon that occurs when two or more independent variables in a regression model are highly correlated with each other. It can cause several issues in the model, including unstable and unreliable estimates of the coefficients, inflated standard errors, reduced statistical power, and difficulty in interpreting the model.

In the presence of multicollinearity, it can be difficult to determine the individual effect of each independent variable on the dependent variable. This is because the correlation between the independent variables can make them appear equally important, leading to confusion about which variables should be included in the model and which should be removed.

There are several methods for detecting multicollinearity, such as the variance inflation factor (VIF) and the correlation matrix. The VIF measures the degree of correlation between each independent variable and all the other independent variables in the model, while the correlation matrix displays the correlation between each pair of independent variables.

To address multicollinearity, one approach is to remove one or more of the highly correlated variables from the model. Another approach is to combine the correlated variables into a single variable using techniques such as principal component analysis (PCA) or factor analysis.

To address the multicollinearity problem, I used VIF measures, Variance inflation factor (VIF) is a measure used to detect multicollinearity in regression analysis. It estimates how much the variance of an estimated regression coefficient increases if we include additional variables in a model. The VIF is calculated using the following equation: $VIF = 1 / (1 - R^2)$ where R^2 is the coefficient of determination for the regression of the predictor variable on the other predictor variables.

I used two techniques to calculate the VIF, the first one is to delete all correlated independents except one then calculate the VIF, and the second one is to delete all independents which have a correlation factor less than 0.7 and more than -0.7. The results were indicated that a very high multicollinearity between all independents as shown in Table 7

Table 7 – VIF values for the independents

Feature	VIF
xauusd_open	2.28e+06
xauusd_high	1.92e+06
xauusd_low	2.980e+06
usdchf_open	3.83e+06
usdchf_high	4.92e+06
usdchf_low	4.51e+06
usdchf_close	5.79e+06

To address this problem, I used PCA - Principal Component Analysis.

PCA - Principal Component Analysis:

Principal Component Analysis (PCA) is a widely used statistical technique that is used to reduce the dimensionality of a large dataset while preserving its original variance. It transforms the original variables into new uncorrelated variables, known as principal components. These principal components are sorted based on the amount of variance they explain and can be used for further analysis or visualization. The PCA equation involves eigenvalue decomposition of the covariance matrix of the original dataset.

To the number of components for PCA, I plot a chart of the number of components and their explained variance %.

Explained variance (%):

In principal component analysis (PCA), explained variance refers to the proportion of the total variance in the original data that is explained by each principal component.

Explained variance is typically expressed as a percentage and can be calculated for each principal component. For example, if the first principal component accounts for 50% of the total variance in the data, that means that half of the variability in the original dataset can be explained by that component. The second principal component may explain, say, 25% of the remaining variance, and so on.

Explained variance is an important concept in PCA because it allows us to understand how much of the original variability in the data is captured by each principal component. This information can help us decide how many principal components to retain for further analysis, or how well the PCA can represent the original data.

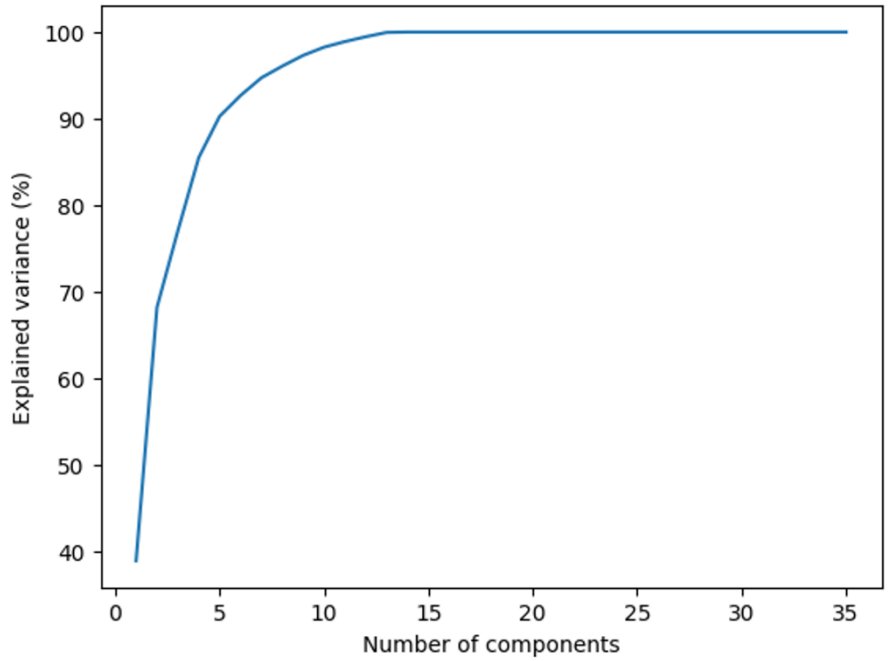

Figure 51 – An example of explained variance % - you should choose the number of components that the curve doesn't increase after it a lot.

As shown in Figure 51 — this example the explained variance freezes on 13 components so this is the number of components we will choose to decrease the dimensionality.

After this process, I operated a multicollinearity test, and the results indicate that no component is correlated with any other component as shown in Table 8.

Table 8 – VIF values for the new components

Features	VIF
Component_1	1.0
Component_2	1.0
Component_3	1.0
Component_4	1.0
Component_5	1.0
Component_6	1.0
Component_7	1.0

2.5.2 Mutual Information:

As we said before mutual information is a measure of the amount of information that two variables share. It quantifies the amount of information obtained about one variable by observing the other variable. Mutual information measures the extent to which the knowledge of one variable reduces the uncertainty of the other variable.

I applied the mutual information between the dependent and all the independents and then choose the top 35 ones.

2.5.3 Data augmentation:

Data augmentation is a technique used in machine learning and computer vision to artificially increase the size of a dataset by creating new variations of the existing data. This can be done by applying transformations to the original data, such as rotating, scaling, flipping, or cropping images, or by adding noise or other forms of distortion to the data. The goal of data augmentation is to improve the performance of machine learning models by increasing their ability to generalize to new data and reducing overfitting to the original training data. Data augmentation is particularly useful when working with limited amounts of data or when the original data is unbalanced or biased in some way.

I considered the smoothing process which I did in the data pipeline as data augmentation data.

3 Implementations and Results:

3.1 Linear regression

3.1.1 Scenario 1:

I applied Linear regression on all my 159 features without any feature selection. As shown in Table 9-1 the model gave good results with 0.28% mean absolute percentage error on test data, with the correlation between actual and predicted values of 0.992.

Table 9-1 – models result

Model	MSE	RMSE	R-squared	MAPE	Correlation	Strategy profit	Win %
Linear regression – s1	57.33 $	7.57 $	0.98	0.28	0.99	83.00 $	47.80%

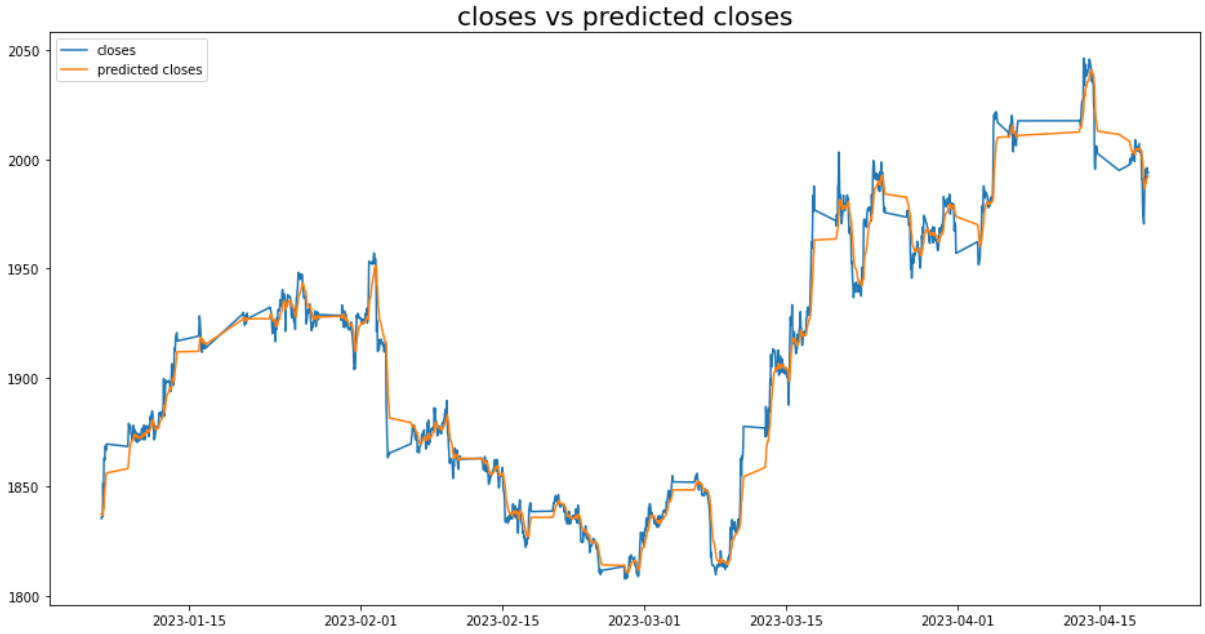

Figure 52 – close vs predicted closes prices linear regression model scenario 1

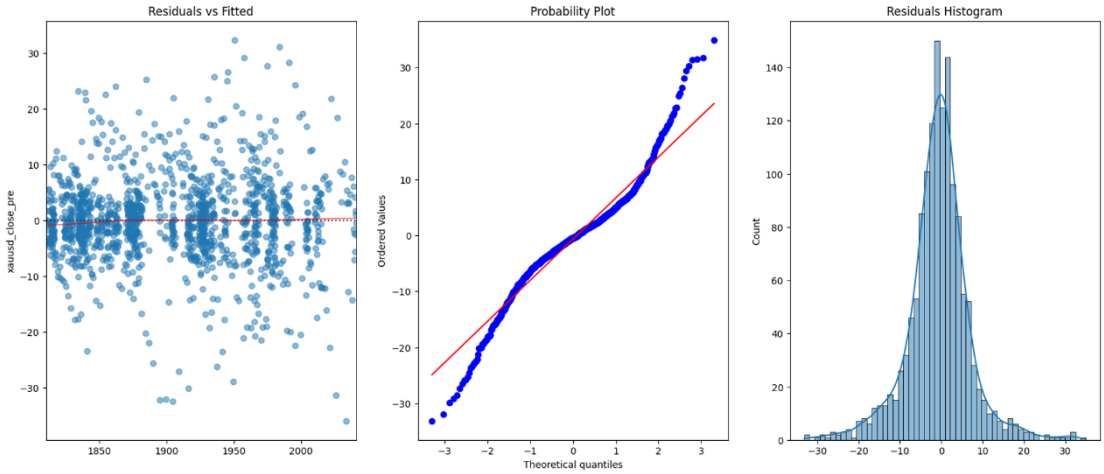

Figure 53 – Residual's analysis

As shown in Figure 53, the residuals are normal and considered as white noise which indicates that my model is optimized and can't be improved more than that.

3.1.2 Scenario 2:

Remove less correlated independents with dependent.

Table 9-2 – model result

Model	MSE	RMSE	R-squared	MAPE %	Correlation	Strategy profit	Win %
Linear regression – s1	57.33 $	7.57 $	0.98	0.28	0.99	83.00 $	47.80
Linear regression – s2	15.47 $	3.93 $	1.0	0.14	0.98	-72.7 $	47.50

3.1.3 Scenario 3:

Apply PCA on all features. using figure 54, we can conclude that about 80 components are required to explained all original features variance.

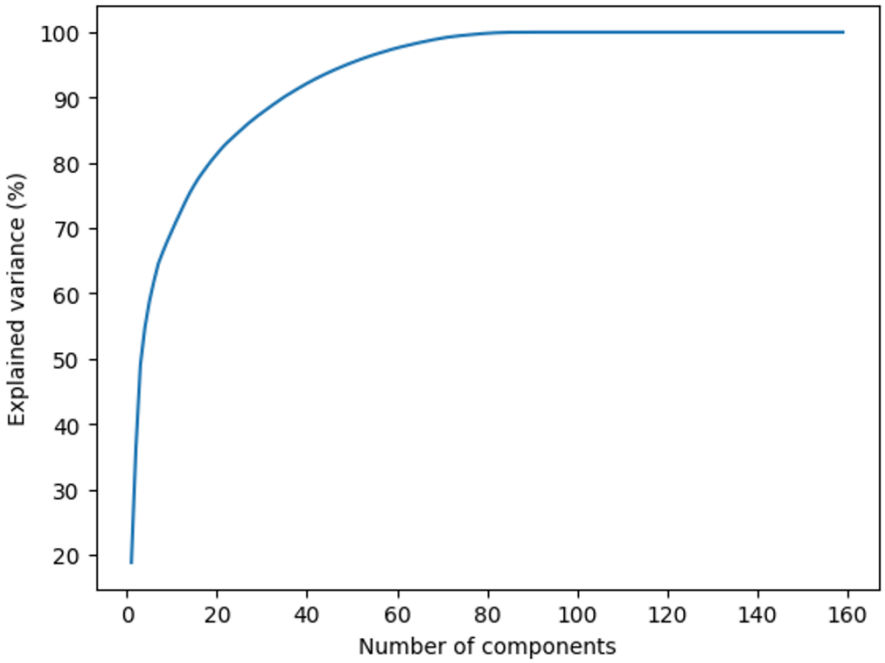

Figure 54 – Explained variance for PCA components

Table 9-3 – Models result

Model	MSE	RMSE	R-squared	MAPE %	Correlation	Strategy profit	Win %
Linear regression – s1	57.33 $	7.57 $	0.98	0.28	0.99	83.00 $	47.80 %
Linear regression – s2	15.47 $	3.93 $	1.0	0.14	0.98	-72.7 $	47.50 %
Linear regression – s3	57.24 $	7.57 $	0.98	0.28	0.99	114.8 $	49.2 %

As Table 9-3 shows, this mode makes more profit with more win %.

3.1.4 Scenario 4:

Using mutual information for feature selection with applying PCA.

Table 9-4 – Models result

Model	MSE	RMSE	R-squared	MAPE %	Correlation	Strategy profit	Win %
Linear regression – s1	57.33 $	7.57 $	0.98	0.28	0.99	83.00 $	47.80 %
Linear regression – s2	15.47 $	3.93 $	1.0	0.14	0.98	-72.7 $	47.50 %
Linear regression – s3	57.24 $	7.57 $	0.98	0.28	0.99	114.8 $	49.2 %
Linear regression – s4	88.60 $	9.41 $	0.98	0.36	0.99	108 $	49.1 %

As we mention before any linear model will need more data than we could obtain to reach a significant edge.

3.2 Lasso & Ridge regression:

Applying Lasso & Ridge regression for all features without any feature selection, the result is shown in Table 9-5

Table 9-5 – Models result

Model	MSE	RMSE	R-squared	MAPE %	Correlation	Strategy profit	Win %
Linear regression – s1	57.33 $	7.57 $	0.98	0.28	0.99	83.00 $	47.80 %
Linear regression – s2	15.47 $	3.93 $	1.0	0.14	0.98	-72.7 $	47.50 %
Linear regression – s3	57.24 $	7.57 $	0.98	0.28	0.99	114.8 $	49.2 %
Linear regression – s4	88.60 $	9.41 $	0.98	0.36	0.99	108 $	49.1 %
Lasso regression	53.57 $	7.32 $	0.99	0.27	0.99	130.75 $	48.7 %
Ridge regression	57.34 $	7.57 $	0.98	0.28	0.99	112.5 $	48.00 %

3.3 Logistic regression:

After using Mutual information for feature selection and PCA I applied the logistic regression on my 7 labels and obtained the results seen in Table 9-6 and the confusion matrix figure 55.

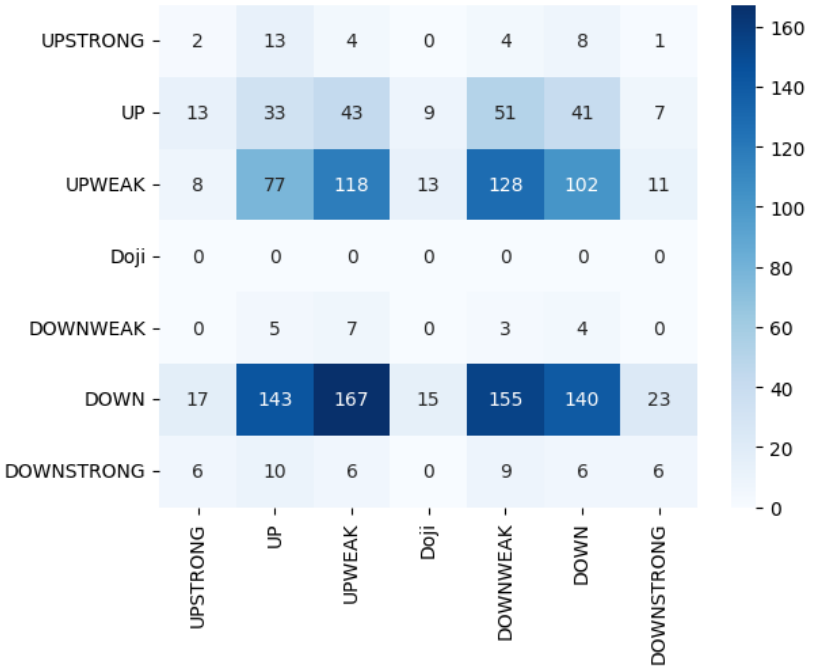

Figure 55 – performance matrix for logistic regression model.

3.4 Random forest:

Grid search:

Grid search is a technique for hyperparameter tuning that is widely used in machine learning. In grid search, a model is trained and evaluated using different combinations of hyperparameters that are specified in a grid. The model with the best performance on a validation set is then selected. Grid search can be time-consuming, but it is an effective way to find the best hyperparameters for a given model. It can also be combined with cross-validation to obtain more reliable estimates of the performance of the model

K-fold cross-validation:

K-fold cross-validation is a technique used in machine learning to evaluate the performance of a model on a given dataset. It involves partitioning the original dataset into k equal parts (or "folds"), then using one of these parts as the validation set and the remaining k-1 parts as the training set. This process is repeated k times, with each of the k folds being used once as the validation set. The results are then averaged to give a more accurate estimate of the model's performance. K-fold cross-validation is a widely used technique for model selection and hyperparameter tuning in machine learning.

After applying grid search to reach the best hyperparameters values the best parameters were `min_samples_leaf` 25, `min_samples_split` 25, `n_estimators` 700.

As I mentioned before, using binary classification in doesn't gave a good result so I increased my classes to 7 but I interested in the up or down signs for the day, I draw the learning curve to check for overfitting, I don't interest in the accuracy but the overfitting problem, and as shown in figure 56, as the sample size increase the overfitting decrease.

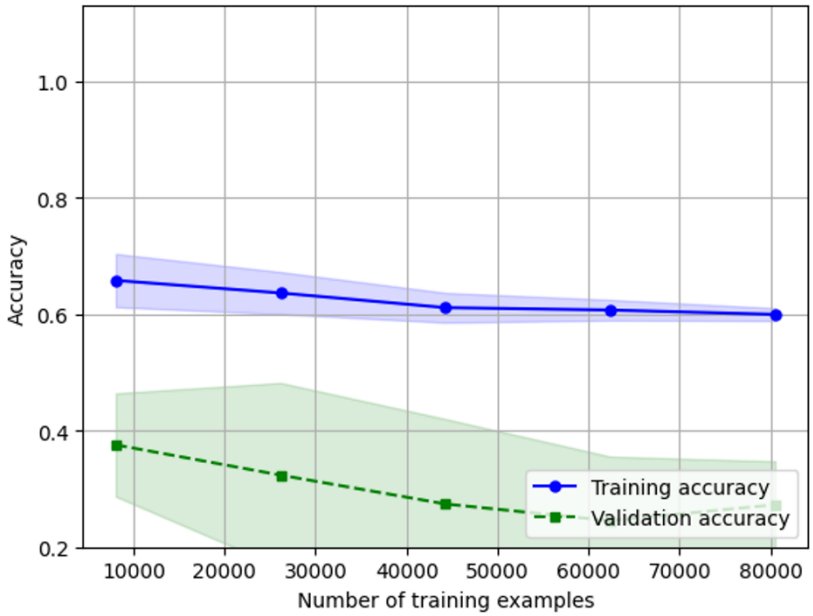

Figure 56 – Learning curve for random forest model

After applying K-fold cross-validation I obtained the results as shown in Table 9-6 and a performance metrics as shown in Figure 57.

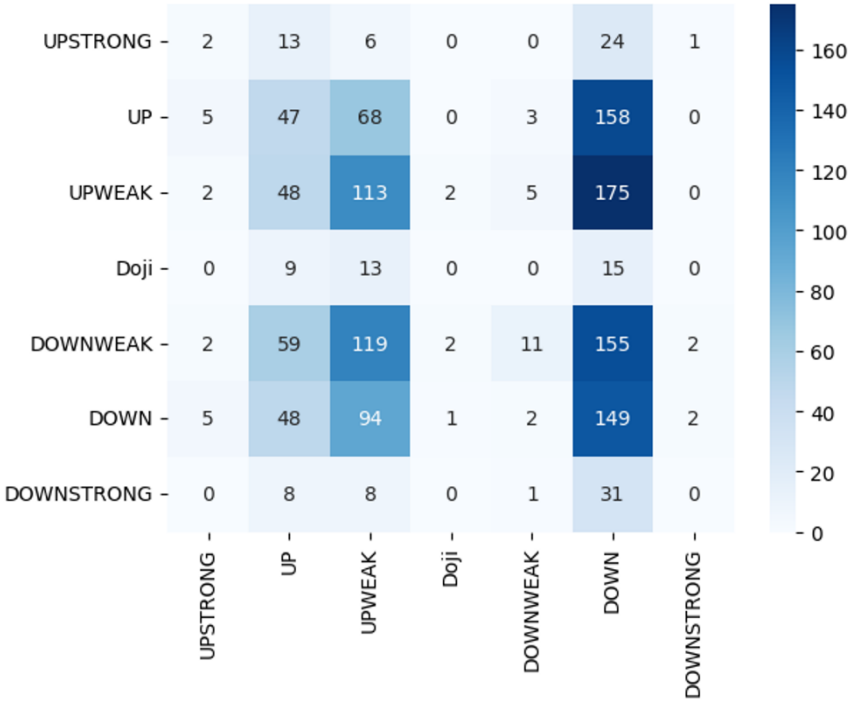

Figure 57 – performance metrics for random forest model

3.5 ARIMA:

I applied auto_arima from pmdarima library on daily return. every new value I obtained I applied auto_arima again. The result is shown in Table 9-6, and from Figure 58 we can that residuals are near to white noise. As shown in figure 59 the model didn't predict the value but can predict the direction.

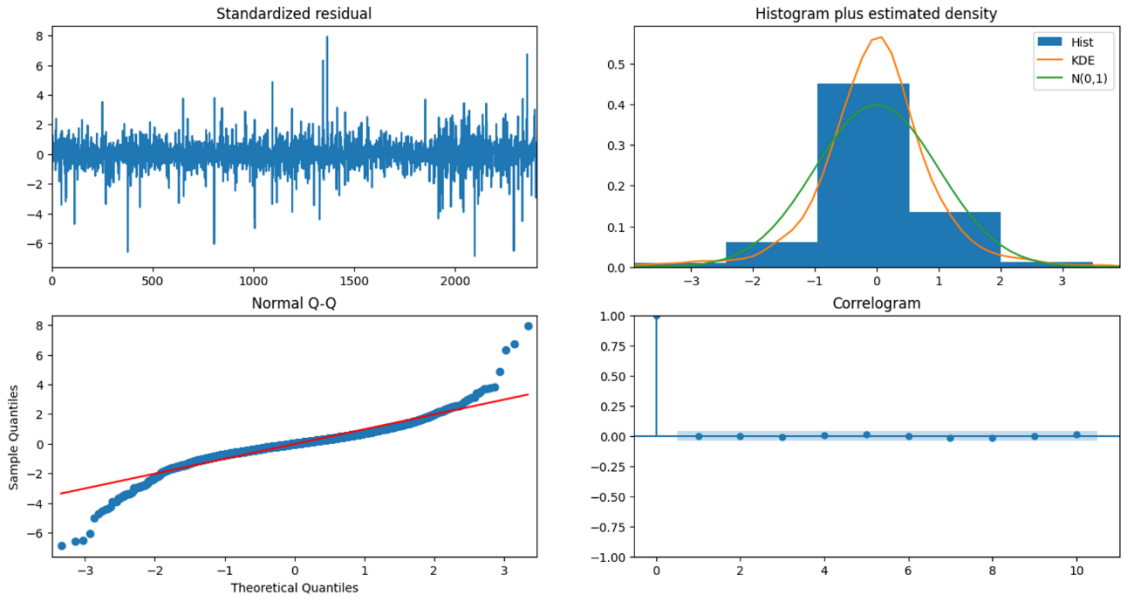

Figure 58 – Residual analysis for ARIMA model

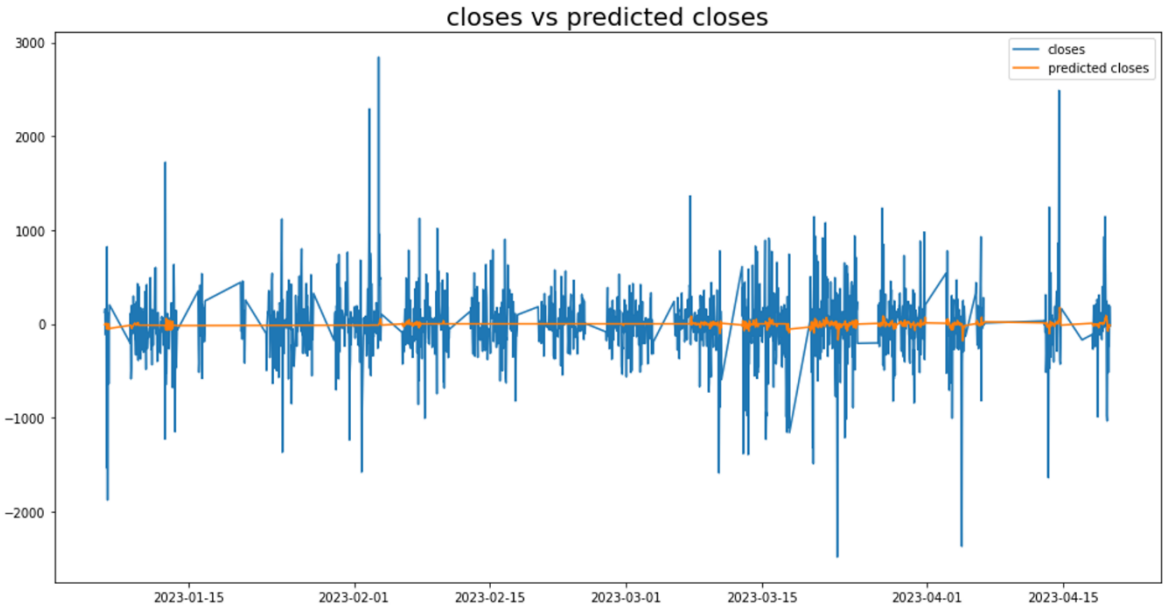

Figure 59 – actual vs predicted returns ARIMA model

With applying the ARIMA model to close prices, as shown in Table 9-6 the result was much worse.

ARIMA 2 (future covariates) :

I applied auto_arima with future covariates (features) as shown in Figure 60 using **darts** library in Python.

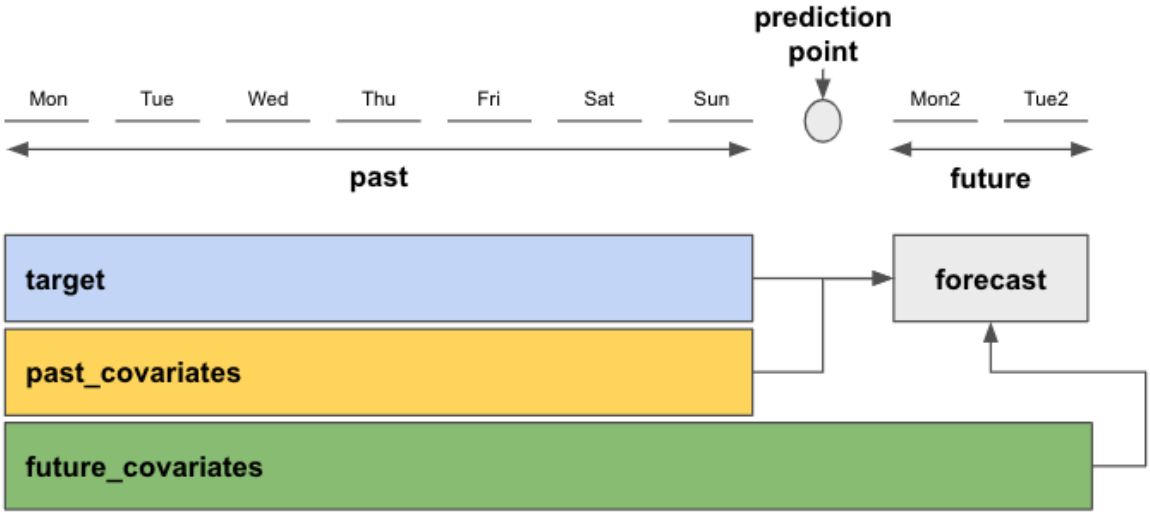

Figure 60 – future covariates explained (darts library)

The most important configuration is the training size. I made a new model every time I obtained a new value and trained on **100** previous values. The results are shown in Table 9-6 and Figure 61.

Figure 61 – Actual vs predicted returns – ARIMA 2 future covariates

ARIMA 2 (past covariates):

I applied auto_arima with past covariates (features) as shown in Figure 60 using darts library in Python.

The most important configuration is the training size. I made a new model every time I obtained a new value and trained on 100 previous values. The results are shown in Table 9-6 and Figure 62.

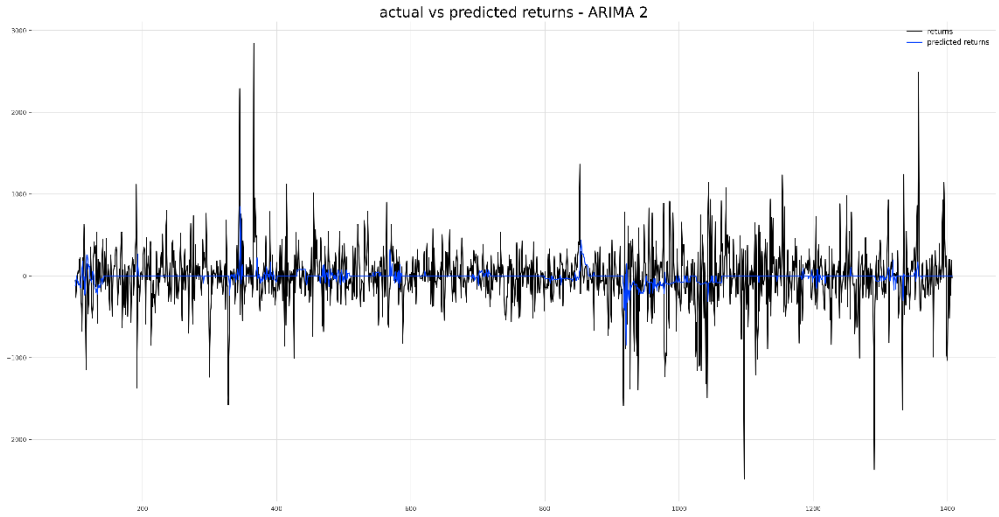

Figure 61 – Actual vs predicted returns – ARIMA 2 past covariates

3.6 LSTM:

For LSTM model I used ***tanh*** activation function as it's sensitive to outliers. with changing the window size, the hidden states and the layers. My best model was on returns with this structure and accuracy loss:

Table 10 – LSTM Model structure

Layer (type)	Output shape	Param #
lstm_90 (LSTM)	(None, None, 640)	1896960
dropout_52 (Dropout)	(None, None, 640)	0
lstm_91 (LSTM)	(None, None, 1280)	9835520
dropout_53 (Dropout)	(None, None, 1280)	0
lstm_92 (LSTM)	(None, None, 2560)	39331840
dropout_54 (Dropout)	(None, None, 2560)	0
lstm_93 (LSTM)	(None, None, 1280)	19665920
dropout_55 (Dropout)	(None, None, 1280)	0
lstm_94 (LSTM)	(None, 640)	4917760
dense_34 (Dense)	(None, 1)	641
Total params:		75,648,641
Trainable params:		75,648,641
Non-trainable params:		0

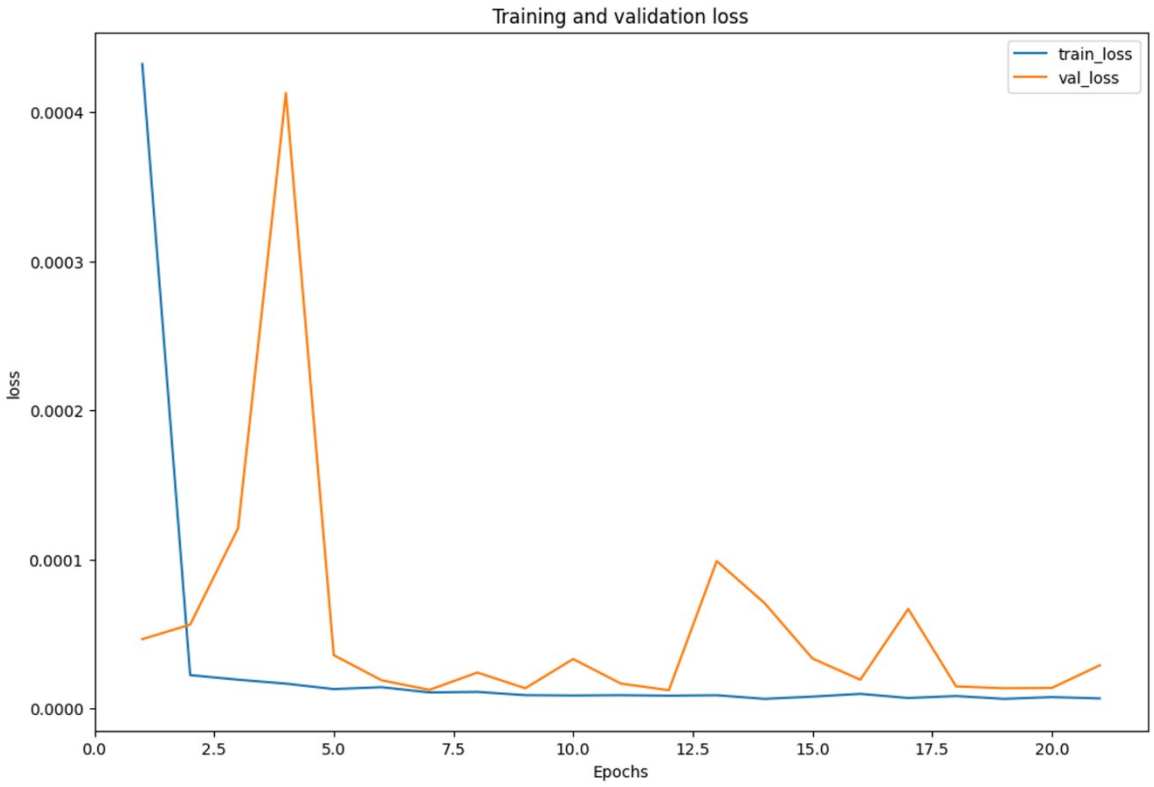

Figure 62 – Training and validation loss for LSTM model – loss function (MSE)

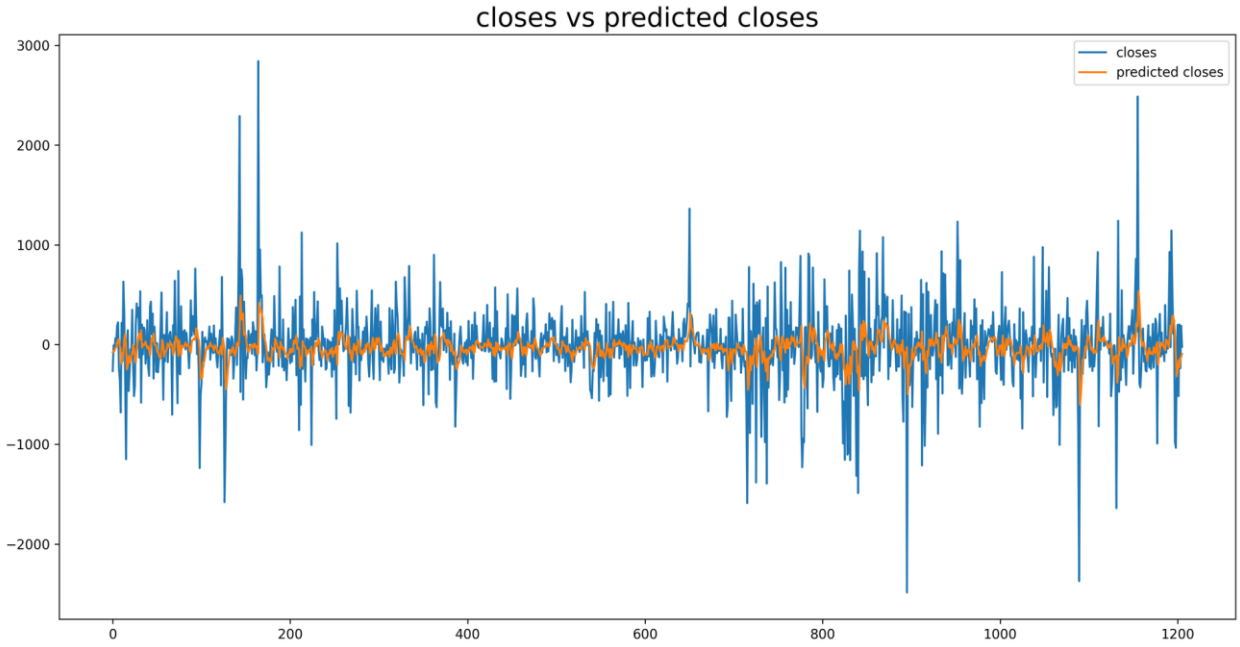

Figure 63 – close and predicted returns using LSTM

Table 9-6 – Models result

Model	MSE	RMSE	R-squared	MAPE	Correlation	Macro-Precision	Macro-Recall	Macro – f`1-score	Accuracy	Strategy Profit	Win %
Linear regression 1	57.33 $	7.57 $	0.98	0.28 %	0.992	-	-	-	-	83 $	47.8 %
Linear regression 2	15.47 $	3.93 $	1.0	0.14 %	0.98	-	-	-	-	- 72.7 $	47.5 %
Linear regression 3	57.24 $	7.52 $	0.98	0.28 %	0.99	-	-	-	-	114.8 $	49.2 %
Linear regression 4	88.6 $	9.41 $	0.98	0.36 %	0.99	-	-	-	-	108 $	49.1 %
Lasso regression	53.57 $	7.32 $	0.99	0.27 %	0.99	-	-	-	-	130.75 $	48.7 %
Ridge regression	57.34 $	7.57 $	0.98	0.28 %	0.99	-	-	-	-	112.5 $	48 %
Logistic regression	-	-	-	-	-	0.14	0.16	0.16	0.21	107.8 $	49.14 %
Logistic regression – binary labels	-	-	-	-	-	0.25	0.50	0.33	0.49	-178.5 $	-
Random forest	-	-	-	-	-	0.62	0.5	0.5	0.5	146.00 $	51 %
ARIMA – returns	152015 $	389.89 $	0.0	231 %	-	-	-	-	-	1083.00 $	67 %
ARIMA – closes	15.76 $	3.97 $	1.0	0.14 %	0.99	-	-	-	-	-1.12 $	52 %
ARIMA 2 – return future	404791.84 $	636.23 $	-0.13	544.22 %	0.23					633.76 $	54.2 %
ARIMA 2 – return past	155079.82 $	393.8 $	-31.66	7.32e+19 %	-0.01					2119.51 $	82.19 %
LSTM	159342.81 $	68.12 $	-1.34	667 %	0.07	-	-	-	-	443.37 $	55 %

3.7 Conclusion:

After applying statistical analysis and hypothesis tests, I concluded that my time series can be predicted and has valuable information. However, my training sample size was not sufficient to lead to significant results. I tested 7 models with different configurations, hyperparameters, and k-fold techniques, and found that the most profitable results were achieved by the ARIMA models. Specifically, my best profitable model was ARIMA 2 (auto_arima) on returns for a training size of 2000, which yielded a $ 2119.51 profit over a 4-month test period, or 211900 points in profit.

Please note that I did not calculate the commission per trade, nor did I employ any risk management techniques. For a significant sample size, I should use smaller periods with the target period and apply augmentation, but I lack the computational power to do so. Additionally, I should have used transformer models, but due to computational limitations, I was unable to do so.

ЗАДАНИЕ К РАЗДЕЛУ
«ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ»

Обучающемуся:

Группа	ФИО
8ПМ1И	Халил Марко Эбрахим Тхабет

Школа	Инженерная школа информационных технологий и робототехники	**Отделение школы (НОЦ)**	Отделение информационных технологий
Уровень образования	магистратура	**Направление/ООП/ОПОП**	09.04.04 «Программная инженерия»

Исходные данные к разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»:	
1. Стоимость ресурсов научного исследования (НИ): материально-технических, энергетических, финансовых, информационных и человеческих	1. Project budget: 540 850 RUB. 2. The cost of purchasing equipment: 87 400 RUB. 3. The cost of salary for supervisor: 26 000 RUB. 4. The cost of salary for ML research: 164 000 RUB.
2. Используемая система налогообложения, ставки налогов, отчислений, дисконтирования и кредитования	According to clause 3 of subclause 16 of Art. 149 of the Tax Code of the Russian Federation, this project is not subject to taxation. Based on Chapter 34 of the Tax Code of the Russian Federation, since 2016, the rate of 30.2% of the wage fund has been used to calculate contributions to extra-budgetary funds.
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. Pre-research analysis	Project Market potential consumers with their segmentation and competitiveness analysis of technical solutions.
2. Project initiation	Project goals and results, project structure, assumptions and limitations, planning, budgeting, etc.
3. Economic model development	Draw Ishikawa Diagram, Appling SWOT analysis, and Calculation of cash flows over 1 year.
4. Determining the effectiveness of projects	Net Present Value Calculation and Sensitivity Analysis
5. Final decision making	Conclude my final decision.
Перечень графического материала:	
1. «Portrait» of the consumer 2. Ishikawa Diagram 3. SWOT matrix 4. Sensitivity analysis chart	

Дата выдачи задания к разделу в соответствии с календарным учебным графиком	01.03.2023 г.

Задание выдал консультант по разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОСГН ШБИП	Спицына Л. Ю.	к.ф.н.		

Задание принял к исполнению обучающийся:

Группа	ФИО	Подпись	Дата
8ПМ1И	Халил Марко Эбрахим Тхабет		

4 Financial management, resource efficiency and resource saving

4.1 Pre-research analysis

4.1.1 Potential consumers of the research

4.1.1.1 Target Market

Gold has been a popular investment option for centuries due to its perceived value and stability. As a result, there are several potential consumers who may be interested in the research. These consumers include investors, traders, central banks, and jewelers as shown in Table 11.

Table 11 – Research potential consumers

Potential consumer of the project	consumer expectations
1. Traders and Investors	Algorithmic trading scripts are often used by traders and investors to automate their trading strategies and execute trades quickly and efficiently. With Expectations of ***(short-term profits).***
2. Financial Institutions	Financial institutions such as banks, hedge funds, and asset management firms may also be stakeholders in algorithmic trading scripts. They can use these scripts to execute large and complex trades or to manage their portfolio risk. ***(long-term profits, liquidations)***
3. Regulators	Regulators such as the Securities and Exchange Commission (SEC) and the Financial Industry Regulatory Authority (FINRA) may also be stakeholders as they monitor and regulate trading activities to ensure fairness, transparency, and compliance with regulations. ***(Fraud analysis)***
4. Software Developers	Software developers who create and maintain the algorithmic trading scripts can also be stakeholders as they are responsible for ensuring that the scripts are accurate, reliable, and secure. ***(Buy solutions)***

4.1.1.2 Segmentation

Market segmentation is a marketing strategy that involves dividing a larger market into smaller groups of consumers who have similar needs, preferences, or behaviors. The goal of market segmentation is to identify specific groups of consumers who are likely to respond positively to a particular product or service, and to tailor marketing efforts to meet the unique needs of each segment.

Market segmentation is based on the principle that not all consumers are the same, and that marketing messages and product offerings should be customized to appeal to specific groups of consumers. By dividing the market into smaller segments, companies can better understand the unique needs and preferences of each group, and develop targeted marketing campaigns that are more likely to be successful.

Here, figure 64. I mentioned my consumers segmentation map.

		Type of Algorithmic Trading Script Type				
		Short-term profit	Long-term profit	Liquidation	Fraud analysis	Development solutions
Type of users	Traders					
	Institutions					
	Regulators					
	Developers					

Figure 64 – Market segmentation map of research results

4.1.2 Competitiveness analysis of technical solutions

A competitiveness analysis is an important aspect of any research study, as it allows you to identify and evaluate your competitors in the market. In the case of your research on predicting gold future prices using machine learning, a competitiveness analysis would involve identifying other companies, products, and solutions that offer similar services or compete with your research.

The goal of a competitiveness analysis is to help you understand the strengths and weaknesses of your competitors, as well as their marketing strategies, pricing, and positioning. By analyzing the competitive landscape, you can identify opportunities to differentiate your research and develop a unique value proposition that sets it apart from your competitors.

A competitiveness analysis can also help you identify potential threats to your research, such as new entrants in the market, changes in consumer preferences, or advancements in technology. By staying informed about the competitive landscape, you can anticipate these threats and develop strategies to mitigate their impact on your research.

Overall, a competitiveness analysis is an important tool for evaluating the market potential of your research and identifying opportunities for growth and differentiation. By understanding the competitive landscape, you can position your research to meet the needs of your target audience and stand out in a crowded market.

4.1.2.1 Competitors

1. Traditional financial institutions: Financial institutions such as Goldman Sachs, JP Morgan, and Bank of America offer research and analysis on gold prices and trends.

2. Commodity trading firms: Firms such as Glencore, Trafigura, and Mercuria are involved in trading and marketing of commodities, including gold futures.
3. Data analytics firms: Companies like Bloomberg and Reuters offer real-time market data and analytics for gold futures trading.
4. Machine learning startups: Startups such as Kensho and Ayasdi offer predictive analytics and machine learning solutions for financial services.

4.1.2.2 Products Offered by Competitors

1. Research reports: Financial institutions offer research reports on gold prices and trends.
2. Trading platforms: Commodity trading firms offer online trading platforms for gold futures contracts.
3. Data analytics software: Data analytics firms offer real-time market data and analytics for gold futures trading.
4. Machine learning solutions: Machine learning startups offer predictive analytics and machine learning solutions for financial services, which can be used to predict gold future prices.

To secure funding for the project, it is essential to establish the commercial value of the work. Conducting an analysis of competitive technical solutions based on resource efficiency and savings enables a comparison of the scientific development's effectiveness. An evaluation card is a useful tool for performing this analysis.

Analysis of competitive technical solutions is determined by the formula:

$$C = \sum P_i \,.\, W_i \tag{50}$$

Where:

- C – the competitiveness of research or a competitor
- $W_i - cirterion\ weight;$
- P_i – Point of i-th criteria

Table 12 - Evaluation card for comparison of competitive technical solutions

Evaluation criteria	Criterion weight	Points					Competitiveness				
	W_i	P_f	P_{i1}	P_{i2}	P_{i3}	$Pi4$	c_f	c_{i1}	c_{i2}	c_{i3}	c_{i4}
Technical criteria for evaluating resource efficiency											
Growth in User' productivity	0.1	2	1	2	3	2	0.2	0.1	0.2	0.3	0.2
Convenience in operation	0.8	4	2	3	1	4	3.2	1.6	2.4	0.8	3.2
Reliability	0.9	4	1	2	3	4	3.6	0.9	1.8	2.7	3.6
Demand for computation power	0.3	2	3	4	4	2	0.6	0.9	1.2	1.2	0.6
Economic Criteria for performance evaluation											
Development cost	0.6	2	5	4	2	1	1.2	3	2.4	1.2	0.6
Grantee results	0.9	4	1	2	2	4	3.6	0.9	1.8	1.8	3.6
After-sales services	0.6	1	1	4	3	1	0.6	0.6	2.4	1.8	0.6
Clear prediction results	0.9	4	1	2	2	4	3.6	0.9	1.8	1.8	3.6

Calculation example:

$$c_f = \sum W_i \,.\, P_i = 0.2 + 3.2 + 3.6 + 0.6 + \cdots + 3.6 = 16.60 \quad (51)$$

$$c_{i1} = 8.9$$

$$c_{i2} = 14$$

$$c_{i3} = 11.6$$

$$c_{i4} = 16$$

By examining Table 12 and the results of competitiveness, we can conclude that based on our evaluation criteria my research schema will return best value for investors.

4.2 Project initiation

Project initiation is a crucial phase in any scientific project, as it sets the foundation for the entire project. Initiation processes involve defining the initial goals and content of the project, and fixing the initial financial resources.

Overall, the initiation phase is a critical first step in any scientific project, and laying a strong foundation during this phase is essential to ensure the project's success.

4.2.1 Project goals and results

In order to ensure the success of any scientific project, it is important to establish a strong foundation during the initiation phase. This phase involves defining the initial goals and content of the project.

The table below provides information about the hierarchy of project goals and the criteria for achieving those goals.

Table 13 — Project goals and results

1. Project goals	A research project aims to predict the future prices of Gold (XAUUSD) using machine learning techniques.
2. Expected results of the project	To reach an acceptable prediction accuracy.
3. Acceptance criteria of the project result	High predicting accuracy for volatility, trend, volume and prices of the future gold prices.
4. Requirement to the project results	Stability of the predicted model.
	Efficient Risk management plan, implemented into the algorithm.
	Giving clear signs for the future predicted values.
	Scaling availability.

4.2.2 Organizational Structure of the Project

The table below presents the organizational structure of the project.

Table 14 — Project Working Group

№	Name	Position	Functions	Hours spent
1	Серкей Аксёнов	Supervisor	Coordination of the research activities and assistance in Model implementation.	100
2	Khalil, Marco	ML scientist (researcher)	Find Data and Build, Train, and test machine learning model.	750
Total:				850

4.2.3 Assumptions and constraints

Limitations and assumptions are summarized in the table below.

Table 15 — Limitations and Assumptions

Factor	Limitations/Assumptions
1. Project budget – Research	~ 540 850 RUB
1.1 Source of Budgeting	Research Donations
2. Project Timeline	21 January 2023 – 15 May 2023
2.1 Date of approval of the project management plan	24 January 2023
2.2 Project completion date	~15 May

After initialized the project, a comprehensive set of goals and expected outcomes were established. This involved identifying the stakeholders involved in the project, as well as devising a robust financial framework. An effective financial framework is crucial in ensuring the successful completion and implementation of the project. Furthermore, a detailed timeline was established to ensure that each of the project's milestones is met within the stipulated time. By setting out clear objectives and timelines, the project team is better able to manage its resources and ensure that the project is completed on time and within budget. Finally, the project team also conducted a thorough risk analysis to identify potential risks that may arise during the project's implementation. This allowed the team to put in place appropriate risk mitigation strategies to minimize the impact of any unforeseen challenges that may arise.

4.2.4 Project planning

The main way to develop a design implementation schedule is called a Gantt chart. A Gantt chart is a horizontal graph that depicts work on a topic in long time periods, described by completion dates and start dates for the assigned work. It is a useful tool for project management because it provides a visual representation of the project's timeline, including the duration of each task and the dependencies between tasks. By using a Gantt chart, project managers can easily identify potential bottlenecks and delays in the project schedule, and adjust the timeline accordingly. Other benefits of using a Gantt chart include the ability to track progress, allocate resources, and communicate the project schedule to stakeholders. Overall, the Gantt chart is an essential tool for project managers and anyone involved in project planning and execution

Table 16 – Design and research timing

Task	**The laboriousness of the task**						**Duration of the task in working days** T_{pi}		Duration of the task in calendar days T_{ki}	
	tmin, person-days		tmax, person-days		$t_{ожi}$, person-days					
	Supervisor	ML scientist	Supervisor	ML scientist	Supervisor	ML scientist	Supervisor	ML scientist	Supervisor	ML scientist
Drawing up the technical assignment	3		8		5		5		7	
Literature review		4		9		6		6		9
Selection of the Research Field		5		8		6		6		9
Calendar planning	10		15		12		6		9	
Searching Historical Data		5		10		7		7		10
Draw statistics analysis for my data		1		2		1		1		1
Data pipeline: Collect features & Collect correlated Currency pairs		2		3		2		2		3
Data pipeline: cleaning my data		4		5		4		4		6
Data pipeline: Handle the outlier's problem		1		2		1		1		1
Apply statistics tests on my data		5		9		6		3		4
Choose ML Models		3		4		3		3		4
Build & Train & Evaluate ML Models	6		10		8		4		6	
Drawing up a final report		6		10		10		10		15

Table 17 – Schedule of the project design

№	Task	Executors	T_{ki} days	Duration of the task																				
				January				February				March				April					May			
				1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	5	1	2	3	4
1	Drawing up the technical assignment	Supervisor	3																					
2	Literature review	ML scientist	5																					
3	Selection of the Research Field	ML scientist	9																					
4	Calendar planning	Supervisor	1																					
5	Searching Historical Data	ML scientist	10																					
6	Draw statistics analysis for my data	ML scientist	5																					
7	Data pipeline: Collect features & Collect correlated Currency pairs	ML scientist	10																					
8	Data pipeline: cleaning my data	ML scientist	10																					
9	Data pipeline: Handle the outlier's problem	ML scientist	10																					
10	Apply statistics tests on my data	ML scientist	7																					
11	Choose ML Models	ML scientist	5																					
12	Build & Train & Evaluate ML Models	Supervisor	5																					
13	Drawing up a final report	ML scientist	10																					

– ML Scientist – Supervisor

The duration of work for an ML scientist is typically longer than that of a supervisor due to the complexity of tasks that ML scientist are responsible for. In general, the duration of work in calendar days for an ML scientist is ***82 days***, while that for a supervisor is only ***8 days***.

4.2.5 Project Budgeting

The project budget must accurately reflect all costs associated with implementation. These costs include:

1. Costs of purchasing computational equipment;
2. Expenses for Data services;
3. Expenses for Physical location and utilities;
4. Expenses fees of outside consultations;
5. Basic salary;
6. Contributions to social funds.

4.2.5.1 Costs of purchasing computational equipment

Table 18 – Calculation of the cost of basic equipment

Name of Equipment	Number	Equipment Cost, RUB
Monitors	2	60 000
Keyboard	2	4 000
Mouses	2	2 000
Motherboard	4	320 000
Graphic cards	8	2 400 000
SSD Hard desk	4	40 000
Linux OS	4	-
Wires & Others	-	10 000
Total		2 836 000 RUB

The cost of specialized equipment is recorded in the form of depreciation charges.

Depreciation is the gradual transfer of costs incurred to purchase equipment to the cost of the finished product.

Let's calculate the amount of monthly depreciation deductions in a linear way. Equipment costs are 2,836 thousand RUB. The operating life of all equipment is 8 years. Then the annual depreciation rate for them,

$$N_E = \frac{1}{8}.100\% = 12.5\% \quad (52)$$

Academic year depreciation (9 months) for equipment.

$$D_E = 2836\,.\,\frac{N_E}{100\%}\,.\,\frac{T}{365} = 2836\,.\frac{125}{100}\,.\,\frac{90}{365} = 87\,400\ RUB \quad (53)$$

4.2.5.2 Expenses for Data services

Data services contains solutions for providing very high accuracy historical data to financial assets, and cloud storage to back up your data and models.

Table 19 - Costs for data services

Service Name	Service Expenses, RUB
High accuracy Data Provider	60 000
Cloud Storage	20 000
Total:	80 000 RUB

4.2.5.3 Physical Location and utilities

To train and implement the ML models, a physical location with always on utilities must be provided. The duration of the project is 5 months, so I expenses is:

$$E_T = E_{1m} * 5 \quad (54)$$

Table 20 – Physical Location and Utilities

Resource Name	Resource Expenses, RUB
Physical location	80 000
Lights / waters	10 000
Fast Internet connection	10 000
Electricity	25 000
Total	125 000 RUB

4.2.5.4 Expenses fees of outside consultations

There is planning to consult with a third party but didn't go through.

Table 21 — Expenses fees of outside consultations.

Third-party consultations	Number	Fees, Thousand RUB
-	-	-

4.2.5.5 Basic Salary

The amount of expenses for wages of employees is determined based on the labor intensity of the work performed and the current system of salaries and tariff rates.

The calculation of the basic salary of the head of a scientific project is based on the sectoral wage system. The branch system of remuneration at TPU assumes the following composition of wages:

1. Salary – determined by the enterprise. In TPU, salaries are distributed in accordance with the positions held, for example, assistant, art, lecturer, associate professor, and professor (see 'Regulations on remuneration' given on the website of the Planning and Finance and Finance Department of TPU).
2. Incentive Payments – set by the head of departments for effective work, performance of additional duties, etc.
3. Other payments, district coefficient.

Since incentive bonuses, other payments and incentives depend on the activities of the manager in particular, we will take the coefficient of incentive bonuses equal to 30%, and the coefficient of incentives for the manager for conscientious work activity is 25%.

The basic salary of a manager is determined by the formula:

$$S_b = S_r \,.\, T_w \tag{55}$$

Where:

- S_r – worker's regular salary;
- T_w – duration of work, work days.

Additional Salary:

$$S_{add} = 0.15\, S_b \tag{56}$$

Average daily salary for a 5-day working week:

$$S_d = \frac{S_m \,.\, M}{F_d} \tag{57}$$

Where:

- S_m – worker's monthly salary, RUB;
- F_d – number of working days in a month, days
- M – number of months of work without vacation during the year

Full salary can be defined as:

$$S_f = S_b + S_{add} \tag{58}$$

Taking into account the document "Regulations on wages", associate professor, candidate of technical sciences, working at TPU has a salary equal to 26 300 rubles. A design engineer with no experience in Tomsk has an average salary of 18 000 rubles. With this in mind, we calculate the size of the total salary of the project manager and design engineer during the study.

Monthly salaries:

- For Project Supervisor:

$$s_b = s_r.\left(1 + K_{pr} + k_d\right).K_r = 26300.(1 + 0.3 + 0.25).1.3 = 52\,995\ RUB \quad (59)$$

$$S_F = S_b + S_{add} = 52\,995 + 0.15\,.52\,995 = 60\,944.00\ RUB \quad (60)$$

- For ML scientist:

$$s_b = s_r.\left(1 + K_{pr} + k_d\right).K_r = 18\,000.(1 + 0.3 + 0.25).1.3 = 36\,270\ RUB \quad (61)$$

$$S_F = S_b + S_{add} = 18\,000 + 0.15\,.\ 18\,000 = 41\,711.00\ RUB \quad (62)$$

Average daily salary:

$$S_D - \text{sup} = \frac{S_b}{F_d} = \frac{52995}{20.58} = 2575.10\ RUB \quad (63)$$

$$S_D - ml = \frac{s_b}{F_d} = \frac{35270}{20.58} = 1752.40\ RUB \quad (64)$$

Where the average number of working days in a month was determined as:

$$F_d = \frac{\mathrm{T}_w}{12} = \frac{247}{12} = 20.58 \quad (65)$$

Let's assume that the project manager spent 8 working days on it, then the ML scientist was engaged in the rest of the time (82 days). Salaries of project participants for the period of work:

$$S_{sup} = S_{\text{D.sup.}} \cdot t_{\text{sup}} = 2575{,}1 \cdot 8 = 23\,175{,}9\ \text{RUB} \quad (66)$$

$$S_{eng} = S_{\text{D.ml.}} \cdot t_{\text{ml}} = 1762{,}4 \cdot 82 = 14\,2754{,}4\ \text{RUB} \quad (67)$$

Additional salaries of project participants:

$$S_{add.sup.} = 0{,}15 \cdot 52995 = 7\,949.30\ \text{RUB}, \quad (68)$$

$$S_{add.ml.} = 0{,}15 \cdot 36270 = 5\,440.50\ \text{RUB} \quad (69)$$

Daily additional salaries:

$$S_{D.add.sup.} = \frac{7949.3}{20.58} = 386.30\ \text{RUB}, \quad (70)$$

$$S_{D.add.eng.} = \frac{5440.5}{20.58} = 264.40\ \text{RUB} \quad (71)$$

Additional salary for the entire project period:

$$S_{add.sup.} = S_{D.add.sup.} \cdot t_{sup} = 386.3 \cdot 9 = 3\,476.70 \text{ RUB}, \quad (72)$$

$$S_{\text{add.ml.}} = S_{D.add.ml.} \cdot t_{\text{ml}} = 264.4 \cdot 81 = 21\,416.40 \text{ RUB} \quad (73)$$

Full salary for the period of the project:

$$S_{F.sup.} = S_b + S_{add} = 23175.9 + 3476.7 = 26\,652.6 \text{ RUB}, \quad (74)$$

$$S_{F.ml.} = S_b + S_{add} = 142754.4 + 21416.4 = 164\,170.8\ RUB \quad (75)$$

4.2.5.6 Contributions to social funds

Here I will consider the obligatory contributions according to the norms established by the legislation of the Russian Federation to the state social insurance bodies (FSS), the pension fund (PF) and medical insurance (FFOMS) from the costs of wages of employees. The number of contributions to extra-budgetary funds is determined by the formula:

$$S_{\text{exb}} = k_{\text{exb}}(S_{\text{b}} + S_{add}), \quad (76)$$

where k_{exb}– contribution rate to extrabudgetary funds.

To date, the following contributions must be made from the amount provided as payment for labor:

- 22% towards the accrual of future pension;
- 5,1% to the Mandatory Health Insurance Fund;
- 2,9% to the Social Insurance Fund;
- from 0,2 to 8,5% for insurance against accidents that may occur at work (the exact amount depends on the risk class, which includes the profession and position of the employee).

The work of a manager and a design engineer belongs to the 1 risk class. Thus, the total deductions amount to 30,2%.

$$S_{\text{exb}} = 0{,}302 \cdot (23\,175{,}9 + 3\,476{,}7 + 142\,754{,}4 + 21\,416{,}4) = 57\,629 \text{ RUB} \quad (76)$$

4.2.5.7 Organization of research costs budget

Table 22 – Research cost budgeting

Name	Cost, RUB	Cost, %
Costs of purchasing computational equipment	87 400	16.16
Expenses of data services:	80 000	14.80
Expenses of Physical Location and utilities	125 000	23.11
Expenses fees of outside consultations.	-	0.00
ML scientist salary costs	164 170	30.35
Supervisor salary costs	26 650	4.93
Contributions to social funds	57 630	10.65
Total	540 850 RUB	

4.3 Economic Model development

4.3.1 Primary project analysis

The price of gold is an important economic indicator that affects many industries and markets around the world. Predicting gold prices can provide significant benefits for investors, financial institutions, and governments. By understanding the factors that drive gold prices and predicting their future movements, investors and institutions can make informed decisions that can lead to higher returns on investment and better risk management strategies.

One of the key benefits of predicting gold prices is that it can help investors and institutions to manage risk. Gold prices are affected by a variety of factors, including economic conditions, geopolitical events, and market sentiment. By analyzing these factors and predicting their future movements, investors can make informed decisions about when to buy and sell gold, which can help to mitigate the risk of losses and maximize returns.

In addition to risk management, predicting gold prices can also provide valuable insights into the global economy. Gold is often seen as a safe haven asset, meaning that it tends to rise in value during times of economic uncertainty or instability. By predicting gold prices, analysts can gain a better understanding of the current economic climate and make predictions about future economic conditions.

Another benefit of predicting gold prices is that it can help governments to manage their monetary policies. Gold prices are often used as a benchmark for currency values and inflation rates, and predicting their movements can help central banks to make informed decisions about interest rates and other monetary policy tools.

Finally, predicting gold prices can also be beneficial for traders and speculators in the gold market. By making accurate predictions about future price movements, traders can take advantage of market opportunities and make profitable trades.

In conclusion, predicting gold prices can provide significant benefits for investors, financial institutions, governments, and traders. By analyzing the factors that affect gold prices and making informed predictions about their future movements, analysts can help to manage risk, gain insights into the global economy, and make profitable trading decisions.

4.3.1.1 Ishikawa Diagram

The Ishikawa Diagram, also known as a Fishbone Diagram or Cause-and-Effect Diagram, is a visual tool used to systematically analyze the potential causes of a problem or issue. It helps to identify and categorize the factors that may contribute to a problem, and to develop effective solutions to address them.

As a research project my I can use Ichikawa diagram to dress my only problem, model accuracy.

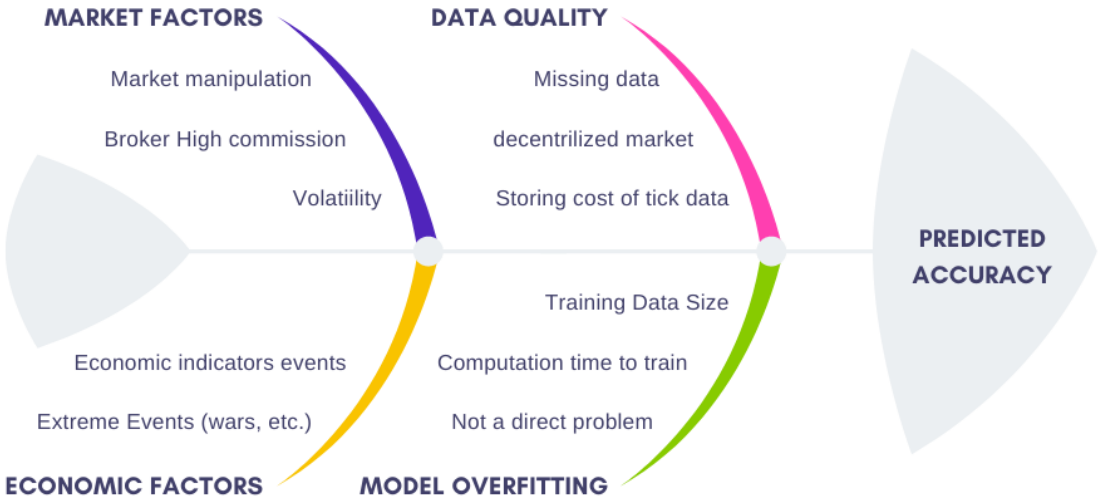

Figure 65 – Ishikawa Diagram for model accuracy

Data Quality:

The relationship between model accuracy and data quality is that model accuracy is highly dependent on the quality of the data used to train and test the model. Poor data quality, such as incomplete, inaccurate, or biased data, can lead to a lower model accuracy and limit the effectiveness of the model. In contrast, high-quality data, such as complete, accurate, and representative data, can lead to a higher model accuracy and better performance. Therefore, it is important to ensure data quality when developing and using predictive models.

Model overfitting:

Model accuracy and overfitting are related because high accuracy can sometimes be achieved through overfitting, which captures noise in the training data. Overfitting occurs when a model is too complex and does not capture underlying patterns useful for predicting new data. This can lead to high training accuracy but poor performance on new data. Maintaining a balance between model complexity and available data is important. Techniques such as cross-validation and regularization can prevent overfitting and ensure good generalization to new data.

Market Factors:

The relationship between model accuracy and market factors is that market factors can impact the accuracy of predictive models, particularly those used for financial forecasting. Market factors such as economic conditions, geopolitical events, and regulatory changes can affect the underlying patterns and relationships between variables that are used in the model. This can result in the model being less accurate or failing to capture changes in the market. Therefore, it is important to consider market factors and adjust the model accordingly to ensure it remains accurate and relevant in changing market conditions.

Economic Factors:

The relationship between model accuracy and economic factors is that economic factors can impact the accuracy of predictive models that are used for economic forecasting. Economic factors such as interest rates, inflation, GDP, and unemployment can affect the underlying patterns and relationships between variables that are used in the model. This can result in the model being less accurate or failing to capture changes in the economy. Therefore, it is important to consider economic factors and adjust the model accordingly to ensure it remains accurate and relevant in changing economic conditions.

4.3.1.2 SWOT Analysis

A SWOT analysis is a strategic planning tool used to evaluate the Strengths, Weaknesses, Opportunities, and Threats of a project, business, or organization. It involves identifying internal and external factors that can affect the success of the project, and using this information to develop strategies and make informed decisions. By analyzing the strengths and weaknesses of the project, as well as the opportunities and threats in the external environment, organizations can gain a better understanding of their competitive position and develop a more effective plan for achieving their goals.

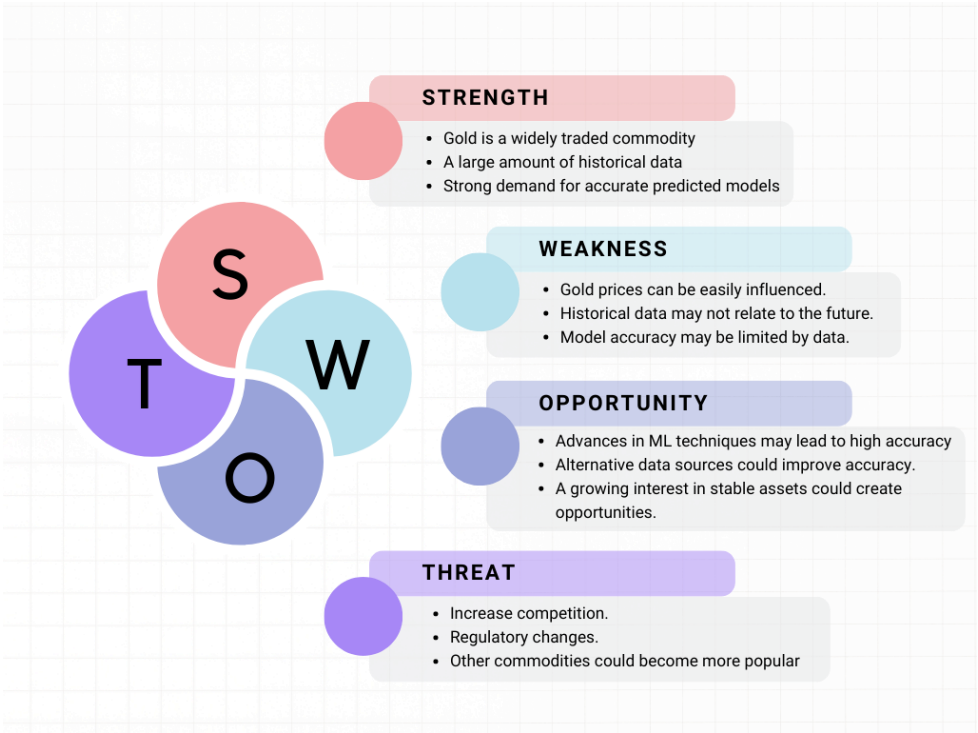

Figure 66 – SOWT Analysis

Strengths:

- S1 -- Gold is a widely traded commodity with a long history, making it a well-researched and well-understood market.
- S2 -- There is a large amount of historical data available for gold prices, which can be used to train and validate predictive models.
- S3 -- There is strong demand for accurate predictions of gold prices from investors, traders, and other market participants.

Weaknesses:

- W1-- Gold prices can be influenced by a wide range of factors, including economic, geopolitical, and market factors, which can make prediction difficult.
- W2 -- Historical data may not be a reliable indicator of future prices, as market conditions and other factors can change over time.
- W3 -- The accuracy of the predictive model may be limited by the quality and quantity of available data.

Opportunities:

- O1 -- Advances in machine learning and artificial intelligence techniques may allow for more accurate prediction of gold prices.
- O2 -- The increasing availability of alternative data sources, such as social media and news articles, could provide additional information to improve prediction accuracy.

- O3 -- The growing interest in sustainable and responsible investing could create new opportunities for predicting gold prices based on environmental, social, and governance factors.

Threats:

- T1 -- Increased competition from other market participants, including large financial institutions and hedge funds, could make it difficult to achieve a competitive edge in predicting gold prices.
- T2 -- Regulatory changes or geopolitical events could cause significant disruptions in the gold market, making it difficult to predict future prices accurately.
- T3 -- Other commodities, such as cryptocurrency or silver, could become more popular as investment options, reducing demand for gold and making predictions less relevant.

In addition, you need to identify the strengths and weaknesses of the research project to the external environmental conditions to determine the need for strategic changes. For this it is necessary to build the project matrices.

Table 23 - Strengths and Opportunities Project Matrix

	S1	S2	S3
O1	+	+	-
O2	-	-	+
O3	-	-	+

Analysis of this interactive spreadsheet showed correlated strengths and opportunities: O1S1S2, O2S3, O3S2.

Table 24 - Weaknesses and Opportunities Project Matrix

	W1	W2	W3
O1	+	+	-
O2	+	+	+
O3	-	-	-

The correlations of weaknesses and opportunities are as follows: O1W1W2, O2W1W2W3. The next step in project analysis is to identify the correlation of strengths and threats.

Table 25 - Strengths and Threats Project Matrix

	S1	S2	S3
T1	+	+	-

T2	-	-	-
T3	-	-	+

The correlations of threats and strengths are as follows: T1S1S2, T3S3. The next step in project analysis is to identify the correlation of weaknesses and threats.

Table 26 - Strengths and Threats Project Matrix

	W1	W2	W3
T1	-	-	-
T2	+	-	-
T3	-	-	-

The correlations of threats and strengths are as follows: T2W1.

4.3.2 Economic comparison of possible option for models

I used some models to predict gold prices for 4 months (01.01.2023 to 01.04.2023). my initial investment was 100,000 USD.

Table 27 – models' description

	Model name	Monthly Return %	Monthly Max drawdown %
Model 1	Linear regression	0.2	3
Model 2	Logistic regression	0.4	9
Model 3	Random Forest	0.3	18
Model 4	ARIMA 1	0.6	2
Model 5	LSTM	1.0	22

NPV formula for an investment with a single cash flow:

NPV (Net Present Value) is a financial metric used to determine the present value of future cash flows of an investment, taking into account the time value of money and discount rate. It helps in evaluating whether an investment is profitable or not by comparing the present value of cash inflows and outflows associated with it. A positive NPV indicates that the investment is expected to generate more cash inflows than outflows, making it a profitable one.

The only variables necessary to calculate the NPV for a short-term project with a single cash flow are the cash flow, period, and discount rate. For a one-year project with a single cash flow, the NPV formula is:

$$NPV = \frac{cash\ flow}{(1+i)^t} \tag{77}$$

Table 28 – NPV values for different i

	Model name	Discount rate	NPV
Model 1	Linear regression	0.2	2400
Model 2	Logistic regression	0.4	4800
Model 3	Random Forest	0.3	3600
Model 4	ARIMA 1	0.6	7200
Model 5	LSTM	1.0	12000

4.3.3 Sensitivity analysis

Sensitivity analysis of an investment project (sensitivity analysis) is an assessment of the impact of changes in the initial parameters of an investment project (investment costs, cash inflows, discount rate, operating expenses, etc.) on its final characteristics, which are usually used as IRR or NPV.

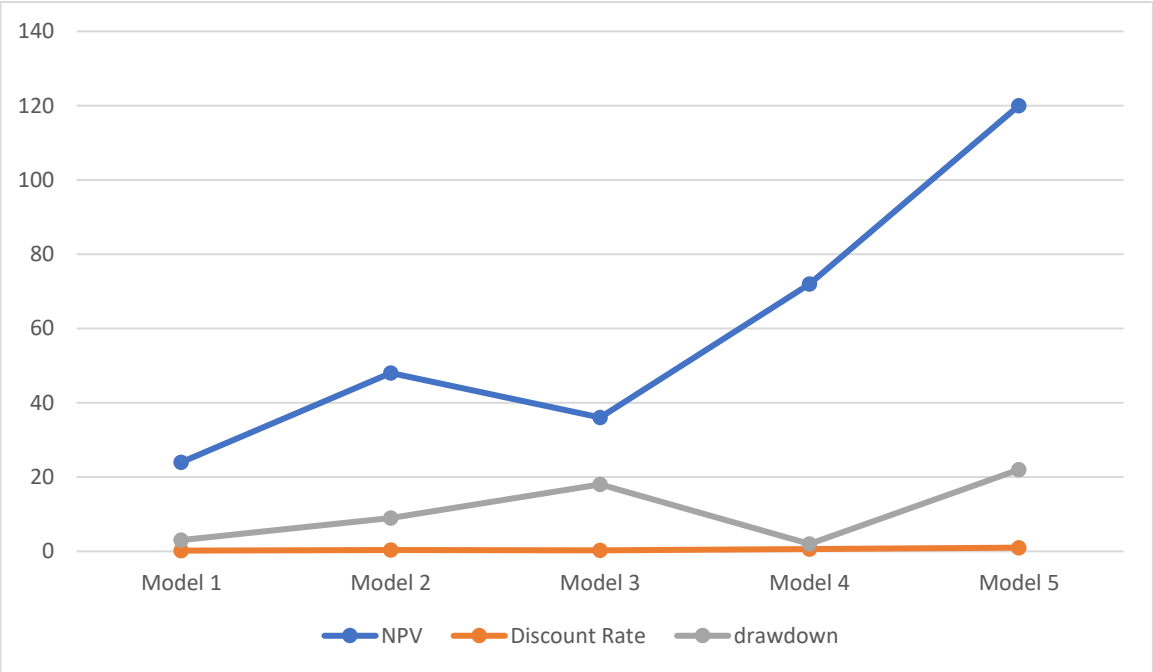

Figure 67 – Sensitivity of NPV & Drawdown to different models

4.4 Final decision making

By examine the sensitivity analysis I can conclude that; model 4 (ARIMA model) is the best model with the least draw down and the second NPV.

ЗАДАНИЕ К РАЗДЕЛУ
«СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ»

Обучающемуся:

Группа	ФИО
8ПМ1И	Халил Марко Эбрахим Тхабет

Школа	Инженерная школа информационных технологий и робототехники	**Отделение школы (НОЦ)**	Отделение информационных технологий
Уровень образования	магистратура	**Направление/ООП/ОПОП**	09.04.04 «Программная инженерия»

Исходные данные к разделу «Социальная ответственность»:	
1. Характеристика объекта исследования (вещество, материал, прибор, алгоритм, методика, рабочая зона) и области его применения	The work area at TPU office 19 is equipped with 4 seats, each of which includes; a chair, a computer with peripherals located on a table. The technological process is working with the python programming language, in the Jupyter notebook.
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. Правовые и организационные вопросы обеспечения безопасности: • специальные (характерные при эксплуатации объекта исследования, проектируемой рабочей зоны) правовые нормы трудового законодательства; • организационные мероприятия при компоновке рабочей зоны.	• GOST 12.2.032-78 SSBT. Workplace when performing work while sitting. General ergonomic requirements; • The labor code of the Russian Federation.
2. Производственная безопасность: 2.1. Анализ выявленных вредных и опасных факторов. 2.2. Обоснование мероприятий по снижению воздействия.	• Increased level of noise; • Lack or lack of natural light, insufficient illumination; • electromagnetic fields; • Increased voltage in an electrical circuit, the closure of which can pass through the human body.
3. Экологическая безопасность:	• Hydrosphere: toxins, chemicals, and heavy metals: lead, cadmium, lithium, alkaline manganese, and mercury; • Lithosphere: glass, metal, plastics, lead, barium, and rare earth metals.
4. Безопасность в чрезвычайных ситуациях:	Fire

Дата выдачи задания к разделу в соответствии с календарным учебным графиком	01.03.2023 г.

Задание выдал консультант по разделу «Социальная ответственность»:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ООД ШБИП	Антоневич О. А.	к.ф.н.		

Задание принял к исполнению обучающийся:

Группа	ФИО	Подпись	Дата
8ПМ1И	Халил Марко Эбрахим Тхабет		

5 Social responsibility

5.1 Introduction

The research project aims to use machine learning techniques to predict the future prices of gold (XAUUSD). The project product is some models with their accuracy report. To development of these models is only carried out with the help of computer computation power.

In this section, harmful and dangerous factors affecting the work of personnel will be consider, the impact of the developed models on the environment, legal and organization issue, measures in emergency situations will be considered.

The work was carried out in the building number 19 of TPU (8^{th} floor). Office 82B was a research execution place.

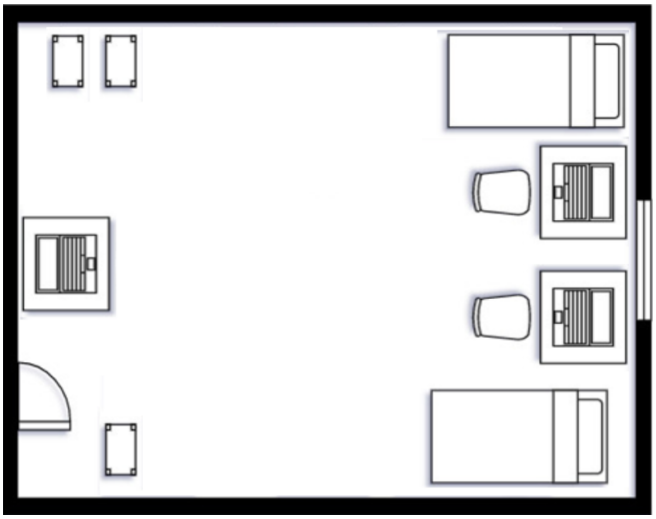

Figure 68 – Office 82B layout

5.2 Legal and organizational issue of occupational safety

According to the labor code of the Russian Federation, the maximum working hours permitted in a week is 40 hours. In recent times, one of the key approaches to significantly improve Preventive Measures aimed at, reducing the overall rate of incidents and work-related illnesses is the extensive implementation of an integrated Occupational Safety and Health (OSH) management system. This system involves combining individual activities into a unified system of focused actions at all levels and stages of the production process. In simpler terms, it means consolidating various safety and health measures into a comprehensive framework that addresses potential risks and promotes a safer working environment.

Preventive measures refer to actions or strategies taken to prevent or reduce the occurrence of undesirable events, risks, or hazards. These measures are implemented proactively to identify and address potential issues before they escalate into problems. In the context of the sentence you provided, preventive measures are initiatives or actions taken to minimize workplace incidents and occupational illnesses. These measures aim to create a safer and healthier work environment by identifying and addressing potential risks and hazards before they result in harm or negative outcomes.

An **Occupational Safety and Health management system** is a structured framework that integrates policies, procedures, and practices to effectively manage and improve workplace

safety and health. It involves systematic planning, implementation, evaluation, and continuous improvement of measures to prevent accidents, injuries, and occupational hazards, ensuring the well-being of employees and compliance with relevant regulations.

According to the Labor Code of the Russian Federation1, employees are entitled to various rights and guarantees related to labor safety, such as:

1. The right to safe and healthy working conditions that prevent harmful effects on workers' health (Article 1).
2. The right to receive information on working conditions and labor protection measures, including the results of assessment of working conditions, the level of occupational risk, the availability and quality of personal protective equipment, the results of medical examinations and other relevant data (Article 212).
3. The right to demand elimination of violations of labor protection requirements, including the right to refuse to perform work or to leave the workplace in case of imminent danger to life or health, without losing wages (Article 214).
4. The right to receive training and instruction on labor protection, including the right to undergo initial and periodic training on safe methods and techniques of work, the rules of using personal and collective protective equipment, the actions in case of emergency situations and accidents (Article 215).
5. The right to participate in the management of labor protection, including the right to elect representatives for labor protection issues, to join trade unions or other public associations that protect workers' interests in labor protection, to participate in inspections and investigations of accidents and occupational diseases (Article 216).
6. The right to compensation for harm caused by injury or other damage to health in connection with the performance of work duties, including the right to receive insurance payments from compulsory social insurance against industrial accidents and occupational diseases, as well as additional payments from the employer in accordance with federal laws and other normative legal acts (Article 217).

These rights and guarantees are further specified and regulated by different articles of Chapter 16 of the Labor Code [19]

The introduction of labor protection rules and safety measures serves the purpose of preventing accidents, ensuring the safety of working conditions for all workers. These rules are mandatory for individuals at all levels, including workers, managers, engineers, and technicians.

To ensure a proper and ergonomic workstation for PC users, the following guidelines should be followed according to GOST 12.2.032-78 SSBT [20]:

1. Adequate Space: The minimum area for a PC workstation should be at least 6 square meters, allowing sufficient room for movement and comfort.
2. Legroom: The legroom should meet specific parameters, including a minimum legroom height of 600 mm, a seat distance to the lower edge of the working surface of at least 150 mm, and a seat height of 420 mm. It's important to adjust the height of the table based on the operator's height for optimal ergonomics.

3. Working Chair Design: The working chair should be designed to maintain a proper working posture while using the PC, allowing for posture changes to reduce static tension in the neck, shoulder muscles, and back, thus preventing fatigue.
4. Chair Selection: The choice of working chair should consider the user's height, the nature of work with the PC, and the duration of PC use. The chair should have lifted and swivel capabilities, adjustable seat and back height and angle of inclination, and the distance between the back and the front edge of the seat. Each parameter should be independently adjustable, easy to use, and securely lockable.

It's important to adhere to the guidelines provided in GOST 12.2.032-78 SSBT [20], a Russian technical standard related to occupational safety and ergonomic requirements.

By following these recommendations, the workplace can be optimized to promote comfort, reduce musculoskeletal strain, and enhance overall productivity for PC users.

5.3 Occupational safety

Occupational safety refers to the efforts and measures taken to ensure the well-being, health, and safety of individuals within the workplace. It involves identifying and mitigating potential hazards, implementing safety protocols, providing proper training and protective equipment, and promoting a culture of safety among employees. The goal of occupational safety is to prevent work-related accidents, injuries, and illnesses, thereby creating a secure and healthy environment for workers. By prioritizing occupational safety, organizations can protect their employees, enhance productivity, and comply with relevant regulations and standards.

Workplace safety is a shared responsibility within the organization, requiring the active participation of all individuals involved.

Occupational hygiene encompasses - a comprehensive system that ensures the well-being of workers throughout their labor activities. It involves a range of measures, including legal, socio-economic, organizational, technical, sanitary and hygienic, treatment and preventive, and rehabilitative efforts.

Working conditions - refer to the combination of factors present in the work environment and the labor process that can influence both the health and performance of individuals.

A harmful production factor - is an environmental or work process-related element that has the potential to cause occupational diseases, result in a temporary or permanent decline in work capacity, increase the occurrence of somatic and infectious diseases, and even have adverse effects on the health of future generations. It is essential to identify and address these factors to safeguard the well-being and long-term health of individuals in the workplace.

A hazardous production factor - refers to an environmental or work process-related element that poses a significant risk of causing injuries, acute illnesses, sudden and severe health deterioration, and even fatalities. Identifying and effectively managing these factors is crucial to ensuring the safety and well-being of individuals in the workplace.

In this subsection, it is necessary to analyze harmful and hazardous factors that can occur during research in the laboratory when the development or operation of the designed solution (in a workplace).

GOST 12.0.003-2015 “Hazardous and harmful production factors. Classification” [21] should be applied to identify potential factors that can affect the health and safety of a worker (employee).

Table 29 – Potential hazardous and harmful production factors

Factors (GOST 12.0.003-2015)	Stages of work			Legislation documents
	developing	manufacturing	operation	
1. Increased levels of noise	+	+		**GOST 12.1.003-2014.** Occupational safety standards system. Noise. General safety requirements
2. Lack or lack of natural light, insufficient illumination	+			**SanPiN 2.2.1/2.1.1.1278-03.** Hygienic requirements for natural, artificial and mixed lighting of residential and public buildings
3. Electromagnetic fields	+	+	+	**SanPiN 2.2.4.1329-03.** Requirements for protection of personnel from the impact of impulse electromagnetic fields
4. Abnormally high voltage value in the circuit, the closure which may occur through the human body		+	+	**Sanitary rules GOST 12.1.038-82 SSBT.** Electrical safety. Maximum permissible levels of touch voltages and currents.

5.3.1 Potential hazardous and harmful production factors

5.3.1.1 Increased levels of noise

Noise can be generated by operating equipment, air conditioning units, daylight illuminating devices, as well as from outside sources. It worsens working conditions and has harmful effects on the human body, specifically the organs of hearing and the whole body through the central nervous system. This results in weakened attention, deteriorated memory, decreased response, and an increased number of errors in work.

When working on a PC, the noise level in the workplace should not exceed 50 dB, according to **GOST 12.1.003-2014** Occupational safety standards system, which outlines general safety requirements for noise. To study in a quiet environment, irrelevant applications on the computer should be closed to reduce power consumption and computer noise. Additionally, windows should be closed to reduce environmental noise.

5.3.1.2 Lack or lack of natural light, insufficient illumination

Light sources can be both natural and artificial. The sun is a natural source of light in a room, while lamps provide artificial light. Working for long periods in low illumination conditions and with improper illumination parameters can lead to decreased visual perception, myopia, eye diseases, and headaches.

According to the **SanPiN 2.2.1/2.1.1.1278-03** standard, the illumination on the table surface in the area of the working document should be between 300-500 lux. Lighting should not create glare on the surface of the monitor, and the illumination of the monitor surface should not exceed 300 lux. The brightness of common light lamps in the area with radiation angles from 50 to 90° should be no more than 200 cd/m, and the protective angle of the lamps should be at least 40°. The ripple coefficient should not exceed 5%.

5.3.1.3 Electromagnetic fields

In this case, the personal computer is the source of the increased intensity of the electromagnetic field. **SanPiN 2.2.4.1329-03** outlines the requirements for protecting personnel from the impact of impulse electromagnetic fields, and deems an intensity of 8 kA/m acceptable. To avoid the negative effects of electromagnetic radiation, an employee's working day at their computer should not exceed one hour with an intensity of no more than 8 kA/m and a magnetic induction level of 0.01 T.

To reduce the level of the electromagnetic field from a personal computer, it is recommended to connect no more than two computers to one outlet, make a protective grounding, and connect the computer to the outlet through an electric field neutralizer. Personal protective equipment when working on a computer includes spectral computer glasses, which improve image quality and protect against excessive energy flows of visible light. These glasses reduce eye fatigue by 25-30% and are recommended for all operators working more than 2 hours a day, and in case of visual impairment by 2 diopters or more - regardless of the duration of work.

System units and monitors of switched-on computers are sources of electromagnetic radiation in the workplace. To reduce exposure to such types of radiation, it is recommended to use monitors with reduced radiation levels, install protective screens, and observe work and rest regimes.

According to the intensity of the electromagnetic field at a distance of 50 cm around the screen, the electrical component should not exceed a certain threshold.

- In the frequency range of 5 Hz - 2 kHz, the value is 25 V/m.
- In the frequency range of 2 kHz - 400 kHz, the value is 2.5 V/m.

The magnetic flux density should not exceed:

- In the frequency range of 5 Hz to 2 kHz - 250 nano T.
- In the frequency range of 2 kHz to 400 kHz - 25 nano T.

There are various ways to protect against EMF exposure:

- Increase the distance between the screen and the user to at least 50 cm.
- Consider using pre-screen filters, special screens, or other personal protective equipment.

5.3.1.4 Abnormally high voltage value in the circuit

When a person comes into contact with an electric current, the physical impact of that current on their body can lead to electrical injuries. Some common types of injuries that can occur as a result of electric accidents are:

- Burns: Electric current can cause burns on the skin due to the heat generated by the flow of electricity through the body.
- Electric shocks: Electric shocks refer to the sensations or signs that a person may experience when they come into contact with an electric current. This can include muscle contractions, numbness, tingling, or even paralysis.
- Skin metallization: This term refers to a phenomenon where the skin comes into contact with a high-intensity electrical current, causing the surface of the skin to melt or become metallic in appearance.
- Tissue tears: When a strong electric current passes through the body, it can cause damage to tissues, leading to tears or lacerations.
- Dislocations of joints: The forceful muscle contractions caused by electric shocks can sometimes result in the dislocation of joints, causing them to come out of their normal positions.
- Bone fractures: In severe cases, the strong muscular contractions induced by electric shocks can lead to fractures or broken bones.

Safety measures for electrical work to prevent accidents and injuries:

- Voltage disconnection from live parts, on which or near which work will be carried out, and taking measures to ensure the impossibility of applying voltage to the workplace.
- Posting of posters indicating the danger at the place of work.
- Electrical grounding of all installations through a neutral wire.
- Coating of metal surfaces of tools with reliable insulation.

- Inaccessibility of current-carrying parts of equipment (including the use of enclosures for current-carrying parts).

5.4 Ecological safety

This section discusses the environmental impacts of the project development activities and the product itself resulting from its implementation in production. The model product itself does not harm the environment during its development or operation. However, the funds required to develop and operate it can harm the environment.

Waste produced in the laboratory includes waste paper, plastic waste, defective parts of personal computers, and other types of computers. Waste paper should be accumulated and transferred to waste paper collection points for further processing. Place plastic bottles in specially designed containers.

Modern PCs are produced practically without the use of harmful substances hazardous to humans and the environment. However, batteries for computers and mobile devices contain heavy metals, acids, and alkalis that can harm the environment if not properly disposed of. Contact

special organizations specialized in the reception, disposal, and recycling of batteries for proper disposal.

Fluorescent lamps used for artificial illumination of workplaces contain from 10 to 70 mg of mercury, which is a dangerous chemical substance that can cause poisoning of living beings and pollution of the atmosphere, hydrosphere, and lithosphere. After 5 years of use, they must be handed over for recycling at special reception points. Legal entities are required to hand over lamps for recycling and maintain a passport for this type of waste. An additional method to reduce waste is to increase the share of electronic document management.

5.5 Safety in emergency

An emergency situation refers to a situation that has arisen due to an accident, natural disaster, catastrophe, or other similar events that cause harm to human health, the environment, or result in significant material losses. In the presented work space, an emergency situation would be a fire. This can occur due to non-compliance with fire safety measures, improper use of electrical devices and PCs, faulty electrical wiring, or other reasons.

According to **FEDERAL LAW-123**, the working space provided for the performance of the WRC (Wireless Radio Communication) is classified as category B (fire hazard). Category B spaces have a moderate fire hazard, meaning that the materials present can burn easily and may produce toxic smoke and gases.

Possible causes of a fire can include the following reasons:

- Short circuit
- Overload of networks, which can cause live parts to heat up and insulation to ignite
- Starting up equipment after incorrect and unqualified repairs

In order to ensure the safety of employees and property from fire, it is necessary to comply with fire safety rules to prevent emergencies.

To protect against short circuits and overloads, it is important to correctly select, install, and use electrical networks and automation equipment.

To prevent fires from occurring, it is necessary to avoid the creation of a combustible environment and to monitor the use of non-combustible or hardly combustible materials in the construction and decoration of buildings.

It is necessary to carry out the following fire prevention measures:

- Organizational measures should be implemented to ensure fire safety in the facility's technical processes. This includes personnel briefing, safety rule training, publication of instructions, posters, and evacuation plans.
- Operational measures should be taken into account when operating equipment. This includes complying with equipment operating standards, ensuring free access to equipment, and maintaining conductor insulation in good condition.
- Technical and constructional measures should be implemented to ensure the proper placement and installation of electrical equipment and heating devices.

This includes complying with fire safety measures when installing electrical wiring, equipment, heating, ventilation, and lighting systems.

To increase the working room's resistance to emergencies, it is necessary to install fire alarm systems that react to smoke and other combustion products, as well as fire extinguishers. Additionally, drills should be conducted twice a year to practice actions in case of fire.

The presented working room has an evacuation plan at the entrance and a fire alarm system installed. Each working area is equipped with two OU-2 type carbon dioxide fire extinguishers. An electrical panel is also available within reach of workers, enabling the complete de-energization of the working room if necessary.

In the event of a fire, call the fire department at 101 and inform them of the emergency's location. Evacuate workers in accordance with the evacuation plan and attempt to extinguish the fire with the available carbon dioxide fire extinguishers if there is no direct threat to health and life. If the fire cannot be controlled, evacuate employees according to the evacuation plan and wait for the fire service specialists to arrive.

5.6 Conclusion

Each employee must carry out their professional activities while taking into account social, legal, environmental, and cultural aspects, as well as issues related to health and safety. They should also be socially responsible for finding solutions and aware of the need for sustainable development.

This section (Social Responsibility) covers the main issues related to the observance of employee rights to work, compliance with labor safety rules, industrial safety, ecology, and resource conservation. It was found that the researcher's workplace satisfies safety and health requirements during project implementation, and the harmful impact of the research object on the environment does not exceed the norm.

Citing & Footnotes:

[1]: **503 error** — An HTTP 503 status code (Service Unavailable) **typically indicates a performance issue on the origin server**. In rare cases, it indicates that CloudFront temporarily can't satisfy a request because of resource constraints at an edge location.

[2]: Top 7 Candlestick Patterns to Use in Trading Forex and Crypto - BabyPips.com

[3]: GitHub - TA-Lib/ta-lib-python: Python wrapper for TA-Lib (http://ta-lib.org/).

[4]: John Bollinger

[5]: J. Welles Wilder Jr. "New Concepts in Technical Trading Systems." Trend Research, 1978.

[6]: George C. Lane

[7]: J. Welles Wilder Jr. - Wikipedia (https://en.wikipedia.org/wiki/J._Welles_Wilder_Jr.))

[8]: Forex Market Hours - Forex Market Time Converter

[9]: "Gold Price Prediction Using Machine Learning Techniques: A Case Study in India" by K. Srinivas and G. K. Patra (2018)

[10]: "Forecasting Gold Price Using Multiple Linear Regression Model and Artificial Neural Network" by M. R. Hasani and M. F. Zarandi (2016)

[11]: "Predicting the Price of Gold Using Machine Learning Techniques" by N. T. Anh and L. T. Trang (2019)

[12]: "Gold Price Prediction Using Ensemble Learning Based on Machine Learning Algorithms" by Y. Wang, L. Liu, and W. Lu (2020)

[13] J. S. Richman and J. R. Moorman, “Physiological time-series analysis using approximate entropy and sample entropy,” American Journal of Physiology-Heart and Circulatory Physiology, vol. 278, no. 6, pp. H2039–H2049, 2000.

[16] “Ridge Regression: Biased Estimation for Nonorthogonal Problems on JSTOR.” *Jstor.org*, 2023, www.jstor.org/stable/1267351. Accessed 11 May 2023.

[17] David A. Freedman (2009). *Statistical Models: Theory and Practice*. Cambridge University Press. p. 26. A simple regression equation has on the right-hand side an intercept and an explanatory variable with a slope coefficient. A multiple regression e right hand side, each with its own slope coefficient

[18] Nik, “How to Calculate MAPE in Python • datagy,” *datagy*, Feb. 11, 2022. https://datagy.io/mape-python/ (accessed May 12, 2023).

[19] “Labor Code of the Russian Federation,” *Cis-legislation.com*, 2023. https://cis-legislation.com/document.fwx?rgn=1811 (accessed May 18, 2023).

[20] “Кодекс,” *Cntd.ru*, 2023. https://docs.cntd.ru/document/1200003913 (accessed May 18, 2023).

[21] “МЕЖГОСУДАРСТВЕННЫЙ СОВЕТ ПО СТАНДАРТИЗАЦИИ, МЕТРОЛОГИИ И СЕРТИФИКАЦИИ (МГС).” Accessed: May 18, 2023. [Online]. Available: https://meganorm.ru/Data2/1/4293754/4293754317.pdf