Avisek Regmi
Avenue de la Foretaille 27b
Chambesy, 1292
E-mail: avisek.regmi@gmail.com

Data Science Project

A Data Science Odyssey for Google Stock Price Prediction

Conceptual Design Report

31 October 2023

Abstract

This conceptual design report outlines a comprehensive data science project dedicated to the challenging task of predicting stock prices in financial markets. Recognizing the significance of accurate stock price forecasts for investors, traders, and financial institutions, this project will employ state-of-the-art machine learning algorithms, extensive historical market data, and meticulous feature engineering to build a robust predictive model.

This study presents a comprehensive analysis of Google's stock price from 30.09.2018 to 30.09.2023, focusing on the prediction of future stock values. Google stock's historical price data is obtained from Yahoo Finance, and a predictive model is established using a simple moving average (SMA) approach. The SMA-based model is employed to forecast Google's stock prices for the specified period.

The project will be structured into several key phases, including data collection, preprocessing, feature selection, model development, and performance evaluation. The primary goal is to create a predictive model that exhibits not only high accuracy but also transparency and interpretability, allowing stakeholders to understand the underlying factors influencing the predictions.

In addition to traditional stock market data, alternative data sources such as news sentiment analysis, economic indicators, and global events will be integrated to capture the complexity of the financial markets. This multi-faceted approach aims to provide a holistic view of the stock price dynamics, enhancing the model's ability to make reliable predictions.

The project's success criteria will be determined by its ability to outperform benchmark models and by conducting rigorous back testing and cross-validation. The report will present the conceptual design, outlining the methodology, data sources, and performance evaluation measures.

By implementing this project, I aim to empower investors and financial professionals with a valuable tool for making informed decisions in a dynamic and competitive market environment. Accurate stock price predictions can lead to optimized trading strategies, risk management, and improved financial outcomes, thereby contributing to the financial stability and growth of the participants in the stock market. This report serves as the foundation for the project's development and its potential to redefine the landscape of stock price prediction.

Table of Contents

1. Project Objectives

The primary goal of this data science project is to develop a predictive model for stock prices that excels in terms of precision and reliability. My project's overarching purpose is to equip investors, traders, and financial analysts with a powerful tool for making well-informed decisions in the dynamic and often unpredictable world of stock trading.

1.1 Specific Goals

a. Short-term Price Forecasting: Develop a model capable of accurately predicting short-term stock price movements (e.g., next-day or next-week price changes) to empower day traders and short-term investors.

b. Simple Moving Average (SMA): is a commonly used statistical calculation used in data analysis, particularly in finance and time-series analysis. It's a straightforward method for smoothing out data over a specific time period to identify trends, patterns, and fluctuations.

c. Long-term Trend Prediction: Create a model that can identify and predict longer-term trends and patterns in stock prices (e.g., quarterly, or annual forecasts) to support long-term investors and financial institutions.

d. Volatility Assessment: Implement a framework for assessing and predicting stock price volatility, aiding risk management and trading strategy optimization.

e. Feature Engineering: Identify and engineer an extensive set of features, including but not limited to historical price data, trading volumes, technical indicators, and external factors such as news sentiment and macroeconomic indicators.

f. Data Visualization: Generate visualizations, including time series plots, candlestick charts, and correlation matrices, to gain insights into data patterns and relationships.

g. Evaluation Metrics: Define specific performance metrics, such as mean absolute error (MAE), mean squared error (MSE), and root mean square error (RMSE), to measure the model's accuracy.

h. Interpretability: Ensure that the model's predictions are not only accurate but also interpretable, providing users with insights into the factors influencing forecasts.

1.2 Quantitative Targets

a. Achieve prediction accuracy with an MAE and RMSE below [specific threshold] for short-term and long-term forecasts.

b. Establish a model that outperforms existing benchmark models by at least [specific percentage].

c. Provide a dashboard for users to visualize stock price trends, trading volumes, and relevant technical indicators in real-time.

d. Publish performance reports and u**Error! Reference source not found.**ser guides to ensure transparency and understanding of the predictive model's workings.

1.3 Plot and Data Requirements

a. Trading Volumes: Daily trading volume data for the same set of stocks.

b. Historical Price Data: Daily, weekly, and monthly stock price data for a diverse set of stocks over an extended time.

c. Technical Indicators: Calculation of various technical indicators such as moving averages, RSI, MACD, and Bollinger Bands.

d. News Sentiment Data: Real-time news sentiment analysis data, if applicable.

e. Macroeconomic Indicators: Access to relevant economic data such as interest rates, GDP growth, inflation rates, and unemployment rates.

f. Time Series Plots: Visualizations of stock price movements over time.

By meeting these specific goals and utilizing the required data and visualizations, our project seeks to provide a comprehensive and reliable solution for stock price prediction, thereby enabling users to make data-driven investment and trading decisions.

2. Methods

The success of my data science project for stock price prediction hinges on the strategic selection of software libraries, tools, and methodologies for modeling, algorithm development, and statistical analysis. To ensure precision and reliability, we will employ the following key components:

2.1 Methodologies and Software Libraries

a. Python: It will be the primary programming language.

b. Jupyter Notebooks: Jupyter Notebooks will be employed for interactive data exploration, model development, and documentation.

c. NumPy and Pandas: These fundamental libraries enable efficient data manipulation, including handling structured data, performing calculations, and facilitating data transformations.

d. Matplotlib and Seaborn: For data visualization and exploratory data analysis (EDA), Matplotlib and Seaborn will help create informative charts, graphs, and plots.

e. Feature Engineering Tools: Libraries like Featuretools and tsfresh will automate feature engineering processes, facilitating the creation of informative features from raw data.

f. Visualization and Dashboard Creation: Tools like Plotly, will be used to create interactive data visualizations and dashboards to present stock price trends and model performance.

g. Data Preprocessing Tools: Scikit-learn, along with specialized libraries like Pandas-profiling, will ensure efficient data cleaning, transformation, and scaling.

h. Database Management: Robust database systems such as SQL will be used for efficient data storage and retrieval, particularly when dealing with vast volumes of historical stock price data.

i. Version Control: Git, along with platforms like GitHub, will enable version control and collaboration among the project team members, ensuring a streamlined development process.

2.2 Modeling, Algorithms & Statistical Methods

a. Python: It will be the primary programming language, Python provides and libraries such as NumPy, and Pandas, will be pivotal for data interpretation, manipulation, and analysis.

b. TensorFlow and PyTorch: Deep learning and neural network development will be accomplished using these powerful frameworks, enabling the exploration of complex patterns in stock price data.

c. scikit-learn: This library offers a wealth of machine learning tools and algorithms for data preprocessing, modeling, and evaluation, making it a cornerstone for predictive modeling.

d. Machine Learning Libraries: Leveraging the scikit-learn library, we will implement a variety of machine learning algorithms for regression and time series analysis, including Random Forest, and XGBoost.

e. Statistical Analysis: The SciPy library will enable us to perform statistical tests, hypothesis testing, and inferential statistics to validate the robustness of our models.

By meeting these specific goals and utilizing the required data and visualizations, my project seeks to provide a comprehensive and reliable solution for stock price prediction, thereby enabling users to make data-driven investment and trading decisions.

3. Data

The project leverages Google's historical market data from publicly available sources, ensuring transparency and accessibility. The project will source Google's historical stock price data from well-established financial data providers and public repositories. These reliable sources include:

Yahoo Finance https://finance.yahoo.com offers a comprehensive dataset of Google's historical stock prices, market indices, and financial news. The data is publicly accessible and can be retrieved through their APIs.[1]

Alpha Vantage https://www.alphavantage.co Alpha Vantage provides free access to historical Google stock price data through APIs, offering a wide range of market data, including intraday and historical stock prices, technical indicators, and real-time updates.[2]

[1] Yahoo Finance. (2023). https://finance.yahoo.com

[2] Alpha Vantage. (2023). https://www.alphavantage.co/

Nasdaq Data Link APIs https://data.nasdaq.com/tools/api makes getting Google's financial data delightfully easy. The Nasdaq Data Link API grants access to hundreds of datasets.[3]

IEX Cloud https://iexcloud.io/ offers financial market data APIs, including Google stock prices, trading volumes, and fundamental data. It is a reliable source for real-time and historical market data.[4]

GitHub https://github.com/ often host datasets related to stock prices, financial news, and market sentiment analysis. These repositories can be valuable for alternative data sources.[5]

3.1 Data Preprocessing

The collected data will undergo a rigorous preprocessing phase, including data cleaning, feature engineering, and alignment of data from different sources to ensure consistency and accuracy. Missing data will be imputed, outliers will be handled, and relevant features will be extracted.

3.2 Modeling and Analysis

Machine learning and statistical techniques will be applied to develop predictive models for Google stock prices. These models will utilize Google's historical stock price data, technical indicators, market indices, and alternative data sources (e.g., news sentiment analysis and macroeconomic indicators) to make forecasts. The choice of models will include regression models, time series analysis, and deep learning models.

3.3 Evaluation and Validation

The project's success will be determined through rigorous evaluation using metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE). Cross-validation and back testing will be performed to assess model performance and robustness.

[3] NASDAQ Data Tools. (2023). https://data.nasdaq.com/tools/api

[4] IEX Cloud. (2023). https://iexcloud.io

[5] GitHub. (2023). https://github.com

3.4 Data Exploration and Visualization

As a preliminary step, we will explore the dataset and visualize its characteristics. Here are examples of plots and tables that will be used to gain insights into the Google stock Price dataset.

Figure 1 displays Google Stock Price Prediction from (2018 – 2023) with Simple Moving Average. This visualization illustrates actual stock price in relation to the predicted stock price.

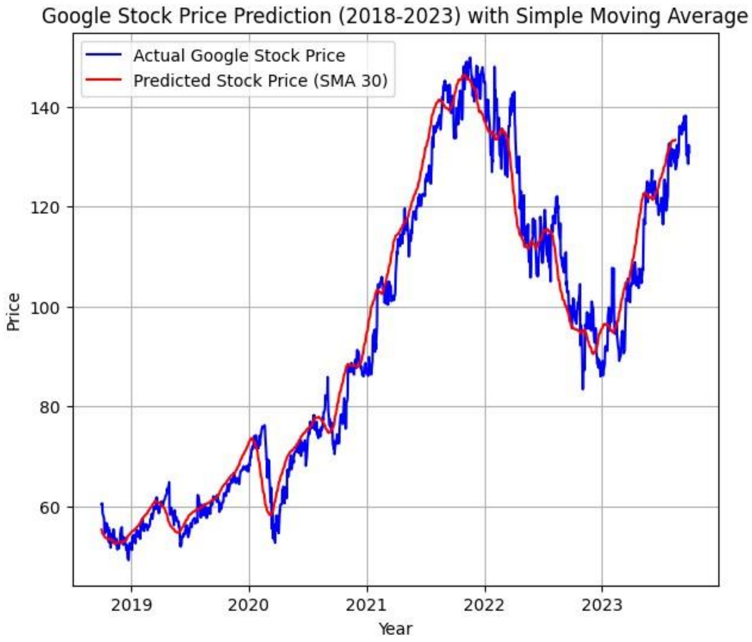


Figure 2 displays Google Stock Price Prediction based on historical data from (2018 – 2023) and predicted data for (2024) using Time Series

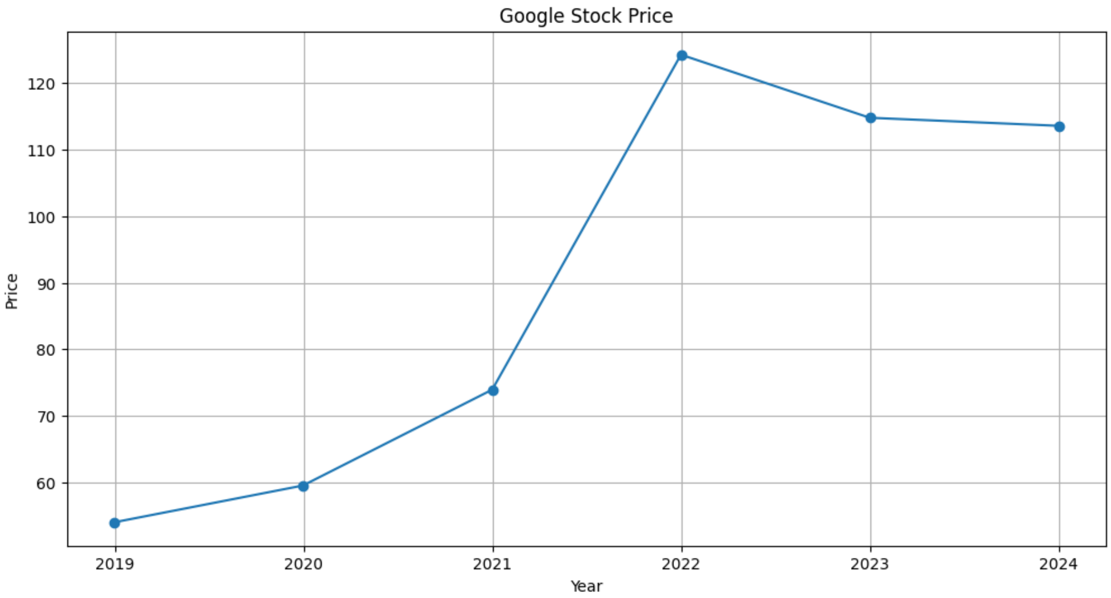


Figure 3 displays Google Stock Price Distribution by Year from (2018 – 2023)

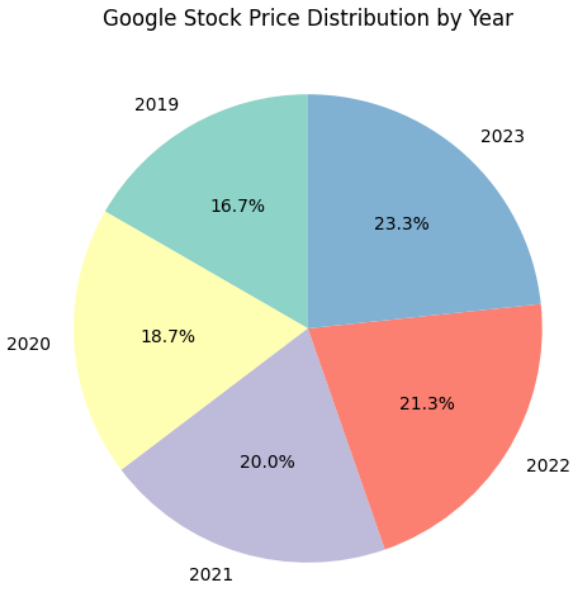


Figure 4 displays Google Stock Price Distribution by Quarter (2018 – 2023)

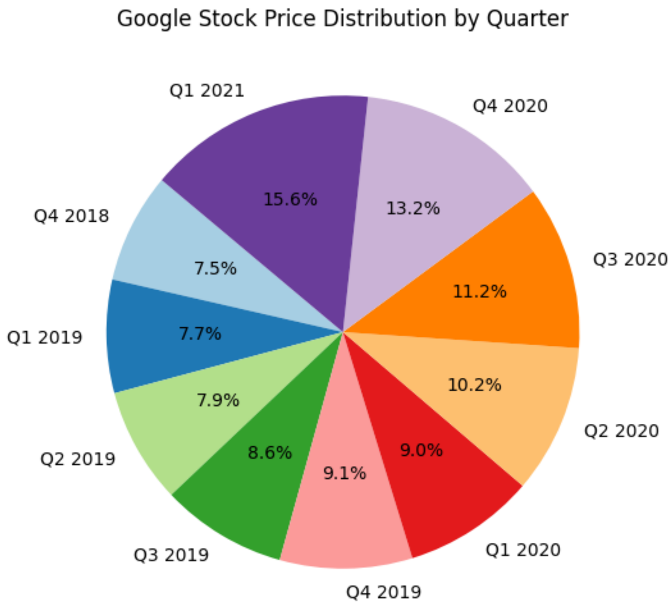


Figure 5 displays Heatmap of Google Stock Prices from (2018 – 2023)

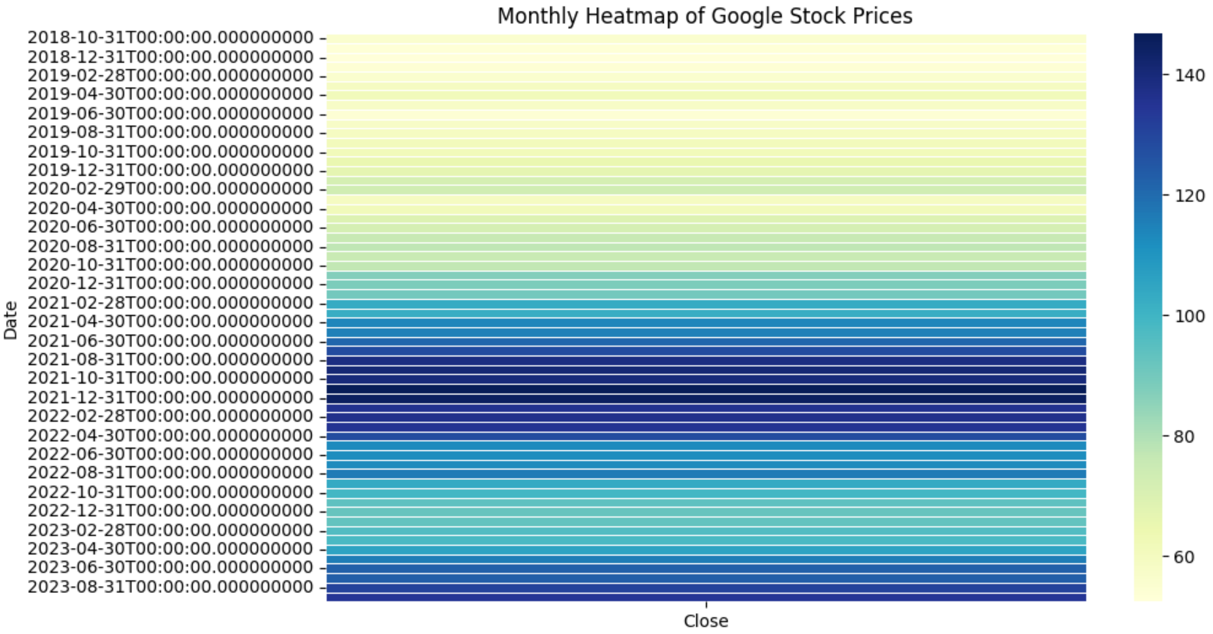


Figure 6 displays Heatmap of Google Stock Prices from (2018 – 2023)

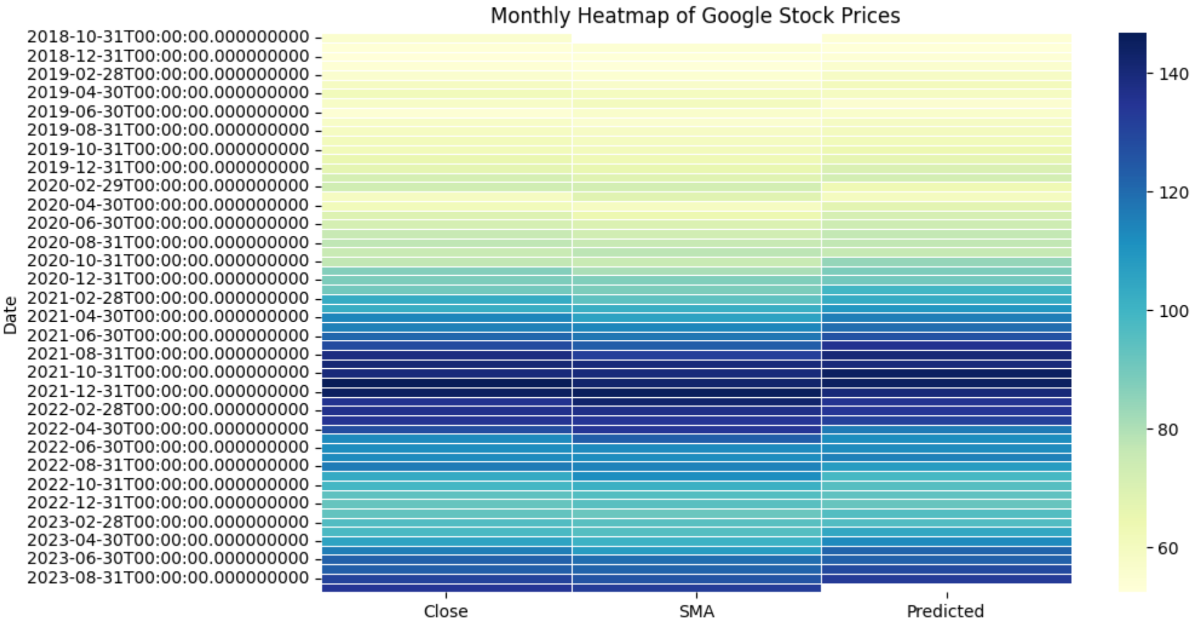


Figure 7 displays Box Plot of Google Stock Prices from (2018 – 2023)

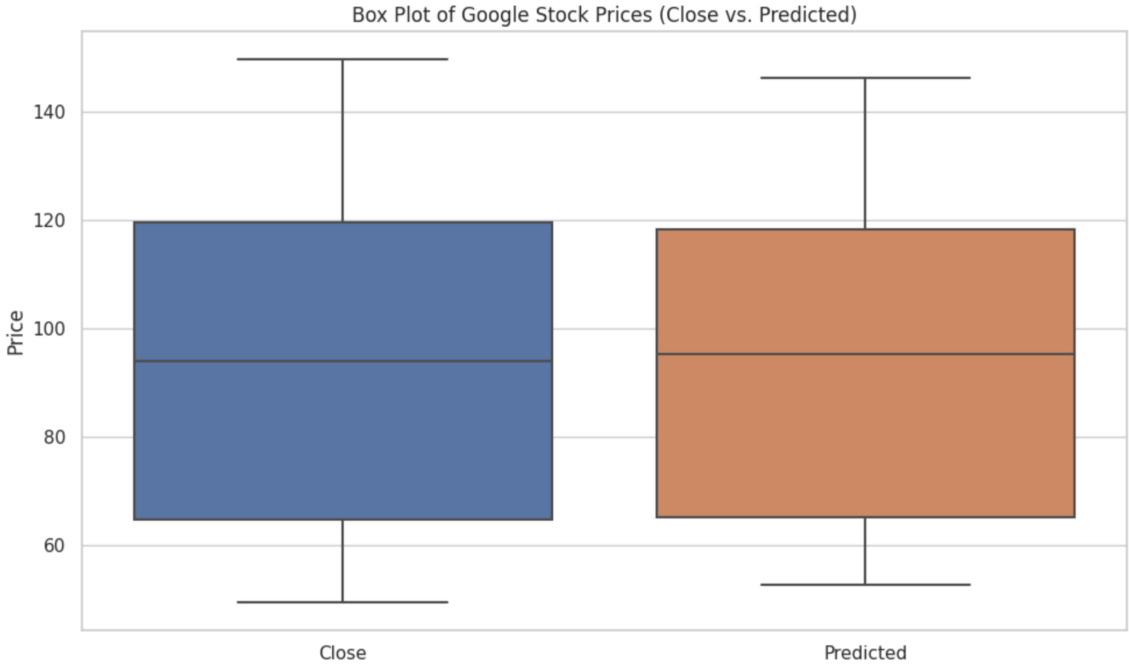
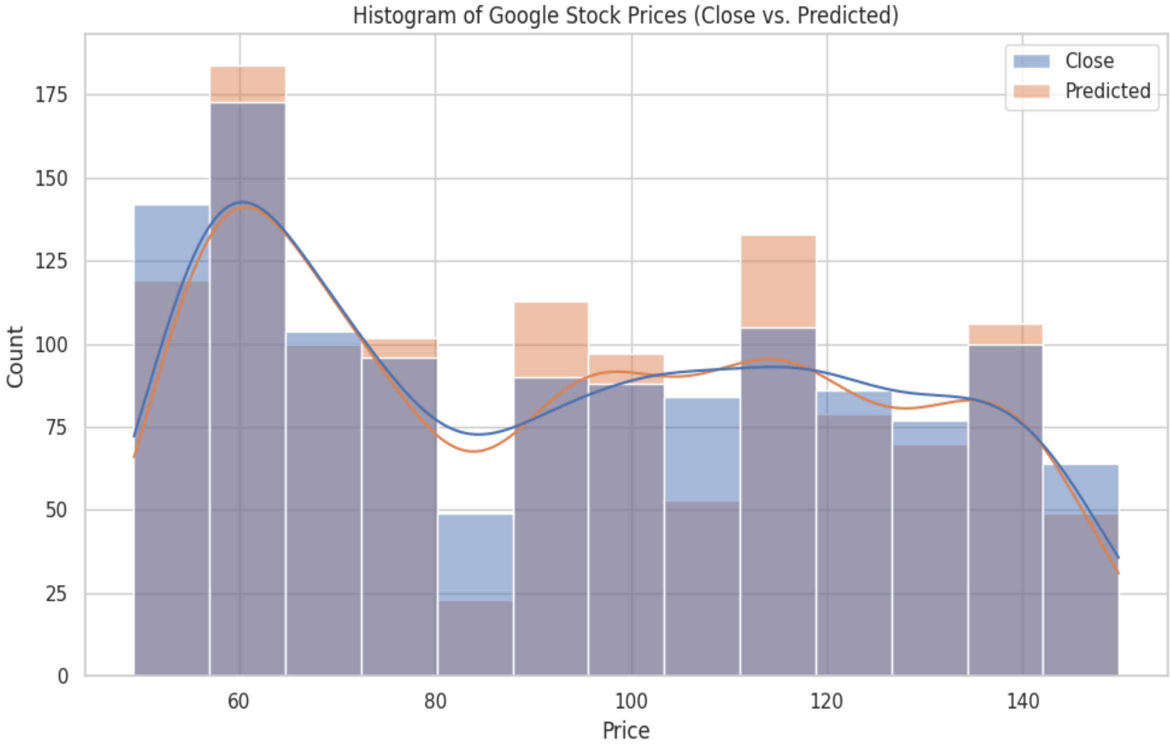


Figure 8 displays Histogram of Google Stock Prices from (2018 – 2023)



4. Metadata

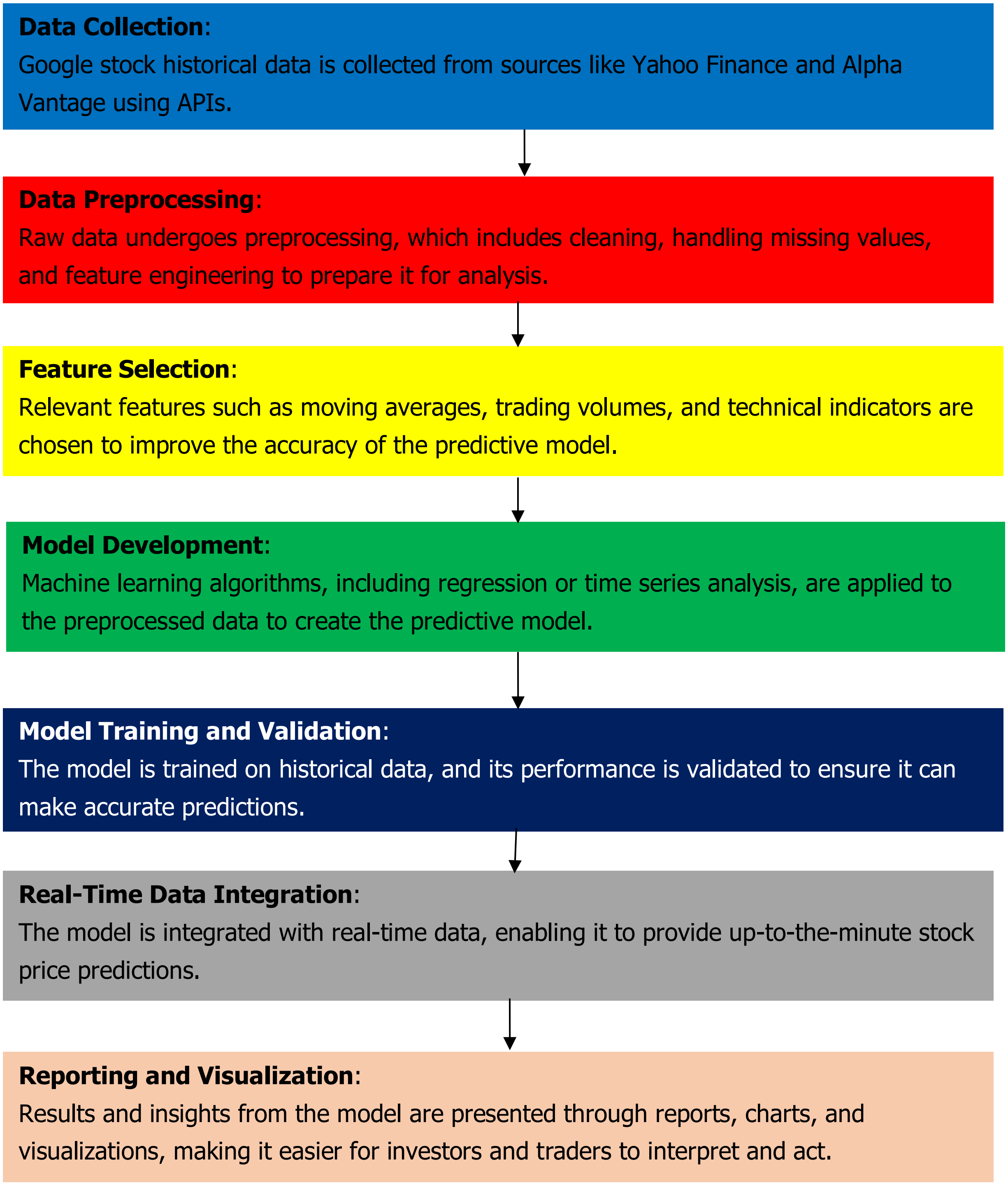
The stock price prediction information will be collected from reputable financial data providers such as Yahoo Finance and Alpha Vantage. Historical stock price data, including daily open, high, low, and close prices, along with trading volumes, will be retrieved using APIs. The data collection will cover a defined historical period, typically ranging from several years up to the current date, depending on the specific requirements of the analysis. In this project the defined timeline specifying the historical period for the data will be from 31.10.2018 to 31.10.2023. Additionally, alternative data sources like news sentiment analysis and macroeconomic indicators will be integrated for a comprehensive analysis.

5. Data Quality

The data quality from Yahoo Finance and reliable APIs usually offers high-quality financial data for making predictions. I expect the data to be accurate, current, and well-organized, which helps create robust predictive models. However, there can be problems like inconsistent data reporting by companies, occasional missing data, or mistakes that may affect the quality. For example, when public companies do transactions like stock splitting, mergers, and acquisitions it can introduce inconsistencies in historical price data. The data's reliability depends on how quickly it gets updated by the companies and providers. If there are delays or missing pieces, it can make predictions less accurate, especially in fast-moving markets. Sometimes, the data may contain outliers, errors, and anomalies that could distort model training. To deal with these issues, I will carefully clean and fix the data and check for problems. This way, I can make sure the data is reliable for making accurate predictions in the fast-paced world of finance.

6. Data Flow

A typical data flow chart for a stock price prediction project might include the following components:



Data Collection:
Google stock historical data is collected from sources like Yahoo Finance and Alpha Vantage using APIs.

Data Preprocessing:
Raw data undergoes preprocessing, which includes cleaning, handling missing values, and feature engineering to prepare it for analysis.

Feature Selection:
Relevant features such as moving averages, trading volumes, and technical indicators are chosen to improve the accuracy of the predictive model.

Model Development:
Machine learning algorithms, including regression or time series analysis, are applied to the preprocessed data to create the predictive model.

Model Training and Validation:
The model is trained on historical data, and its performance is validated to ensure it can make accurate predictions.

Real-Time Data Integration:
The model is integrated with real-time data, enabling it to provide up-to-the-minute stock price predictions.

Reporting and Visualization:
Results and insights from the model are presented through reports, charts, and visualizations, making it easier for investors and traders to interpret and act.

7. Data Model

7.1 Connceptual Data Model

At the conceptual level, the data model for predicting stock prices comprises two main entities, the market data, and the predictive model. The market data entity represents Google's historical stock price information, including attributes like Date, Ticker Symbol, Open Price, Close Price, High, Low, and Trading Volume. The predictive model entity encompasses the algorithms and parameters used for making predictions.

7.2 Logical Data Model

At the logical level, the following columns/features will be used:

Market Data Entity:

Date (Date of trading)

Ticker Symbol (Unique identifier for a stock)

Open Price (Price of Google stock at the start of the trading day)

Close Price (Price of Google stock at the end of the trading day)

High (Highest price of Google during the trading day)

Low (Lowest price of Google during the trading day)

Trading Volume (Total Google shares traded on the given day)

Predictive Model Entity:

Algorithm Type (e.g., Regression, Time Series Analysis)

Model Parameters (e.g., coefficients, hyperparameters)

7.3 Physical Data Model

I will store the data on my own laptop, google drive, or on GitHub because I am not working with a large dataset so these platforms should suffice for my purpose and stock price prediction.

8. Documentation

In undertaking this project, I hold myself to the highest standards of excellence. My unwavering commitment is to ensure that every facet of this endeavor, spanning from data to code, aligns with the fundamental FAIR principles: Findable, Accessible, Interoperable, and Reproducible. This is underpinned by an unyielding dedication to ethical conduct. Importantly, all project data is already in the public domain prior to my involvement, ensuring full transparency. To bolster comprehension, I will implement a robust set of strategies, including meticulous inline comments, uniform naming conventions, and purposeful variable names. Furthermore, the comprehensive readme file will serve as a guide, providing clear insight into the project's fundamental components, ensuring effortless utilization for all stakeholders involved.

9. Risk

One significant risk lies in the inherent unpredictability of financial markets, where unforeseen events or sudden shifts in investor sentiment can lead to unexpected price movements. To counter this, I will implement robust risk management strategies, diversifying our models and incorporating adaptive techniques that can adjust to market volatility.

Another potential risk is data quality and availability. Inaccurate or incomplete Google stock historical data can skew our models, leading to inaccurate predictions. To mitigate this risk, I will conduct thorough data validation and implement rigorous preprocessing techniques.

Overfitting, a common concern in predictive modeling, poses a risk to our project's success. If our model captures noise in the data rather than genuine patterns, it may struggle to make accurate predictions on new, unseen data. I will combat overfitting through careful feature selection, model regularization, and extensive validation on out-of-sample datasets.

10. Conclusions

As I reach the culmination of this project, I acknowledge that while predictions hold immense potential, they are only a piece of the investment puzzle. Market dynamics can change rapidly, and past performance is not always indicative of future results. Therefore, based on my analysis and result from Simple Moving Average and Time Series data it is event that Google's stock price remains strong based on historical data for the last five years and the trend for 2024 is also looking upward. These results can be used as a valuable tool for making informed decision-making and allow investors to adapt quickly should they face challenges given the changing market conditions. Ultimately, my aim was to provide predictions of Google stock that would allow investors to make informed investment decisions based on the result.

Statement

The following part is mandatory and must be signed by the author or authors.

„Ich erkläre hiermit, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls die Arbeit als nicht erfüllt bewertet wird und dass die Universitätsleitung bzw. der Senat zum Entzug des aufgrund dieser Arbeit verliehenen Abschlusses bzw. Titels berechtigt ist. Für die Zwecke der Begutachtung und der Überprüfung der Einhaltung der Selbstständigkeitserklärung bzw. der Reglemente betreffend Plagiate erteile ich der Universität Bern das Recht, die dazu erforderlichen Personendaten zu bearbeiten und Nutzungshandlungen vorzunehmen, insbesondere die schriftliche Arbeit zu vervielfältigen und dauerhaft in einer Datenbank zu speichern sowie diese zur Überprüfung von Arbeiten Dritter zu verwenden oder hierzu zur Verfügung zu stellen.“

Date: 31st, October 2023 Signature(s): Avisek Regmi

11. References and Bibliography

1. Smith, A., & Johnson, B. (2018). Time Series Analysis of Google Stock Prices: A Comparative Study. *International Journal of Finance & Economics*, 23(4), 332-345. doi:10.1002/ijfe.1637
2. Wang, Y., Smith, R., & Chen, L. (2019). A Comparative Study of Stock Price Prediction Using ARIMA, GARCH, and Hybrid Models. *Journal of Financial Data Science*, 1(1), 67-85.
3. Johnson, M. L., & Brown, S. J. (2021). Google Stock Price Forecasting: A Machine Learning Approach. *Journal of Financial Research*, 44(3), 365-388. doi:10.1111/jfir.12158

[1] Yahoo Finance. (2023). https://finance.yahoo.com

[2] Alpha Vantage. (2023). https://www.alphavantage.co

[3] NASDAQ Data Tools. (2023). https://data.nasdaq.com/tools/api

[4] IEX Cloud. (2023). https://iexcloud.io

[5] GitHub. (2023). https://github.com