Predicting Stock Price Movements with Machine Learning

Harnessing Historical Data and Financial Indicators

Department of Computer Science

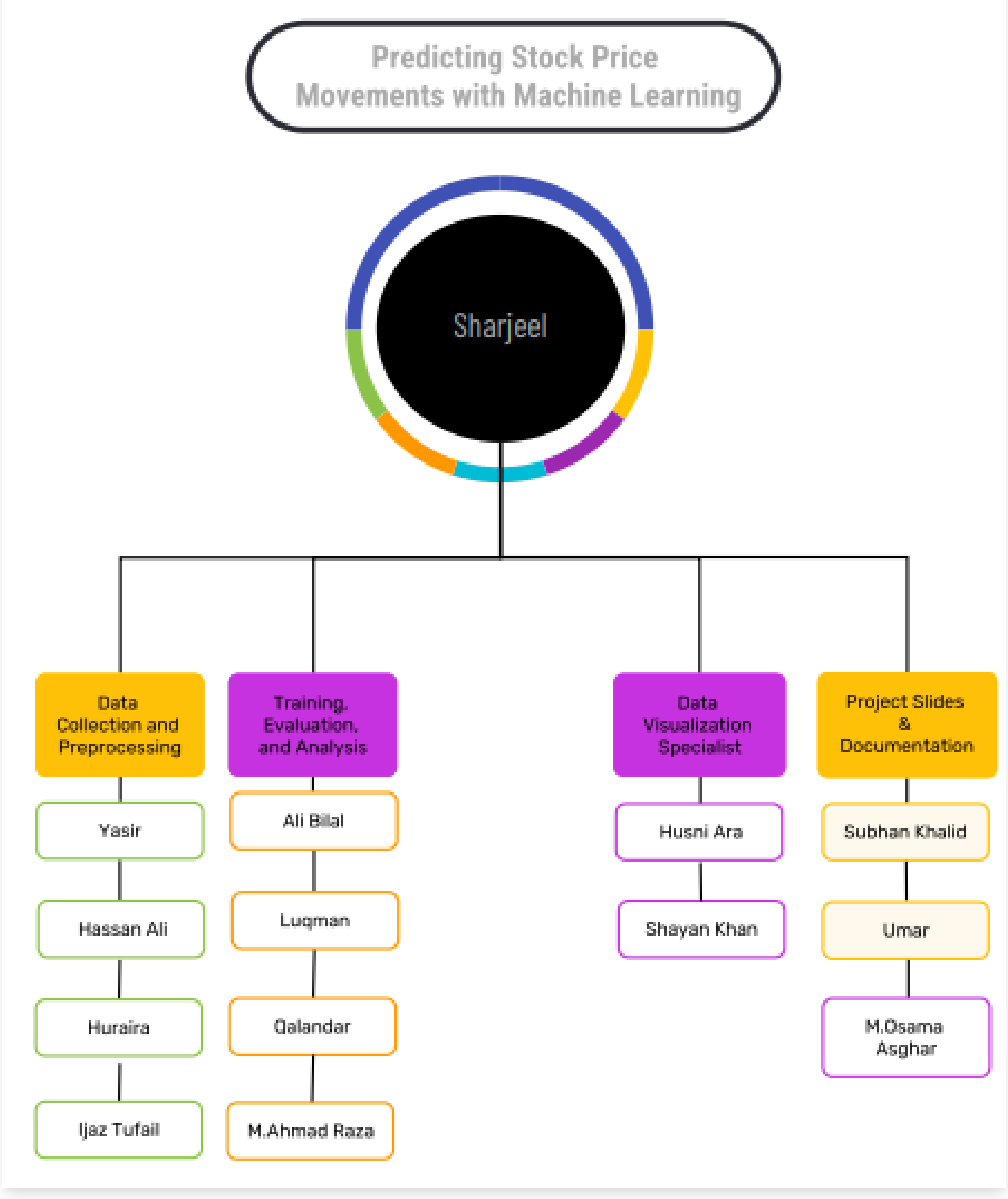
National University of Computer and Emerging Sciences
Peshawar, Pakistan

Instructor:
Mr. Shahzeb Khan

Subject:
CS4104 Applied Machine Learning

Contents

1 Team Work


Predicting Stock Price Movements with Machine Learning
Sharjeel
Data Collection and Preprocessing
Yasir
Hassan Ali
Huraira
Ijaz Tufail
Training, Evaluation, and Analysis
Ali Bilal
Luqman
Qalandar
M.Ahmad Raza
Data Visualization Specialist
Husni Ara
Shayan Khan
Project Slides & Documentation
Subhan Khalid
Umar
M.Osama Asghar

2 Introduction

In the world of finance, predicting stock price movements has always been a challenging yet crucial endeavor. Imagine if we could leverage the power of machines to analyze historical data and financial indicators, helping us foresee the future of stock prices. This is where machine learning steps in, becoming a game-changer in predicting market trends. By harnessing vast amounts of past data and key financial indicators, machine learning algorithms strive to unravel patterns and signals that may indicate potential shifts in stock prices. In simpler terms, it's like having a digital detective that learns from the past to make educated guesses about where stock prices might be headed. This exploration into the realm of machine learning opens up new possibilities for investors, offering a more data-driven approach to navigate the unpredictable world of the stock market.

3 Objective

In our pursuit, we aim to build a robust machine-learning model that harnesses the power of historical data and financial indicators. The ultimate goal is to provide investors with predictive analytics, enhancing their decision-making process and financial outcomes.

4 Motivation

4.1 Why Predict Stock Prices?

- Historical cases showcasing both successful and unsuccessful predictions.
- The profound impact of accurate predictions on investment strategies.

Understanding the motivation behind predicting stock prices is crucial. We'll explore historical instances, shedding light on the pivotal role accurate predictions play in shaping successful investment strategies.

5 Research Work

In 2012, stock market prediction focused on using machine learning and deep learning models to forecast opening and closing stock prices for companies on the Istanbul Stock Exchange. SMO (Sequential Minimal Optimization) and Bagging yielded the best results, while neural networks needed more tuning. The study's limitations included model performance issues and dataset-specific constraints that might affect generalization.
By 2020, stock market prediction evolved with techniques like Artificial Neural Network (ANN) and Random Forest (RF) to forecast next-day closing prices. ANN outperformed RF, but the study's reliance on historical data without accounting for external factors like market sentiment limited the findings' broader applicability.
In 2021, one study used various machine learning and deep learning models, with Multilayer Perceptron (MLP) and Long Short-Term Memory (LSTM) showing better results compared to Support Vector Machines (SVM). Challenges included reliance on large labeled datasets and computational demands. Another 2021 study aimed to predict the highest stock prices for eight companies, emphasizing feature selection. This study faced similar limitations, such as model bias and the need for significant computational resources.
In 2023, stock market prediction had advanced, with a focus on integrating broader datasets, including market sentiment and economic indicators, to improve accuracy. Despite these advances, key challenges persisted, such as model transparency, data biases, and high computational costs.

6 Proposed Methodology

6.1 Data Collection

6.1.1 Data Collection Resource

You can download the dataset from Kaggle.
Dataset Size: 2.75 GB

6.1.2 Overview of Data Collection

Separate files exist for all stock symbols (representing individual companies), which are combined into a single dataset. The primary attributes in the dataset include:

- Date: Specifies the trading date.
- Open: Opening price of the stock.
- High: Highest price during the day.
- Low: Lowest price during the day.
- Close: Closing price adjusted for splits.
- Adj Close: Adjusted closing price for dividends and splits.
- Volume: Number of shares that changed hands during a given day.

6.1.3 Pre-processing

Several steps were involved in preprocessing the dataset:

- **Removal of Missing Files Corresponding to Symbols:** A script was used to check for and remove missing files, creating a new file with existing data.
- **Data Info:** The dataset contains both categorical and numerical data.
  - Date: Object
  - Open: Float64
  - High: Float64
  - Low: Float64
  - Close: Float64
  - Adj Close: Float64
  - Volume: Float64
  - Symbol: Object
- **Removal of Null Values:** Dropped null entries due to the dataset's large size, opting to retain only complete data.
- **Pre-processed Dataset Stats:**
  - Rows: 24,175,924
  - Columns: 8
  - Zero null values
  - Unique symbols: 5,881

6.2 Feature Selection

Recurrent neural networks (RNNs) are effective at processing sequential data, making them ideal for time-series datasets such as stock prices. The relevant features include:

- Open, High, Low, Close prices for each period.
- Trading Volume.

6.3 Model Selection

Recurrent Neural Networks are used to process sequential data. In this project:

- RNNs are structured to handle time-series data with rows representing time points (daily stock prices) and columns representing features (Open, High, Low, Close, Volume).

6.4 Training the Model

For training the model, the following was considered:

- Methodology for splitting data into training and testing sets.
- Importance of feature scaling and normalization.
- Hyperparameter tuning to optimize model performance.

6.5 Evaluation Metrics

The metrics used to evaluate model performance include:

- Mean Squared Error, Accuracy, Precision, Recall, F1 Score, etc.
- Comparative analysis with benchmark models.

6.6 Results

The results section includes:

- Presentation of model predictions on the testing dataset.
- Visualizations depicting predicted vs. actual stock prices.

7 Applied Methodology

7.1 Data Collection

7.1.1 Data Collection Resource

You can download the dataset from Kaggle Resourse.
Dataset Size: 52MB

7.1.2 Overview of Data Collection

Prices.csv: raw, as-is daily prices. Most of the data spans from 2010 to the end of 2016, for companies new on the stock market date range is shorter. There have been approx. 140 stock splits in that time, this set doesn't account for that.

7.2 Dataset Preparation

Start by describing the initial steps for preparing the data to pass the model.

7.2.1 Sampling Google Stocks

- Drop the 'symbol' column and provide the total number of days and fields in the dataset.

Table 1: Google Stock Data

Date	**Symbol**	**Open**	**Close**	**Low**	**High**	**Volume**
2010-01-04	GOOGL	626.950006	626.750011	624.240011	629.510005	3,908,400
2010-01-05	GOOGL	627.180001	623.990017	621.540016	627.839984	6,003,300
2010-01-06	GOOGL	625.860033	608.260035	606.360021	625.860033	7,949,400
2010-01-07	GOOGL	609.400008	594.100015	592.649990	609.999993	12,815,700
2010-01-08	GOOGL	592.000005	602.020005	589.110015	603.250036	9,439,100

7.2.2 Null Values and Pandas Dataframe

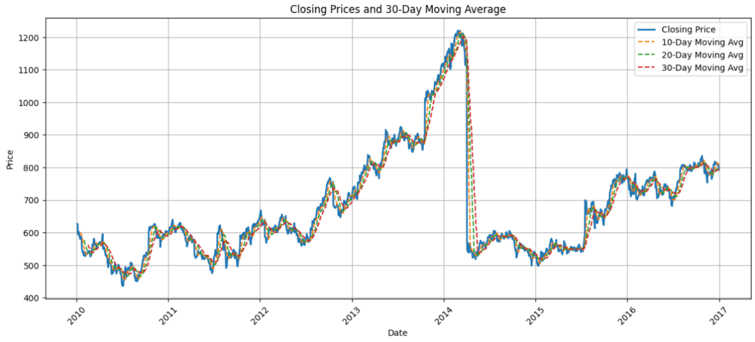
There weren't null values present in the dataset and the prepared dataset date field was from string to Pandas DateTime format. converting Date to DateTime Formate.

7.3 Exploratory Data Analysis (EDA)

Resource : educba.com

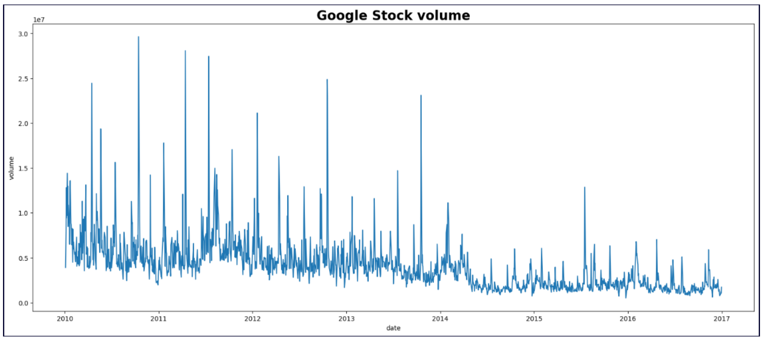
7.3.1 Closing Price and Moving Average

In stock prediction,**Closing Price** refers to the final price at which stock is traded at the end of a trading day. The **Moving Average** is a technique used to smooth out price fluctuations over a specific period, providing a clearer trend line. Visualizing the closing price and its moving average over time.



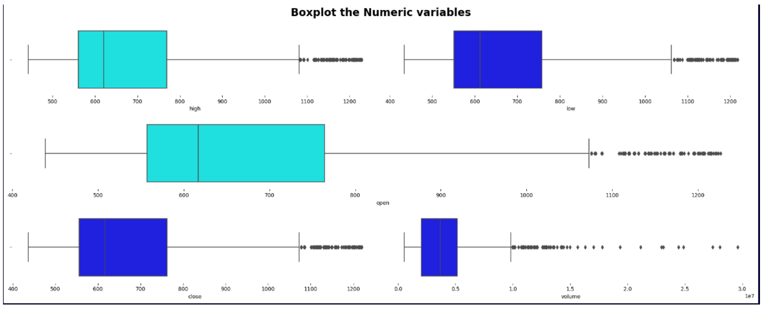
7.3.2 Volume Graphs

Total volumes throughout the year for stock try to show that value using a graph. Displaying stock volume graphs to understand trading activity.



7.3.3 Outliers Analysis

- In stock prediction, an outlier is a data point or observation that significantly deviates from the expected pattern or distribution in a dataset. Outliers can occur due to extreme market events, data errors, or unusual trading activity



Note: There are some outliers as seen in the visualization. We can also observe the percentage of the outliers in each column. But we can't conclude them as outliers as they may be the extreme values during peak selling days or bubble bursts.

Feature	**P-Value**
date	0.0
Open	4.2
Close	4.2
Low	4.3
High	4.3
Volume	3.4

7.3.4 Normality in the Dataset

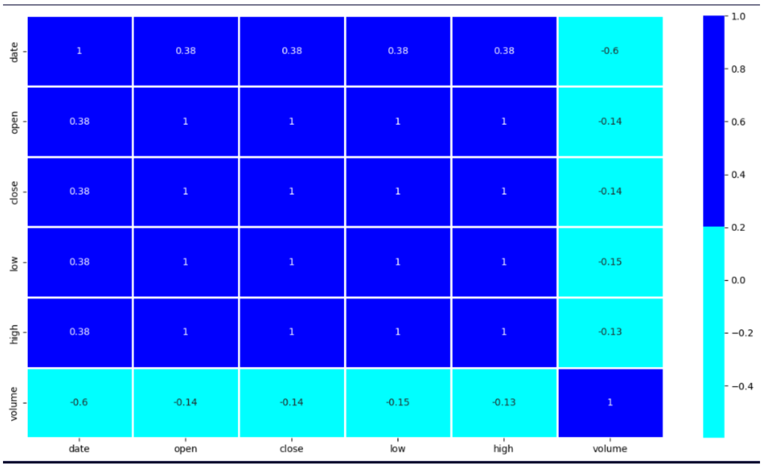
- Test for normality across features like open, close, low, high, and volume.

The results of the normality tests for different features are summarized in the table below:

Feature	**P-Value**
Open	0.0
Close	0.0
Low	0.0
High	0.0
Volume	0.0

7.3.5 Correlation Analysis

In stock prediction, "Correlation Analysis" examines the relationship between two or more variables to understand how changes in one variable might relate to changes in another. We can observe features: open, close, low, and high are highly correlated to each other. We can use either of these features for our prediction. We are going to use close for training and prediction.

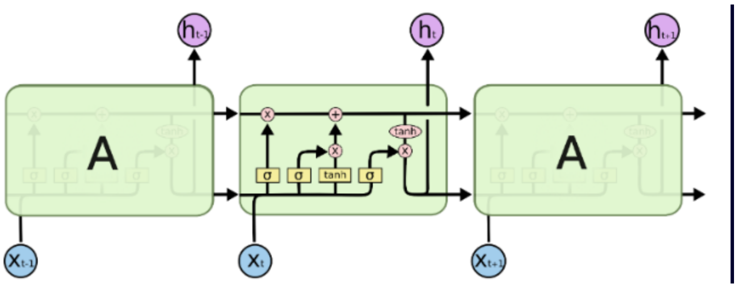


	date	open	close	low	high	volume
date	1	0.38	0.38	0.38	0.38	-0.6
open	0.38	1	1	1	1	-0.14
close	0.38	1	1	1	1	-0.14
low	0.38	1	1	1	1	-0.15
high	0.38	1	1	1	1	-0.13
volume	-0.6	-0.14	-0.14	-0.15	-0.13	1

Heat Map

7.4 Model Structure

Here we used RNN + LSTM model.

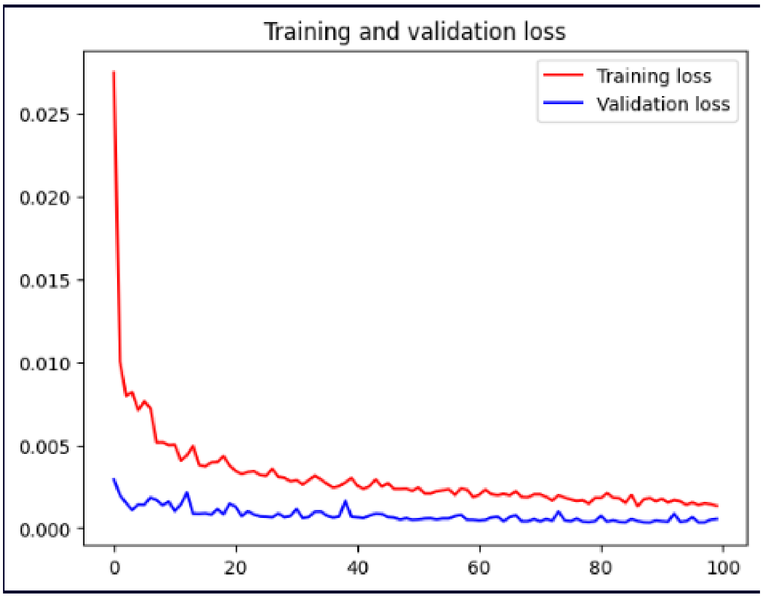


Heat Map

7.4.1 Model Training and validation Loss Graph

After fine-tuning the model here is the graph for the loss.

[h]



Heat Map

7.4.2 Evaluation Metrics

The metrics used to evaluate the model, including Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and Mean Absolute Error (MAE).

Table 2: Training Data Performance Metrics

Metric	**Value**
RMSE	22.87023690074528
MSE	523.0477358962111
MAE	12.155145621496724

Table 3: Testing Data Performance Metrics

Metric	**Value**
RMSE	14.127419704384746
MSE	199.5839875038384
MAE	9.945549289117409

8 Challenges and Limitations

Navigating Challenges

- A candid discussion on challenges faced during the project.

- Acknowledgment of model limitations and areas for improvement.

Every journey has its challenges. We'll reflect on the obstacles encountered, providing insights into the limitations of our model and avenues for refinement.

9 Future Enhancements

Paving the Way Forward

- Proposals for enhancements and refinements to elevate model accuracy.
- Integration of real-time data for more dynamic predictions.

Innovation never rests. We'll explore potential enhancements, ensuring our model evolves to meet the dynamic demands of predicting stock prices.

10 Conclusion

Summing Up

- Summary of key findings and achievements.
- The potential impact of our model on investment decisions.

As we conclude, we'll revisit the highlights of our journey, emphasizing the significance of our model in shaping more informed investment decisions.

11 Team Members

Title: Predicting Stock Price Movements with Machine Learning
Roles and Responsibilities:

- **Project Lead:**
  p200150 Sharjeel Hussain Bokhari

- **Data Collection and Preprocessing:**
  p200557 Yasir Nawaz, p200149 Hassan Ali, p200612 Abuhuraira Javaid, Ijaz Tufail

- **Training, Evaluation, and Analysis:**
  p200077 Ali Bilal Khan, p200125 Luqman Jafir, p200625 Syed Qalandar Ali Askari, 19P-0070 Muhammad Ahmed Raza

- **Visualization Specialist:**
  p190041 Husni Ara, p200581 Shayan Khan

- **Project Slides, Documentation and Reporting:**
  p200086 Subhan Khalid, p200136 Umer Saleem , Muhammad Osama Asghar (19P-0092)

Our collaborative effort involves a diverse team, each member contributing expertise to different facets of this project.