An Exploration of Traditional Machine Learning Algorithms in Predicting Stock Market Trends.

Submitted

By

Group No…….

NAME	ID, NO	EMAIL
Majed Alhazmi	11603092	majedalhazmi@my.unt.edu
Raja Vinay Kumar Vaka	11713344	rajavinaykumarvaka@my.unt.edu
Divya Anusha Chandrupatla	11642682	Divyanushachandrupatla@my.Unt.edu
Palanati Anvesh Reddy	11710518	anveshreddypalanati@my.unt.edu

An Exploration of Traditional Machine Learning Algorithms in Predicting Stock Market Trends.

Introduction

The stock market is a complex and dynamic system that has fascinated investors and financial experts for centuries. Predicting the future performance of the market is a challenging task, as it is influenced by a multitude of factors, including economic indicators, news events, investor sentiment, and external shocks. Traditional methods of stock market prediction, such as fundamental analysis and technical analysis, have limitations and often fail to accurately forecast market movements.

In recent years, the field of machine learning has emerged as a powerful tool for analyzing and predicting complex data. Machine learning algorithms can learn from historical data and identify patterns that can be used to make predictions about future events. This has led to a growing interest in using machine learning for stock market prediction.

Motivation

There are several reasons to motivate the use of machine learning for stock market prediction. First, machine learning algorithms can extract patterns and insights from large datasets that would be difficult or impossible to identify using traditional methods. Second, machine learning algorithms can adapt to new information and improve their performance over time. Third, machine learning algorithms can be used to build predictive models that are both accurate and interpretable.

Significance

Accurate stock market prediction has the potential to provide significant benefits to investors, traders, and financial institutions. By being able to forecast market movements, these stakeholders can make informed investment decisions, reduce risk, and maximize returns. In addition, accurate stock market prediction can be used to develop investment strategies and risk management tools.

Objectives

The objective of this project is to investigate and apply traditional machine learning algorithms to predict stock market trends. We will develop predictive models that leverage historical stock data, news sentiment analysis, and economic indicators. Our goal is to create a user-friendly tool that provides real-time predictions and insights into stock market trends.

Previous Research

Extensive research has been conducted in stock market prediction, predominantly using historical price and trading volume data. Researchers have explored various predictive models, including statistical models, technical indicators, and machine learning algorithms. (Liu, H., & Long, Z.2020).

One of the most common approaches to stock market prediction using machine learning is to use supervised learning algorithms. Supervised learning algorithms are trained on a dataset of labeled data, where each data point is associated with a known outcome. In the case of stock market prediction, the data points would be historical stock prices, and the outcomes would be the future stock prices. Once a supervised learning algorithm has been trained, it can be used to predict the future stock prices for new data points.

Another common approach to stock market prediction using machine learning is to use unsupervised learning algorithms. Unsupervised learning algorithms are trained on a dataset of unlabeled data, where the data points are not associated with any known outcomes. Unsupervised learning algorithms can be used to identify patterns and relationships in the data, which can then be used to make predictions about future events.

Related Work (Background):

The landscape of stock market prediction has witnessed extensive research, employing various methodologies and algorithms to unravel the complexities of market trends. Understanding the background and related works in this domain is crucial for contextualizing the proposed project. Previous studies have explored diverse approaches, ranging from statistical models to sophisticated machine learning techniques. The following discussion delves into notable works that have paved the way for predicting stock market trends.

Numerous researchers have delved into the realm of machine learning for stock market prediction, recognizing the potential of algorithms to decipher patterns and trends from historical data. Ahmed et al. (2019) introduced an ant colony optimization-based algorithm for stock market forecasting, showcasing the significance of optimization techniques in enhancing predictive models. Their work highlights the utilization of nature-inspired algorithms in financial prediction, providing a foundation for algorithmic approaches in stock market analysis.

Bustos and Pomares-Quimbaya (2020) conducted a systematic review of stock market movement forecasting. Their comprehensive survey covered a range of methodologies, from traditional statistical models to advanced machine learning techniques. This work serves as a valuable resource for understanding the evolution of forecasting techniques and the challenges inherent in predicting the dynamic nature of stock markets.

The study by Yuan et al. (2020) explored integrated long-term stock selection models, emphasizing feature selection and machine learning algorithms. This work underscores the importance of selecting relevant features and integrating them into predictive models, aligning with the proposed project's focus on feature engineering. Feature selection has proven to be a critical aspect of accurate stock market prediction, influencing the performance of machine learning models.

Zubair et al. (2021) contributed to the field by introducing a multiple regression model for stock market prediction. Their use of multiple r-square as a measure of accuracy and evaluation across different datasets showcases the diversity of approaches in assessing predictive models. While their focus was on individual stocks and specific indices, their methodology provides insights into the challenges and opportunities in predicting stock prices.

Liu and Long (2020) presented an improved deep learning model for predicting stock market price time series. The emphasis on deep learning signifies the shift towards more complex models, addressing the intricate patterns present in financial time series data. Deep learning, particularly Long Short-Term Memory (LSTM) networks, has gained prominence in recent years for its ability to capture long-term dependencies in sequential data.

In summary, the background in stock market prediction is rich with diverse methodologies, ranging from traditional statistical models to advanced machine learning algorithms. The proposed project aims to contribute to this landscape by exploring the effectiveness of traditional machine learning algorithms, such as linear regression, decision trees, and support vector machines, in predicting stock market trends. Building upon the insights gained from related works, the project seeks to enhance the interpretability and performance of predictive models for informed decision-making in financial markets.

Dataset

The dataset for this project comprises stock market data from 20 companies, including Apple (AAPL), Amazon (AMZN), Cisco (CSCO), Google (GOOGL), Intel (INTC), AMD (AMD), Microsoft (MSFT), NVIDIA (NVDA), Kraft Heinz (KHC), Coca-Cola (KO), Kellogg (K), Kimberly-Clark (KMB), Campbell Soup (CPB), General Mills (GIS), Mondelez (MDLZ), and Unilever (UL). The dataset includes historical stock prices, news sentiment data, and economic indicators for analysis.

Detail Design of Methods

The project employs traditional machine learning algorithms, including linear regression, logistic regression, decision trees, random forest, support vector machines (SVM), k-Nearest Neighbors (k-NN), Naive Bayes, Principal Component Analysis (PCA), AutoRegressive Integrated Moving Average (ARIMA), Exponential Smoothing Methods (e.g., Exponential Moving Average), Ridge

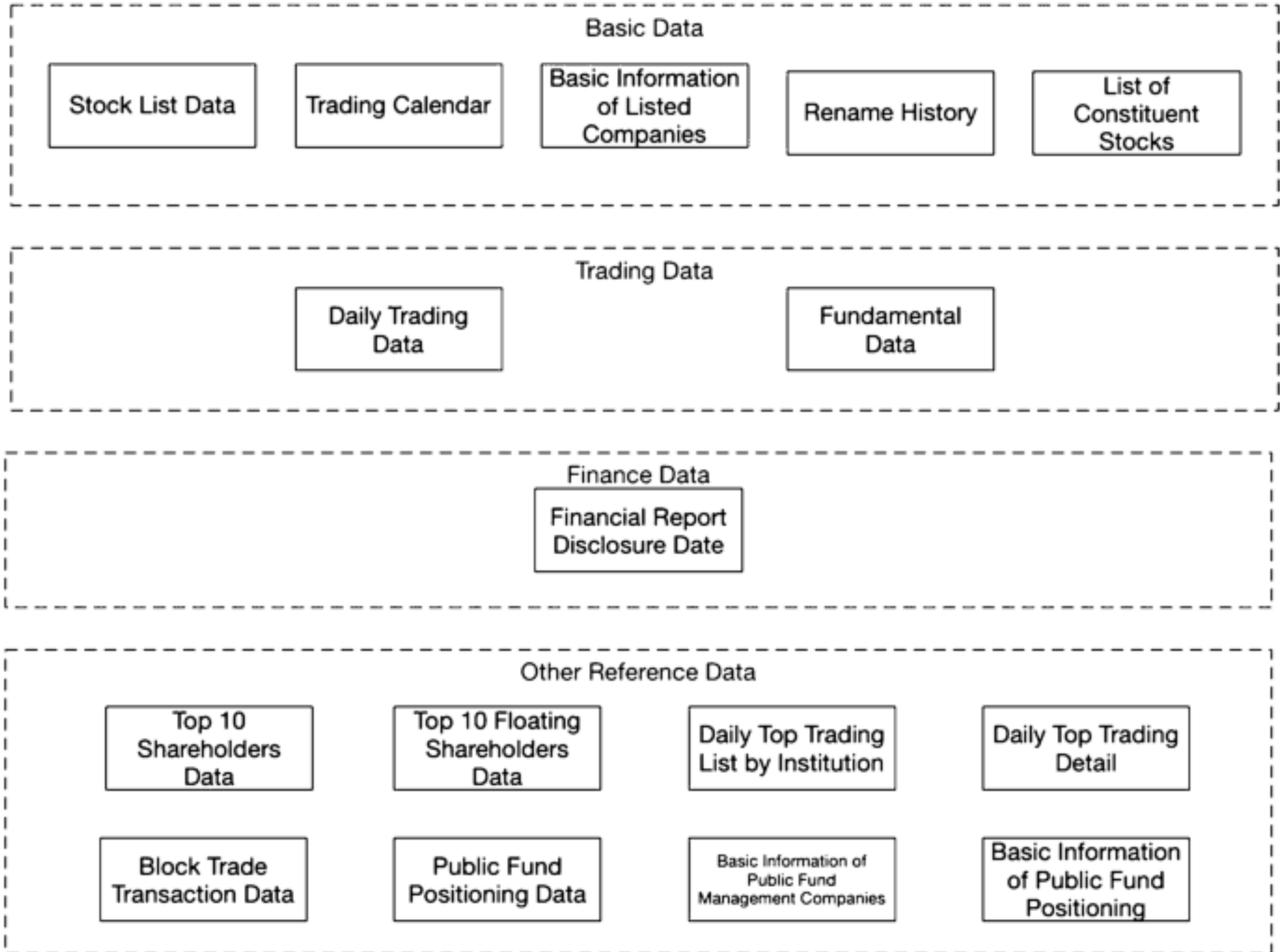
Regression, Lasso Regression, and Long Short-Term Memory (LSTM) networks. Time series analysis, gradient boosting machines (e.g., XGBoost, LightGBM), and other techniques are considered. The use of feature engineering techniques, such as Recursive Feature Elimination (RFE), and pre-processing methods like PCA, contributes to model development.

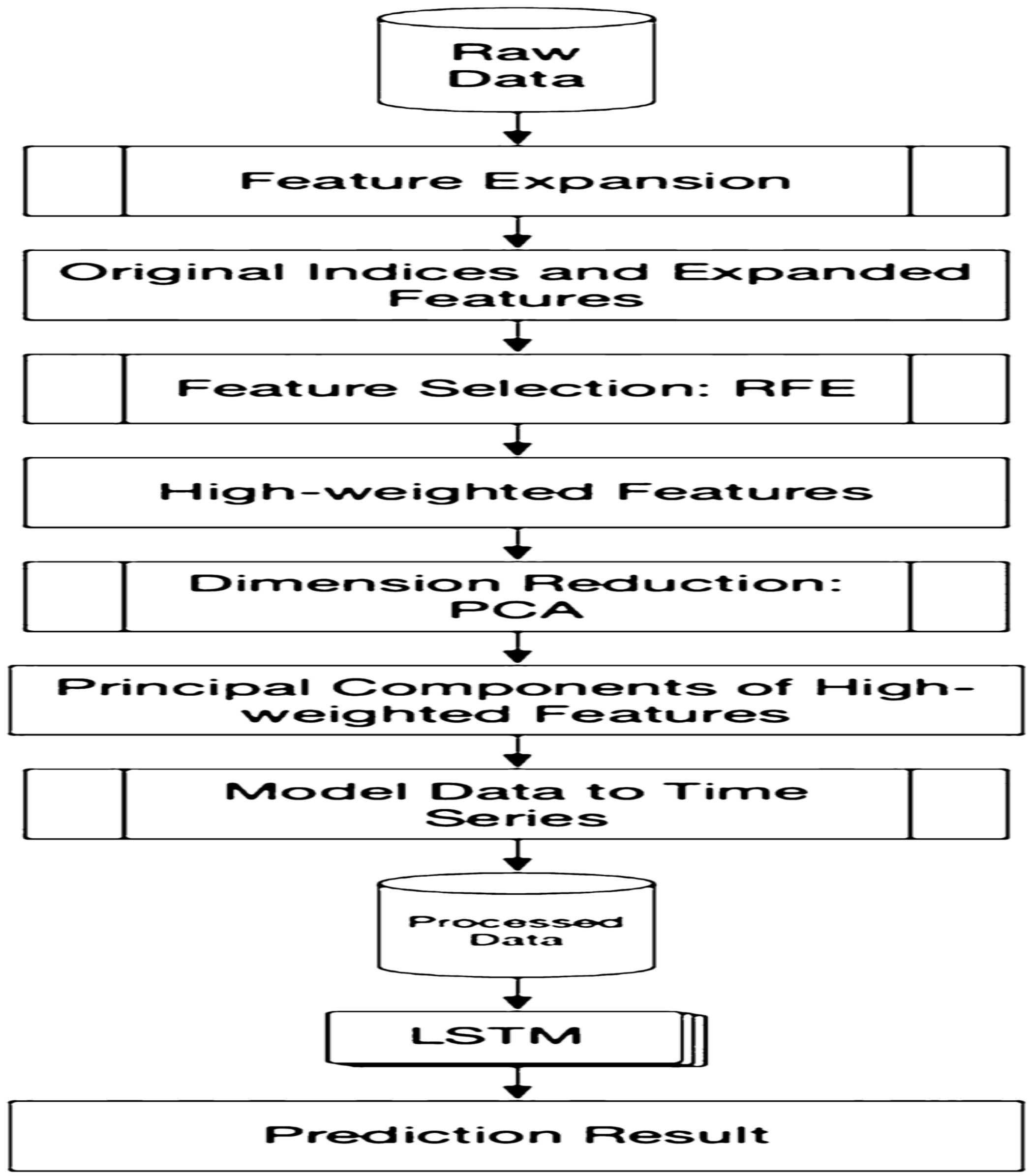
Analysis

The analysis involves exploring the effectiveness of various machine learning algorithms in predicting short-term stock market trends. Feature engineering methods, such as RFE and PCA, are applied to enhance predictive models. The project evaluates the impact of different algorithmic approaches on bi-weekly price trend predictions.

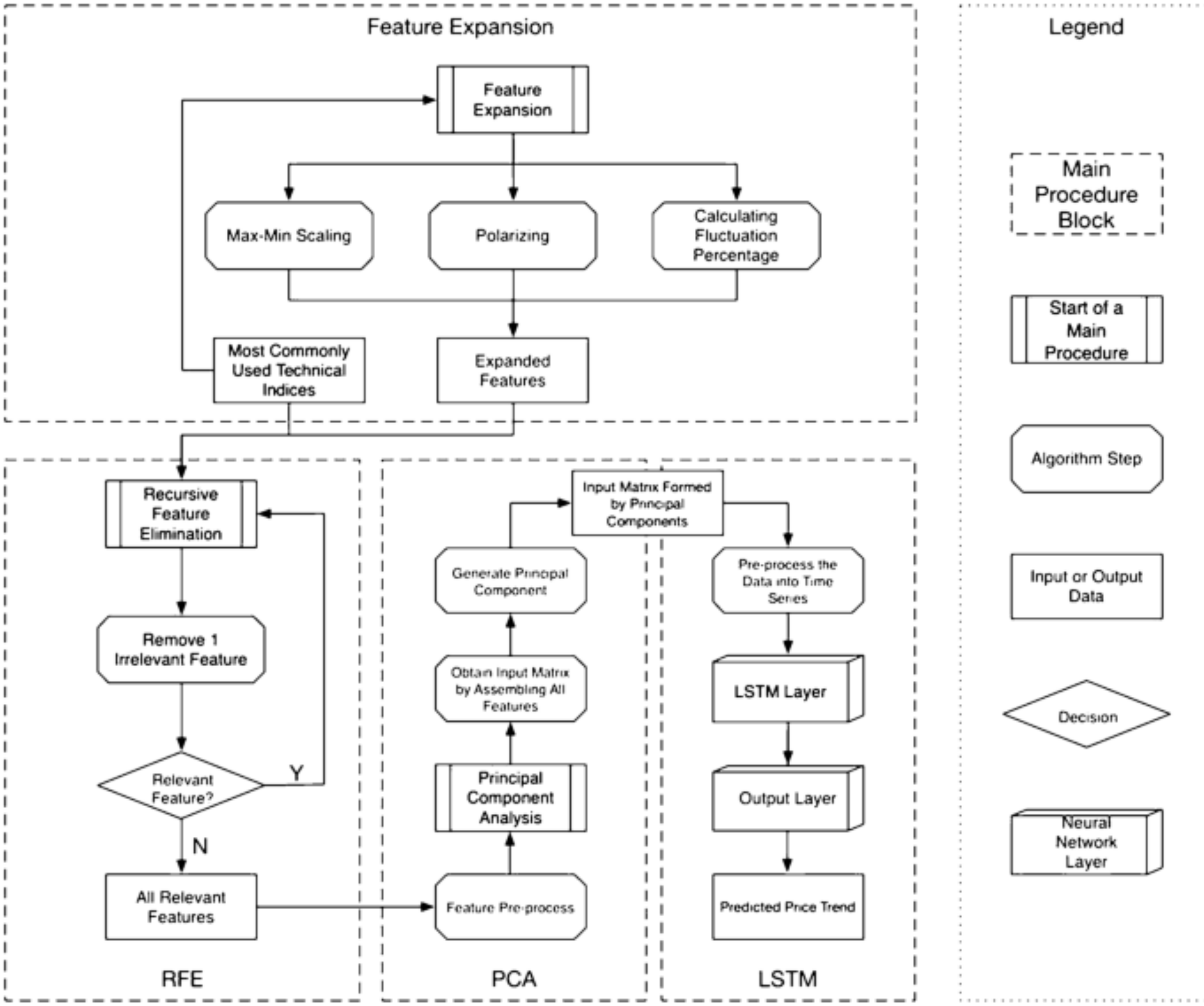
Implementation

The implementation phase includes the development of a user-friendly stock market prediction tool using the selected machine learning algorithms. The tool incorporates the trained models and provides real-time predictions and insights into stock market trends.


Basic Data
Stock List Data
Trading Calendar
Basic Information of Listed Companies
Rename History
List of Constituent Stocks
Trading Data
Daily Trading Data
Fundamental Data
Finance Data
Financial Report Disclosure Date
Other Reference Data
Top 10 Shareholders Data
Top 10 Floating Shareholders Data
Daily Top Trading List by Institution
Daily Top Trading Detail
Block Trade Transaction Data
Public Fund Positioning Data
Basic Information of Public Fund Management Companies
Basic Information of Public Fund Positioning



High-level architecture of the proposed solution



Short-term stock market price trend prediction—applying feature engineering using FE + RFE + PCA

The function FE is corresponding to the feature extension block. For the feature extension procedure, we apply three different processing methods to translate the findings from the financial domain to a technical module in our system design.

Algorithm 1

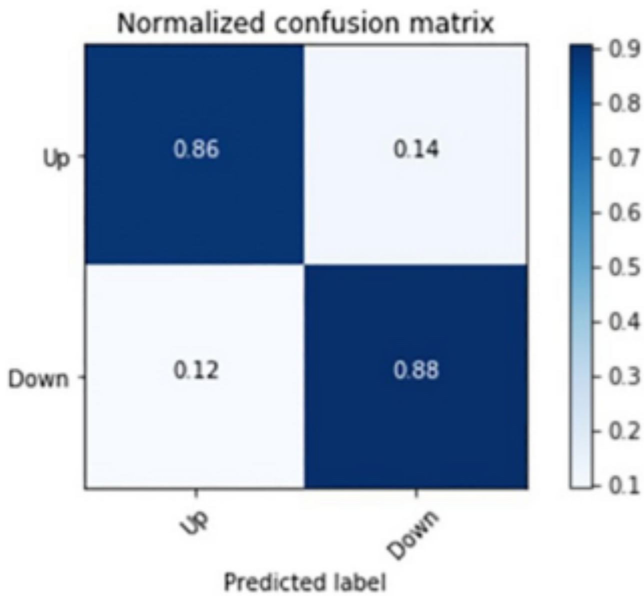
Algorithm 1: Short-term Stock Market Price Trend Prediction - Feature Engineering using FE + RFE + PCA

```
function FE(df)
  # Apply only the meaningful methods on data
  df_expandedfeatures  = Max-MinScaling(df)
  df_expandedfeatures  = Polarizing(df)
  df_expandedfeatures  = CalcFluctuationPercentage(df)
  return df_expandedfeatures
end function

function RFE(df)   # (Utilizing Recursive Feature Elimination function)
  Train the model on all the features of the training dataset in df
  Calculate performance of the model with samples from the test data
  Rank the weights of different features based on testing the model
  for each subset do
    Retain i most weighted features
    Train the model on all the features of the training dataset
    Calculate performance of the model with samples from the test data
  end for
  Calculate the overall performance profile for each feature over samples from the test data
  Rank and select top ranked features
  Train the model on the selected features using the training dataset in df
  return df_RFE    # (df_RFE is the processed data frame after RFE algorithm)
end function

function PCA (df)     # (Utilizing PCA to reduce dimension from i to j)
   df_PCA = applyPCA (n_components=j, whiten=False, copy=True, batchsize = 200)
   return df_PCA     # df_PCA is the optimized data frame after applying PCA algorithm)
end function

function MAIN()                    # (Main function)
  df_alldata = load data
  df_partition = DataPartition(df_alldata, method = resampling)
  df_FE = FE(df_partition)
  df_RFE = RFE(df_FE)
  df_PCA = PCA(df_RFE)
  return df_PCA
end function
```

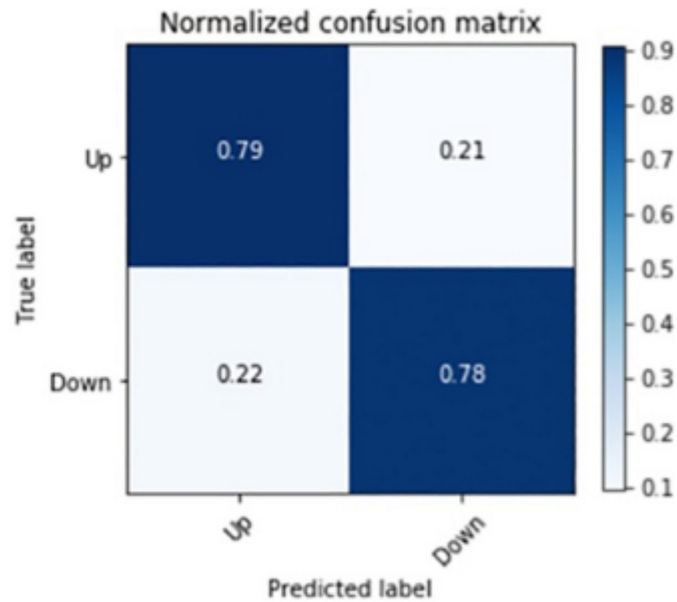
Algorithm 2: Price trend prediction model using LSTM

Algorithm 2

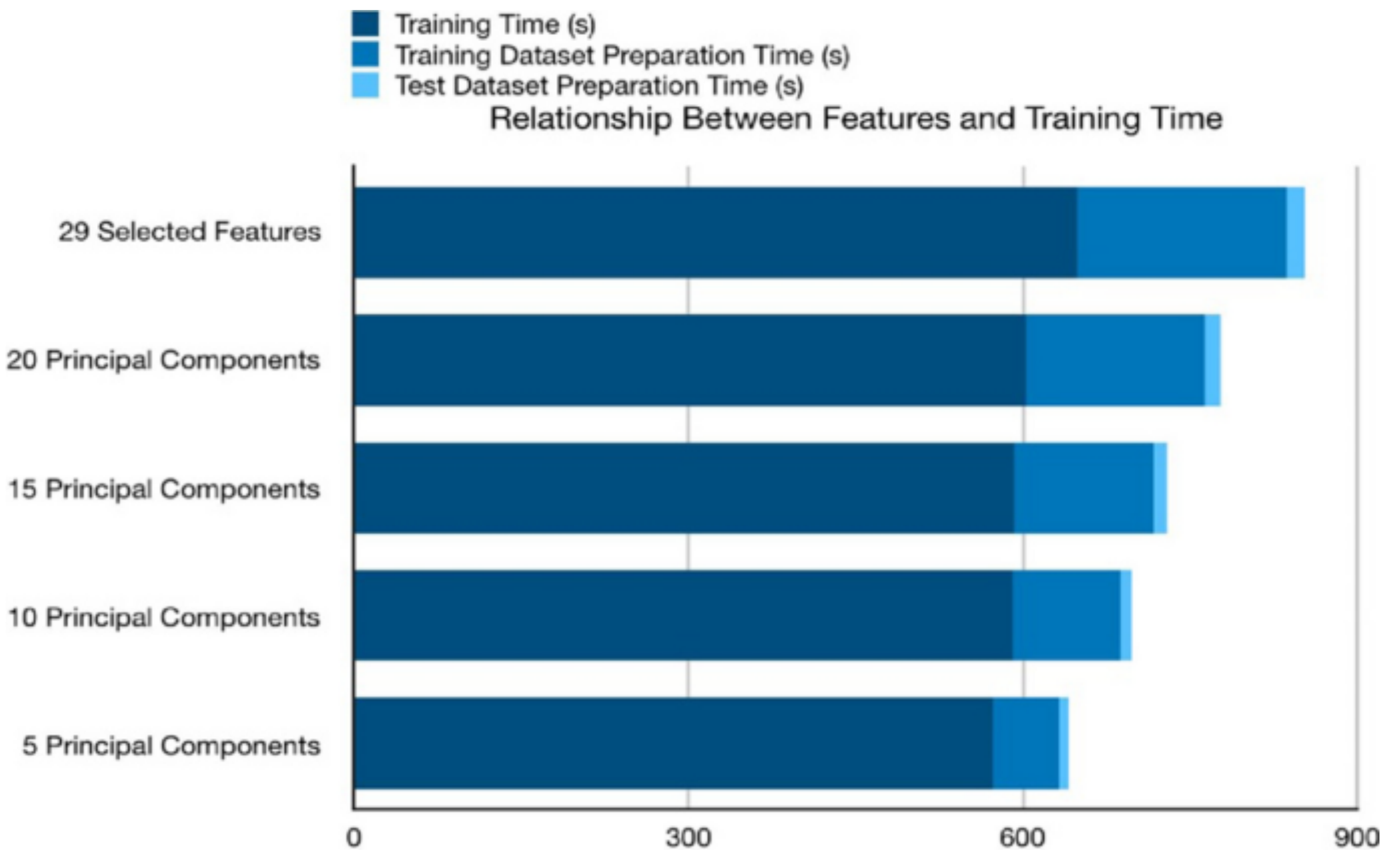
```
Algorithm 2: Price Trend Prediction Model using LSTM

function TimeSeriesConversion (df, term_length, lag)
   # Utilizing time series conversion technique to convert the training data matrix, after applying PCA
from Algorithm 1, to time series
   cols = list()
   for i in range (term_length, 0, -1) do                  # Input sequence
      shift df by i
      append shifted df to cols
   end for
   for i in range (0, lag) do                              # Forecast sequence
      shift df by -1
      append shifted df to cols
   end for
   df_TS = concat(cols, axis = 1)                          # Put all sequences together
   return df_TS
end function

function ModelCompile()                                    # Applying LSTM model with given structure
and compiling it
   Stack_method = Sequential()
   Layer_1 = LSTM(50, input _shape=(train_X.shape[1], train_X.shape[2]))
   Layer_2 = Dense(1)
   Loss_Function=mae
   Optimizer=adam
   Metrics=f1, metrics.binary_accuracy, metrics.mean_squared_error, metrics.mean_absolute_error
   return LSTMmodel
end function

function MAIN()                                            # Main Function
   df_TS = TimeSeriesConversion(df_PCA, N_TIME_STEPS, LAG)
   DataPartition(df_TS, method = resampling)
   ModelCompile(j)
   FitModel(X, y, epochs=50, batch_size=3000)              # Train and fit the model
   EvaluateModel(X_test, y_test)                           # Calculate evaluation metrics on the trained
model using test data
end function
```

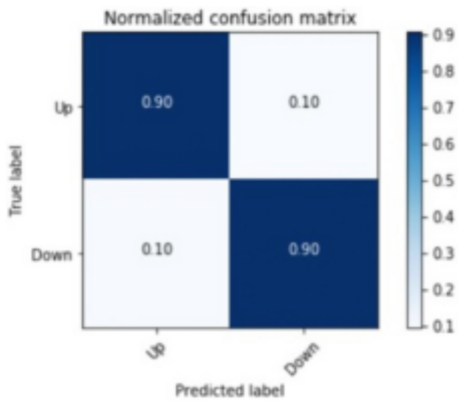


Trained with expanded features

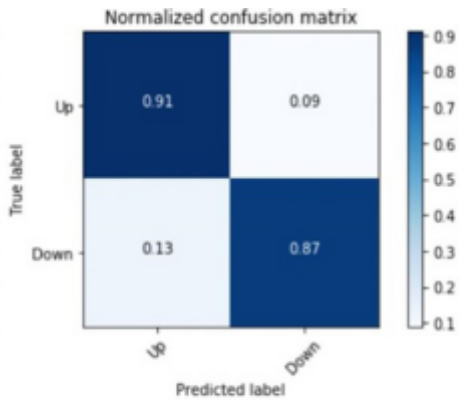


Trained with original features

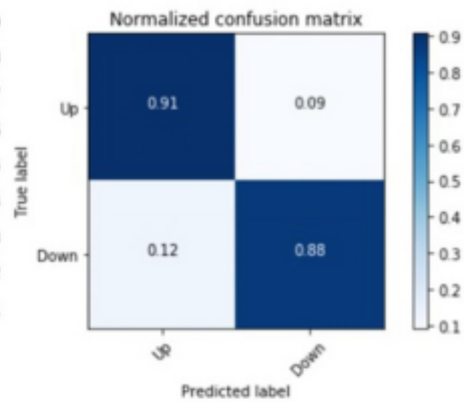

Training Time (s)
Training Dataset Preparation Time (s)
Test Dataset Preparation Time (s)
Relationship Between Features and Training Time
29 Selected Features
20 Principal Components
15 Principal Components
10 Principal Components
5 Principal Components
0
300
600
900



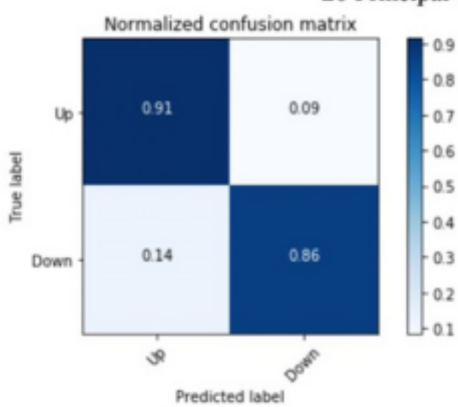
Trained by 29 Features



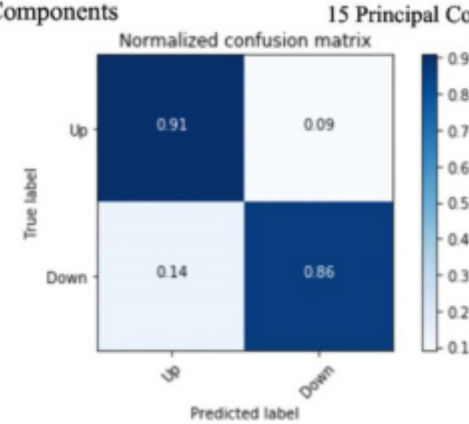
Trained by
20 Principal Components



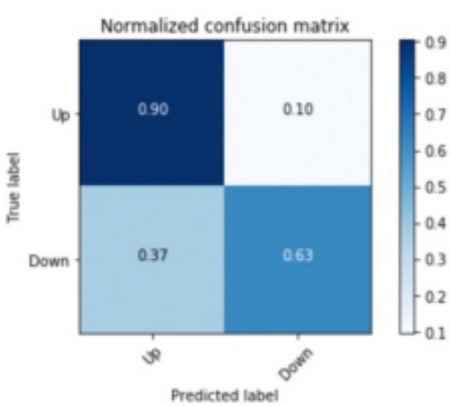
Trained by
15 Principal Components



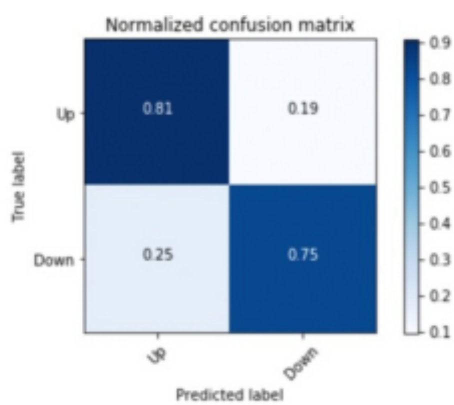
Trained by
10 Principal Components



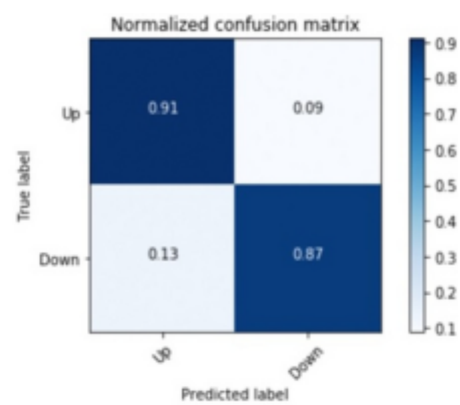
Trained by
5 Principal Components



Without Feature Pre-processing



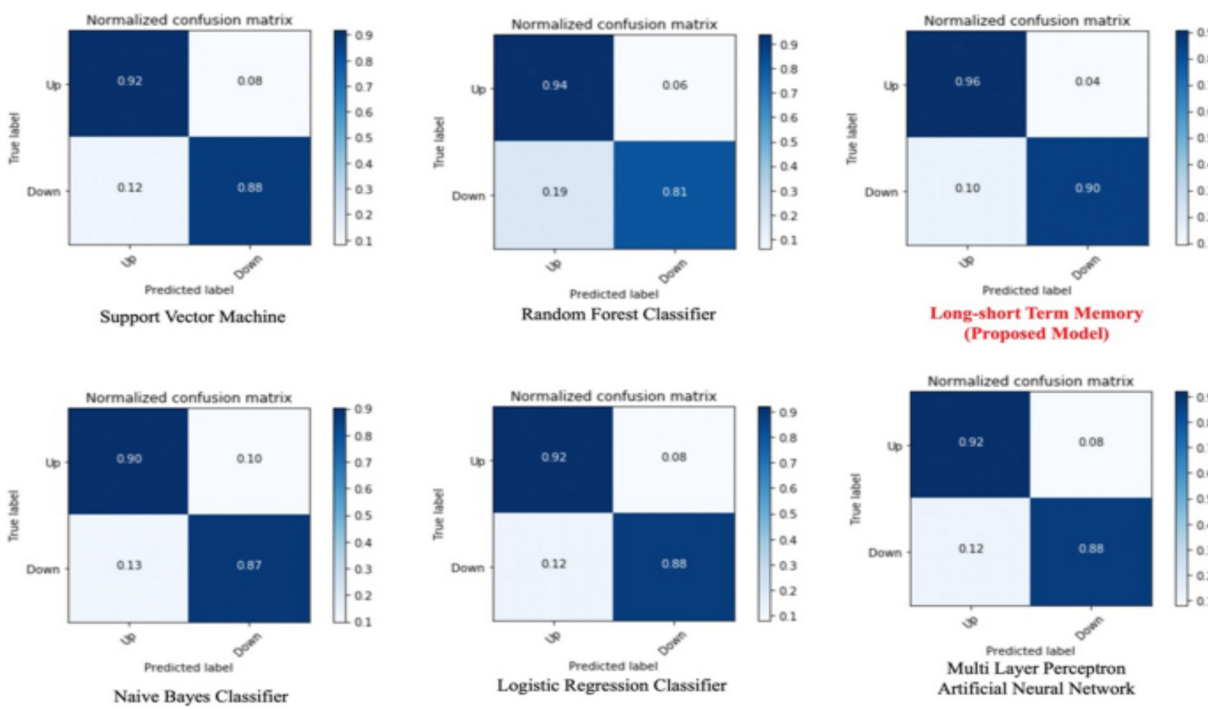
Normalization



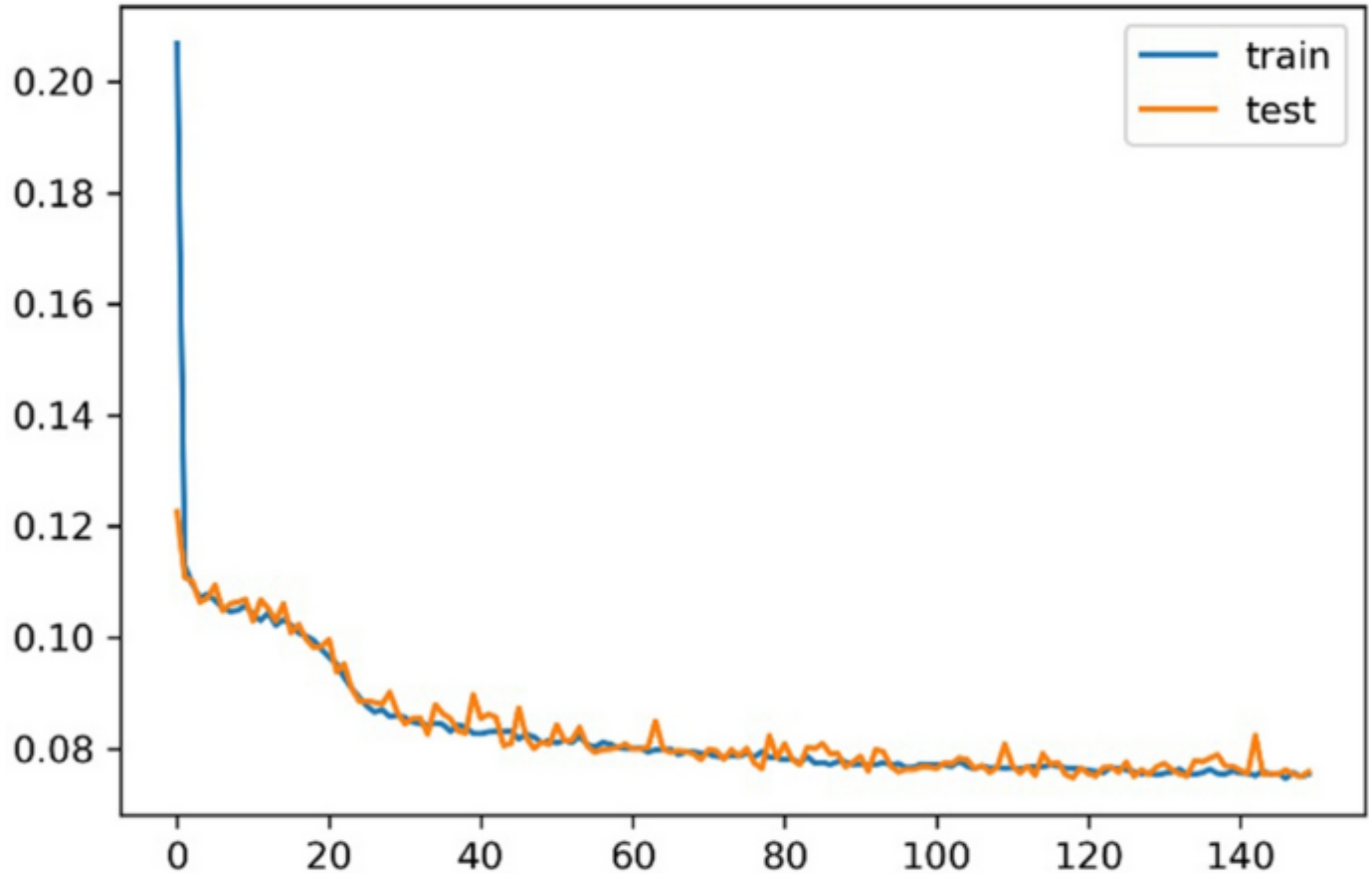
Max-min Scaling

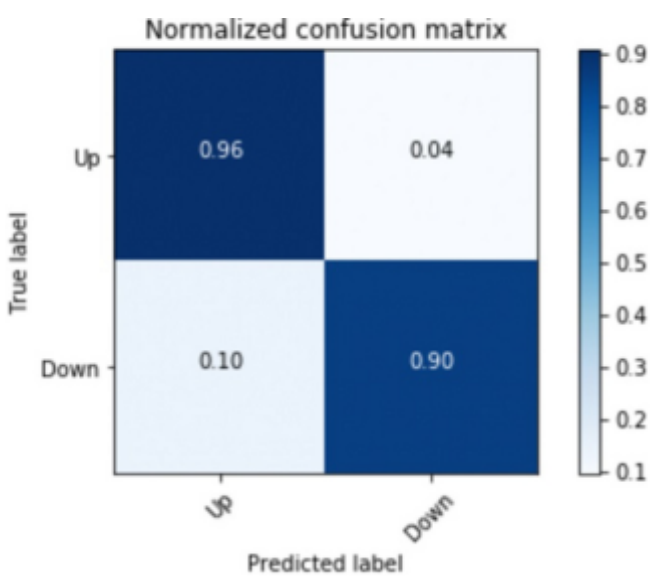
Comparison with related works

In the previous works, we found that mostly common exploited models for stock market price trend prediction are support vector machine (SVM), multilayer perceptron artificial neural network (MLP), Naive Bayes classifier (NB), random forest classifier (RAF), and logistic regression classifier (LR). The test case of comparison is also bi-weekly price trend prediction, to evaluate the best result of all models, we keep all 29 features selected by the RFE algorithm.

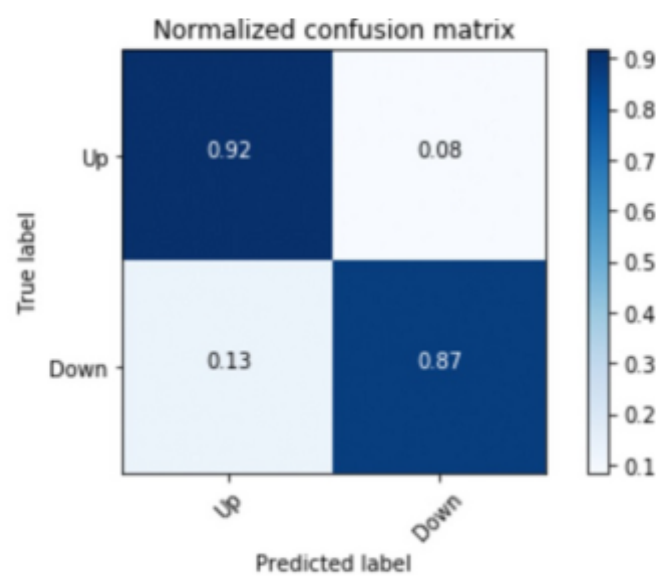


Model prediction comparison—confusion matrices





LSTM Model
Trained by 29 Features



LSTM Model
Trained by 5 Principal Components

Preliminary Results

Preliminary results indicate promising performance, with the proposed LSTM model achieving a binary accuracy of 93.25% in predicting bi-weekly price trends. The inclusion of PCA in the pre-processing stage has significantly improved the training efficiency of the LSTM model by 36.8%.

Project Management

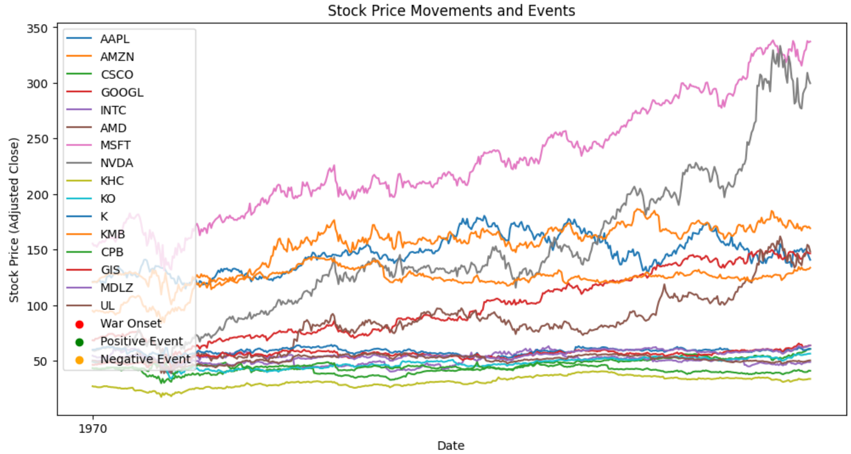
Implementation Status Report

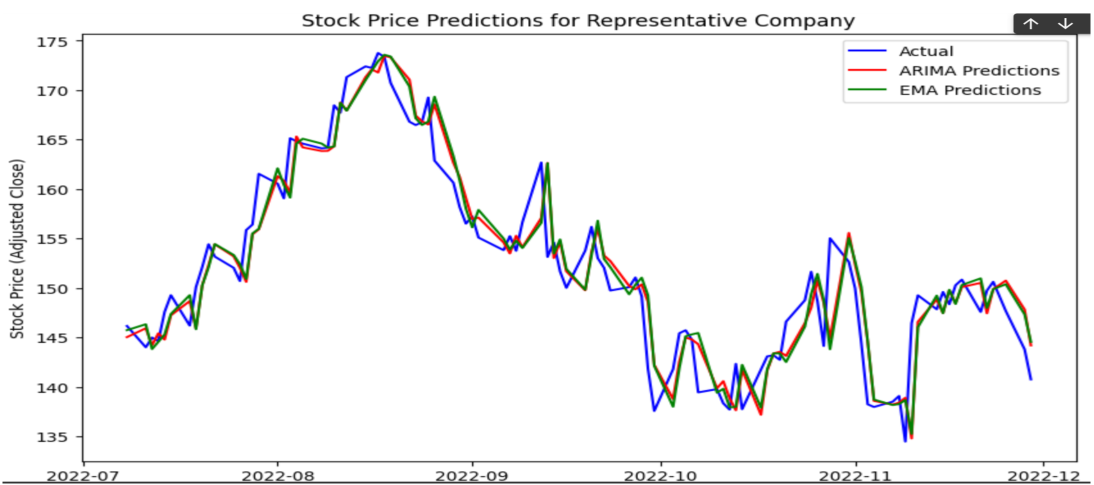
Work Completed:

- **Description, responsibilities, contributions:**
    - P Anvesh Reddy:
        - Worked on Data Preprocessing, the data set for our project comprises stock market data from 20 companies which includes historical stock prices, news events, and economic indicators for analysis each data set is stored in separate CSV file,
          I then worked on model evaluation, training, testing of two algorithms named as linear regression , decisom trees and calculated their mse values.
          I plotted the stock price prediction for respective companies using ARIMA and EMA

UL Linear Regression MSE: 0.11529696942279467

UL Decision Tree MSE: 0.24512012062775299





- Majed Alhazmi :
  - `Worked on training testing evaluation of Randon forest, support vector machines, ridge regression, lasso regression, KNN model.
  - I also worked on recursive feature eliminations and plotted the visualizations

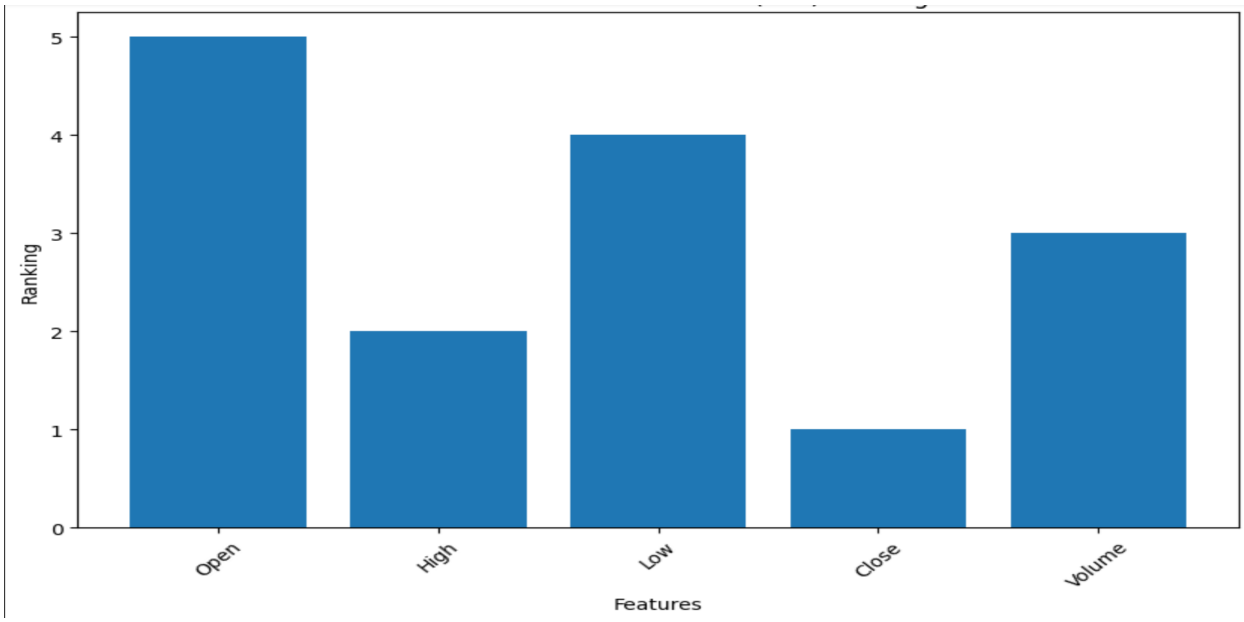
Random Forest MSE: 0.15116601479747122

SVM MSE: 262.1829331986679

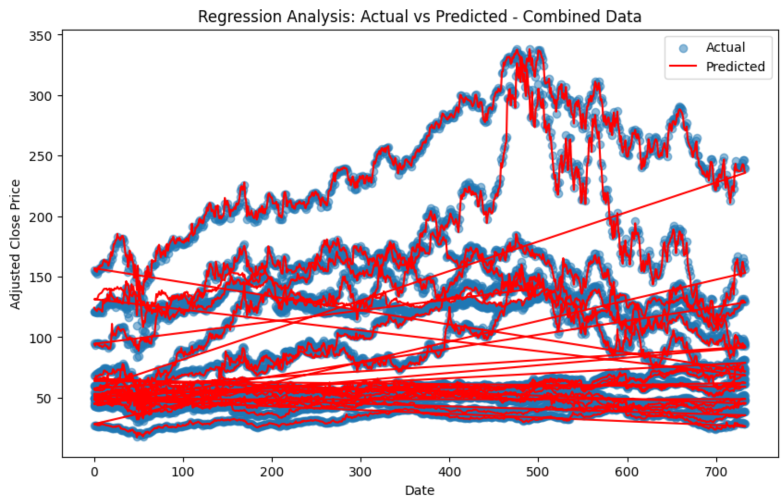
Ridge Regression MSE: 0.11522150167826271

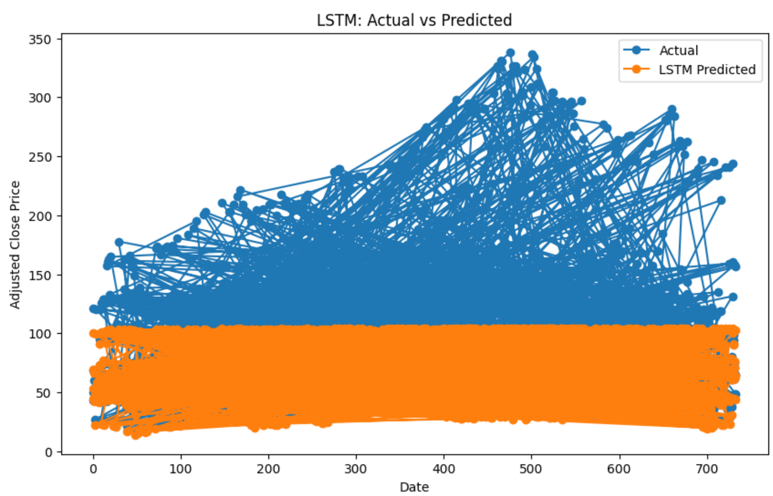
Lasso Regression MSE: 0.14559704109246294

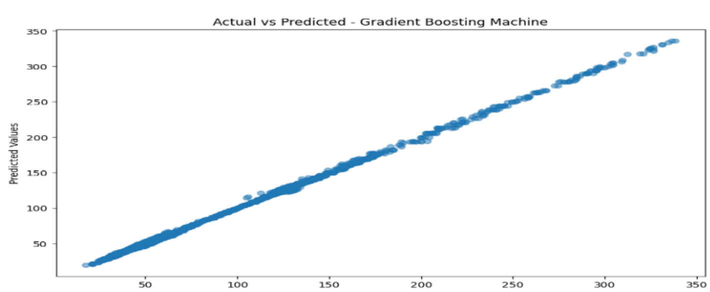
k-NN MSE: 309.26966924654744

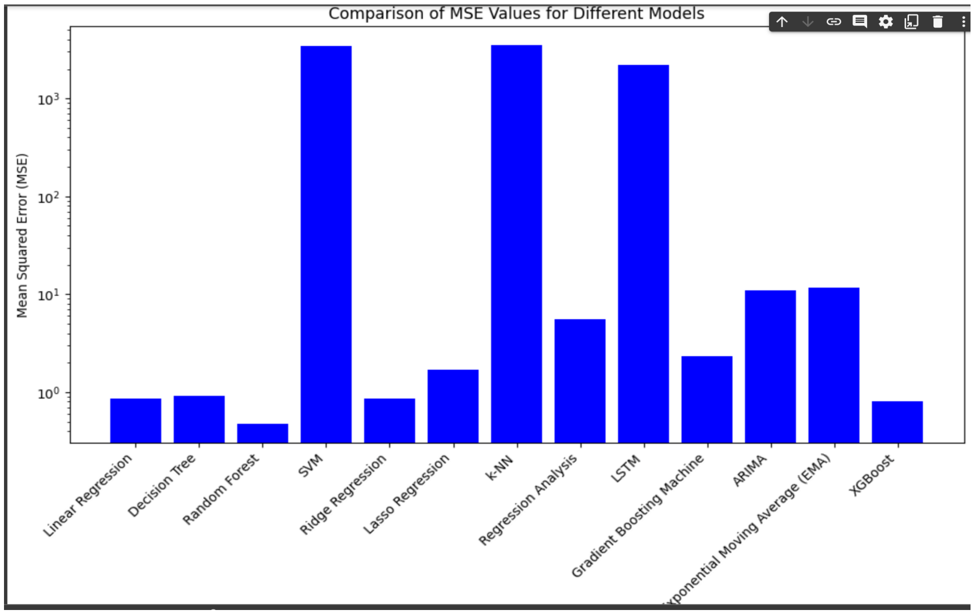


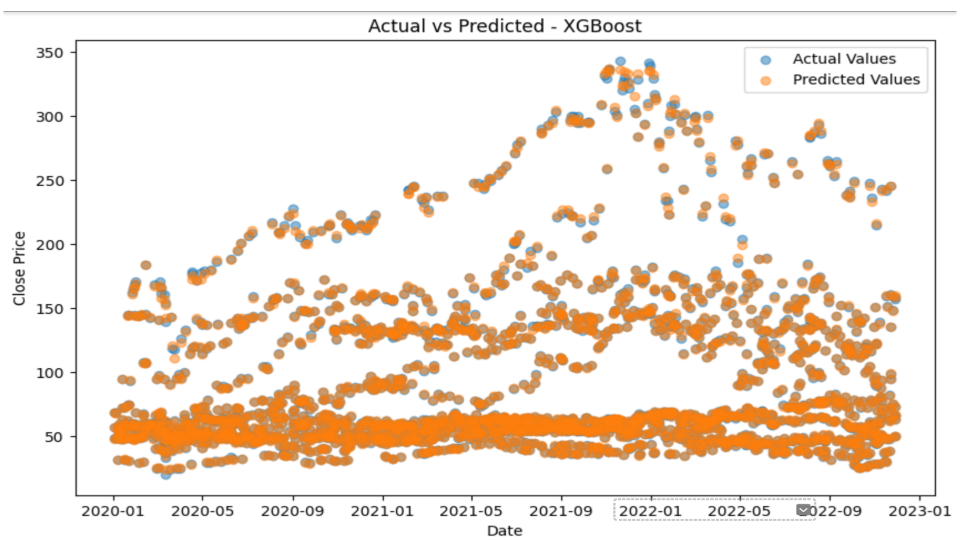
- Divya Anusha Chandrupatla:
    - Worked on Regression analysis, LSTM and Gradient Boosting
    - Apart from them I even worked on other models of the project like XG boost where I plotted the actual vs predicted values using close price and date
    - In the end I compared the mse values of all the algorithms listed and visualized it in a bar graph


Regression Analysis: Actual vs Predicted - Combined Data
Actual
Predicted
Adjusted Close Price
Date
0
100
200
300
400
500
600
700
50
100
150
200
250
300
350


LSTM: Actual vs Predicted
Actual
LSTM Predicted
Adjusted Close Price
Date
0
100
200
300
400
500
600
700
0
50
100
150
200
250
300
350

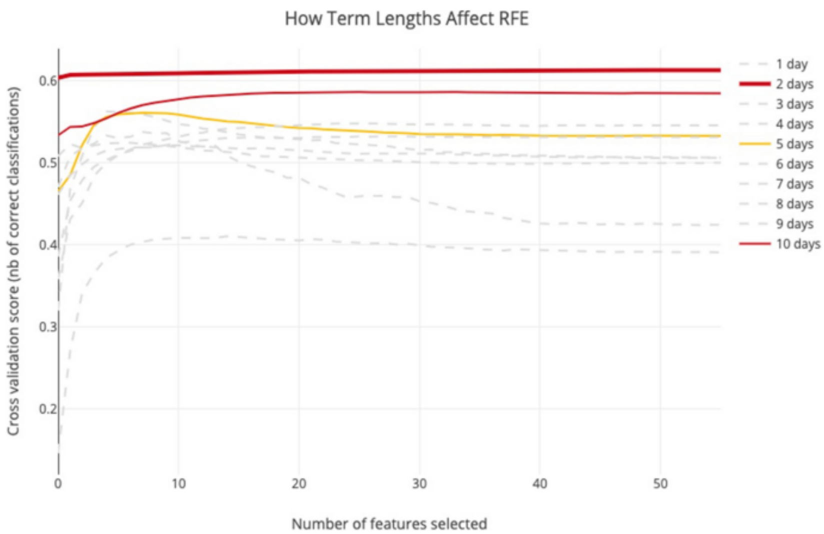

Actual vs Predicted - Gradient Boosting Machine
Predicted Values
Actual Values
50
100
150
200
250
300
350



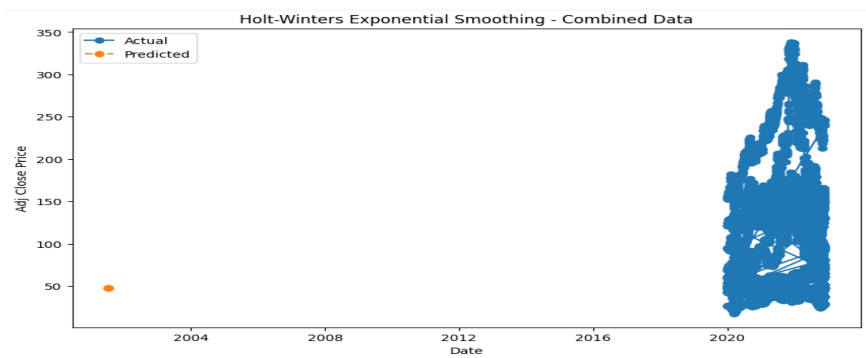


Vaka Raja vinay kumar:

- **Worked on RFE and plotted results**
- **I also worked on holt winters exponential smoothing and calculated the mse value**
- **I contributed working on confusion matrix and comparing with past results**



○



:

Conclusion:

In conclusion, our project proposal outlines a comprehensive approach to explore the potential of traditional machine learning algorithms in predicting stock market trends. With a well-defined team and clear responsibilities, we aim to develop a user-friendly tool that can provide valuable insights for investors and financial analysts. This research has the potential to substantially impact the financial industry, aiding in more informed investment decisions and risk management strategies. We are confident in our ability to deliver a successful project and look forward to contributing to the field of stock market prediction.

References:

Ahmed, M. K., Wajiga, G. M., Blamah, N. V., & Modi, B. (2019). Stock market forecasting using ant colony optimization-based algorithm. Am J Math Comput Model, 4(3), 52-57.

Bustos, O., & Pomares-Quimbaya, A. (2020). Stock market movement forecast: A systematic review. Expert Systems with Applications, 156, 113464.

Gandhmal, D. P., & Kumar, K. (2019). Systematic analysis and review of stock market prediction techniques. Computer Science Review, 34, 100190.

Gandhmal, D. P., & Kumar, K. (2019). Systematic analysis and review of stock market prediction techniques. Computer Science Review, 34, 100190.

Javed Awan, M., Mohd Rahim, M. S., Nobanee, H., Munawar, A., Yasin, A., & Zain, A. M. (2021). Social media and stock market prediction: a big data approach. MJ Awan, M. Shafry, H. Nobanee, A. Munawar, A. Yasin et al.," Social media and stock market prediction: a big data approach," Computers, Materials & Continua, 67(2), 2569-2583.

Liu, H., & Long, Z. (2020). An improved deep learning model for predicting stock market price time series. Digital Signal Processing, 102, 102741.

Nabipour, M., Nayyeri, P., Jabani, H., Mosavi, A., & Salwana, E. (2020). Deep learning for stock market prediction. Entropy, 22(8), 840.

Nabipour, M., Nayyeri, P., Jabani, H., Shahab, S., & Mosavi, A. (2020). Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis. IEEE Access, 8, 150199-150212.

Nikou, M., Mansourfar, G., & Bagherzadeh, J. (2019). Stock price prediction using DEEP learning algorithm and its comparison with machine learning algorithms. Intelligent Systems in Accounting, Finance and Management, 26(4), 164-174.

Shah, D., Isah, H., & Zulkernine, F. (2019). Stock market analysis: A review and taxonomy of prediction techniques. International Journal of Financial Studies, 7(2), 26.

Singh, S., Madan, T. K., Kumar, J., & Singh, A. K. (2019, July). Stock market forecasting using machine learning: Today and tomorrow. In 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT) (Vol. 1, pp. 738-745). IEEE.

Yuan, X., Yuan, J., Jiang, T., & Ain, Q. U. (2020). Integrated long-term stock selection models based on feature selection and machine learning algorithms for the Chinese stock market. IEEE Access, 8, 22672-22685.

Zhong, X., & Enke, D. (2019). Predicting the daily return direction of the stock market using hybrid machine learning algorithms. Financial Innovation, 5(1), 1-20.