Milestone 3
STOCK PRICE
PREDICTION

Team 10
Srikanth Parvathala
Vijay Arni
Ya-Ting Yang

Table of contents

Research Question

Can future stock prices be predictable using historical stock price of specific company and macroeconomic indicators like the nominal GDP index, real GDP index, unemployment rate, and federal funds rate?

SVM

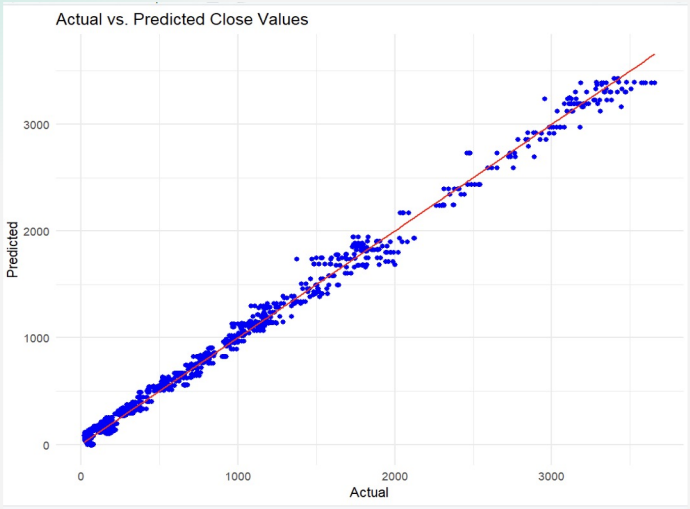
Objective

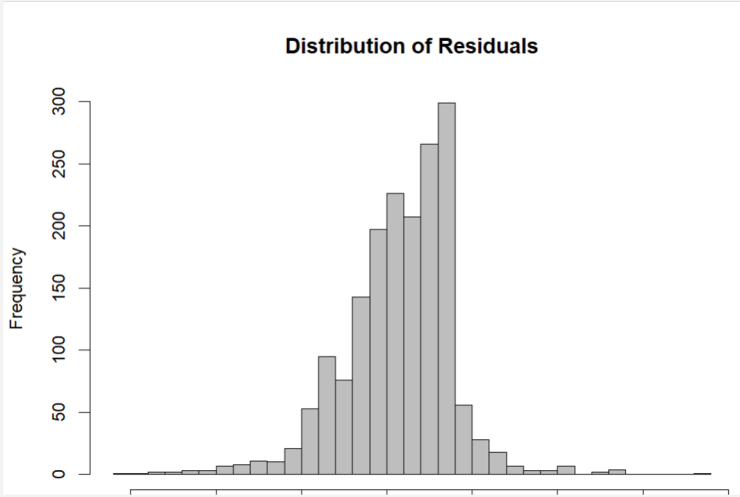
- To forecast the 'Close' price using historical stock data and macroeconomic indicators.
- Data collection completed in milestone 1, covering stock prices and indicators like GDP indexes, unemployment rate, and federal funds rate.
- Dummy variable encoding, missing value handling, feature scaling.

SVM

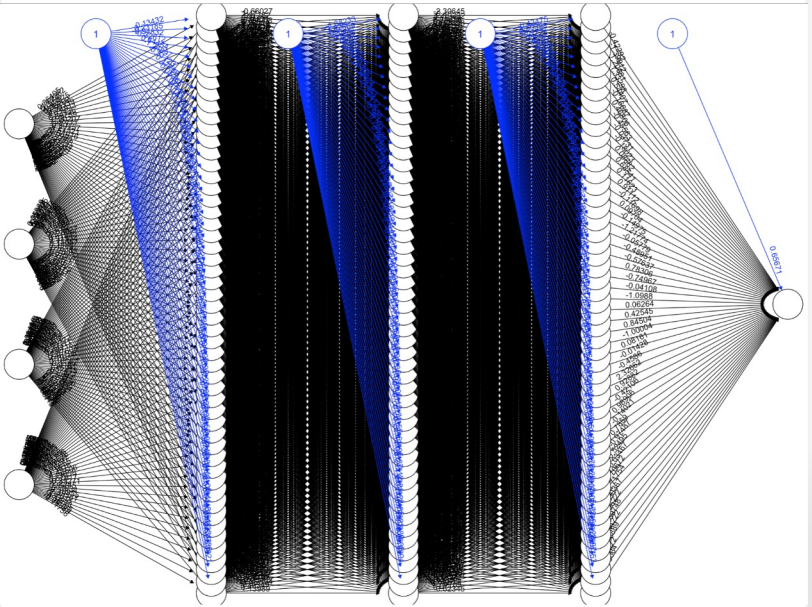
- **Model training**:
  - Linear SVM model established baseline; non-linear SVM models (RBF, Polynomial, Sigmoid) explored complex relationships.
  - RBF kernel showed best performance, indicating non-linear trends in the data.

SVM





- RMSE Results: Linear (433.97), RBF (66.39), Polynomial (74.26), Sigmoid (50,923.34).
- Visualization of Actual vs. Predicted values indicates a strong fit for the RBF kernel model.
- Residual histogram suggests areas for model improvement, especially for underprediction.



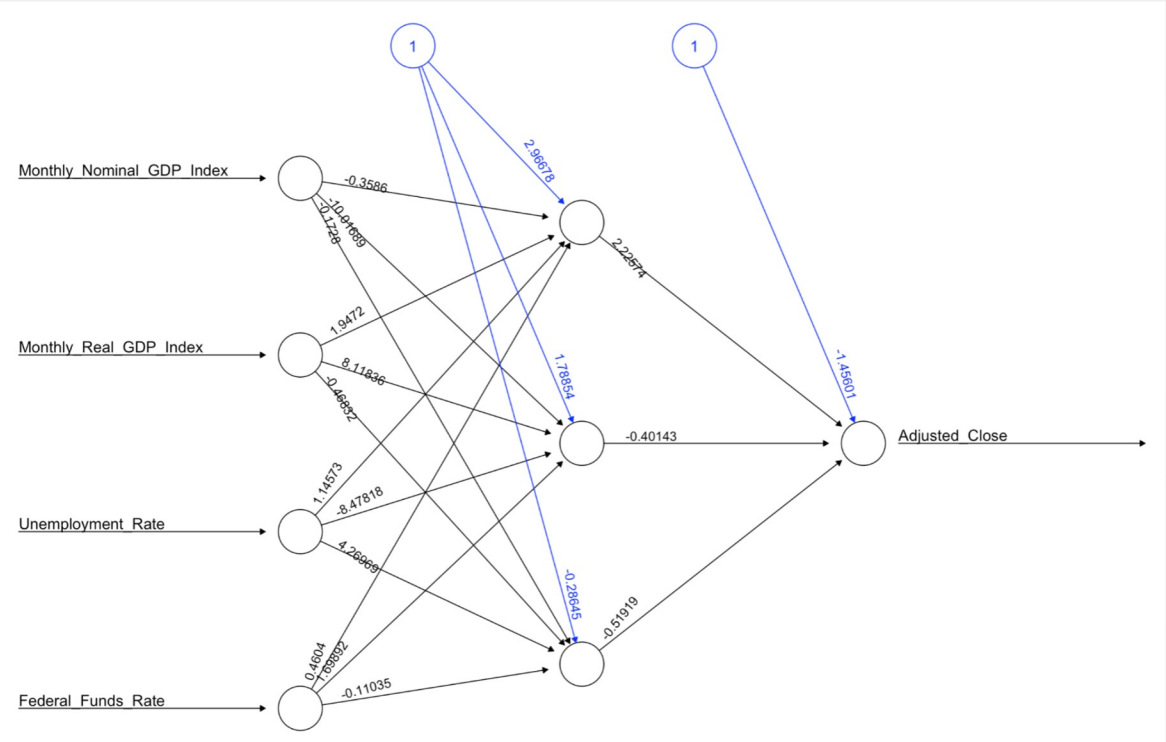
02

Neural Networks

Data Preprocessing & Configurations

- **Normalization of Input Variables**
  - Scaling continuous variables such as GDP indexes, unemployment rate, federal funds rate to a [0, 1] range
- **Dataset**
  - Stock prices from Amazon, Google, Facebook, Netflix, and Apple
- **Training Parameters**
  - Setting learning rate at 0.01
  - Conducting 100 training epochs

Training Combined Data Model



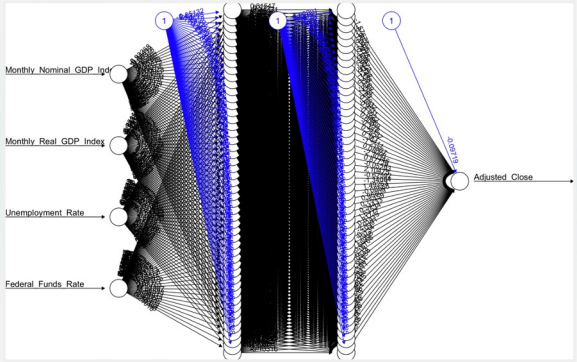
- Hidden layer: 1~4
- Neurons: 1~30
- Activation function: logistic, tanh
- Data: Incorporating stock prices from five major companies

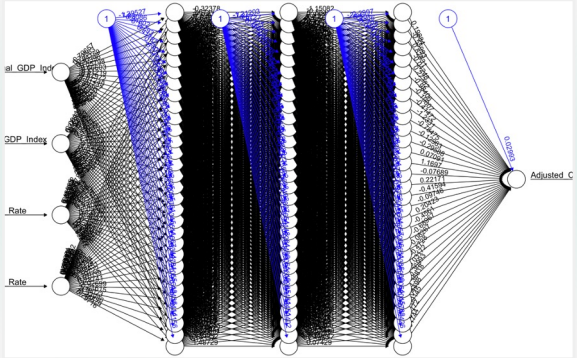
BEST MODEL

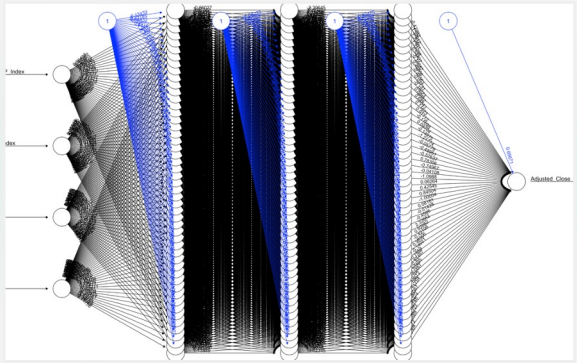
1 hidden layer with 3 neurons using the logistic activation function

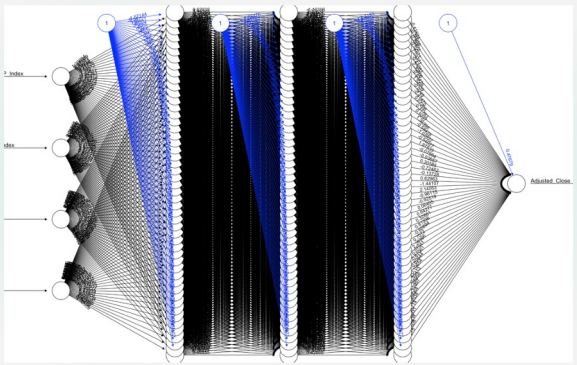
0.417

Training Separate Models For Each Company









- Hidden layer: 1~3
- Neurons: 1~50
- Activation function: logistic, tanh
- Data: using their specific stock price data

BEST MODEL

2 or 3 hidden layer with 50 neurons using the tanh activation function

0.998

Key Findings

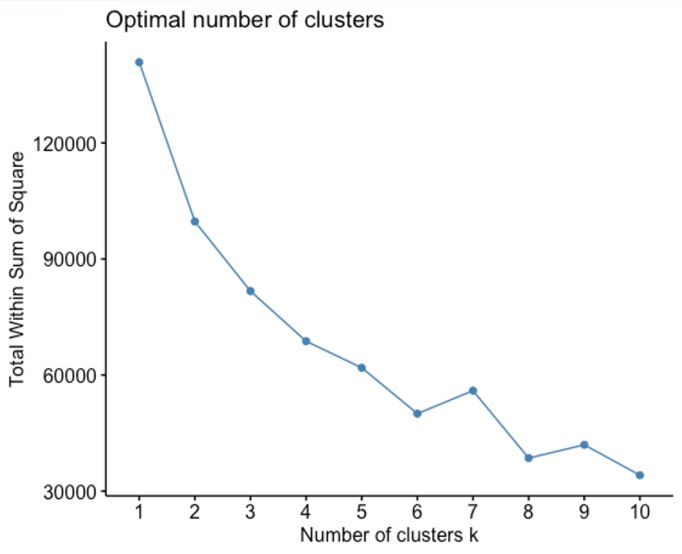
Best Performing Models

- Combined Data Model
  - Correlation : 0.417
  - Simpler models might be sufficient to capture the underlying patterns in the data
- Individualized company model
  - Correlation: 0.998
  - Better results with more neurons and layers; tanh activation function slightly outperforms logistic.

03

Clustering

K-Means Clustering



Optimal number of clusters are **6**

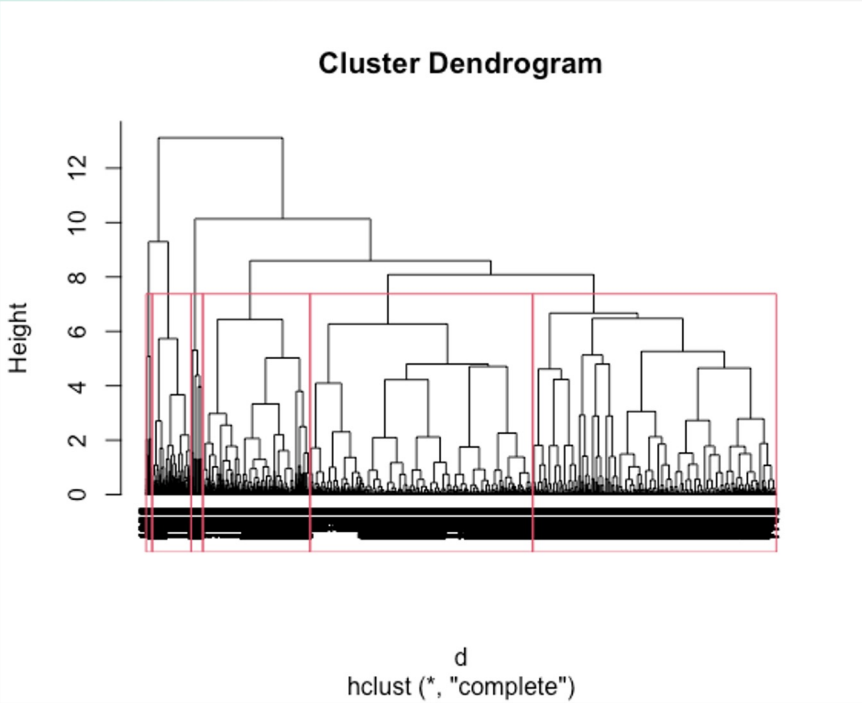
K-Means Clustering

Number of elements in each cluster are

- Cluster 1: 1761
- Cluster 2: 1761
- Cluster 3: 1532
- Cluster 4: 1321
- Cluster 5: 669
- Cluster 6: 1761

Each centroid of the cluster represents a group of data points (days) with similar characteristics in terms of stock prices, economic indicators, and news sources.

Hierarchical Clustering



- Based on Dendrogram, optimal number of clusters are **6**.
- The number of elements per cluster are :

  Cluster 1: 3111

  Cluster 2: 1489

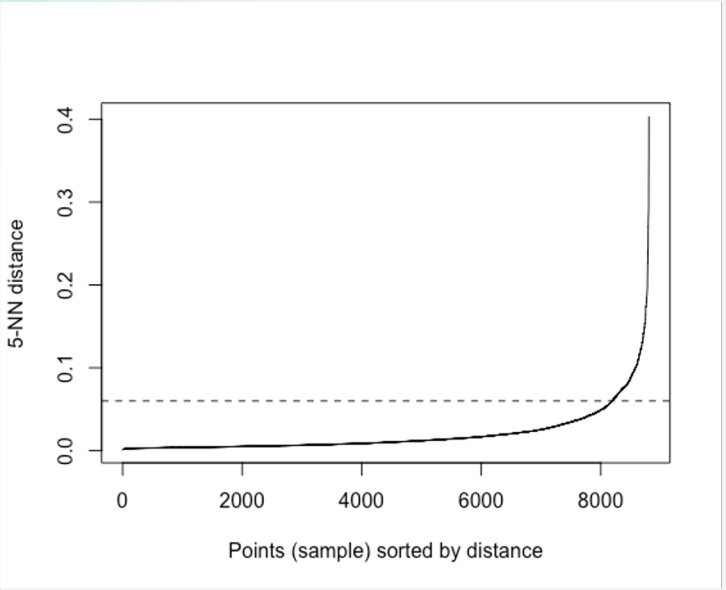
  Cluster 3: 85

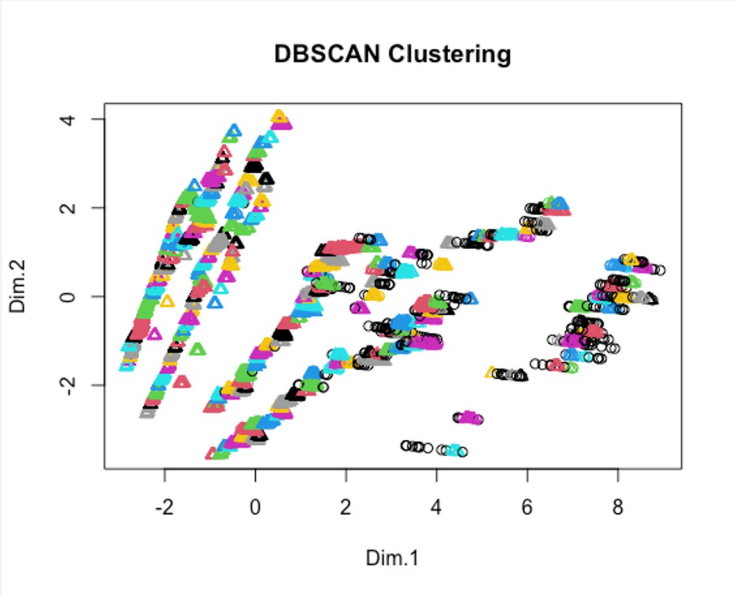
  Cluster 4: 544

  Cluster 5: 3403

  Cluster 6: 164
- The largest clusters 1 and 5 , likely represent standard market conditions. In contrast, the smaller clusters capture the more unique and nuanced patterns within the data.

Density-Based Clustering (DBSCAN)





- Based on above plot, we found 0.06 is the best epsilon value to train the DB model

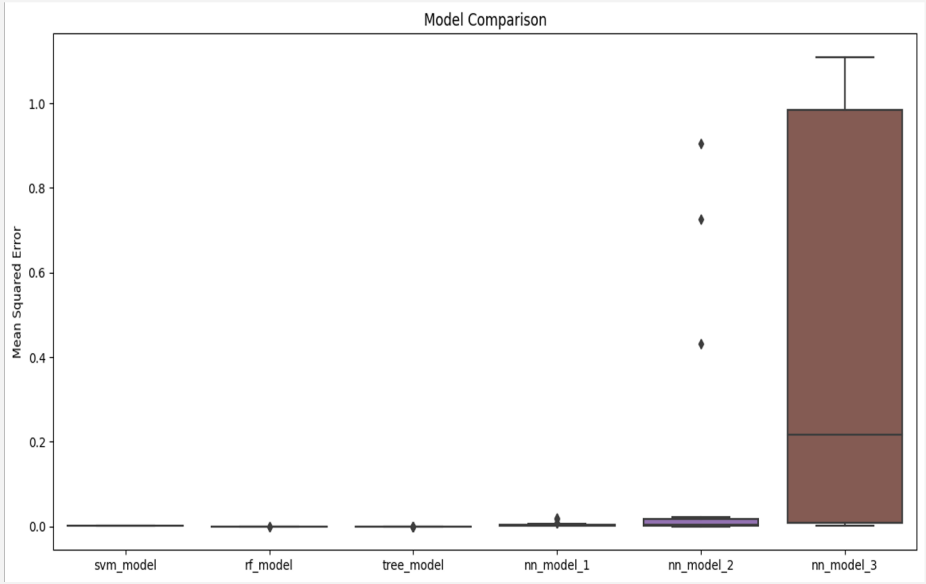
- Total number of Outliers are 289 points.

Comparative Analysis

- **Data Set preprocessing:**
  - Features and Size
  - Normalization and encoding
  - Conversion of Close into classes

- **Analytical Methodology:**
  - Machine learning models used: SVM, RF, NN, DT
  - Cross-validation techniques: K-Fold, Random Subsampling, Bootstrapping
  - Performance metrics: MSE

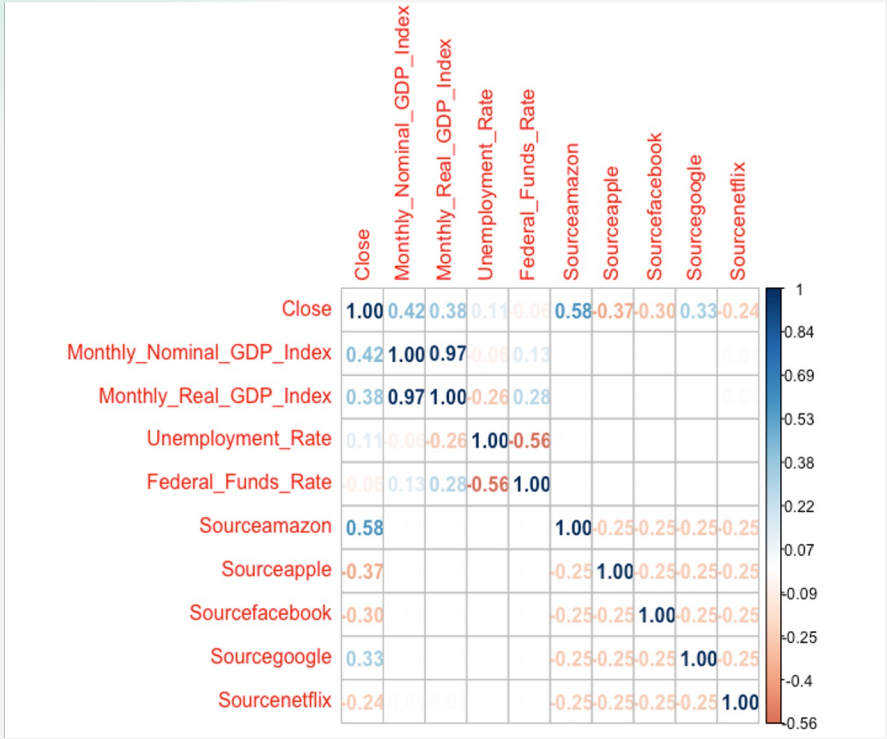
Comparative Analysis



Comparative Analysis

- **Conclusion:**
  - SVM and RF models are recommended
  - The potential of NN models with further tuning

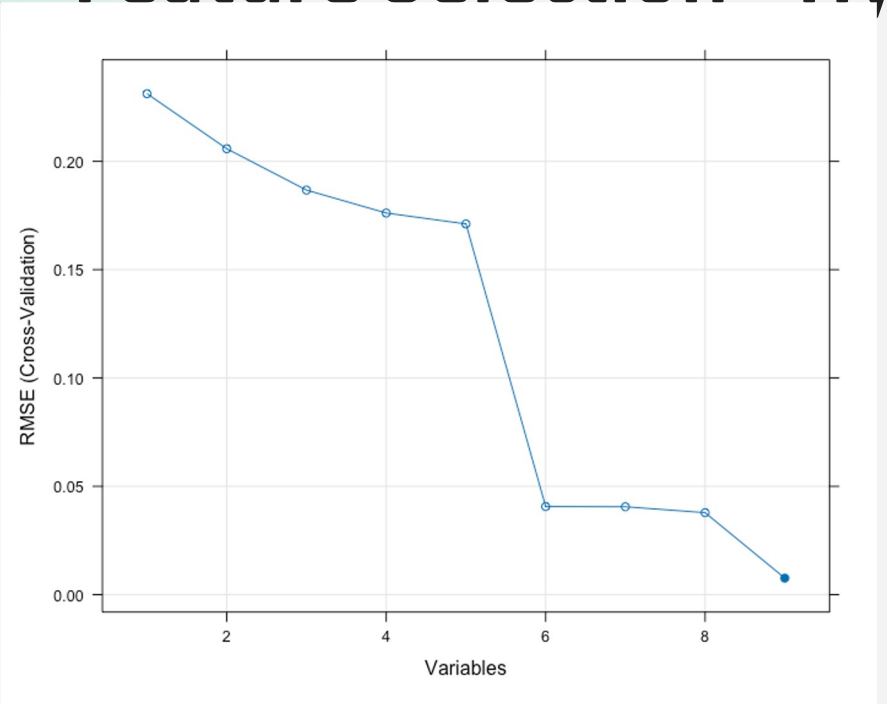
Feature Selection - Filter Method



Filter method on SVM model:

- Based on the matrix, "Unemployment_Rate" and "Federal_Funds_Rate"" appear to have the least correlation values with respect to the "Close". So we eliminated both.
- Sourceamazon, Monthly_Nominal_GDP_Index, "Monthly_Real_GDP_Index" and "Sourcegoogle" are positively correlated with Close price.
- "Sourceapple", "Sourcefacebook", and "Sourcenetflix" are negatively correlated with Close price.
- SVM model has low MSE value (0.012) and a high R-Squared value (0.701), demonstrating a strong ability to predict the Close prices.

Feature Selection - Hybrid Method (RFE)



RFE on RandomForest model:

- From the plot, we can see the number of features increases from 2 to around 5, there's a noticeable decrease in the RMSE.
- This suggests that adding more features up to this point improves the model's predictive accuracy.
- Selecting more than 5 features does not appear to provide a substantial decrease in RMSE.
- Sourcenetflix, Sourcefacebook, Federal_Funds_Rate, Monthly_Nominal_GDP_Index and Sourceamazon are the best features for this RandomForest model.

Feature Selection - Wrapper Method (SFS)

Wrapper method on Linear regression model:

- Using a forward stepwise algorithm, a refined model is identified by progressively adding predictors to enhance model accuracy.
- The final model included Sourceamazon, Sourcegoogle, Monthly_Nominal_GDP_Index, Unemployment_Rate, Sourceapple, Federal_Funds_Rate, Sourcefacebook, and Monthly_Real_GDP_Index.

Model Performance:

- This model demonstrated a high degree of predictive accuracy for stock close prices, as indicated by a high R-squared value(0.7861).

Ethical Issues

Data Collection

- Historical stock price data: web scraping and data integrity in financial datasets
- Excluding unquantifiable factors such as emotional influence in stock price data

Model Evaluation

- Industry bias and its impact on stock market representation
- Predictive model influence on investment behavior

References

Alexius, A. (2018). *Stock prices and GDP in the long run*. https://econpapers.repec.org/article/sptapfiba/v_3a8_3ay_3a2018_3ai_3a4_3af_3a8_5f4_5f7.htm

Farsio, F., & Fazel, S. (2013). *The Stock Market/Unemployment relationship in USA, China and Japan. International Journal of Economics and Finance, 5(3)*. https://doi.org/10.5539/ijef.v5n3p24

Historical Stock Price of (FAANG + 5) companies. (2021, December 30). Kaggle. https://www.kaggle.com/datasets/suddharshan/historical-stock-price-of-10-popular-companies/code?select=Microsoft.csv

Kumbure, M. M., Lohrmann, C., Luukka, P., & Porras, J. (2022). Machine learning techniques and data for stock market forecasting: A literature review. *Expert Systems With Applications*, *197*, 116659. https://doi.org/10.1016/j.eswa.2022.116659

Sharing insights elevates their impact. (n.d.). S&P Global. https://www.spglobal.com/marketintelligence/en/mi/products/us-monthly-gdp-index.html

Shiblee, L. S. (2009). The Impact of Inflation, GDP, Unemployment, and Money Supply On Stock Prices. *Social Science Research Network*. https://doi.org/10.2139/ssrn.1529254

Unemployment rate. (2023, September 1). https://fred.stlouisfed.org/series/UNRATE