# Stock Prediction: Agglomerative approach with News and Tweets Sentiment

Abhijeet Dhakane, *Bredesen Center, UT Knoxville*

## Abstract

**Stock prediction is a major trending topic in the field of data science. In this project machine learning models are applied to data gathered through the Quantopian Research platform[1] to make predictions on buy and sell signals. The Moving Average Distance (MAD) strategy between 21 and 252 days, along with sentiment analysis of news and tweets are taken into account. Features are based on previous events from a certain anchor day. The machine learning models applied include Logistic Regression, Random Forest, and Naive Bayesian.**

## Background

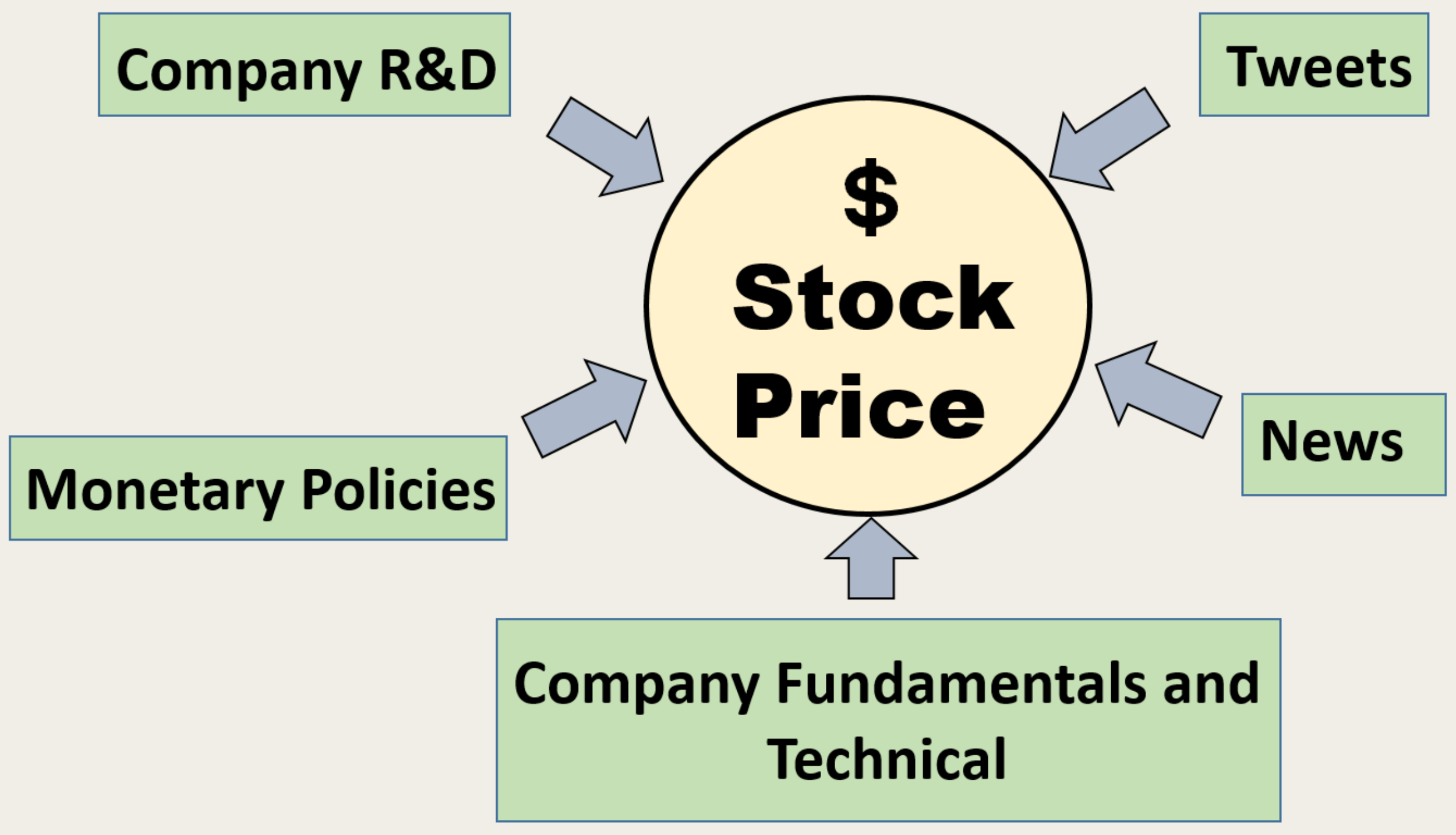

**Figure 1.** Factors affecting Stock Price.

## Dataset

**The Quantopian research platform[1] was used extensively in this project. The dataset is built from stock data between 2018-01-01 to 2020-01-1.**

**The data contains the closing prices, news, and tweets sentiments for all equities listed in the US stock market. The features used include the following. {"Market Capital", "sentiment_score", "Sentdex_lag", "MAD", "Trading signal", "Stock classification", "Sentdex", "returns"} Figure 2 outlines feature engineering. The ground truth buy or sell decisions are based on the returns.**

## Methodology

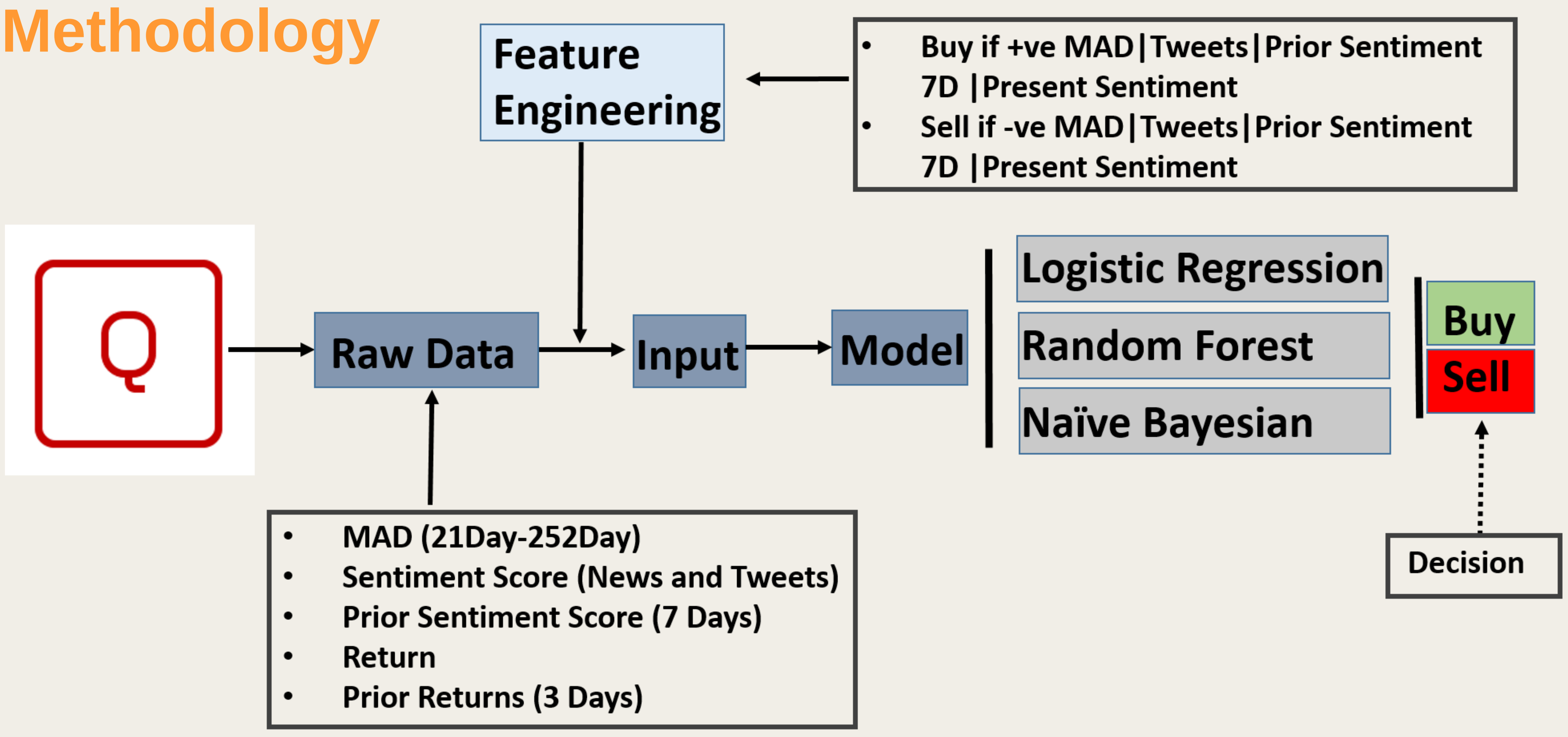
Feature Engineering
Buy if +ve MAD|Tweets|Prior Sentiment 7D |Present Sentiment
Sell if -ve MAD|Tweets|Prior Sentiment 7D |Present Sentiment
Q
Raw Data
Input
Model
Logistic Regression
Random Forest
Naïve Bayesian
Buy
Sell
Decision
MAD (21Day-252Day)
Sentiment Score (News and Tweets)
Prior Sentiment Score (7 Days)
Return
Prior Returns (3 Days)

**Figure 2.** Framework of prediction algorithm pipeline.

## Results

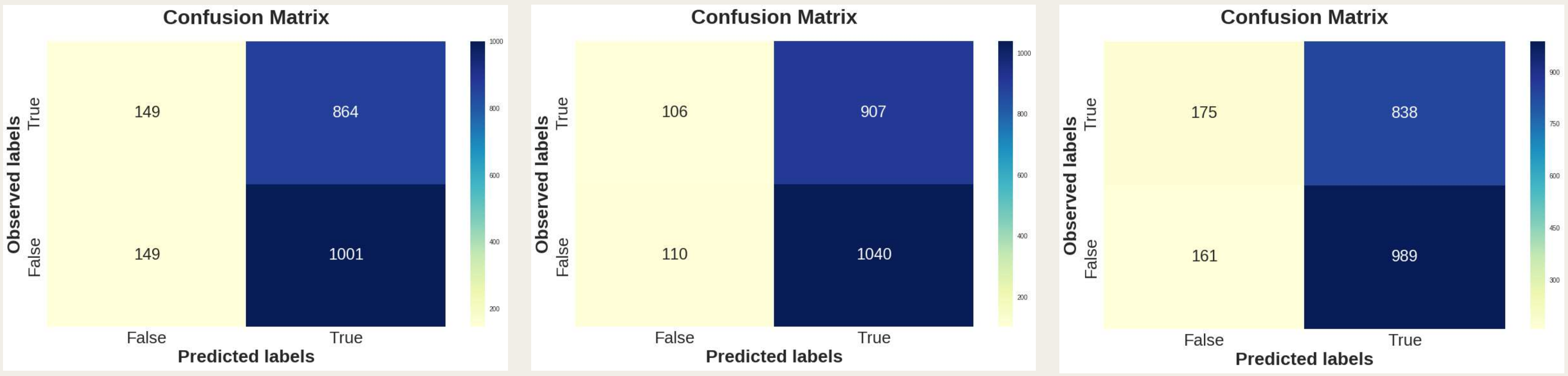

**Figure 3.** Confusion matrixes for Logistic Regression, Random Forest and Naïve Bayesian(L-R).

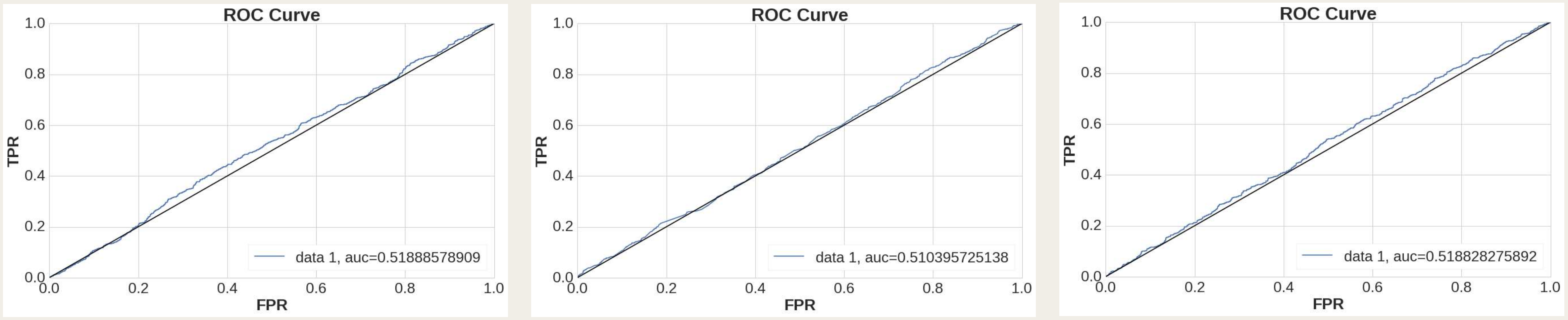

**Figure 4.** ROC curves for Logistic Regression, Random Forest and Naïve Bayesian(L-R).

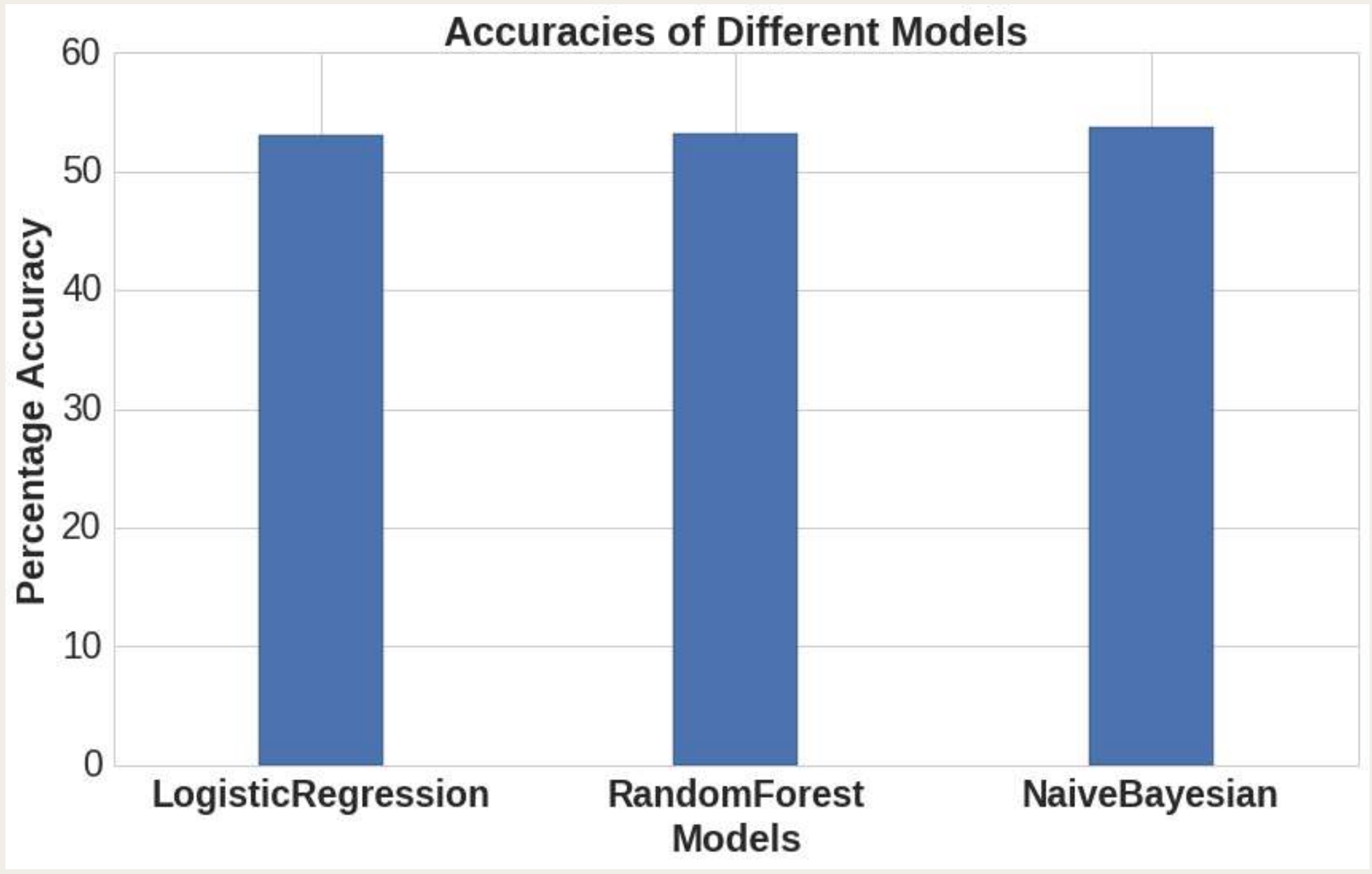

**Figure 5.** Percentage accuracies for the used models.

## Conclusion

- **Logistic regression and random forest models works poorly on unseen data, while Naive Bayesian models work slightly better.**
- **Feature selection is an important aspect of data analysis and other features such as company fundamentals and monetary policies should be investigated to build more rigorous prediction algorithms.**

## Reference

1) www.quantopian.com/research
2) Avramov, Doron, Guy Kaplanski, and Avanidhar Subrahmanyam. "Stock return predictability: New evidence from moving averages of prices and firm fundamentals." Available at SSRN 3111334 (2018).

## Acknowledgement

**I would like express my gratitude to Prof. Taufer for this wonderful opportunity and also thankful for Nigel Tan and Michael Wyatt for the helpful guidance throughout the semester.**

THE UNIVERSITY OF TENNESSEE KNOXVILLE