Predicting IPO Excess Returns: A Sentiment Analysis Approach

Problem Identification Overview:

In this case study, we aimed at predicting IPO excess returns for the years 2019 to 2021 across different time intervals using a range of predictor variables. We were particularly interested in exploring whether a broad sentiment score, derived from Twitter sentiment analysis concerning tech giants Amazon, Apple, Google, Microsoft, and Tesla, can help predict IPO excess returns.

Data Sources:

Our primary data sources are:

University of Florida's IPO Database: This provides us with historical IPO data for the predictor variables, specifically variables such as: whether the IPO was a Rollup, whether it involved multiple share classes, whether it was an internet company, and the year of founding.
Kaggle - Company IPOs (2019 - 2021): We used this dataset to obtain the IPO offer price.
Kaggle - Tweets about the Top NASDAQ Companies (2015 - 2020): This dataset provides the raw tweets for sentiment analysis.
Yahoo Finance API: This API provides us with stock price data necessary for our analysis.

Steps followed:

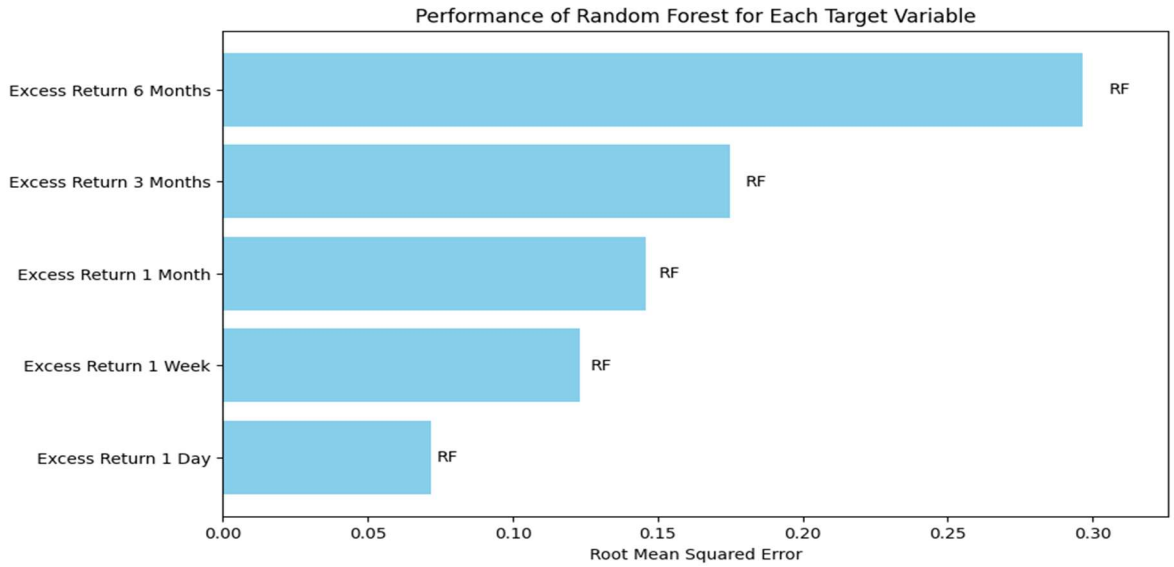
1. Data Cleaning and Pre-processing: We ensured that the data was appropriately cleaned, and missing values were handled.
2. Feature Selection: Selected features were used as predictors, including sentiment analysis data.
3. Model Selection: We utilized Lasso Regression and Random Forest Regressors to predict excess returns.
4. Hyperparameter Tuning: We employed RandomizedSearchCV to find the optimal hyperparameters for each target variable.
5. Model Training and Validation: The optimal models were then trained on the training data and validated on the test data.

Model Performance:

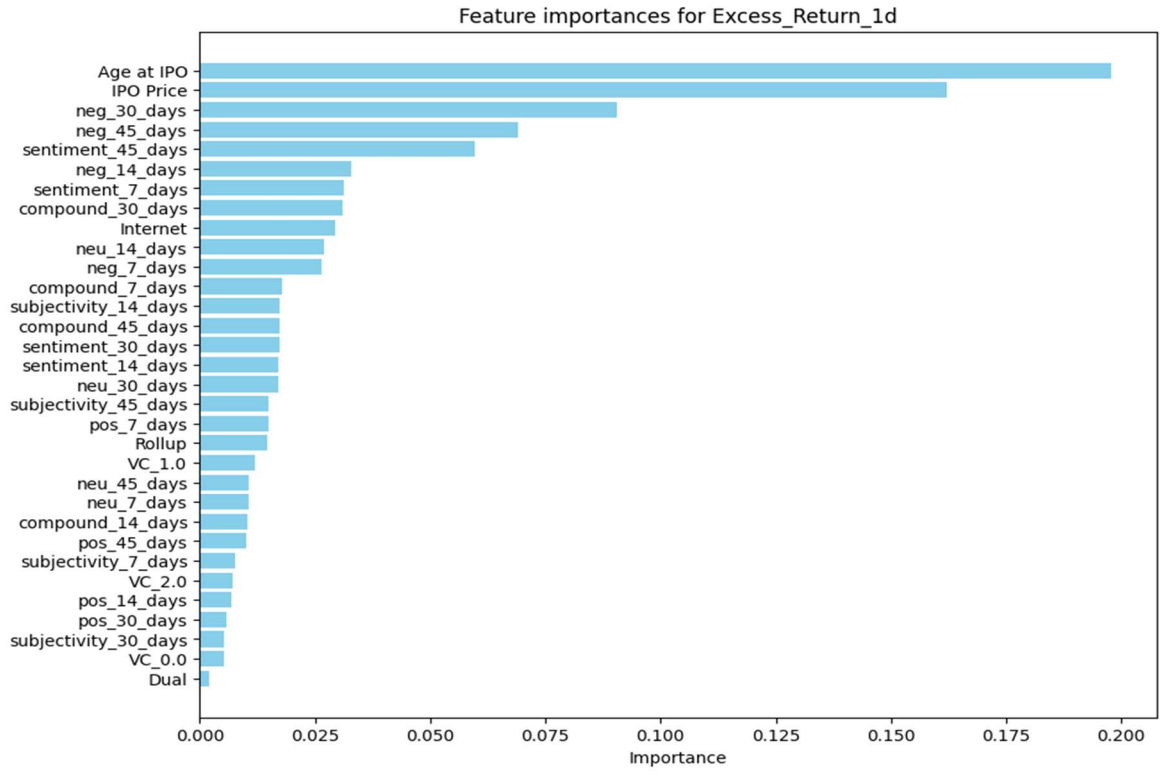
Our analysis revealed that the Random Forest model performed well with varying degrees of accuracy in predicting excess returns for the different time horizons. We observed an increasing RMSE pattern as the time horizon for the excess returns extended. This suggests that predictive accuracy diminishes for longer time horizons due to the inherent volatility of the market and numerous other influencing factors, which complicate long-term predictions.

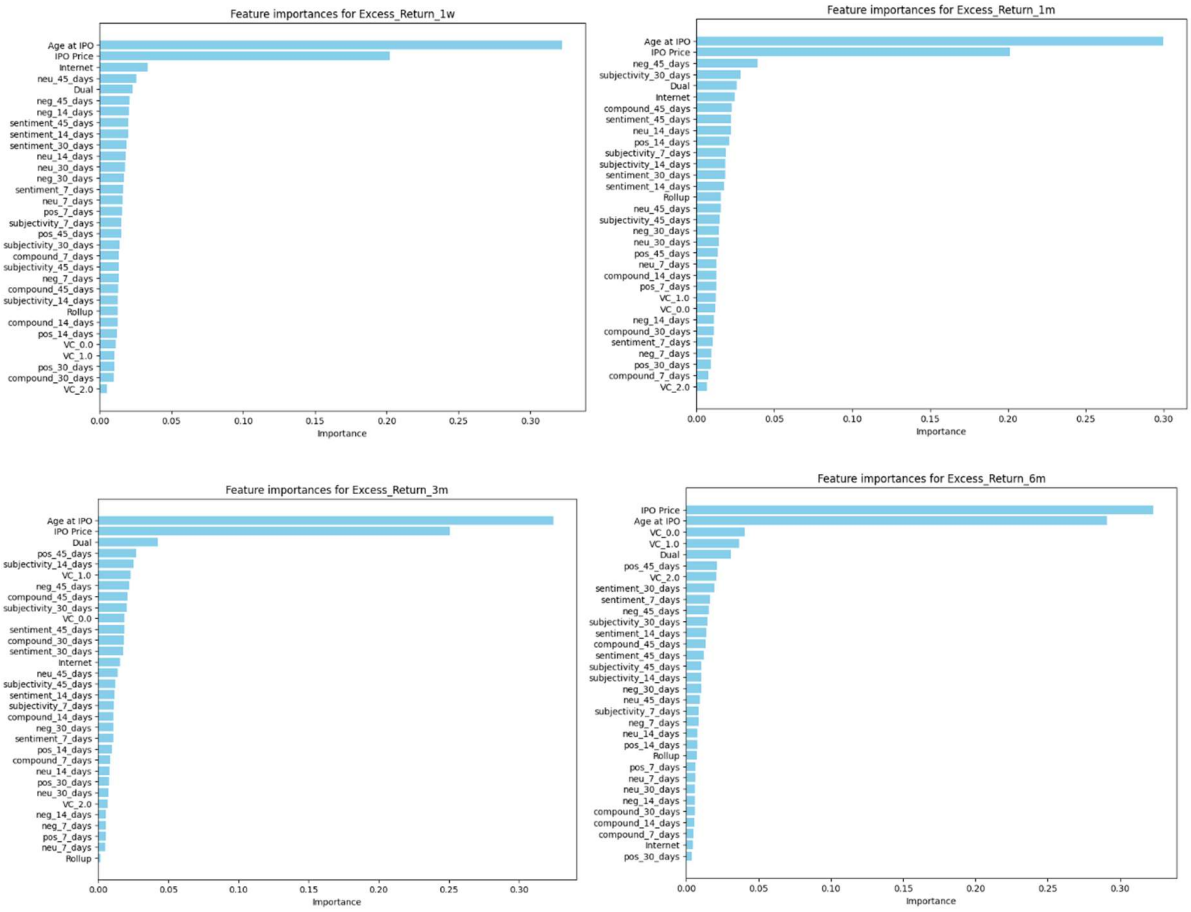
The following are the RMSE values for the optimal models:

Excess Return 1 Day: 0.0716, Excess Return 1 Week: 0.1230, Excess Return 1 Month: 0.1460, Excess Return 3 Months: 0.1750, Excess Return 6 Months: 0.2966.



Feature Importance:





Impact of Sentiment Scores: The sentiment variables derived from TextBlob and VADER provided inconsistent results across different time intervals. They sometimes emerged among the top influential features, while at other times, their impact was relatively lower. For example, sentiment scores from the past 45 days ('sentiment_45_days') were significantly influential in predicting 1-day and 1-week excess returns, but less so for longer periods.

Most Important Features: Most notably, the two dominant features across all models were 'Age at IPO' and 'IPO Price.' This underscores the role the company's characteristics play in influencing IPO performance.

Conclusions:

While we successfully implemented predictive modelling, the results are not considered robust. Overall, it is evident that a mix of traditional financial indicators and contemporary data like sentiment analysis from social media platforms can indeed predict IPO excess returns to a certain degree. However, the accuracy decreases over longer-term horizons, indicating the complexity and dynamism of financial markets.

Future Directions:

Expanding Data Sample: Looking at a larger dataset of IPOs, considering more features, or integrating a larger collection of Twitter data could lead to more insightful findings.

Advanced Sentiment Analysis: Applying more advanced sentiment analysis techniques that consider the context and content of tweets might improve prediction accuracy.

Focus on Short-Term Predictions: Given the diminished accuracy for longer-term horizons, prioritize predictions for shorter periods.

While the models developed in this case study can offer some degree of insight and predictive power, it is important to understand their limitations and use them in conjunction with other tools and information. Future work could consider additional predictors, such as more detailed financial information, broader macroeconomic indicators, and more sophisticated sentiment analysis techniques that take into account the context and source of the sentiment.