Pre-processing & Initial Model Combined Report

Predicting Next-Day S&P 500 Direction Using Price and Volume Data

Group A

Team Lead: Chase Golden

Recorder: Ross Schanck

Spokesperson: Auriana Anderson

December 3, 2025

Background & Question (Recap)

Our study examines whether one can predict short-term fluctuations in the S&P 500 using features created from past price and volume. We will investigate how effective machine learning models will be classifying next day market direction better than random guessing. Our main hypothesis is that rolling returns, volatility indicators, and volume indicators have enough informational content for basic classification to outperform an accuracy of 50%. We also anticipate that Logistic Regression or Random Forest may reach an approximate accuracy rate of between 52% and 60%.

Methods

The Preprocessing and Feature Engineering of our study was guided by the exploratory analysis done last week. Our focus is to develop time-series features based on the patterns in the rolling returns as well as how to impute missing values that result from rolling windows. Next, we will encode the Direction variable as a binary classification variable. Finally, we plan to filter features based on variance, remove features on a correlation basis, and use unsupervised learning approaches to understand more about the underlying structure of the dataset prior to fitting initial models. Decisions made regarding our study were informed by the EDA, particularly the high degree of noise present with all financial data.

Pre-processing

We started with the Historical S&P 500 data and built a target variable indicating if the next day the market closed or went down. As machine learning models need numerical outputs, we converted “Up” to “1” and “Down” to “0” to fix previous errors where a model rejected strings for labels. Creating rolling-window features like percentage changes, rolling means, and volatility results in rows with missing values, typically at the beginning of data, so we dropped those rows to keep data valid. We also applied low variance filtering to drop any feature with very low variance, as those predictors provide little value and can add in more noise when run through models. To reduce multicollinearity, we looked at our correlation matrix and dropped redundant predictors when their correlations were above an acceptable threshold. Once the data was preprocessed, we did a chronological split to avoid any data leakage which is an important process for time series models and performed a standard scale on our predictors, so all predictors were at the same scale.

Feature Engineering & Selection

The main motivation behind our feature-engineering approach was the development of a method that could better capture the short-term patterns that occur within the market's activity that could potentially improve prediction capabilities. We constructed lagged return values, rolling averages, volatility measures, and volume-based indicators for this purpose. The ratios or indicators used within the engineered features reflect well-known financial signals that have been studied extensively. We wanted to better understand the overall structure of the engineered features; therefore, we conducted a PCA which enables us to measure the amount of variance of each of the features and whether a dimensional reduction of them is possible. Through the job done by PCA, we noted that, although many of the predictors displayed a strong overlap, the components provided us clarity in determining which engineered features provided the greatest amount of information. We also attempted to conduct an analysis utilizing K-Means clustering techniques to identify potential cycles that were formed from the engineered features. While the boundaries established through this clustering were not sufficient to utilize them as predictors within a model at this time, the K-Means analysis suggested that the engineered features do contain regime-like patterns, albeit with subtle variations within the short-term time series data. Therefore, we elected not to add cluster labels as additional features to our prediction model, but we will explore the consideration of including them in the feature selection process during model refinement.

Initial Model

As our first model, we chose Logistic Regression because it provides a good, interpretable baseline for binary classification problems and is frequently employed within financial predictive analyses. In addition, Logistic Regression will help us establish whether engineered features can be linearly separated. In order to contrast Logistic Regression's linearity with Random Forest's ability to capture interactions and nonlinear structures, both models will be trained using our scaled predictors and tested on a test set that preserves the chronological order of financial data. However, we will apply cross-validation during model tuning. Assumptions made by Logistic Regression include linearity of log-odds, independence of predictors and absence of perfect multicollinearity. Independence among individual values in financial data is usually violated, but preprocessing steps addressed the most critical assumptions. Random Forest allows for fewer assumptions to be made than Logistic Regression but has a high likelihood of overfitting, which we will explore in depth next week.

Results: Pre-processing & Feature Engineering

The final dataset we created after preprocessing contained 3,763 records after the removal of missing values caused by using rolling windows. We were able to successfully encode the Direction variable into a binary format, which allowed for correct execution of the classification models. The process of filtering and removal of logically associated features resulted in a reduction of redundancy in the dataset. The scaling process ensured that all features had the same range of values. PCA confirmed that there is significant overlap between the engineered features in the dataset, but it did also provide us with an indication of which components were the most informative. K-Means clustering indicated that there are certain groupings of behavior in the market, however these groupings were not sufficiently strong for inclusion as engineered labels now.

Results: Initial Model

The Logistic Regression model accuracy was 0.5259. The model had a high recall of "Up" during the prediction phase. The Logistic Regression model did very well relative to the baseline accuracy but did not perform the best in regard to detecting the direction of the changes in price. The results of the Random Forest model had an accuracy of 0.49 which is close to the Logistic Regression's. With this being the case, it shows that predicting the direction of short-term market movements is challenging and may require additional feature refinement or even more advanced algorithm implementation. This project has only done surface testing for overfitting by performing a comparison of the training and test data's model performances. This leaves us completing a detailed analysis and cross-validation testing of the models in the next phase of the project.

Discussion & Next Steps

One of the things that has stood out this week in the analysis of financial time series data has been the difficulties associated with preparing financial time series data to be utilized with machine learning methods. Feature engineering created several variables that had a significant degree of multicollinearity. Building direction variables also created difficulties in encoding the direction. Clustering and PCA analyses showed us that financial time series datasets have distinct structure and are heavily influenced by noise. The preliminary predictive statistics of the models we have developed have been relatively modest. The results indicate that Logistic Regression performed slightly better than Random Forest; however, neither model came all that close to what we were expecting. As the results show, predicting daily market direction using these methods may require more extensive feature engineering or a more complicated set of models.

Our plan for the next steps will consist of us focusing mainly on our models. We will be trying to expand and refine the modelling pipeline through implementing cross-validation to improve the reliability of results being produced. We will also be evaluating other algorithms such as Gradient Boosting or Regularized Logistic Regression to better tune model performance. The completion of this next phase will help finalize our analysis plan and allow us to determine if short-term market direction can be accurately predicted from this dataset.