Model Evaluation & Interpretation Report

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

Group A

Team Lead: Auriana Anderson

Recorder: Chase Golden

Spokesperson: Ross Schanck

December 8, 2025

The most recent activity in our project has been the evaluation, interpretation, and hyperparameter tuning of models used to predict whether the S&P 500 index will rise or fall on the following day. We used a binary Direction target variable and the same, previously developed Technical Indicators which incorporate Lagged Returns, Rolling Volatility, Moving Averages, Bollinger Bands, RSI, MACD, Momentum, and On-Balance Volume. Our data set is chronologically separated into an 80% - 20% split to avoid using future data for validation. Since our data set is unbalanced, with more UP than DOWN days, we have chosen to use ROC-AUC as our primary scoring method in addition to accuracy.

We evaluated four supervised learning algorithms: Logistic Regression, Random Forest, XGBoost, and LightGBM. To train and optimize all our supervised models, we used TimeSeriesSplit cross-validation with RandomizedSearchCV, and all imputations and scaling for all our models were conducted with sklearn pipelines to avoid data leakage. The class weighing for both Logistic Regression and Random Forest was adjusted and XGBoost used the scale_pos_weight to correct the imbalance within the dataset. The factors we looked at included Tree Depth, Number of Estimators, Learning Rate, and Regularization Terms. Each of the tuned models was tested only once during the holdout period.

The naive baseline method predicted UP every day with 55.4% accuracy. In addition to showing superior performance relative to the naive baseline method, our tuned models did not outperform the daily UP only baseline method in accuracy test scores and ROC-AUC scores

- Logistic Regression: Accuracy: 43.6% /ROC-AUC Score: 0.48
- Random Forest: Accuracy: 44.8% /ROC-AUC Score: 0.48 (best overall)
- XGBoost: Accuracy: 46.8% /ROC-AUC Score: 0.48

- LightGBM: Accuracy: 47.4%/ROC-AUC Score: 0.47-0.48

The Logistic Regression and Random Forests models had limited success improving their overall accuracy; however, they did not outperform the naive UP only model. Moreover, accuracy and ROC-AUC scores near 0.50 indicate that the models were likely to be randomly predicted, and confusion matrices reveal that these models were less effective at accurately predicting DOWN day events than UP day events, even when the classes were balanced.

We have conducted an examination of the different assumptions that constitute our logistic regression, and based on our findings and analysis, we believe that the linear decision boundaries of Logistic Regression models may not adequately represent the complexities associated with our data set. Although we tuned the hyper-parameters of our tree-based models, they did not provide any measurable improvements over time series cross validation results. A combination of cross validation using various periods, hyperparameter tuning, and class weight adjustments limited the occurrence of overfitting. Our analysis of the performance results showing rolling 50-day accuracy illustrates that our model performance fluctuates around the baseline based on the underlying market conditions.

Feature importance across the models utilized in this study have shown that the following features contribute the most in terms of feature importance: Volatility, Lag Returns, SMA Difference, MACD, Momentum. This indicates that our feature engineering techniques employ traditional market signals. The Random Forest Model was identified with ROC-AUC as the most optimal model to use across our daily prediction pipeline utilizing threshold tuning.

In conclusion, although our current evaluation and models have demonstrated a utility in predicting last day movements of the S&P 500, it is still a very difficult process due to the

random nature of movements in the market, as indicated by finance theory. To gain additional predictive power, future research should investigate additional predictors, extend the forecast lengths, and develop models specific to the respective regimes of the markets. Our results indicate that traditional technical indicators provide limited advantages for out of sample daily forecasting