451 Feature Engineering: Programming Assignment 1

Prepared by Albert Lee

July 13, 2025

Background

The new administration has brought quite a bit of instability to the Government Contracting industry – more specifically the contractors that provide Professional Services. The industry has not seen this kind of turmoil ever. The closest incident of this kind of federal budget changes is most likely from federal budget Sequestration in 2013. Because of the volatility seen through the first 6 months of the year, I decided to prepare a model to predict Booz Allen Hamilton's stock.

Booz Allen Hamilton is a public fortune 500 company headquartered in McLean, VA. According to Booz Allen's FY2025 annual investor report, their revenue is at $12.0 billion; making them one of the largest company within this industry. Within their 10-K, Booz Allen generates ~98% of their revenue from the Federal Government with the remaining coming from their Commercial and Other business.

Not only has Booz Allen Hamilton seen contract cuts from DOGE./ new administration cost-cutting efforts, but it's also been a frequent target by the administration. Their CEO recently did an interview with Fortune Magazine discussing how they're adapting and how it's trying to be resilient. The stock though has taken a significant hit. The stock is down ~16% YTD which at one point was at a high of $186 on close.

Dataset and Feature Engineering

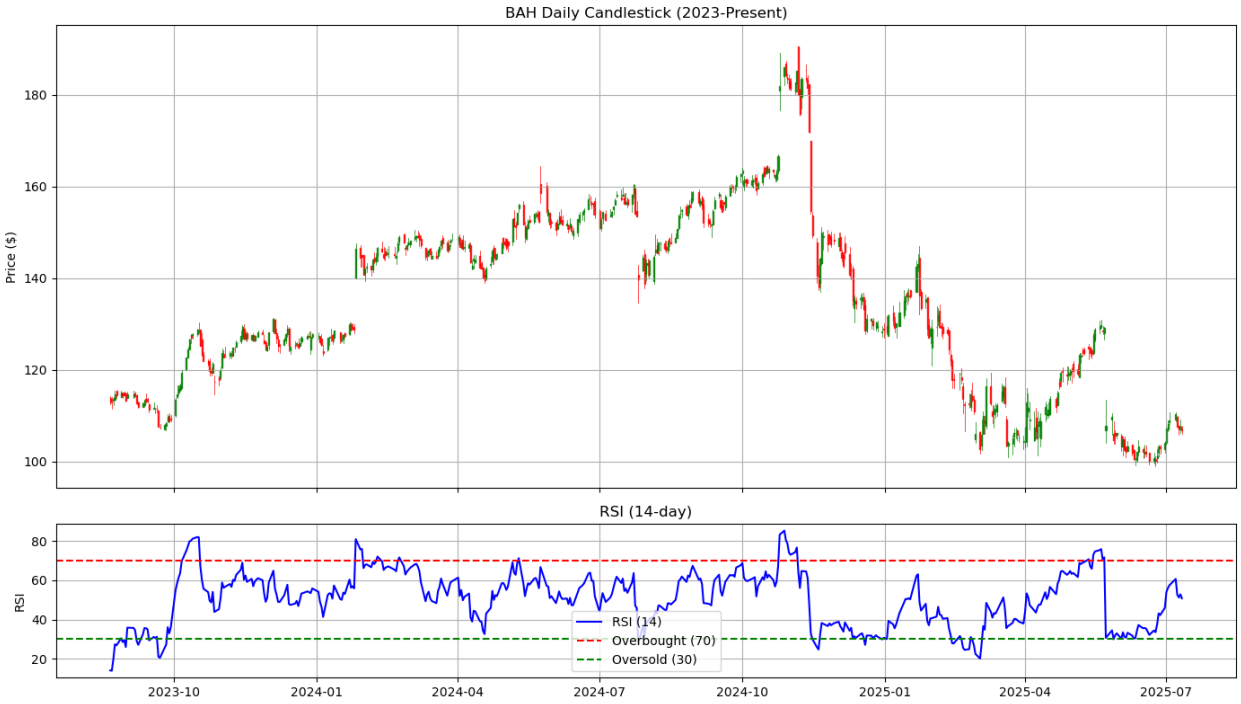
To build a predictive model, the first step was to bring in historical information on the stock. The time range was from 01/01/2023 – 06/30/2025. The data was connected via a live API using polygon.io. Polygon provides market data for all but the enhanced features require a paid subscription. The free version is also limited to only 2 years of historical data. I decided to use Polygon because in most financial stock predictions, you would want to have as much real-time data as possible. Having the API allows for that flexibility and prevents the strict limitations of yfinance.

The ancillary data for companies within this industry is wide including the Federal Procurement Data System (FPDS). FPDS provides historical contract award data given to companies and across the US Government. There are also other companies within the industry that could have been used as potential indicator to Booz Allen's stock price.

For this assignment, I just focused on the stock price. Before diving into feature engineering, the first step was to do some exploratory data analysis. One of the key financial indicators and

marquee chart is the candlestick market chart for a stock. The candlestick chart is a powerful visualization that help users understand the bullish and bearish patterns along the stock price movement. It displays the price movement of a stock, fund, or currently over time. The top chart indicates the price movement of Booz Allen over the last 2 years.

Right below the candlestick chart is another helpful indicator for stock performance. The Relative Strength Index (RSI) is a momentum indicator to detect whether there are overbought or oversold conditions in the price of the security. When the RSI is above 70 then it's overbought whereas if it's under 30, that means it's oversold. The RSI trends well with the candlestick chart above and helps provide potential indications of how a stock might perform. The average RSI is 51.54.



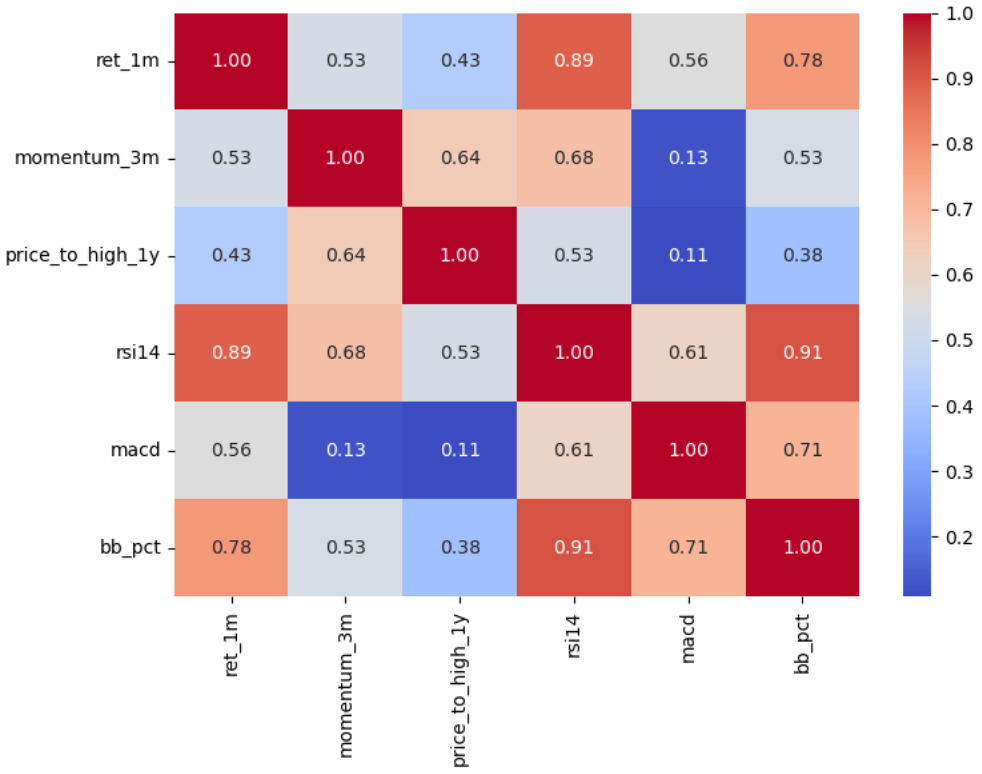
For Feature Engineering, we're focused on close prices, momentum, returns, volatility, price positioning, oscillator (or RSI), trend indicators, and Bollinger band percentile.

Close Prices	Stock market closing price
Momentum	The speed or price changes in stock
Returns	Gain / loss on investments
Volatility	How varied the returns are for a stock
Price Positioning	Short and Long positions
RSI	Detection for overbought or oversold conditions
MCAD	Trend-following momentum indicator that shows the relationship between two moving averages of a asset price
Bollinger Band Percentile	Determines where the stock price is relative to Bollinger Band

Next, I needed to create target variable engineering by creating the classification labels for the model to try and predict. Since daily predictions is very volatile and I only had 2 years of data, I went with a 5-day return prediction. The classification for the model is as follows:

1 = Price will **go up**
0 = Price will **go down or stay flat**

For speed and simplicity in the Feature Engineering and Selection process, I decided to use the filtering method. The filter method focused on low-variance threshold and high correlation filters provide speed and simplicity to the model. There were also a small set of technical indicators used in the feature engineering which didn't require complex methods such as Wrapper or Embedded. Once we removed the low-variance (or near-zero) thresholds and highly correlated features, I was left with the below correlation matrix.



The ones with the strongest correlations were rs14 and bb_pct at 91% and rs14 and ret_1m at 89%.

Model – XGBoost

Now that the feature engineering and selection is done, I can now start creating the model. The first step was to prepare the training and testing datasets for time series classification, while being able to address class imbalance.

The model is using XGBoost classification method to try and predict whether the stock returns are going to go up or down / stay flat. In the untuned XGBoost model, I'm getting the following results:

```
Test Accuracy: 0.7291666666666666
              precision    recall  f1-score   support

           0      0.882     0.577     0.698        26
           1      0.645     0.909     0.755        22

    accuracy                         0.729        48
   macro avg      0.764     0.743     0.726        48
weighted avg      0.774     0.729     0.724        48
```

While a ~73% is strong, I wanted to do better and perform hyperparameter tuning on the XGBoost model. For the hyperparameter tuning, I used a time-series cross-validation and randomized search cross-validation. The parameter grid for the new model was as follows:

```
    'n_estimators': [200, 400, 600, 800, 1000],
    'max_depth': [3, 4, 5, 6, 8],
    'learning_rate': [0.01, 0.05, 0.1, 0.2],
    'subsample': [0.6, 0.8, 1.0],
    'colsample_bytree': [0.6, 0.8, 1.0],
    'gamma': [0, 0.1, 0.2],
    'min_child_weight': [1, 3, 5]
```

The randomized search cross-validation then samples 40 from the parameter grid scoring each using accuracy across 5 time-based folds. The best model was then selected and came out with a stronger accuracy score of ~75%. This is an increase of ~2%. Although it may not seem like a lot, it's still helpful to try and improve the accuracy of any model.

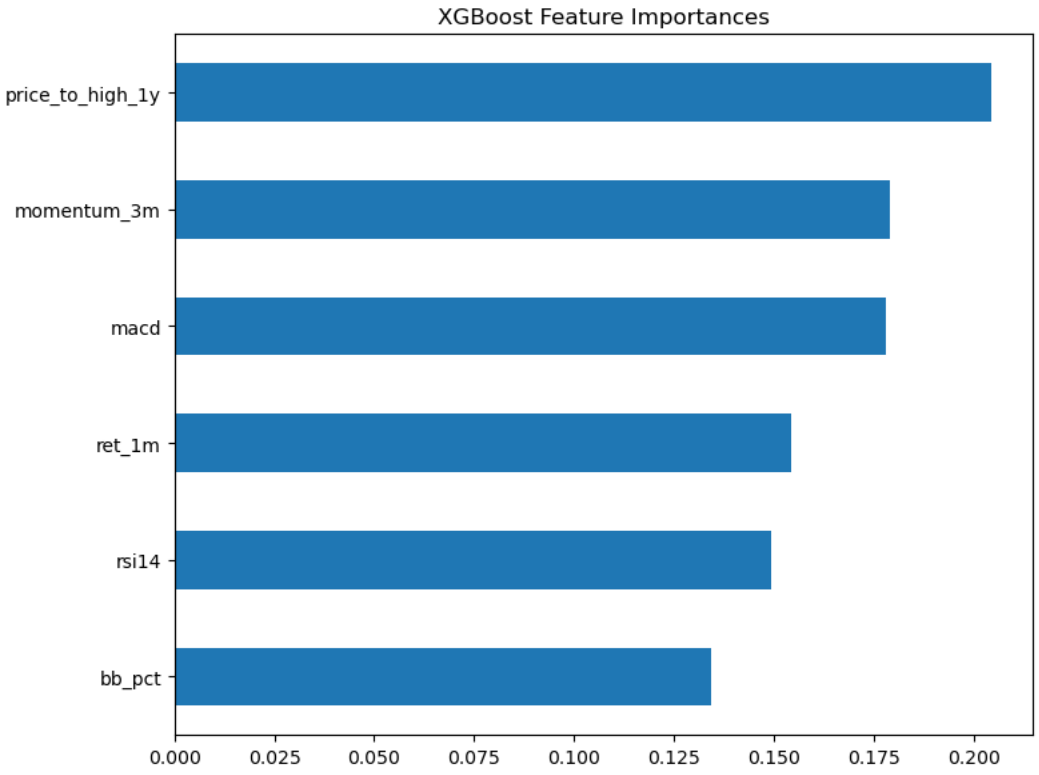
```
Test Accuracy (XGB tuned): 0.75
              precision    recall  f1-score   support

           0      0.938     0.577     0.714        26
           1      0.656     0.955     0.778        22

    accuracy                         0.750        48
   macro avg      0.797     0.766     0.746        48
weighted avg      0.809     0.750     0.743        48
```

To evaluate the model prediction accuracy, a confusion matrix was used. The confusion matrix resulted in the following:

Confusion Matrix		**Predictions**	
		Sell	**Buy**
Actual	**Sell**	15	11
	Buy	1	21

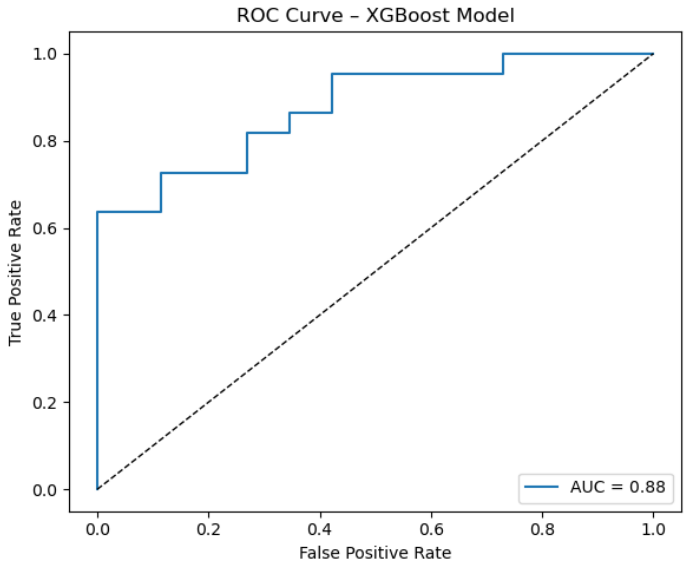
Accuracy: (15 + 21) / (15 + 11 + 1 + 21) = 75%
Precision (Buy): 21 / (21 + 11) = 65.6%
Recall (Buy): 21 / (21 + 1): 21 / (21 + 1) = 95.5%

This shows that the model is good at identifying buying opportunities but it may struggle more with false positives. The model across a single week (5-day period) across 6 features with an accuracy score of 75% is still a win. The most important feature in the model was the price_to_high_1y.



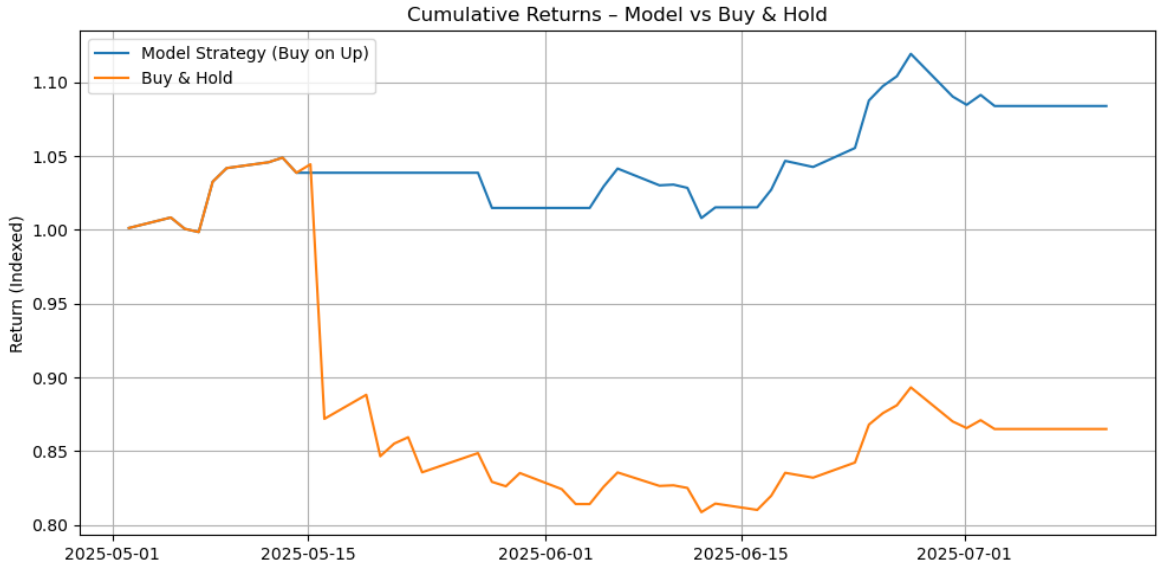
Model Evaluation

I performed one more additional model evaluation and then compared that to how the stock would perform if we had used this model to predict vs just a buy and hold strategy. A ROC was calculated and the AUC (area under curve) was 88%. The 88% is strong and the model will most often then not choose “buy” instances higher than “sell”.



To accurately determine the model, I always like to implement the model back to historical to see how it would perform. To do that, I created a "Model Strategy" which represents returns from investing only when the model predicts an upward move. This results in a 1.10 cumulative return of ~10%. The "buy & hold" strategy buying the Booz Allen Hamilton stock and holding it throughout the same time period. This resulted in a cumulative return of 0.86 or -14%. That is ~24% swing and when you're talking about stocks and potential financial portfolios holding millions if not billions of dollars, this is significant. The sharpe ratio for the model strategy is 2.56 while the sharpe ratio for the buy and hold strategy is -1.47. The 2.56 sharpe ratio indicates excellent returns per unit of risk.

Although Booz Allen Hamilton has been hammered by the recent administration, based on the model, it still seems like the stock is still a potential buy and may be undervalued.



References

0001443646-25-000076: 10-K. Booz Allen Hamilton. (n.d.). https://investors.boozallen.com/sec-filings/sec-filing/10-k/0001443646-25-000076

About. Polygon. (n.d.). https://polygon.io/about

Booz Allen Hamilton Holding Corporation (BAH) stock price, news, Quote & History - Yahoo Finance. (n.d.). https://finance.yahoo.com/quote/BAH/

Brady, D. (2025, July 9). *Booz Allen Hamilton may have been a doge target-but its CEO is still bullish on his biggest client*. Fortune. https://fortune.com/article/booz-allen-hamilton-ceo-horacio-rozanski-interview-us-china-ai-quantum-doge-cuts/

Corporation. (2025, April 28). About Us. https://www.boozallen.com/about.html

Dolan, B. (2024, September 16). *What is MACD?*. Investopedia. https://www.investopedia.com/terms/m/macd.asp#:~:text=MACD%20is%20often%20displayed%20with,generate%20overbought/oversold%20trade%20signals

Dundas, R. (2022, May 5). *Calculate relative strength index (RSI) and chart with candles using python, pandas and Matplotlib*. Medium. https://rbdundas.medium.com/calculate-relative-strength-index-rsi-and-chart-with-candles-using-python-pandas-and-matplotlib-f58d926249ac

Federal Procurement Data System. fpds.gov. (n.d.). https://www.fpds.gov/fpdsng_cms/index.php/en/

Fernando, J. (2024, November 19). *Relative strength index (RSI) indicator explained with formula*. Investopedia. https://www.investopedia.com/terms/r/rsi.asp

Filter methods. Codecademy. (n.d.). https://www.codecademy.com/article/fe-filter-methods

Hayes, A. (2024, May 31). *Position definition-short and long positions in financial markets*. Investopedia. https://www.investopedia.com/terms/p/position.asp

Hayes, A. (2025a, May 11). *Volatility: Meaning in finance and how it works with stocks*. Investopedia. https://www.investopedia.com/terms/v/volatility.asp

Hayes, A. (2025b, June 24). *What are returns in investing, and how are they measured?*. Investopedia. https://www.investopedia.com/terms/r/return.asp

Office, U. S. G. A. (2014, March 6). *2013 sequestration: Agencies reduced some services and investments, while taking certain actions to mitigate effects*. 2013 Sequestration: Agencies

Reduced Some Services and Investments, While Taking Certain Actions to Mitigate Effects | U.S. GAO. https://www.gao.gov/products/gao-14-244

Omarzai, F. (2025, March 25). *XGBoost classification in depth*. Medium. https://medium.com/@fraidoonomarzai99/xgboost-classification-in-depth-979f11ef4bf9

Team, T. I. (2024, October 18). *The basics of bollinger bands®*. Investopedia. https://www.investopedia.com/articles/technical/102201.asp

Team, T. I. (2025, June 1). *Momentum indicates stock price strength*. Investopedia. https://www.investopedia.com/articles/technical/081501.asp

Thompson, C. (2025, March 12). *Understanding basic candlestick charts*. Investopedia. https://www.investopedia.com/trading/candlestick-charting-what-is-it/