

Angelica Hussar  
MSDS 451 Financial Engineering  
Programming Assignment 1  
July 8, 2025

## **Barnes and Noble Stock Direction Prediction**

### **Introduction**

This assignment aims to predict whether Barnes & Noble Education stock will go up or down on the next trading day using machine learning. I want to answer the question of: can we find patterns in historical stock prices that help predict future price movements? This is valuable because successful prediction could help with trading decisions and portfolio management. This is a challenge because stock markets are noisy and unpredictable. I am trying to test whether machine learning can identify these patterns in BNED stock and make accurate predictions about daily stock direction.

### **Data Preparation and Pipeline**

I collected daily BNED stock data from Yahoo Finance from January 2020 to July 2025, giving me 1,381 trading days with opening prices, daily highs and lows, closing prices, and trading volume. Following Professor Miller's framework, I created 15 features from this raw data:

- Three price lag features - yesterday's prices, two days ago, three days ago
- Three volatility measures - daily high minus low for past three days
- Three gap features - open minus close for past three days
- Three volume features - trading volume for past three days
- Three trend indicators - moving averages over different time periods

The target variable is: if the stock price increased from the previous day, I label it as "up" (1), and if it decreased or stayed the same, I label it as "down" (0).

I cleaned the data by removing any rows with missing values. This mostly occurred in the first few days where I couldn't calculate lagged features.

### **Research Design**

My approach used AIC to select the best features from my initial set of 15 features. AIC helps find the right balance between having a model that fits the data well and keeping it simple enough to avoid overfitting (Scikit-learn Developers). I tested all possible combinations of features and the AIC process selected just 2 features: OMCLag1 and VolumeLag2. This demonstrates that a parsimonious model with carefully selected features can be more effective than using all available features.

For the prediction model, I chose XGBoost because it works well with financial data and can handle complex relationships between features and outcomes (XGBoost Developers). I also tried to optimize the model by testing 100 different parameter combinations to find the settings

that give the best performance. The hyperparameter tuning process improved cross-validation performance by 0.009, validating the importance of model optimization.

## **Programming**

Key tools include yfinance for getting stock data, Polars for fast data processing, XGBoost for machine learning, and matplotlib for creating visualizations. The feature engineering pipeline efficiently creates all 15 features using data operations.

The implementation includes functions for calculating AIC, testing of all feature combinations, and hyperparameter optimization using randomized search. I created comprehensive visualizations showing feature importance, model performance curves, confusion matrices, and data distributions.

## **Conclusion**

The analysis demonstrates that machine learning can find predictable patterns in BNED stock data. The AIC feature selection identified 2 optimal features from 15 candidates: OMCLag1 (importance: 0.489) representing overnight gap effects, and VolumeLag2 (importance: 0.511) capturing delayed volume patterns (Hariom Tatsat et al., 2020). The XGBoost model achieved a training accuracy of 95.7% and an AUC score of 0.994, showing the ability to distinguish between up and down days. However, the cross-validation accuracy of 52.5% suggests overfitting.

The substantial gap between training performance (95.7%) and cross-validation performance (52.5%) indicates that while the model learns training patterns very well, it struggles to generalize unseen data (Hariom Tatsat et al., 2020). This overfitting is a limitation that suggests the model may not perform well in real trading situations (Hariom Tatsat et al., 2020). Future improvements should focus on expanding the dataset and incorporating additional data sources like market sentiment.

## References

Hariom Tatsat, Puri, S., & Lookabaugh, B. (2020). *Machine Learning and Data Science Blueprints for Finance*. O'Reilly Media.

Ross, S. M. (2011). *An Elementary Introduction to Mathematical Finance*. Cambridge University Press.

Scikit-learn Developers. (2025). *Lasso model selection: AIC-BIC / cross-validation*.

Scikit-Learn.

[https://scikit-learn.org/stable/auto\\_examples/linear\\_model/plot\\_lasso\\_model\\_selection.html](https://scikit-learn.org/stable/auto_examples/linear_model/plot_lasso_model_selection.html)

XGBoost Developers. (2022). *XGBoost Documentation — xgboost 1.5.1 documentation*.

Xgboost.readthedocs.io. <https://xgboost.readthedocs.io/en/stable/>