MSDS 451: Assignment 1
Forecasting Stock Direction and Price for Constellation Brands (STZ) Using XGBoost

Ben Caldwell

July 11, 2025

Abstract

This project applies financial machine learning techniques to forecast the daily direction and price of Constellation Brands (NYSE: STZ) stock. Using historical price data from Yahoo Finance, I engineered time-series lag features and popular technical indicators, then trained an XGBoost model using time-series-aware cross-validation. Two models - a baseline with lagged features and an extended version with technical indicators - were trained and evaluated using directional accuracy and mean absolute error. The model was also tested in a rolling backtest window to simulate real-world deployment. While directional accuracy in backtesting hovered around 45%, the extended model showed promise with cross-validated accuracy exceeding 67%.

1 Introduction

Stock price forecasting is a foundational problem in quantitative finance. Although markets are not entirely predictable, machine learning models can uncover short-term patterns in price movements, especially when aided by engineered features from historical data. This project seeks to forecast the next-day direction (up/down) and implied price for STZ stock using a classification model and probability-based return estimation.

XGBoost, an efficient implementation of gradient-boosted decision trees, was selected due to its ability to model non-linear relationships and handle noisy financial data. All modeling was conducted in Python, leveraging Polars for data manipulation and Scikit-learn for cross-validation.

2 Data and Feature Engineering

Daily OHLCV (Open, High, Low, Close, Volume) data for STZ was downloaded from 2000 to May 2025. Several engineered features were created, grouped as follows:

Lagged Price and Volume Features

- CloseLag1, CloseLag2, CloseLag3: Prior days' closing prices
- HMLLag1-3: High minus Low, lagged
- OMCLag1-3: Open minus Close, lagged
- VolumeLag1-3: Volume, lagged
- EMA2, EMA4, EMA8: Exponentially weighted moving averages

Technical Indicators

For the extended model, additional indicators were added:

- RSI (14): Relative Strength Index
- MACD, MACD Signal Line: Trend-following momentum
- Bollinger Bands (20): Upper and lower price volatility bands
- OBV: On-Balance Volume

3 Modeling Strategy

The classification target was binary: 1 if the next-day close was higher than the prior day, 0 otherwise. A log return was also computed to estimate expected price changes.

Cross-Validation and Hyperparameter Tuning

To respect temporal ordering, TimeSeriesSplit was used with a gap of 10 to prevent lookahead bias. A randomized grid search across 50 configurations of XGBoost hyperparameters was performed. The following were tuned:

- `max_depth`, `min_child_weight`
- `subsample`, `learning_rate`
- `n_estimators`

4 Evaluation and Backtesting

Metrics

Two primary metrics were used:

- Directional Accuracy: % of times the model correctly predicted up/down movement
- Mean Absolute Error (MAE): Dollar error between predicted and actual next-day prices

Single-Day Prediction

Using the final model, the last known close of $184.42 was forecasted to decline, with an expected return of -1.24% and a predicted price of $182.14.

Rolling Backtest

A 60-day rolling window simulated realistic forecasting. For each day, the model was retrained using all prior data, then used to predict the next day's movement. Results:

- Directional Accuracy: 45.00%
- MAE: $21.65

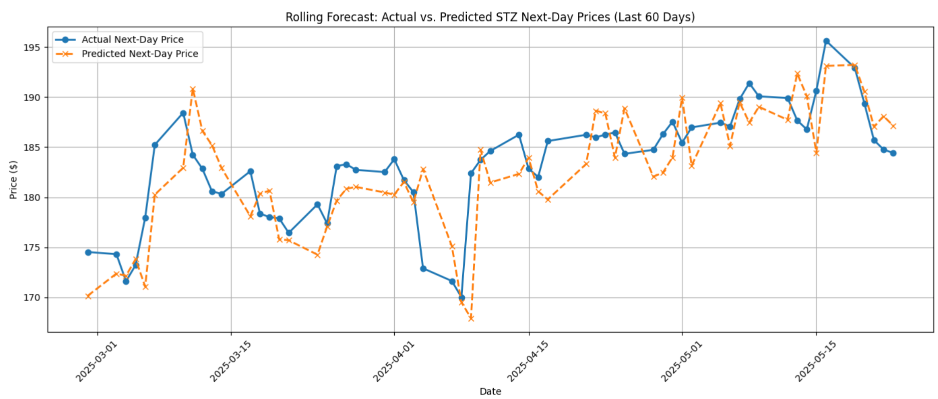

Figure 1: Rolling Backtest: Actual vs. Predicted STZ Prices (Last 60 Days)

5 Extended Model Results

After adding technical indicators, the model's cross-validated accuracy improved substantially to 67.28%. The most recent forecast predicted a sharp drop to $154.20, suggesting the model captured bearish sentiment through technical signals like falling MACD and overbought RSI.

6 Discussion

While the rolling backtest directional accuracy is close to random, the extended model showed strong promise. The lag between cross-validation and real-world performance is expected due to the non-stationary nature of financial markets. However, the incorporation of volume- and volatility-based indicators notably improved predictive power.

Despite some overfitting risks and the inherent randomness in market data, the project showcases how structured time-series features, probabilistic forecasting, and rolling evaluation can together inform a tactical model.

7 Conclusion and Future Work

This project applied XGBoost to short-term stock forecasting for STZ, with results suggesting modest predictive power. Future directions could include:

- Incorporating macroeconomic variables (e.g., interest rates, CPI)
- Testing regime-switching models (e.g., Hidden Markov Models)
- Adding ensemble techniques or stacking with linear/logistic models

Additionally, improving the robustness of predictions via calibrated probabilities or bootstrapped confidence intervals may help translate model output into practical trading signals.