Quantitative Machine Learning Pipeline for Predicting Stock Outperformance

Author: Justin Upson
Date: 23/07/2025

---

1. Objective

This project explores whether machine learning can help predict which stocks are likely to outperform the short-term risk-free rate over a 21-day window. The idea was to build a basic alpha model — similar to what's used in quantitative trading — using historical price data and technical indicators. The aim is to find patterns that might suggest outperformance, using a practical and modular pipeline.

---

2. Data Overview

Stock data was pulled from Yahoo Finance, covering daily open, high, low, close, volume (OHLCV) data for a group of tickers. The 13-week Treasury Bill index (^IRX) was used as a stand-in for the risk-free rate. The dataset spans several years to allow for meaningful signal extraction and testing.

The project follows a clear structure:

- **Data Ingestion:** Historical prices are downloaded using yfinance.
- **Storage:** Files are saved as CSVs in folders: raw/, features/, and labelled/. This setup makes it easy to rerun or update any part of the pipeline.

---

3. Feature Engineering

To give the model enough useful information, a wide range of features was created using the ta library and custom functions. These include:

- **Returns:** 1-day, 5-day, and 21-day percent and log returns
- **Volatility:** Rolling standard deviation over 5 and 21 days
- **Momentum & Trend:** Ratios of current price to past prices, moving averages (MA10, MA21), MACD and signal line
- **Relative Strength:** RSI (14), 5-day rank, and how close price is to its 52-week high/low
- **Bollinger Bands:** Upper/lower/mid band values, and bandwidth

To clean up the data and reduce noise, features with very low correlation (|corr| < 0.05) to the forward return were removed, and highly correlated features (correlation > 0.9) were pruned to avoid redundancy.

---

4. Labelling Strategy

The idea was to predict whether a stock would beat the risk-free rate over the next 21 days. Labels were assigned like this:

- **1:** Stock outperforms the 21-day compounded risk-free return
- **0:** Stock underperforms or performs equally

This creates a binary classification problem that mirrors a simple long-only stock selection setup.

---

5. Modelling and Evaluation

The model used is an XGBoost Classifier, chosen because it tends to work well with structured/tabular data and handles class imbalance internally.

- **Cross-Validation:** A 5-fold TimeSeriesSplit was used to keep the temporal order of data intact.
- **Hyperparameter Tuning:** Used RandomizedSearchCV to search for the best model settings.

Metrics Reported:

- Accuracy
- Precision, Recall, F1-score
- Confusion Matrix
- Top 10 features by importance (based on model gain)

These help evaluate how well the model predicts outperformance, and which features actually mattered.

---

6. Results & Takeaways

- After running the full pipeline and training the model, several points stood out:
- Most raw price or trend indicators had low (or even negative) correlation with future returns.

- However, a few features like price_52w_low, vol_21d, and volume showed stronger positive correlations, hinting that volatility or price rebounds from lows might matter.
- Technical indicators like bb_bbl, ma_10, and ma_21 had slight negative correlations, possibly due to mean-reverting behaviour.
- The XGBoost model handled the data well. It clearly separated positive vs. negative cases better than chance, and feature importance plots confirmed that only a few features were doing most of the work.
- The final model reached **63% accuracy**, with a **precision of 0.67** and **recall of 0.60** on the positive (outperformance) class.

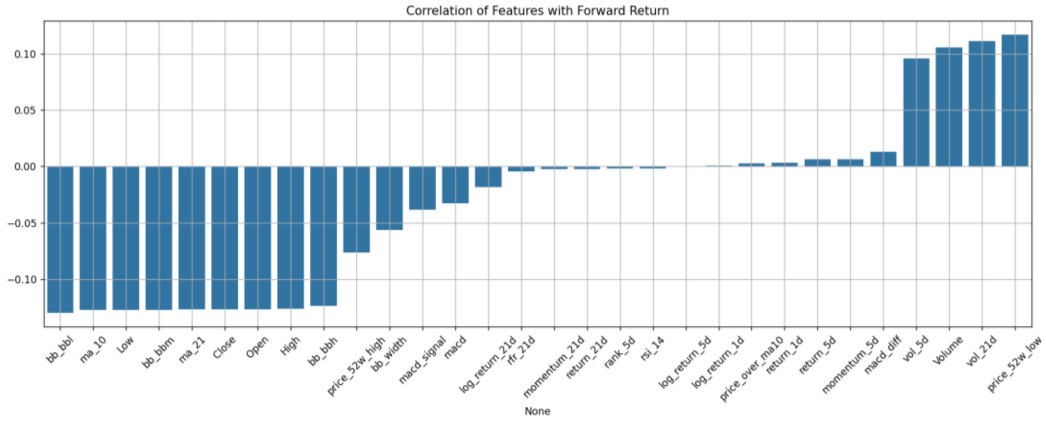

Figure 1: Correlation of Engineered Features with 21-Day Forward Return

7. Example Outputs

The pipeline generated several helpful visual and numerical outputs:

- **Correlation Bar Plot**
  Shows which features had useful correlation with future returns.
- **Feature Importance (XGBoost)**
  Ranks the top 10 signals that the model relied on most.
- **Classification Report**
  Lists precision, recall, and F1 scores for both labels.
- **Confusion Matrix**
  Helps spot class imbalance or common misclassifications.
- **Model Accuracy**
  Reported across CV folds and on the test set (if held out), giving an idea of generalization.

8. Conclusion

This project builds a working machine learning pipeline for predicting stock outperformance using technical features. It’s modular and can be extended to include more complex models, different types of data (like fundamentals or news), or even deployed in a live setting.

It’s a good starting point for experimenting with systematic trading ideas, especially for those exploring the intersection of data science and finance.