Machine Learning Based Algorithmic Trading

Md Kauser Ahmmed

March 9, 2025

1 Data Collection

We collect historical data of eleven assets (e.g., stocks, ETFs). The following table shows the assets and associated data parameters:

Asset Name	**Time Period**	**Frequency**	**Source of Data**
TSLA	2 years	1 min	Alpha Vantage
AAPL	Jan 2023 – Dec 2024		
NVDA			
AMD			
AMZN			
MSFT			
NFLX			
XOM			
META			
NKE			
S&P 500			

Table 1: Asset Data Parameters

Cluster (Sector)	Assets	Rationale
Technology	AAPL, AMD, MSFT, NVDA	Companies in semiconductors, consumer electronics, and software.
Communication Services	META, NFLX	Focus on digital media, social networking, and streaming services.
Consumer Discretionary	AMZN, TSLA, NKE	Industries tied to consumer spending: retail, automotive, and apparel.
Energy	XOM	Major oil & gas company with distinct sector dynamics.
Diversified ETF	SPY	Tracks the S&P 500, representing the overall market.

Table 2: Clusters of Assets Based on Sectors

2 Observation Exclusions

Only one exclusion principle is applied upto now. Work in progress for additional exclusion logics.

Table 3: Summary of Observation Exclusion Techniques

Step	Description
Time-Based Exclusion	Rows with timestamps outside the range `[10, 15]` (10 AM to 3 PM) are excluded.

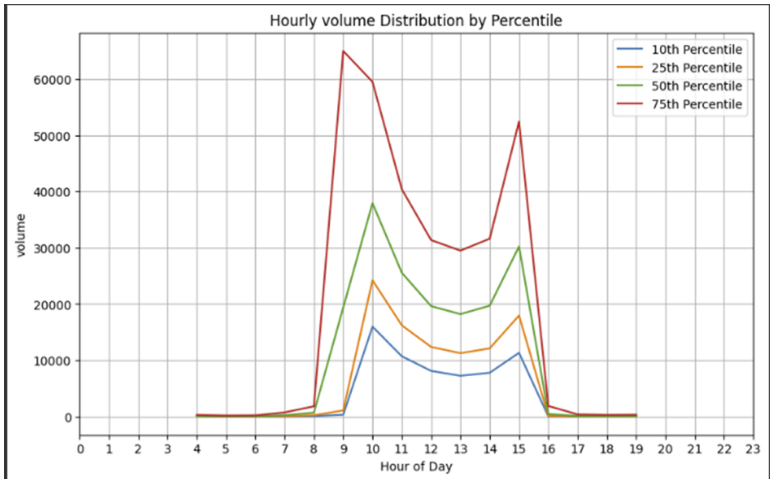

Figure 1: Exclusion Criteria Based on Hour of Day. Data outside the desired trading hours are excluded. (< 10 AM or > 3 PM)

3 Feature Generation

Feature Types

The features are categorized into two types:

- **Binary Features**: These are indicator variables (e.g., flags for crossovers, thresholds) that take on values of 0 or 1.
- **Continuous Features**: These are numerical variables (e.g., moving averages, volatility measures, RSI) that capture quantitative aspects of the data.

Feature Category	Period	Feature Type
Binary_SMA_Crossover	Short|Long: 120|240, 240|480, ...	binary
Price_Above_VWAP	120, 240, 480, 960, 1920	binary
MACD_Bullish	fast|slow|signal: 120|240|90, 240|480|120 ...	binary
RSI	120, 240, 480, 960, 1920	continuous
RSI_Threshold	lower|upper: 15|85, 20|80, 25|75, 30|70, 35|65	binary
BB_Breakout	120, 240, 480, 960, 1920	binary
SMA_cont	120, 240, 480, 960, 1920	continuous
ATR	120, 240, 480, 960, 1920	continuous
Momentum	120, 240, 480, 960, 1920	continuous
Average_Return	120, 240, 480, 960, 1920	continuous
Rate_Close_Greater_Open	120, 240, 480, 960, 1920	continuous
Downside_Deviation	120, 240, 480, 960, 1920	continuous
Sortino_Ratio	120, 240, 480, 960, 1920	continuous
Max_Close	120, 240, 480, 960, 1920	continuous
Min_Close	120, 240, 480, 960, 1920	continuous
Open		continuous
Close		continuous
High		continuous
Low		continuous
Volume		continuous
Total Number of Features	103	

Table 4: Feature Categories with Periods and Feature Types

4 Train–Test Split

- Train: 70%
- Test1: 15%
- Test2: 15%

5 Feature Reduction

Use XGBoost to get feature importance. First used a default xgboost model, filter feature importance with a threshold of 0.0145. Then filter a XGBoost model with parameter and same threshold. Then from the combined feature importance, get the significant features in both of the filtered important features. Rest of the features are removed.

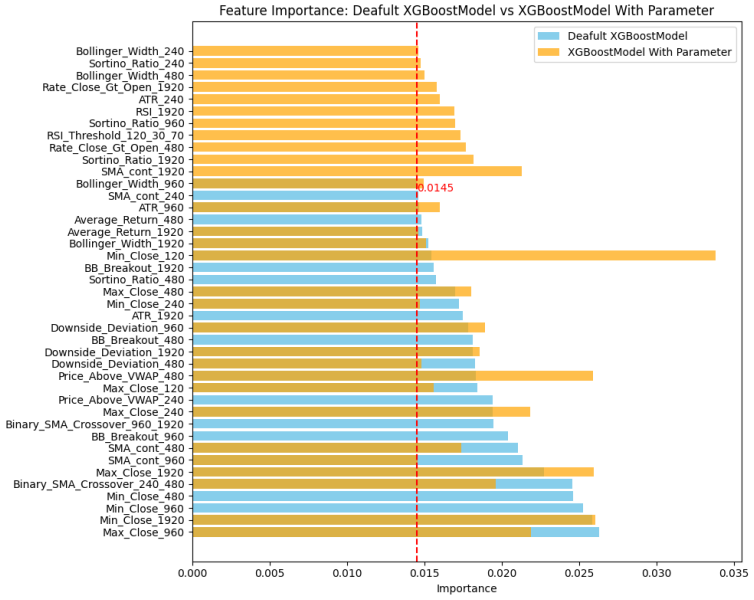

Figure 2: Feature Importance using XGBOOST Model

6 Grid Search

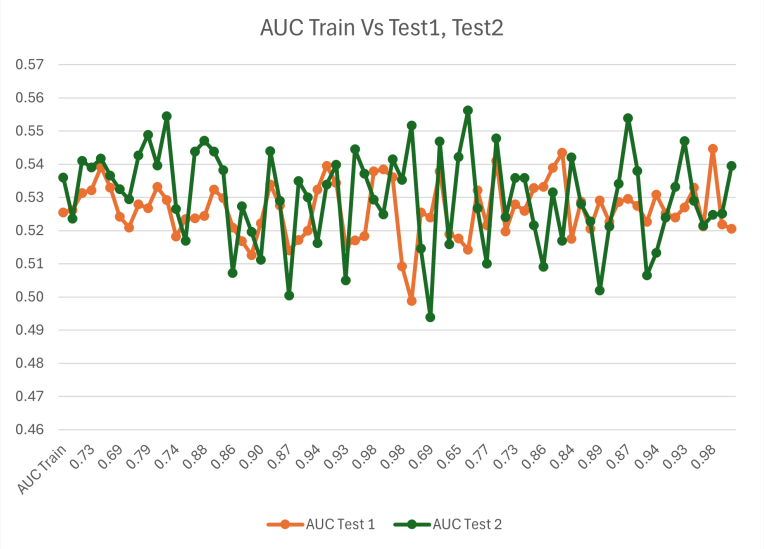

Figure 3: Grid Search Results XGBOOST Model

7 Shap Analysis

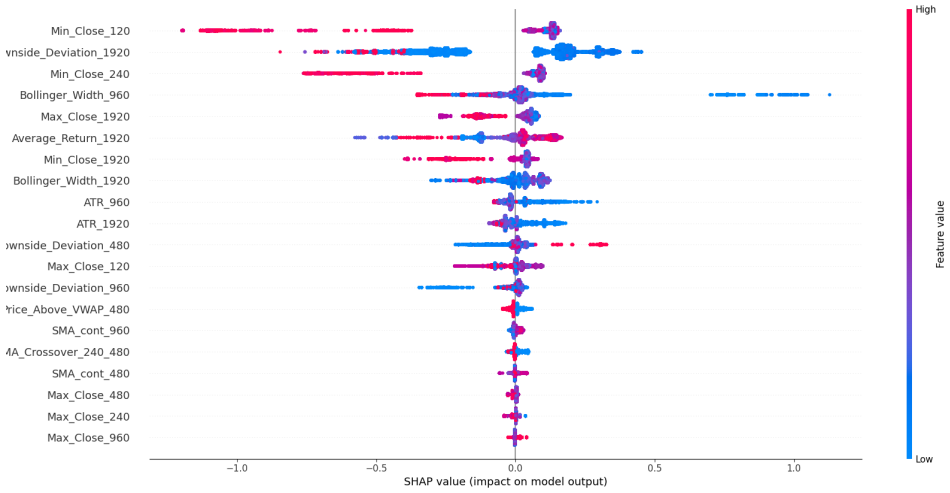

Figure 4: Shap Analysis: beeswarm for Test1 dataset

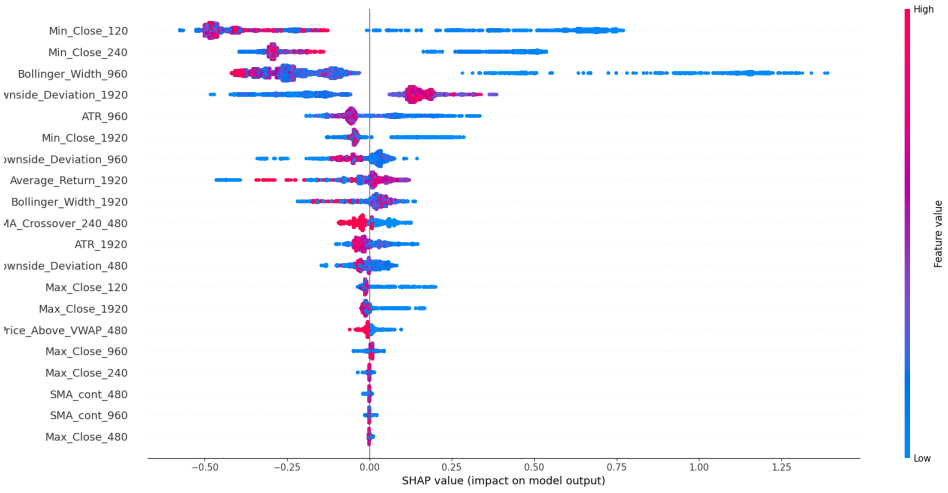

Figure 5: Shap Analysis: beeswarm for Test2 dataset

8 Strategy Definition

The goal is to identify the optimal trading strategy by selecting the best threshold in conjunction with a defined trade setup. Note that strategy performance depends on both the predictive model (and its threshold) and the trade setup parameters. The key components of our methodology are as follows:

Trade Setup Definition

- **Holding Period:** The maximum number of periods a trade is held before closing.
- **Take Profit:** The profit target (e.g., 0.5% or 1%) that triggers an early exit.
- **Stop Loss:** The loss limit (e.g., -0.5% or -1%) at which the trade is terminated to control risk.

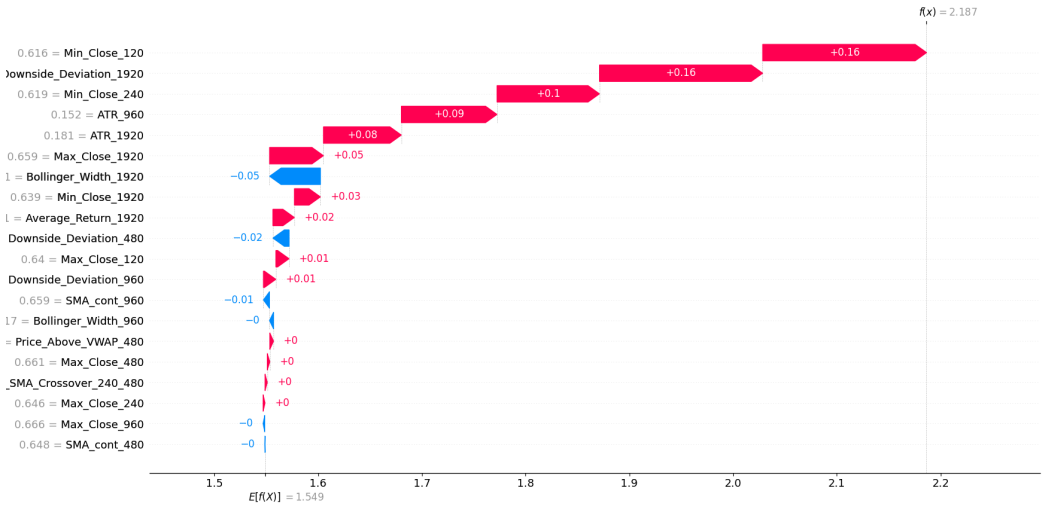

Figure 6: Shap Analysis: Waterfall for obervation number 1000 on TEST1 dataset

Strategy Simulation

- A trade is initiated when the predicted signal (derived from the model) meets or exceeds a specified threshold.
- Once entered, the trade is held for up to the defined holding period.
- The trade is exited early if:
  - The return reaches or exceeds the **Take Profit** level, or
  - The return falls below the **Stop Loss** level.
- If neither condition is met, the trade is closed at the end of the holding period.

Performance Metrics

For each strategy configuration, the following metrics are computed:

- **Number of Trades:** Total trades executed.
- **Win Rate:** Proportion of trades with positive returns.

- **Total Return:** Cumulative return over all trades.
- **Sharpe Ratio:** Ratio of the average trade return to the standard deviation of returns, representing risk-adjusted performance.

Grid Search and Back-testing

- A grid search is conducted over various combinations of the trade setup parameters and threshold values. **Holding Periods:** 60, 120, 240 periods. **Take Profit:** 0.005, 0.01. **Stop Loss:** -0.005, -0.01. **Thresholds:** 0.5, 0.6, 0.7, 0.8, 0.9.
- For each parameter combination, the strategy is simulated using the development sample (Test + Train), and the performance metrics are recorded.
- The results are compiled into a table for further analysis.

Best Strategy Selection

- The best configuration is chosen based on the highest Sharpe.
- This configuration is then adopted for further model deployment and strategy implementation.

9 Back-Testing & Results

holding_period	take_profit	stop_loss	threshold	num_trades	win_rate	total_return	sharpe_ratio
240	0.01	-0.01	0.9	28	0.68	0.11	0.31
120	0.01	-0.01	0.9	30	0.67	0.11	0.29
60	0.01	-0.01	0.9	37	0.59	0.09	0.24
240	0.005	-0.01	0.9	47	0.77	0.09	0.22
120	0.01	-0.005	0.9	43	0.47	0.09	0.21
240	0.01	-0.005	0.9	43	0.47	0.09	0.20
60	0.01	-0.005	0.9	49	0.55	0.08	0.20
120	0.005	-0.01	0.9	47	0.74	0.08	0.19
60	0.005	-0.005	0.9	66	0.62	0.07	0.17
60	0.005	-0.01	0.9	53	0.70	0.07	0.16
120	0.005	-0.005	0.9	64	0.58	0.04	0.11
240	0.005	-0.005	0.9	64	0.58	0.04	0.11
240	0.01	-0.01	0.7	301	0.53	0.14	0.03
240	0.01	-0.01	0.8	270	0.52	0.10	0.03
240	0.01	-0.01	0.5	373	0.53	0.13	0.03
240	0.01	-0.005	0.7	469	0.37	0.09	0.02
120	0.01	-0.01	0.7	354	0.51	0.07	0.02
60	0.01	-0.01	0.7	478	0.52	0.08	0.02
240	0.01	-0.005	0.8	421	0.37	0.07	0.02
240	0.01	-0.01	0.6	363	0.52	0.06	0.02
240	0.01	-0.005	0.5	574	0.37	0.09	0.02
60	0.01	-0.01	0.8	431	0.52	0.07	0.02
120	0.01	-0.01	0.8	315	0.50	0.04	0.02
240	0.005	-0.01	0.8	451	0.65	0.05	0.02
60	0.01	-0.01	0.5	560	0.53	0.06	0.01
120	0.01	-0.005	0.7	504	0.38	0.05	0.01

Figure 7: Backtesting: Performance of strategy on Test1 dataset