S&P 500 Future Contracts: Predicting Short Term Price Action

Introduction

This project focuses on addressing a longstanding challenge in predicting stock market movements, specifically within the context of S&P 500 future contracts represented by the symbol "/ES." While previous attempts to forecast market trends have been met with limited success, the allure of substantial profits continues to drive exploration in this domain. Recognizing the complexity of forecasting long-term price actions influenced by a multitude of variables, this project takes a targeted approach. It aims to predict the short-term (5 minutes) price action of S&P 500 future contracts. By doing so, the objective is to streamline the prediction process, reducing the number of independent variables required for accurate forecasts.

Over the last decade, intra-day futures trading has witnessed a surge in popularity, due to brokerages expanding their client base. Future contracts, such as the "/ES" contract, typically comprise multiple equities of the underlying asset, in this case, 50 shares of the S&P 500 index valued at approximately $240,000 ($4,800 per share). The dynamics of day trading futures involve aiming for a two-point or $2.00 share increase before selling, coupled with implementing risk management strategies such as cutting losses at 2 ticks ($0.50 per share). This approach allows traders to operate profitably with a smaller edge over the market, typically ranging from 2-5%. The goal of this predictive model is to consistently outperform the traditional trading edge, making it suitable for integration into a trading program. Due to the final goal of implementing this model, into a trading application the numeric price action variable will be transformed into a binary variable. The purpose of this is to enable the use of multiple models collectively as logic gates to determine when to execute buying and selling decisions.

Data

The data source utilized for this project was sourced from "Kaggle.com" and is titled "Intraday market data." It encompasses weekly CSV files containing price action and volume data for various stocks and futures, captured at three-second intervals from February 2020 to July 2023. The original data was obtained from the trading platform "Think or Swim" before being made available on Kaggle. Each of the 201 CSV files comprises a total of 22 features. For this project, all features were removed except for the following three: 'Timestamp,' '/ES,' and '/ES Volume.'

After eliminating non-usable features, the remaining three features were converted into five-minute candlesticks of data using the "grouper" package. Each candlestick/row includes a 5-minute timestamp, minimum and maximum prices, last price over the time period, and the sum of volume during that period. Then the price difference between the next candlestick (5 minutes) was calculated, the distribution of the price movements is shown in figure 1 below. Afterwards the price difference was used to create a binary column stating if the price increased by at least two points in the five-minute increment.

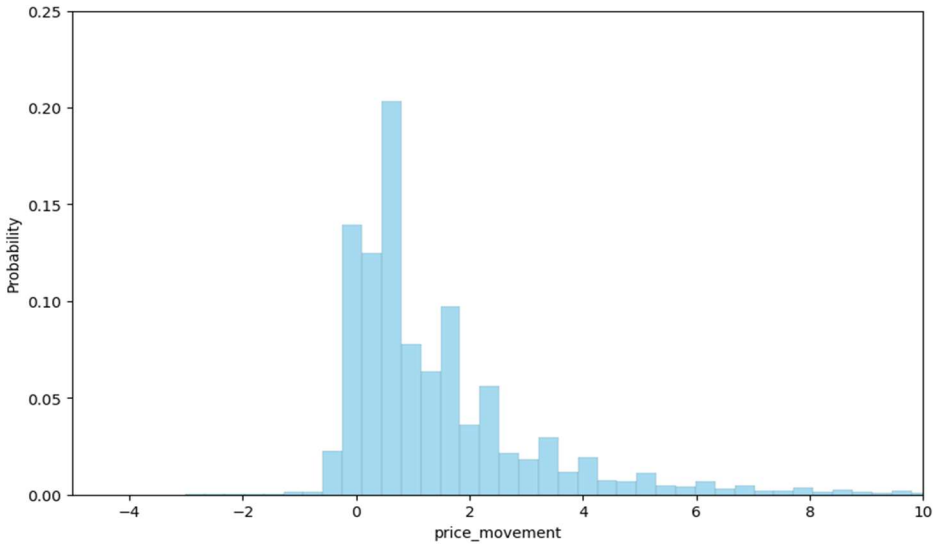

Figure 1

Subsequently, five types of technical indicators were computed based on these candlesticks. These indicators encompass various period length simple moving averages (SMA), exponential moving averages (EMA), Bollinger Bands, moving average convergence divergence (MACD), Stochastic Oscillator, and Fibonacci retracements. During the process, additional columns not mentioned were created to help with calculating these indicators but were dropped after the calculation to remove non useful features from the model.

Analysis

Due to the complex and nonlinear relationships between the independent and dependent variables. Mutual information from the "Scikit-learn" package was chosen to better understand these relationships. Mutual information, a measure of the relationship between two quantities, indicates the extent to which knowledge of one reduces uncertainty about the other. A mutual information of 0.0 signifies independence, where knowing one quantity provides no information about the other. Theoretically, there is no upper limit to mutual information, but in practice, values above 2 are uncommon due to its logarithmic nature. The results of the analysis can be seen in figure 2 below. After determining the most useful features, the least useful were removed to increase the accuracy and precision of potential models.

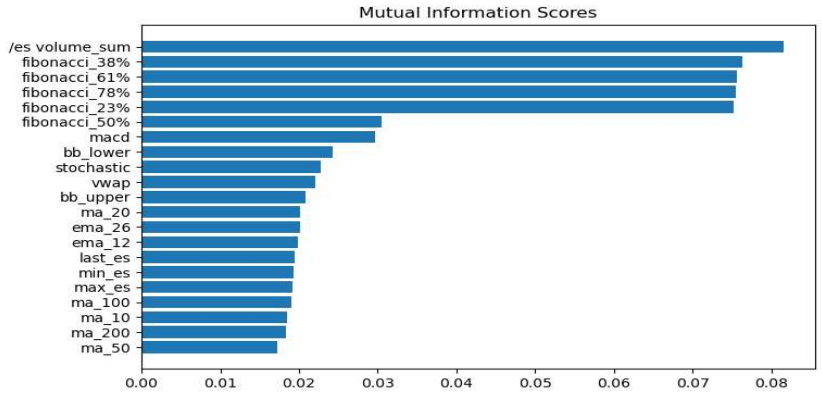

Figure 2

Methods

Due to transforming the dependent variable into a binary representation, logistic models were used during the testing and analysis. Grid searches were used to streamline the testing of the three different models and their hyper parameters. A pipeline was decided against, due to concerns of the compute time of testing all three models and hyperparameters in a single run. The three models used were the logistics regressions from the following packages: “Scikit-Learn”, “XGBoost”, and “CatBoost”. Multiple hyperparameters were tested for each regression. For “Scikit-Learn”; “penalty”, “C”, and “max-iterations” were all tested with a range of variables. While for “XGBoost”, and “CatBoost” the following hyperparameters were tested; “learning_rate”, “n-estimators”, and “depth”.

After training the models with the test data and optimizing the hyperparameters, we calculated both accuracy and precision using both the training and test datasets for each model. This approach helped identify any potential "overfitting" issues with the models. While accuracy provided insights into the overall predictive capabilities of the models, precision emerged as the primary criteria for selecting the best model. The emphasis on precision stems from the intended application in a trading scenario. In practice, the occurrence of false negatives (missed trades) does not significantly impact success, whereas false positives (incorrect trades) can have more substantial consequences. Therefore, precision was prioritized to ensure the selected model minimizes the risk of making incorrect trading decisions.

Conclusion

All three models demonstrated commendable performance in terms of both accuracy and precision. The logistic regression model emerged as the top performer, achieving a precision of 61.9% and an accuracy of 75% on the test dataset. The optimal hyperparameters for this model included a "penalty" of L2, "max_interations" of 500, and a "C" of 0.000695. The "XGBoost" and "CatBoost" models closely followed, with precisions of 60.0% and 59.8%, respectively. There were no apparent concerns of overfitting with the logistic regression, as evidenced by the minor differences between the predictions on the training and testing datasets. For a comprehensive view of the models' predictions, the confusion matrix for each model is below.

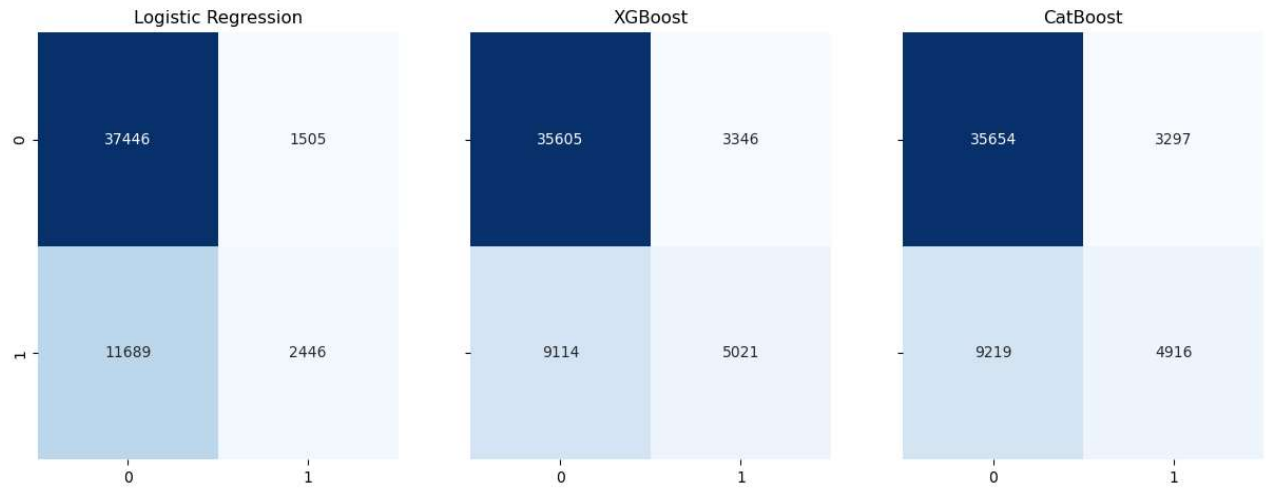

Figure 3

The generated model demonstrates superior predictive abilities compared to typical day trading strategies that rely on technical indicators. Interestingly, a significant number of these indicators proved to be less effective than simply considering the volume of each candlestick. This likely explains the model's ability to outperform the mentioned trading strategies. While

the model serves as a solid foundation for a day trading application, it is not standalone. It predicts only a two-point movements within the next five minutes, introducing several pitfalls that need addressing. To enhance its functionality, additional models must be developed to create logic gates, mitigating some of these pitfalls. Furthermore, the implementation of risk management is crucial to minimize potential losses and enhance the overall trading strategy's edge.

Ethical Considerations

In general, the development and utilization of a model for predicting stock market price movements does not raise any significant ethical concerns. However, a specific area of scrutiny within this project is the sourcing of data, which, while not illegal, may potentially violate the terms of use for 'Think or Swim.' Further investigation is warranted now that the model demonstrates signs of initial success and may be intended for deployment in a production environment.

References

1. Barton-Smith, D. (2023, December 1). Intraday market data. Kaggle. https://www.kaggle.com/datasets/brtnsmth/intraday-market-data

2. Hayes, A. (n.d.). *Bollinger Bands®: What they are, and what they tell investors*. Investopedia. https://www.investopedia.com/terms/b/bollingerbands.asp

3. Maverick, J. B. (n.d.). How to calculate Moving average convergence divergence (MACD). Investopedia. https://www.investopedia.com/ask/answers/122414/what-moving-average-convergence-divergence-macd-formula-and-how-it-calculated.asp

4. Hayes, A. (n.d.-b). Stochastic oscillator: What it is, how it works, how to calculate. Investopedia. https://www.investopedia.com/terms/s/stochasticoscillator.asp

5. Murphy, C. (n.d.). What are Fibonacci retracements and Fibonacci ratios?. Investopedia. https://www.investopedia.com/ask/answers/05/fibonacciretracement.asp

Appendix

Simple Moving Average (SMA)

The Simple Moving Average (SMA) is a widely used trend-following indicator that calculates the average price of a security over a specified period, smoothing out price fluctuations and helping identify trends.

$$SMA = \frac{P_1 + P_2 + \ldots + P_n}{n}$$

Exponential Moving Average (EMA)

The Exponential Moving Average (EMA) is a type of weighted moving average that gives more weight to recent prices. It reacts more quickly to price changes compared to the SMA.

$$EMA = \alpha \times (P_t - EMA_{t-1}) + EMA_{t-1}$$

Bollinger Bands

Bollinger Bands are volatility bands placed above and below a moving average. They dynamically adjust to price volatility, expanding during volatile markets and contracting during stable markets.

$$\text{Upper Band} = SMA + (2 \times \text{Standard Deviation})$$
$$\text{Lower Band} = SMA - (2 \times \text{Standard Deviation})$$

Moving Average Convergence Divergence (MACD)

The Moving Average Convergence Divergence (MACD) is a trend-following momentum indicator that shows the relationship between two moving averages of an asset's price.

$$\text{MACD Line} = EMA_{12} - EMA_{26}$$
$$\text{Signal Line} = EMA(\text{MACD Line}, 9)$$
$$\text{Histogram} = \text{MACD Line} - \text{Signal Line}$$

Stochastic Oscillator

The Stochastic Oscillator measures the relative position of a closing price within a range over a specified period. It is used to identify overbought or oversold conditions.

$$\%K = \frac{(C - L14)}{(H14 - L14)} \times 100$$

Fibonacci Retracements

Fibonacci Retracements are used to identify potential reversal levels in a price trend by highlighting key support and resistance levels based on the Fibonacci sequence.

$$\text{Retracement Level} = \text{Low} + (\text{High} - \text{Low}) \times \text{Fibonacci Ratio}$$