*PREDICTING ETF AND STOCK PRICE TRENDS*

By: Michael Alicea, Talia Andrews, Darsh Chandura, Rutuja Dongre, Mikaela Milch

*Table Of Contents*

* Introduction
* Background and Motivation
* Data Collection and Analysis
* Model Development
* Results
* Conclusion and Future Work

# **Introduction**

Artificial intelligence has made significant strides in various fields, including image, text, speech recognition, self-driving vehicles, and games such as chess and Go. The financial industry has also adopted self-learning algorithms to improve investment decisions. Quantitative hedge funds have a history of using algorithms to define systematic trading strategies, making them an ideal place for new machine-learning techniques.

The report begins with the necessary background and motivation for this project, followed by data collection and analysis, and addressing challenges. Specific applications in model development, highlighting results, are then discussed before reaching conclusions.

# Background and Motivation

A financial security is a financial instrument which has monetary value (Kenton, 2023). Some examples of securities are stocks, bonds, and options (Kenton, 2023). An ETF, or Exchange Traded Fund, is a financial security which tracks a collection of assets such as stocks, commodities, or indexes (Chen, 2024; What Is an ETF?, 2024). ETFs are intended to be a lower risk investment which is more stable than the individual assets it tracks and which allows the investor to diversify their portfolio (What Is an ETF?, 2024). As of March 2024, there were 3,457 ETFs in the U.S., totaling $8.87T (ETFGI Reports Assets Invested in ETFs, 2024).

The study of ETFs is important because of their effect on the securities contained within them and, in turn, their effect on the market overall. In a 2014 study on the impact of ETFs, Ben-David et al. found that “one standard deviation increase in ETF ownership is associated with an increase of 16% in daily stock volatility” and higher stock turnover. Market volatility is the variance of a security within a short period of time and high volatility generally means higher risk for investors (Wagner, 2022).

In this project, we aim to create tools to predict the prices of ETFs over time, in order to help inform ETF and non-ETF investment strategies. Existing methods for predicting ETF prices include deep learning approaches with the use of LSTMs and CNNs (Zheng, 2021) and statistical models such as ARIMA and linear regression (Bollapragada et al., 2013). ARIMA (Autoregressive Integrated Moving Average) is a model designed to use time series data, such as stock prices, to predict future values (Hayes, 2024). We plan to use ARIMA models, along with XGBoost and Prophet, to predict future ETF prices/price directions. We will compare the regression models using RMSE and MAPE.

Our project aims to leverage these machine learning models to predict stock market trends with unparalleled accuracy and reliability. Our goal is to build a system that can predict opportune moments for buying and selling stocks to generate profits. To achieve this, we propose creating a comprehensive model that integrates several advanced techniques to capture various patterns and signals in the stock market data. By providing accurate forecasts, we enable investors to make more informed decisions, enhancing their ability to capitalize on market movements and mitigate risks.

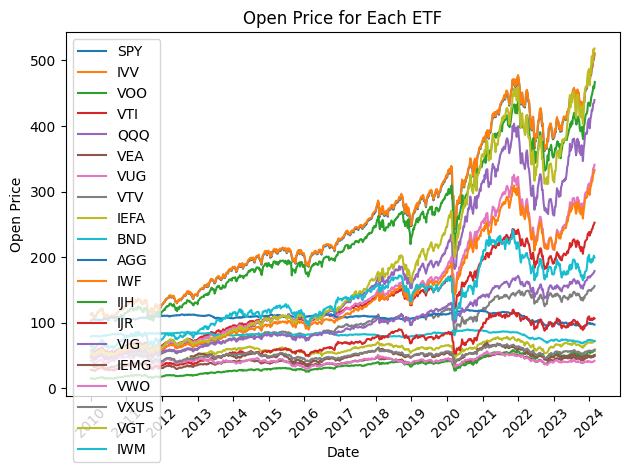
# Data Collection and Analysis

The data was collected using the YahooFinancials Python library. YahooFinancials is open source and geared towards gathering historical data for stocks, currencies, ETFs, mutual funds, U.S. Treasuries, cryptocurrencies, commodities, and indexes. We selected this library for its ability to work with ETFs as well as its recommendation from online forums. When historical data is requested for a specific stock symbol, the following information is returned: high, low, open, close, volume, adjusted close. We also added an additional column, delta\_price that indicates how much the price changed from the open price to close price.

We elected to gather data for the top 20 ETFs in the US stock market. We chose to do this because it allowed us to target the model on predicting the prices of these ETFs as opposed to considering the price of stocks. SInce ETFs are groups of stocks, their price is less affected by company specific events, and should therefore be more accurately predicted by trends in the market at large. Additionally, we gathered data from January 1st, 2010 to March 1st 2024 as the training set. We did not gather data from earlier dates because ETFs were not commonly sold until the late 2000’s. We saved the data from March 1st to today’s date as a testing validation set. The data was gathered in daily buckets.

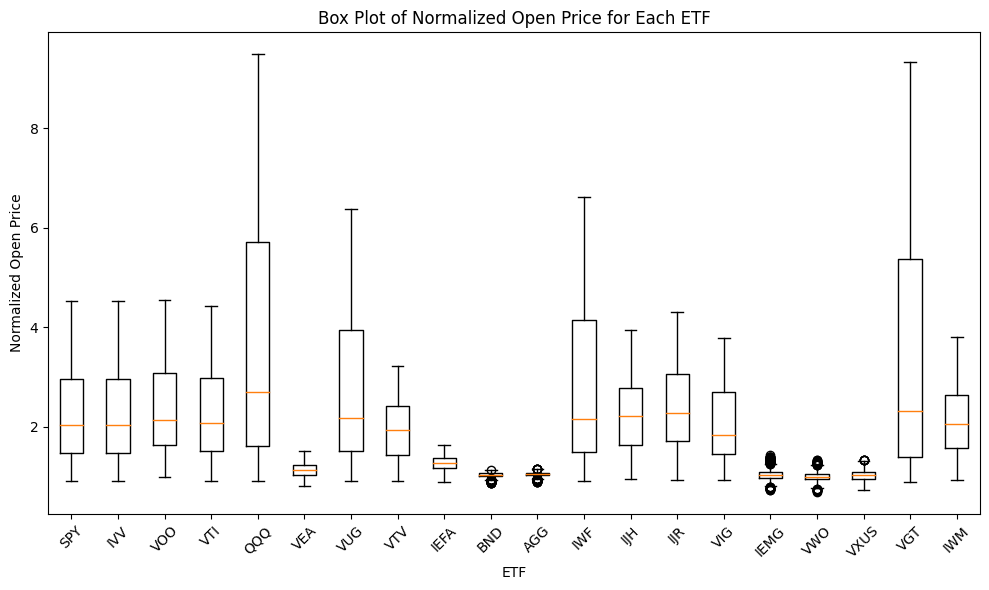
The data analysis is focused on preprocessing and exploring ETF price data. Initially, the data was preprocessed, including fetching historical price data for the top 20 ETFs, splitting it into training and testing datasets, and saving it into CSV files.

The analysis then delved into visualizing the data. Line plots were utilized to illustrate the open price trends for each ETF over time, providing insights into their historical performance.

****

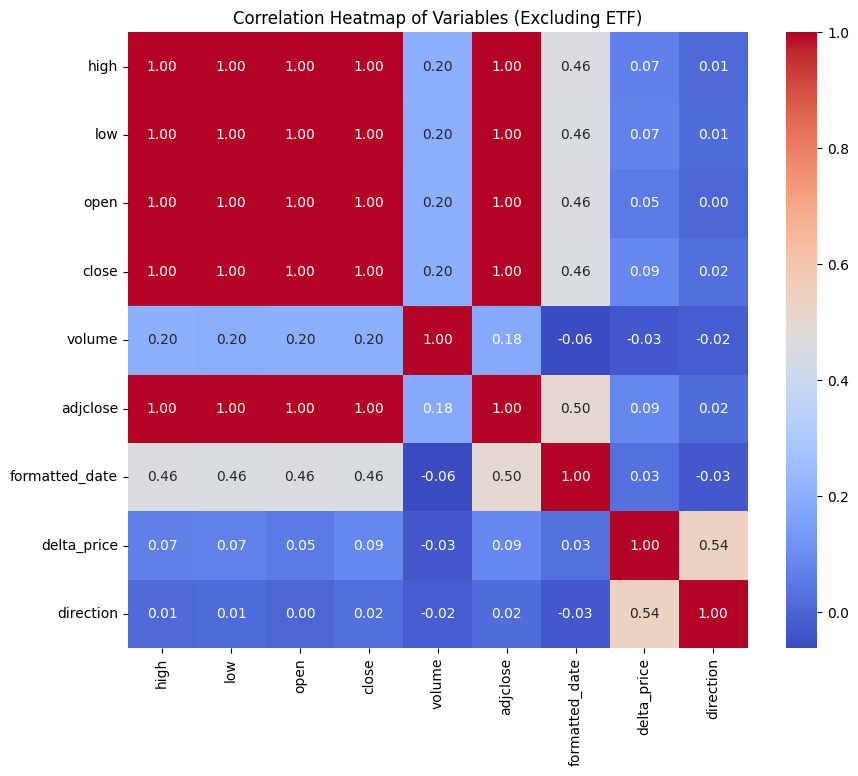
**Figure 1**

Box plots were employed to examine the distribution of normalized open prices across different ETFs, facilitating comparisons of their relative performance.

****

**Figure 2**

Furthermore, correlation heatmaps were generated to explore the relationships between various variables within each ETF and across the entire dataset. These heatmaps offered valuable insights into potential correlations between different features, aiding in understanding the interdependencies within the dataset.

****

**Figure 3**

Overall, the data analysis phase provided a comprehensive exploration of ETF price data, laying the groundwork for subsequent modeling and prediction tasks.

# Model Development

XGBoost is a powerful implementation of gradient boosting that is designed for solving supervised learning problems. It is widely known for its remarkable performance and speed in both classification and regression tasks. XGBoost works by building an ensemble of weak predictive models, typically decision trees, sequentially. In this process, each subsequent model corrects the errors made by the previous ones, leading to a more accurate final model (XGBoost Documentation, n.d.). Optuna was used to find the best set of hyperparameters for the final XGBoost model.

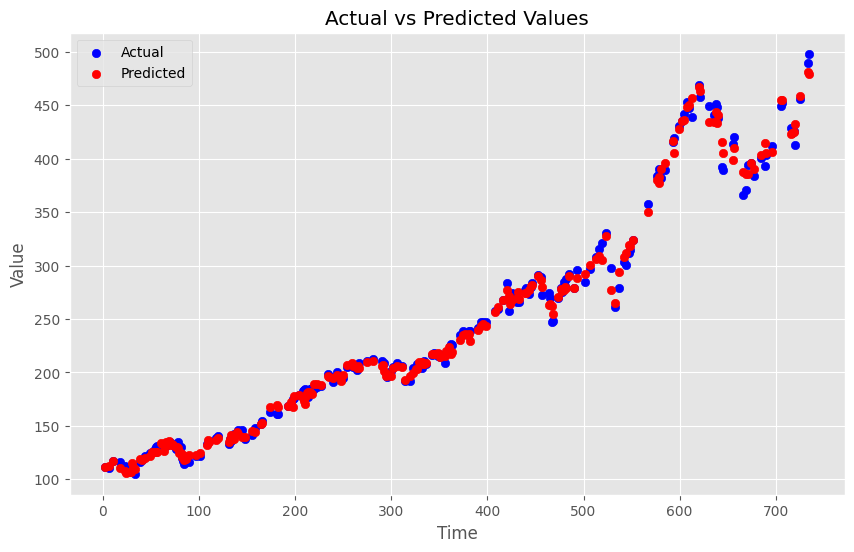
Prophet is a forecasting tool created by Facebook that is ideal for analyzing time series data with different patterns on various time scales. It is robust to missing data and shifts in the trend, and typically handles outliers well. Prophet is particularly useful for forecasting data where seasonality is strong and highly customizable, such as website traffic or sales data. It works well with time series that have several seasons of historical data and can handle strong seasonal effects (Prophet, 2023).

ARIMA is a classic statistical model that is used for forecasting time series data. It is a combination of autoregressive (AR) and moving average (MA) models and integrates differencing of the data to make the time series stationary. ARIMA is highly flexible and can model various types of time series data. ARIMA can be used to forecast a wide range of variables, including stock prices. The auto ARIMA functionality of the pmdarima library was used to find the best hyperparameters for the final model.

# Results

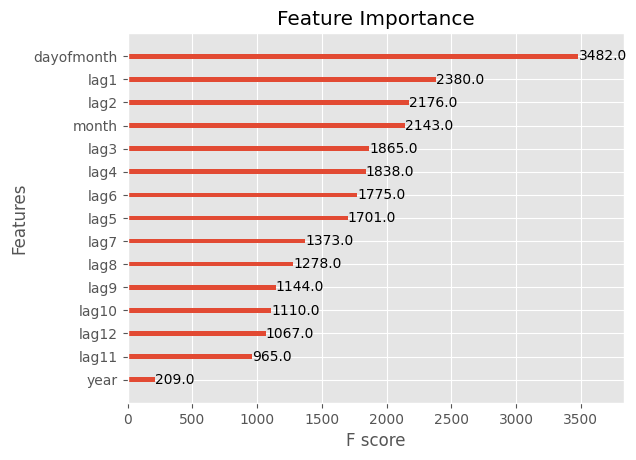
One ETF’s price information was chosen to act as a case study for the viability of using these models to predict ETF prices. The chosen ETF has the ticker SPY.

The final XGBoost model used 851 estimators, had a max depth of 7, a learning rate of 0.0726, an alpha of 3.534, and a lambda of 8.118. It achieved an RMSE of 6.446, with a mean absolute error percentage of 1.891%. Figure 4 below shows the predictions made by the final XGBoost model.



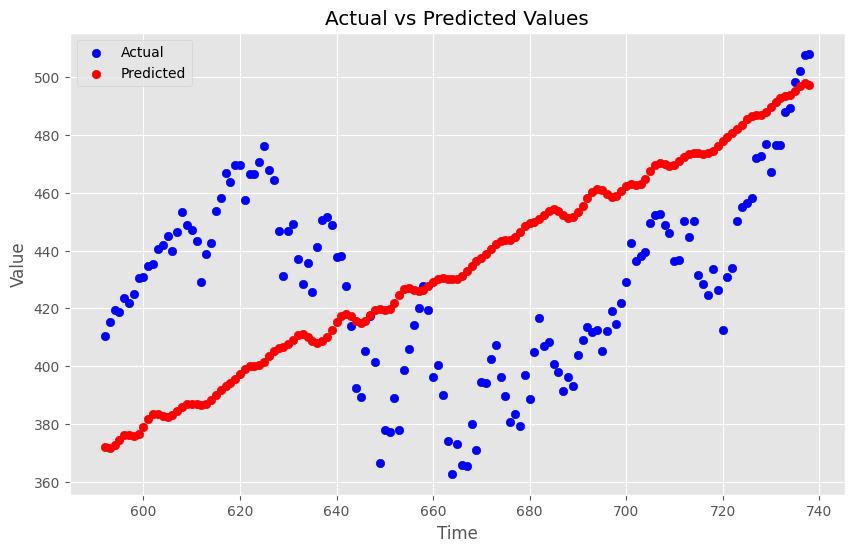
**Figure 4**

The XGBoost model also was able to provide information about the most important features that it used when predicting the close prices. Figure 5 below shows these importance scores. The day of the month was by far the most important feature, followed by lag1 and lag2, which represent the previous day’s close price and the second previous day’s close price, respectively. Year was relatively unimportant, and future models might be able to be created without its use entirely.



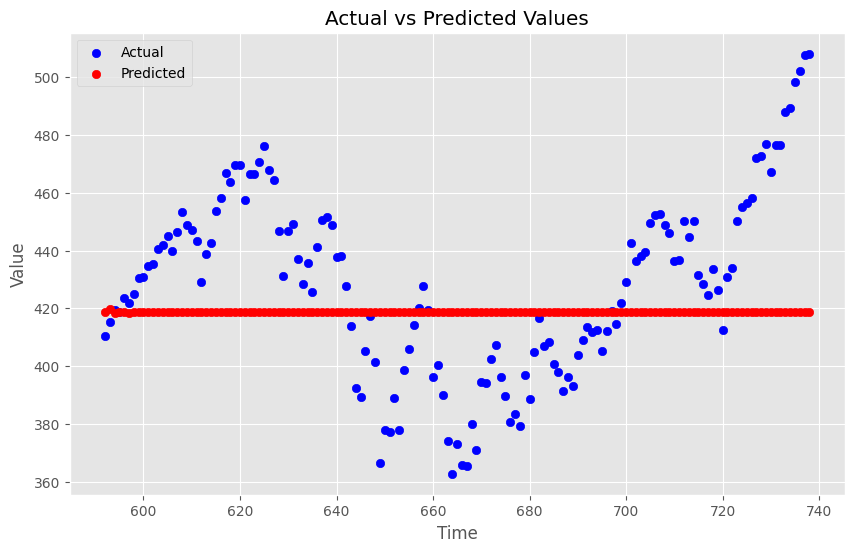
**Figure 5**

The Prophet model showed somewhat promising results. The Prophet model achieved an RMSE of 44.592 and a mean absolute error percentage of 9.558%, both higher than the XGBoost model. Figure 6 below shows the prices that were predicted by the model.



**Figure 6**

Finally, the ARIMA model had the worst results, as evidenced by Figure 7. This model was unable to capture the true pattern and variance in the ETF data. It achieved an RMSE of 33.051 and a MAPE score of 6.244%, but it was unable to generalize for future data as well as the XGBoost or even the Prophet models could.



**Figure 7**

# Conclusion and Future Work

In conclusion, this project aimed to build machine learning models to predict the future price of stocks and ETFs. Three machine learning models were created towards this end goal: XGBoost, Prophet, and ARIMA. One ETF’s historical price data was used as a case study to explore whether these methods could be used for price prediction. XGBoost was the most successful model of the three in capturing the patterns in the data.

Future work can use this small case study to explore other models, such as LSTMs or ARIMA-GARCH models, which may be more successful in capturing the variance of the data. Additionally, future work can explore combining these models to build one, stronger final model.

# References

Ben-David, I., Franzoni, F., & Moussawi, R. (2014). Do ETFs Increase Volatility? National

Bureau of Economic Research. https://doi.org/10.3386/w20071

Bollapragada, R., Savin, I., & Kerbache, L. (2013). Price Forecasting and Analysis of Exchange

Traded Fund. Journal of Mathematical Finance, 3(1A). https://doi.org/10.4236/jmf.2013.31A017

Chen, J. (2024, April 12). Exchange-Traded Fund (ETF): Pros and Cons. Investopedia.

https://www.investopedia.com/terms/e/etf.asp

ETFGI reports assets invested in ETFs industry in the United States reached a new record of

US$8.87 trillion at the end of Q1 2024. (2024, April 15). ETFGI. https://etfgi.com/news/press-releases/2024/04/etfgi-reports-assets-invested-etfs-industry-united-states-reached-new

Hayes, A. (2024, April 5). Autoregressive Integrated Moving Average (ARIMA) Prediction

Model.Investopedia. https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp#:~:text=An%20autoregressive%20integrated%20moving%20average%2C%20or%20ARIMA%2C%20is%20a%20statistical,values%20based%20on%20past%20values.

Kenton, W. (2023, December 20). What are Financial Securities? Examples, Types, Regulation,

and Importance. Investopedia. https://www.investopedia.com/terms/s/security.asp

Prophet: Forecasting at scale. (2023). https://facebook.github.io/prophet/

Wagner, H. (2022, June 2). Why Volatility Is Important for Investors. Investopedia.

https://www.investopedia.com/articles/financial-theory/08/volatility.asp

What is an ETF? (2024). Charles Schwab. https://www.schwab.com/etfs/understand-etfs

XGBoost Documentation. XGBoost. https://xgboost.readthedocs.io/en/latest/

Zheng, W. (2021). Exchange-Traded Fund Price Prediction Based on the Deep Learning Model.

7429–7434. https://doi.org/10.1109/CAC53003.2021.972776

# 