Stock Price Prediction

Authors: Yijia Song, Daniel Korn, Nicole Davis, Yidi Wang

DS3000: Foundation of Data Science, Group 52

ABSTRACT

This project aims to develop a predictive model for stock market performance using Python, leveraging its ability to process large amounts of stock market data. Using the real-time dataset on Yahoo Finance, the objective is to create a more accessible and transparent stock forecasting tool. By analyzing historical data, industry forecasts, and overall market health data, the project seeks to identify factors with the highest correlation to accurate stock price prediction and investment potential. This project will use ETFs and stock market data from the NYSE, focusing on exploratory data analysis, indicator identification, and concise visualizations. The final goal is to develop an algorithm that assists users in making informed investment decisions based on historical data and industry trends.

INTRODUCTION

There are many factors involved in predicting stock market performance, including industry trends, the external environment, etc. Therefore, analyzing stocks is not an easy task. This is compounded by the volatility of stock prices and the overall unpredictability of the market, which makes it very hard to forecast stock performance and make appropriate investments. In this project, we aim to use Python's powerful ability to quickly process large amounts of stock market data and learn about which factors most heavily influence a stock's price and build predictive models to try to forecast and analyze stock trends. We are hoping to be able to gain insight into which stocks are going to be positive investments, and where one will make money.

OBJECTIVES

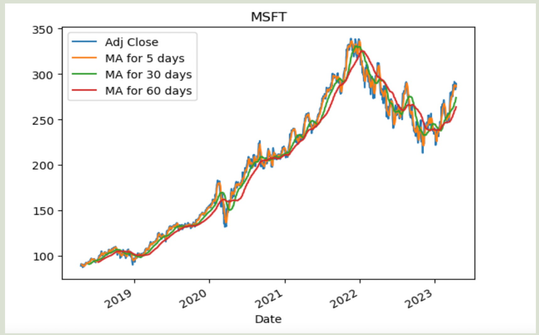
We will likely be using ETFs as a part of our analysis, this stands for “Exchange-Traded Fund”, which is a group of stocks that are all in the same industry. They are usually considered a low-risk investment but they can be good predictors of how a certain industry is going to do/is doing on the market. The goal of this project is to use stock market data for the NYSE to try to predict the market’s behavior in a way that can help us decide what the most profitable investments would be. Unlike other online analysis tools we want to be straightforward and transparent about what our model is doing, and state our recommendations in layman's terms, instead of convoluted language. Our first objective is to successfully explore and clean our data using what we learned in class, especially regarding the exploratory data analysis section. To do this we will likely have to scrape it from Yahoo Finance. We also want to find out which would be the best indicators for when it's a good time to invest in a stock and, from that create concise visualizations that will help show our results in an easier-to-understand manner, this will likely include lots of line charts to show stock growth. The next step is to use the selected indicators and test them to see if they could accurately predict trends in the stock market. When all of that is done, our final objective is to create an algorithm that would help the user to decide whether or not they should invest in a given stock/industry based on historical data.

MOTIVATION

What motivated us to do this project is the fact that our group is composed mainly of Economics/Business majors so we decided to pick something we are very familiar with hearing about in our other classes and we were interested in approaching it from the Data Science point of view. This led us to our project surrounding stocks. There are hundreds if not thousands of models of stock forecasting and predictors that exist in the market, and investing in stocks is a lot like gambling. The stock market is inherently unpredictable and many people need to learn to use logic when picking stocks to invest in. Additionally, you can hire companies like JP Morgan or other investment banks to manage your money for you, but the fees are high, and there is no visibility or guarantee that you are going to make money. There is also a significant barrier to entry into investing because of these models, and general know-how. Throughout this project, we want to see if it is possible for us to build our own model, therefore potential for other people, using historical data, industry forecasts, and overall market health data to predict whether or not someone should invest in a stock, and use this project as a chance to learn more about what factors have the highest correlation to correctly predicting a stock price and whether or not to invest.

METHODOLOGY

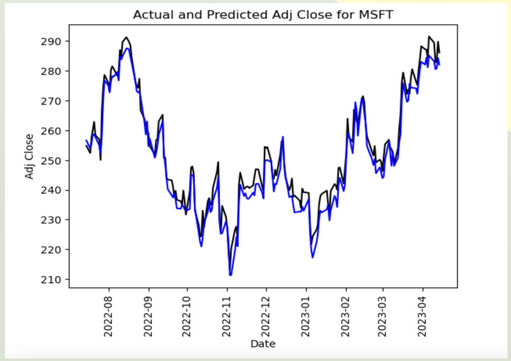
we gathered our data from Yahoo Finance. We are scraping our own data based on the S&P 500 which we are getting a list of from Wikipedia. The timeframe is within 5 years from this year, which is from April 2019 to April 2023. After obtaining the data, we load it into a data frame to better perform other data preparation steps. We first took a look at the basic information about the data frame, in order to get an overall understanding of the data. Since the format of this project is to let the user input the ticker they want to predict, we decided to calculate the moving average and add it to the data frame and plot it to let users have a basic understanding of the ticker. For example, if a user wants to make a prediction for the ticker ‘MSFT’, then a graph (shown below) will be presented.



Next, to continue with data preparation, we changed all data to suitable types (such as type string, integer, or float). Then, we checked to see if there are any null values in the data frame, and none were found. We then removed the tickers that are not on yahoo finance and have likely been delisted.

Once the preparation was complete, we began to build our model with our target variable, 'Adj Close', and predictors, 'Open', 'High', 'Low', and 'Volume'. Based on the fact that "Adj Close" is a continuous value, we finally chose the random forest tree regulator as our final model due to several reasons. First, this model has a high accuracy score, and it is good at handling large datasets since our dataset contains data within 5 years. Next, Stock price data often exhibits non-linear relationships between predictor variables and 'Adj Close', and Random Forest Regressor is a good choice for handling non-linear relationships. Finally, overfitting could be a common problem in machine learning where a model performs well on training data but poorly on test data. Random Forest Regressor is designed to be robust to overfitting and can perform well on unseen data. We then performed hyper tuning of the model by using GridSearchCV(). For each of the parameters, we selected 3 values. The number of folds is 3. We fit the grid using the x and y training data and displayed the best parameters and the score it produces. Finally, we used the best parameters to create a new regression model

To evaluate the performance of the model, we used a model. score() to get the regression confidence score, and also used the mean squared error (MSE) value. We also made a plot to enable users to visualize the accuracy of predictions. An example is shown below.

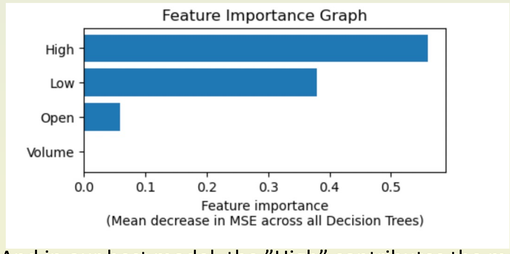


We also think of some investors who may be more concerned with short-term returns, therefore, we also made a model with RandomForestRegressor() to predict daily returns. We used the same methods as above, including GridSearchCV(), and used regression confidence and MSE to measure accuracy. However, we find that it is super hard to predict daily returns because of the market's volatility.

To evaluate how important features were when making the predictions, we used the feature_importances_ function.

RESULTS AND EVALUATION

The best ‘max_depth’ value was 10. This means the maximum depth of the tree in the random forest produced the best accuracy score. The regression confidence score was 0.96, which means our model performed well in predicting the adjusted closing price.



And in our best model, the ”High” contributes the most to the model's performance, while volume affects the less.

IMPACT

The stock market is inherently unpredictable, and many people need to learn to use logic when picking stocks to invest in. As a result, individual investors, as well as students and researchers, can benefit from our models and results. For an individual investor, hiring a stock manager can be expensive. And using this model can help investors make informed decisions about buying, selling, or holding specific stocks, as well as risk management and long-term investment planning. For students and researchers, the model can help them better understand how the stock market works. In addition, by analyzing the model's predictions and performance, researchers can gain insight into stock market trends and identify potential research and analysis opportunities, while developing data analysis skills.

CONCLUSION

Through this project, we were able to use NYSE stock market data to predict market behavior and help users to find the most profitable portfolio as possible. This was successful because we implemented a RandomForestRegressor model. By finding the best values for the parameters, we were able to use our model to predict the adjusting closing price of the stock ticker entered by the user, and in doing so, advise investors on the most profitable investments. Thus, our project succeeded in achieving its goal, satisfying our motivation to create our own model to give investment advice to investors.

The first limitation is limited features since now we are using only Yahoo Finance data to make stock price predictions. Additional features that may be useful include social media sentiment, economic indicators, financial ratios, and market indicators. These features can provide insights into market trends, investor sentiment, and the overall health of a company, which can be important factors for predicting stock prices. By using a wider range of features, prediction models can potentially capture more of the complexity and variability of the stock market and make more reliable predictions.

The second limitation is background knowledge. In order to use this model, users need to have some background knowledge of the stock market and financial concepts, such as the specific terms showed in the model.