Project 3: Milestone 2

Proposal

Jocelyn Disla

DSC 680-T302 Applied Data Science

Amirfarrokh Iranitalab

June 1st, 2024

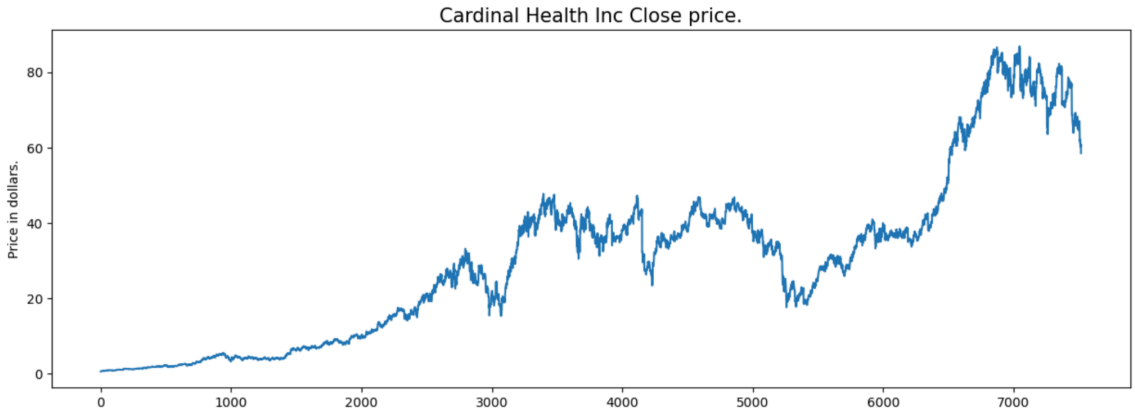
Business Problem and Background

The stock market is very important in the economic world that all companies and businesses live in to raise their capital. By having a predictive system that can provide a heads up to companies, it can prepare for large economic losses and wins. This will keep the companies that have access to this system with an advantage over others that would create an imbalance in knowledge and preparation among the competition. The stock market is also an indicator of the state of the company's capital which will allow users to use the predictions towards decisions regarding prices and products on shelves.

Data Explanation

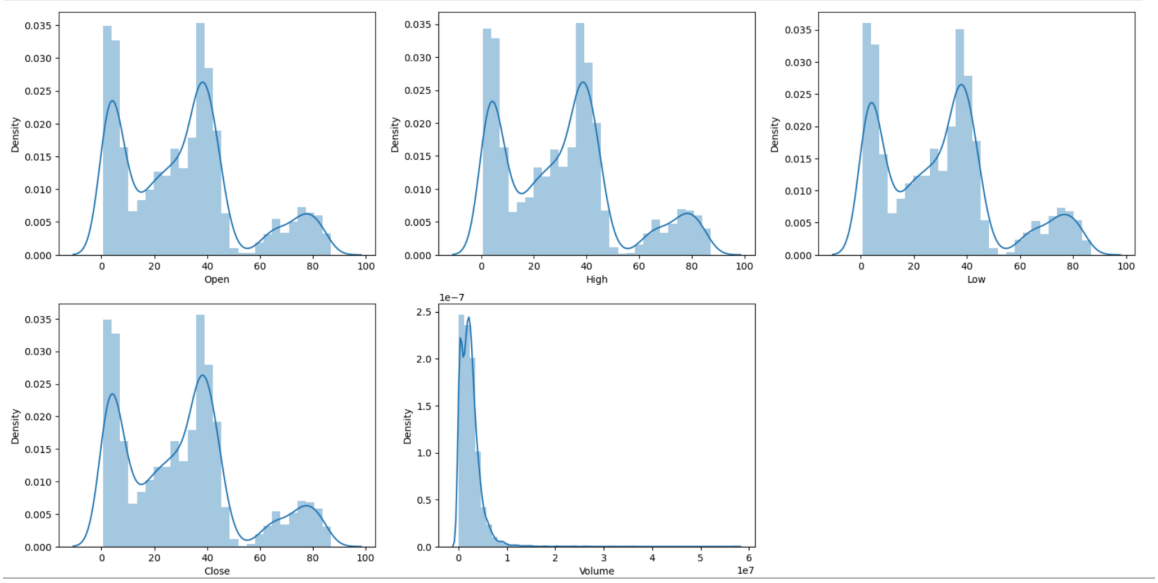
The data set is coming from Kaggle.com, the chosen stock is Cardinal Health's. With health care there are government regulations and reimbursement rates, they also deal with getting government approval of products and services. These impacting factors, as well as, rapid obsolescence and patent expiration can have significant effects on the company's stock price and availability. The data set carries 7 columns and 7519 rows that have the date and OHLCV features spanning over 20 years for this stock . As mentioned, the data will be OHLCV meaning the features will be open, high, low, close, and volume. Open and close represent the starting and ending price, high and low represent the highest and lowest reached price during that interval of time.

The data did not have any null values, which allowed for the data preparation process to take up minimal time consumption. In the data exploratory analysis, trends were found from performing graphical representations, the first graph demonstrates the stocks' price over the course of time. Presented below, the stock rises over time with a few dips along the way.

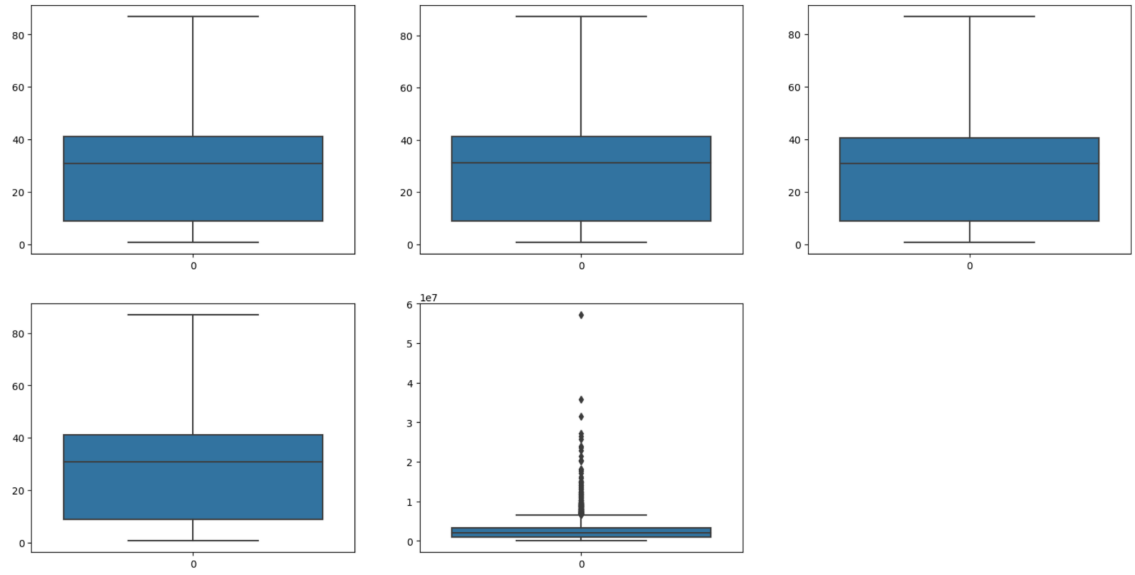


A distribution plot was graphed using the OHLCV values, which shows 2 peaks meaning the data significantly varied within those two regions, a smaller peak did present itself yet due to it'insignificant size, it was disregarded.

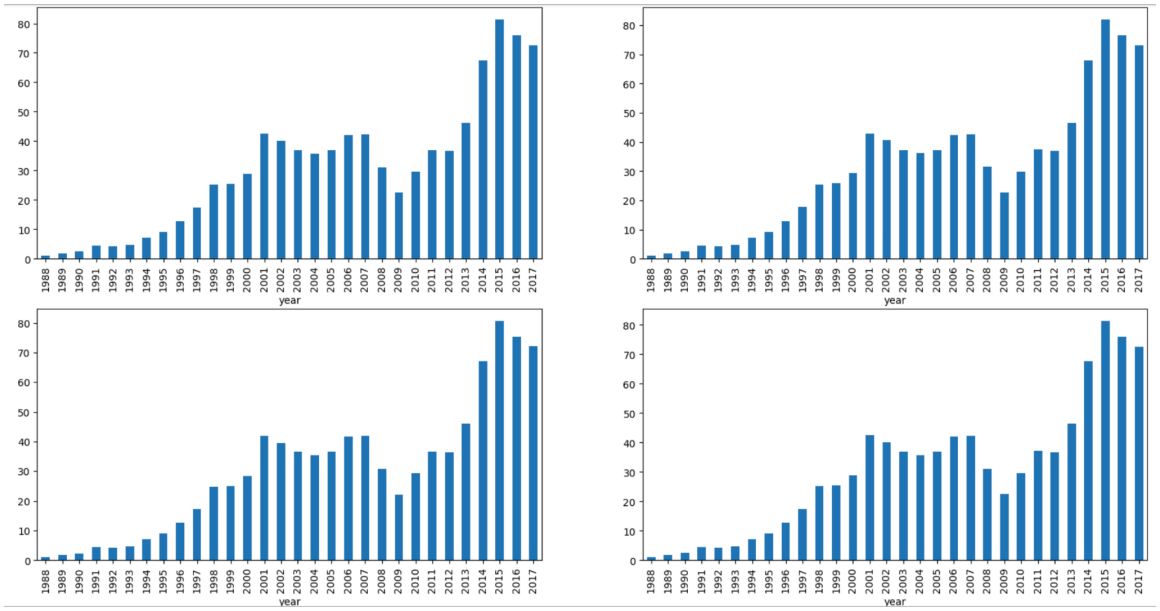
‘



From the box plots presented below, it is evident that outliers were only present for the volume feature and the other feature did not carry such outliers.



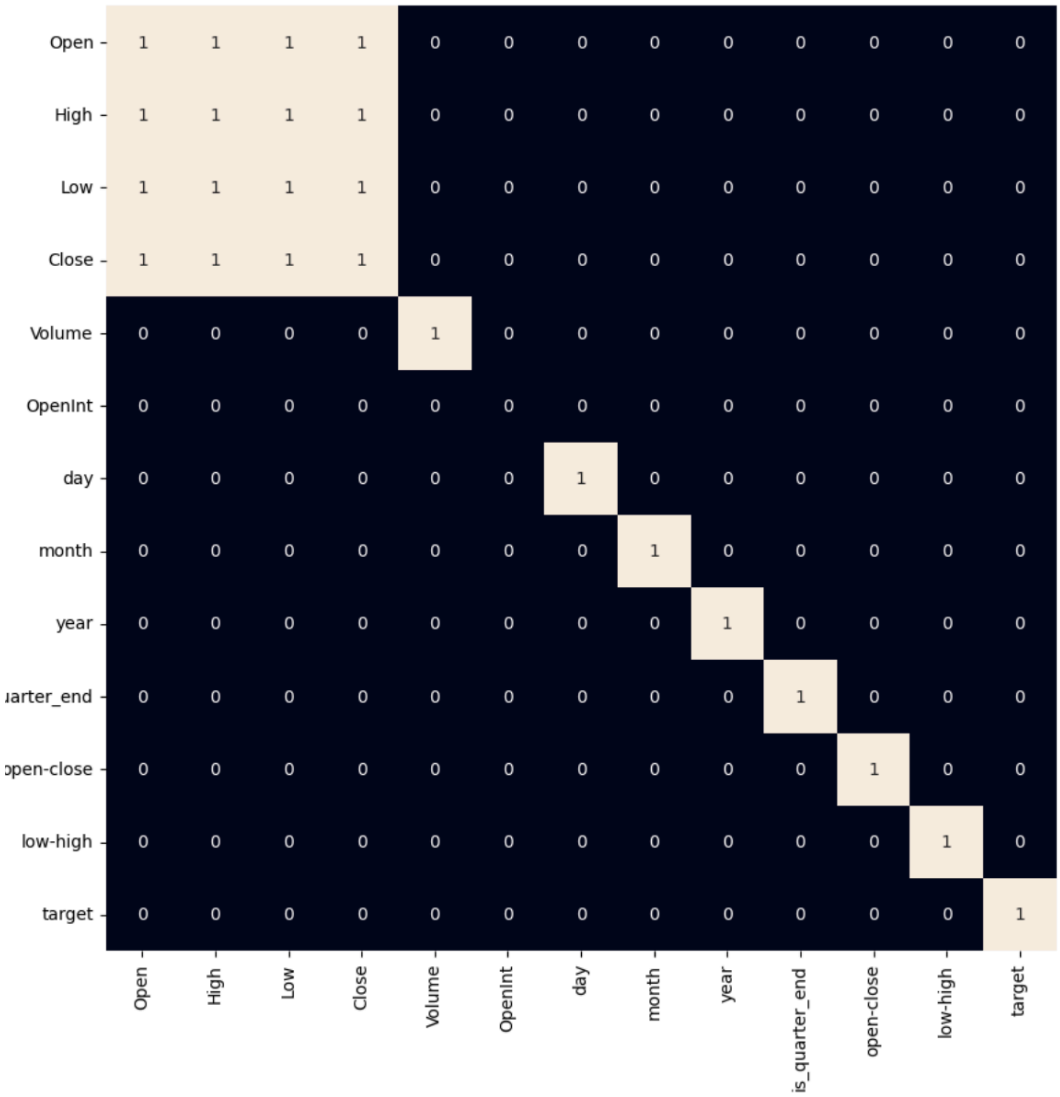
Features were added specifically to use the date column more efficiently to create bar graphs that show the yearly growth clearer. There were years that Cardinal Health lost a bit of capital around 2008 and 2009 that can be explained with the 2008 financial crisis.



Method and Analysis

The data was split by the target variable and the features. The target variable which is the signal that lets the user know whether to buy or not and this will allow the

model to be trained in order to predict this signal. A heatmap was plotted to identify any correlations between the newly created features and the already existing ones. As seen from the map below, there is a high correlation among the OHLCV features but not with the new ones which is a green light to move on to building the model.



The data was then fit with a standard scaler and split into a training and testing set into a 90/10 ratio. The data was fit into 3 models, a logistic regression, support vector machine, and XGBclassifier. The models were evaluated by using a ROC-AUC curve which also helps past the binary target variable, by predicting soft probabilities which are the values

that lie between 1 and 0. The SVC model received a similar score with its training and validation accuracy of about 54%, which means the data is less likely to be overfit. This is more reliable compared to XGBClassifier that had a high training accuracy of 84% with a lower validation accuracy of 54%, which shows the likelihood of the data being overfit. Lastly, the logistic regression performed the least accurate out of all 3 with a score of 53%.

Conclusion and Assumptions

Overall, the models performed just a little bit better than a 50/50 guess, making financial decisions based solely on this model would not be optimal. The insights from the patterns and trends found can help in making the decision to buy stock or not. Some key take away messages from this project are to add research for the healthcare industry for further information of the current state of the stock and also where it will go, as well as predicting how others will potentially pull out or buy in. Since most of the models performed near 54% accurate for validation. This model is 4% better than a random guess, by using one's personal intuition based of knowledge and research, this system can be used for further reassurance or understanding previous trends in the data.

Future Uses and Implementations

Moving forward with this model, applying different stocks to this model would create similar results, by adding impacting factors such as the overall economic state of the current time. The news is also very important with the health care stock in making sure, real time analysis can be more accurate. Adding a news dataset will further advance the predictions, as well as, understanding impacting, real-world factors. Another

implementation would be focusing on the risk of overfitting the data for the XGBboost model by adding more data and implementing cross-validation.

Ethical Considerations, Limitations, and Challenges

An ethical point to consider would be the risk of basing decisions solely on algorithmic analysis could be perceived as unfair in the competition in the stock market. Some challenges that may be faced with this project is choosing the model, XGB boost for the predictive results because of the insufficient and imbalance data. The noise and non-stationarity of the data also brings challenges since they are known for their random fluctuations that are insignificant and unpredictable. Any machine that can predict the stock market would be of extreme value since it would be similar to cheating in a gabling game, this can lead to legal issues and negative interactions with the media. The possibility of manipulating the market is also an important factor to consider when creating such a model.

References

"An Easy Guide to Stock Price Prediction Using Machine Learning." *Simplilearn.com*, www.simplilearn.com/tutorials/machine-learning-tutorial/stock-price-prediction-using-machine-learning.

Andy Acker, C., & Daniel Lyons, P. (2024, April 23). *Why Healthcare stocks could catch a break this election year*. US Institutional. https://www.janushenderson.com/en-us/institutional/article/why-healthcare-stocks-could-catch-a-break-this-election-year/#:~:text=Health%20care%20industries%20are%20subject,rapid%20obsolescence%20and%20patent%20expirations.

Machine Learning for Engineers. XGBoost Classifier. (n.d.). https://apmonitor.com/pds/index.php/Main/XGBoostClassifier#:~:text=It%20is%20an%20implementation%20of,to%20produce%20a%20strong%20prediction.

"OHLCV." *Amberdata API*, docs.amberdata.io/docs/ohlcv-1. Accessed 13 May 2024.