Ethan Mathew

Evaluation of Fill Rate and Latency Given Option Order Data

1. Procedures
    a. For the vast majority of this report, I used Jupyter Notebooks and python programming language to evaluate general trends, find correlations and create visualizations for the data. Furthermore, I utilized pandas, matplotlib and plotly as a means to visualize and parse the csv file. At the end of this report, I used the tensorflow and sklearn libraries to create a machine learning algorithm to detect whether or not an order would be filled using segments of the given csv.
2. Findings
    a. **Factors influencing Fill Rate**
        i. When analyzing the fill rate, I first wanted to consider the fact that the average fill rate amongst all the data that was not rejected is around 44% meaning that any trends that are substantially above that value are worth exploring.
        ii. Exchange - My first action was to encode all the categorical data(exchange, order type, etc) into numerical data as a way to see whether or not certain values had correlation. Immediately, I noticed that the attribute "exchange" had the highest correlation with the fill rate which is where I began my analysis. Examining these values, I found that the Philadelphia Exchange had the highest fill rate, averaging at 55% which I plotted in my notebook and included below. This was interesting as I would have expected newer exchanges like CBOE and ISE to have higher filler rate rather than the nation's oldest exchange PHLX.
        iii. Size - Next, I saw that the attribute "size" had the next highest correlation so analyzed that with a bar chart. We see from the bar chart below that as the size of our order increases that generally our orders are filled less. This is consistent with patterns we observe in the market today since larger orders are generally harder to fill as buyers and sellers are hard to match as order size increases. In general liquidity decreases as the order sizes increase as there will be few buyers who will be takers.
        iv. Stock Symbol - I wanted to also examine whether or not the attribute "stock" contributed to the percentage of filled orders, however, there wasn't a very strong correlation. Nevertheless, I did examine that certain stocks such as AMZN, GS and USO had a much smaller fill rate than most stocks which might be a feature to consider for future strategy. AMZN is almost 4% of the SPX market cap so I found it strange the fill rates were low. It is possible that AMZN was trading at higher prices before it split in the market. Amazon split 20:1 on June 6th before which it was trading at about $2800. So it is very likely the liquidity was low at this price and hence the fill rates were also lower.
        v. Lastly, I wanted to see whether or not the time of day had an influence. We saw that as the market opened, the number of filled orders generally decreased. However, at around noon(from 11:45 to 12:30) the fill percentage was consistently higher than the rest of the day. I would likely attribute this to a higher number of active traders on the market during this time which means there are more orders available to be traded. This I found quite surprising as generally

markets are very active during the open as everyone participates in the markets. So one would expect the fill rates to be much higher as soon as the market opens between 9:30 to 10AM.

b. **Factors affecting Latency**

    i. The major factor that I saw impacted the latency of the order(time between when the order manager receives the order to when the order is placed) was the exchange the order was received in. The CBOE and ISE had a latency of just over 0.0006 seconds while PHLX and AMEX was around 0.00100 seconds. The ISE exchange is well known to be one of the most modern exchanges and reputed for its use of advanced tech. These technology investments may have most likely translated to better latency rates in execution for ISE.

3. Machine Learning Model
    a. For the last part of this project I attempted to create a binary classification model that could successfully predict whether or not an order would be filled depending on the features that were given to us. While I was unable to get my desired accuracy, my code does have the architecture for the model that is capable of successfully outputting this prediction. For the data preprocessing element of this task, I had to one-hot-encode my categorical features to pass them through the model. This is similar to what I did above to generate the correlation matrix. After I processed this data, I was left with the Strike, Size, Price, MarketSize, FILL, TimeOfOrder and Exchange as the inputs of my model.
    b. My model architecture consisted of four dense layers and with the last layer using a sigmoid activation function to determine the likelihood an order would be filled or not. The loss function used is binary cross entropy, the standard function for a binary classification problem provided by tensorflow, however, this produced a very low accuracy. My next step was to regularize my loss function by punishing false positive values in a function called get_bce().
    c. Even after this, the function still didn't output the test accuracy I wanted, which can be seen through the confusion matrix. I had a true positive and true negative rate of around 65% which I would consider about average. I attempted to plot an ROC curve which measures the performance of the model at various threshold values. Ideally the model would hug the y-axis and appear like a rotated L but our trained model is bowed out indicating the model is not that strong.
4. Next Steps and Trading Strategy Insights
    a. If I were given more time for this assignment, I would dive deeper into why orders processed in the PHLX were on average filled much higher than any other exchange. I would see whether or not certain stocks of certain indexes or industries were sent to that exchange more than others and why they might be processed higher. Furthermore, I would do some more tuning on the machine learning model. I would consider balancing the data in different ways and through different variables so that the model could learn the trends in the data easier.
    b. If my objective of my trade was high frequency and I was required to take advantage of split second market information to gain an edge over other participants, I would certainly route my orders into ISE or CBOE to ensure that I got into the market ahead of the others and get price advantage.

c. If on the other hand I was basically trying to just trade and get to trading in and trading out positions where timing was not of high concern, I would use PHLX as it has the highest fill rate of all the exchanges. On the same token, I would also try to split my larger order size into smaller chunks to make sure that orders would be executed and filled as opposed to getting rejected in the market or getting an efficient price. It is best to split orders anyway so as not to alert the market and cause large price swings.
d. Lastly, I would also use the time of the day to my advantage. It appears from the results that the best time to trade would be between 11:30 to 12:00pm where most of the orders had a higher fill rate.