Introduction

Preface: For generalize conclusion, visit the Results Summary page. This page is for specific results analysis.

Our results will be presented in two ways. Firstly, we will compare the accuracy of the Logistic Regression Classifier on our data sets to that of ZeroR, which is a classifier that assigns all target attribute values to being that of the majority. Essentially, we will be comparing the accuracy of Logistic Regression on our data set compared to just saying that every stock will do the same thing.

In addition to reporting our accuracies, we will present graphs of the odds ratios for some of the attributes. This will be done to try and see which attributes give the most information in a logistic regression model. For a given attribute in a classification model, the odds ratio is an alternate representation of its coefficient in the regression function, which represents the change in odds of a positive outcome given a one unit change in the given attribute. In other words, it is an indicator of how much information is gained from the attribute in question. An odds ratio of 1 implies that the attribute doesn't really give any information, whereas values above and below 1 imply that an increase in that attribute's value increases the probability of a positive or negative outcome, respectively. Since the odds ratio is on a logarithmic scale, a value of 1 indicates that the attribute is not that important to determining the target attribute values.

For this project, we used three different data sets which are described in the table below (details will be given in the data set's respective section). It is important to note that not all attributes are contained in all data sets. The intention of this variation is to see if our model is able to remain consistent among different timeframes and when give varying amounts of data.

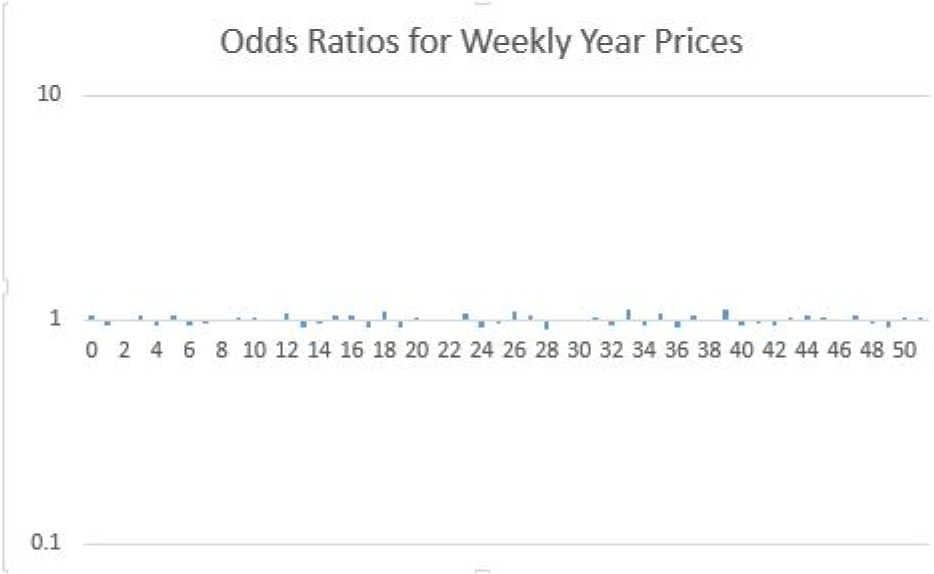
Dataset Name	Prices	Attributes
Weekly Year	Weekly Close Price	Returns, Vol, Avg Returns,
Weekly Year +	Weekly Close Price	Weekly, plus beta, Volume, Sector
Daily 3 Months	Daily Open/Close Prices	Returns

Data Set 1: Weekly Year

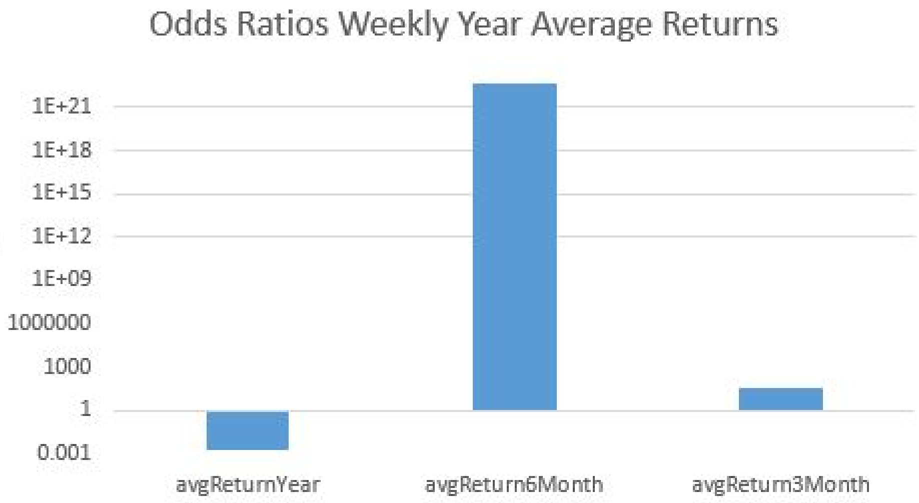
ZeroR Accuracy: 51.09%
Logistic Accuracy: 58.98%

This was our initial data set. The data was derived from a year's worth of weekly price data. The core data set consists of the Wednesday closing price for every stock instance between 1/1/2014 and 1/1/2015. As such, returns are calculated on a weekly basis. For this training set, the target attribute was determined based on whether the stock went up or down in the month after the time period of the training data.

The graph below shows that the odds ratios for the prices themselves were very close to 1. This makes sense since there is nothing to be learned about a stock's future movement from just the price itself, but rather all the valuable information is contained in the movement of the price. A 5 dollar stock has no greater or less of a chance of going up as a 200 dollar stock.



This next graph shows that there are a number of weeks which an increase in price in that time period indicated greater probability of the stock rising in the two months after the period of the graph. This gives actually a rather predictable result – if the stock has been going up consistently over the past 6 months prior to the testing period, then the stock will continue to go up. One year has too much fluctuation, and while 3 months is important is not long enough to give a conclusive result.



In this model, volatility had an odds ratio of 0.98. This implies that volatility does not have too much of an influence as to the direction of a stock. This result calls for further investigation. Intuitively, this makes sense because standard deviation does not have much of an indication on the direction of trends, just the existence of a trend. However, it is relatively common knowledge in the finance world that volatility of a stock has an indication of whether or not trends hold true to that particular stock.

In summary, while this model outperforms ZeroR by about 8%, there is potential for improvement, as a 60% success rate is not practically important. This is to be expected, as there are very few different attributes included in this model. A model that only has straight returns, volume, and volatility data is not likely have a great accuracy. Including more attributes will lead to a higher success rate which is shown in the next section.

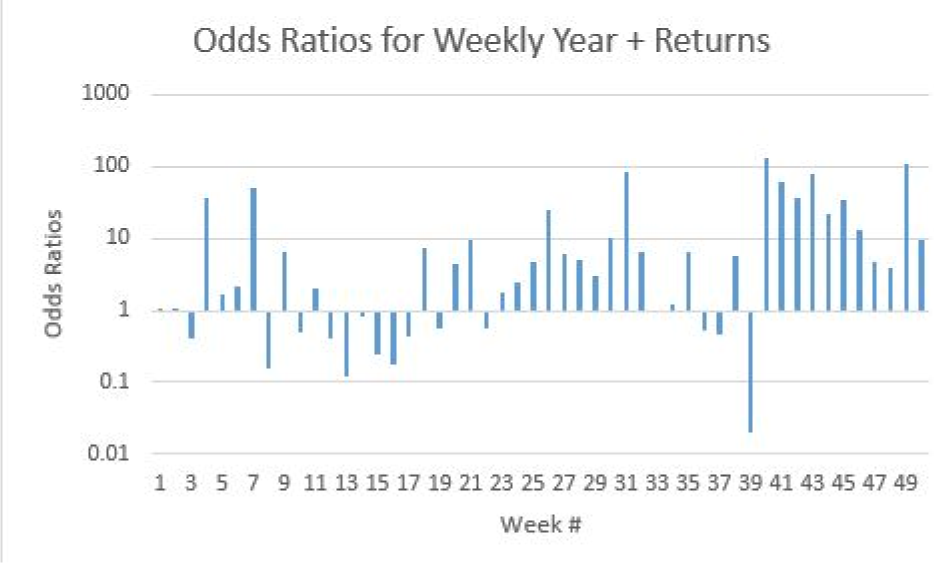
Weekly Year +

ZeroR Accuracy: 51.09%

Logistic Accuracy: 66.92%

For this data set, we used the same time frame and base attributes of Weekly Year, but with additional attributes. The common characteristic between the attributes withheld from Weekly Year but present in Weekly Year + is that they give information about the stocks general health (see data tab for more information). The predicted outcome was that there would be a higher accuracy with these attributes included, and that is what happened.

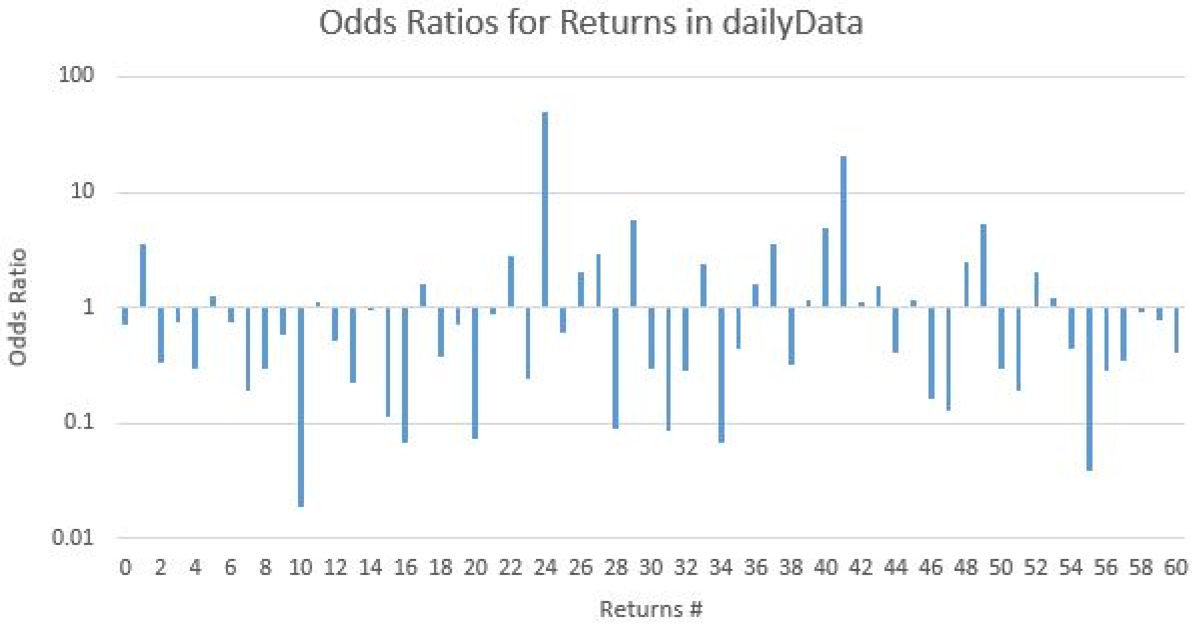
The graph below is of the odds ratio for Weekly Year +. The interesting characteristic here is that with the addition of more attributes, previously helpful attributes now have little influence on the target attribute. While we expected there to be less importance on these attributes now that more important attributes are included, such a difference is surprising. Inconsistencies across odds ratios of models on similar datasets are signs of overfitting, so one possible explanation for this could be overfitting.



In addition, another surprising result was the inverse relationship of beta to the outcome of the target attribute. If beta was high (indicating health of the stock relative to the general market), then the stock was more likely to have a downward trend the two months after the analysis period. This is especially interesting, as the intuition behind this result contradicts that of the result of Weekly Year, but both are popular investing strategies that compete with each other. One popular strategy is that if the stock is exhibiting an upward trend, then continue to invest, while another states that a stock that has gone up recently is likely to come back down as more investors sell their shares and inherently lower the price. The fact that our results obtained both methodologies is an indication that indeed, the market cannot be predicted using solely price and volume data, and that it is indeed efficient. That being said, we also wanted to observe if there were any conclusions to be made on a smaller time scale.

Daily 3 Months

ZeroR Accuracy: 51.62%
Logistic Accuracy: 55.44%



Unlike the previous graphs, we found a distinct pattern in the odds ratios for this data set. According to the graph, a greater value in returns for days 24 and 41 is more likely to indicate an increase in the stock price in the time period after this data. What?s special about these days is that these are the days on which stock options expire. Too make sure that this was the reason, we looked closer into the volume change attribute around those two days. Indeed, there was much higher change from day 23 to 24 and 24 to 25 (similarly with 41) than any other days. This is an interesting result, and could be the foundation for a future project.

However, that being said, this model returned a fairly low accuracy. We believe that the reason for this is that intraday pricing fluctuates a lot, rarely exhibiting a clear trend. This was confirmed in further exploration, as we noticed that approximately 62% of the stocks had an intraday volatility greater than the difference between the open and close prices, for more than one half the analyzed days. This is a signal that there are not very many clear trends on such a small time scale.