Seth Emery
CS5565-05
HW3 (Classification)
Feb 12, 2024

(a) Produce some numerical and graphical summaries of the Weekly data. Comment on the summaries. Are there any patterns?

data.shape() is a numerical summary of the data that shows there are 1089 rows and 9 columns of data.

data.describe() is another numerical summary of the data that shows the column headers (Year, Lag 1-5, Volume, Today) and the rows are (Count, Mean, STD, Min, 25%, 50%, 75%, Max)

Using matplotlib I made a scatter plot that plots 'Year' on the x-axis and 'Volume' on the y-axis. From this plot I can see a pattern that as years have gone on, volume has increased. This suggests that more trading is happening in the markets as time has passed.

Using matplotlib I made a bar chart plotting the means of the Lag (1-5) returns. From this plot I can't see any patterns that show much of a difference between the returns in different lag weeks.

I used this site to change my y-axis range to zoom in on the data:

https://www.scaler.com/topics/matplotlib/matplotlib-set-axis-range/

Next I make a box plot to show the median and quartile ranges and outliers of the lag returns data. This plot shows that there is always a positive mean for the returns of the lags, which shows that stocks go up (generally) over time.

Lastly I make a correlation map of the variables by using Seaborn. The main takeaways from the map are noting the high correlation between 'Year' and 'Volume' and this pattern was already detected in the scatter plot I made. Another pattern to note is that there is low correlation between the lag returns and the 'Year' variables, which means that these returns don't depend on a specific 'Year' or 'Year' range. So it could be reliable to see returns in the stock market on average no matter what 'Year' it is.

In stocks, some years are good and some years are bad, but this map shows a negative correlation between weekly lag returns and years, so it seems that in general stocks tend to go up regardless of the year.

link to chatGPT conversation for help to make the correlation map:

https://chat.openai.com/share/d1765b06-cdd8-4363-b110-19f8a2bb4054

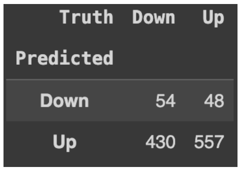
Seth Emery
CS5565-05
HW3 (Classification)
Feb 12, 2024

(b) Use the complete data set to perform a logistic regression with Direction as the response and the five lag variables plus Volume as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?

The lag 2 predictor has a p-value of .030 (<.05) and is therefore statistically significant. This might be a good predictor to make a model around.

(c) Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix tells you about the types of mistakes made by logistic regression.

The confusion matrix explains the number of true and false positives and negatives. In this context it displays the number of times the model's prediction was consistent with the true value. The matrix predicted 'Down' 54 times when the truth was 'Down' and predicted 'Down' 48 times when the truth was 'Up'. The model predicted 'Up' 430 times when the truth was 'Down' and predicted 'Up' 557 times when the truth was 'Up'



Truth	Down	Up
Predicted		
Down	54	48
Up	430	557

(h) Which of these methods provides the best results on this data?

Accuracy: (TP + TN) / (TP + TN + FP + FN) - the proportion of true results among the total number of cases examined.

Logistic regression model has an accuracy of **.4615** on the test dataset. LDA has an accuracy of **.4615** on the test dataset. KNN has an accuracy of **.5096** on the test dataset. Naive Bayes has an accuracy of **.4519** on the test dataset.

These accuracy numbers show that the KNN method has the best results on this dataset. You may be able to do calculations on precision, recall, and F1 score to get a bigger picture of how each method performs.

Seth Emery
CS5565-05
HW3 (Classification)
Feb 12, 2024

Sources:

y-axis zoom:

https://www.scaler.com/topics/matplotlib/matplotlib-set-axis-range/

chatGPT thread for making the correlation map:

https://chat.openai.com/share/d1765b06-cdd8-4363-b110-19f8a2bb4054

ISLP dataset:

https://islp.readthedocs.io/en/latest/datasets/Weekly.html

Slides and In-Class Lecture Notes:

Dr. Baffour

Notebooks from In-Class Lab on Classification and Resampling Methods