Predicting Weekly Returns with KNN and Comparing to Logistic Regression, LDA, and QDA

Data Analysis Assignment Lesson 9

David Chen

March 31st, 2020

Introduction

In this report, we will be using the `Weekly` dataset to predict if the S&P 500 markets had positive or negative returns for the week. Since we are trying to predict a binary response (‘Up’ or ‘Down’ for weekly returns), we will begin by applying a KNN (K-Nearest Neighbors) model, and then compare it to our previous results when applying logistic regression, linear discriminant analysis (LDA), and quadratic discriminant analysis (QDA). Since `Lag2` was the only significant predictor previously, we used `Lag2` as the only predictor again for the initial KNN model. We found that while this model did not perform better than the logistic regression / LDA model in terms of accuracy or sensitivity, it had a higher specificity. In the end, the overall best model was determined to be the KNN with K = 10 and `Lag2` as the only predictor.

Data

Taken from DA7

For this analysis, we will be analyzing the `Weekly` dataset. As listed from the online documentation, this dataset provides the S&P 500 weekly percentage returns from 1990 to 2010. The following variables are included:

1. `Year` - Year the data was observed.
2. `Lag1` - Percentage return for the previous week
3. `Lag2` - Percentage return from two weeks ago
4. `Lag3` - Percentage return from three weeks ago
5. `Lag4` - Percentage return from four weeks ago
6. `Lag5` - Percentage return from five weeks ago
7. `Volume` - Volume of shares traded
8. `Today` - Percentage return for the week.
9. `Direction` - Whether the market had positive (‘Up’) or negative (‘Down’) returns for the week.

Exploratory Data Analysis

We begin by examining a summary of all the variables. From the summary statistics, we observe that all the `Lag` variables are fairly consistent with each other. We can also note the distribution of `Direction`, how 55.6% of the data is `Up`.

```
##       Year           Lag1             Lag2             Lag3
##  Min.   :1990   Min.   :-18.20   Min.   :-18.20   Min.   :-18.20
##  1st Qu.:1995   1st Qu.: -1.15   1st Qu.: -1.15   1st Qu.: -1.16
##  Median :2000   Median :  0.24   Median :  0.24   Median :  0.24
##  Mean   :2000   Mean   :  0.15   Mean   :  0.15   Mean   :  0.15
##  3rd Qu.:2005   3rd Qu.:  1.40   3rd Qu.:  1.41   3rd Qu.:  1.41
##  Max.   :2010   Max.   : 12.03   Max.   : 12.03   Max.   : 12.03
##       Lag4             Lag5            Volume          Today
##  Min.   :-18.20   Min.   :-18.20   Min.   :0.09   Min.   :-18.20
##  1st Qu.: -1.16   1st Qu.: -1.17   1st Qu.:0.33   1st Qu.: -1.15
##  Median :  0.24   Median :  0.23   Median :1.00   Median :  0.24
##  Mean   :  0.15   Mean   :  0.14   Mean   :1.57   Mean   :  0.15
##  3rd Qu.:  1.41   3rd Qu.:  1.40   3rd Qu.:2.05   3rd Qu.:  1.40
##  Max.   : 12.03   Max.   : 12.03   Max.   :9.33   Max.   : 12.03
##  Direction
##  Down:484
##   Up  :605
##
##
##
##
```

Next, we examine a correlation plot between all the variables except Distribution (given that it is binary). From Figure 1, we can clearly see that only `Volume` and `Year` has a large correlation coefficient, the rest of the values seemingly below 0.1. Investigating further, we see in Figure 2 that `Volume` and `Year` appear to have an exponential relationship, and thus a log transformation may be beneficial. We examine this relationship in Figure 3, applying a log transformation to Volume and comparing it to `Direction`, the variable we will attempt to predict. This transformation does seem to be beneficial.

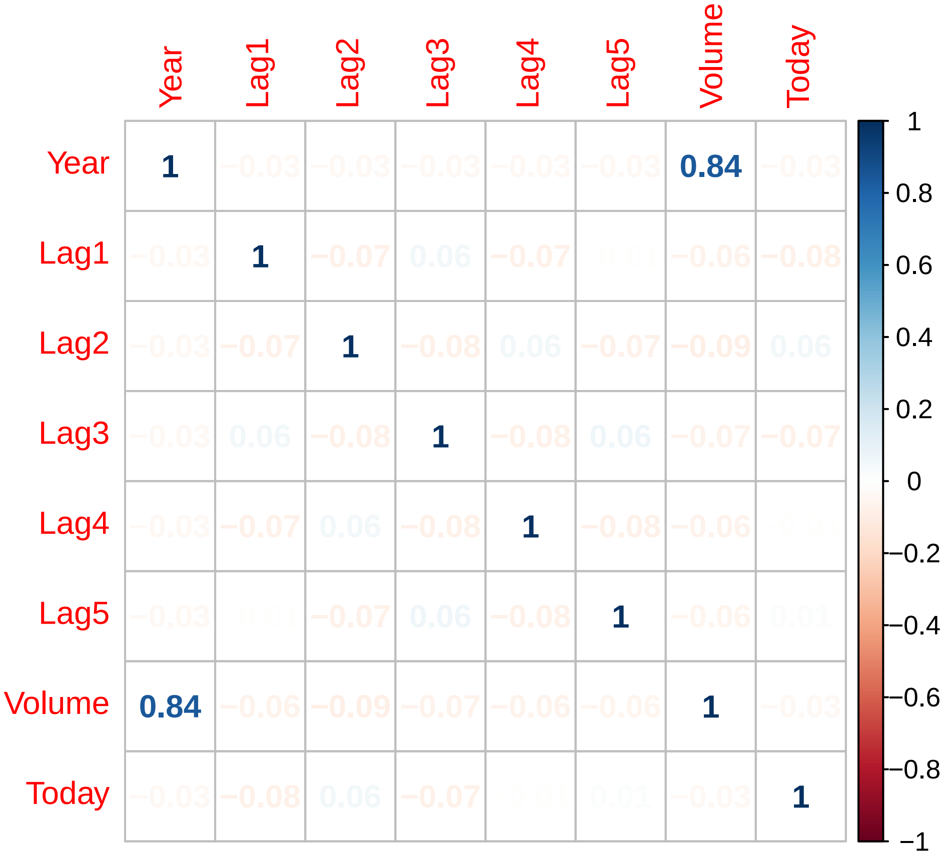

Figure 1: Weekly Dataset Correlations

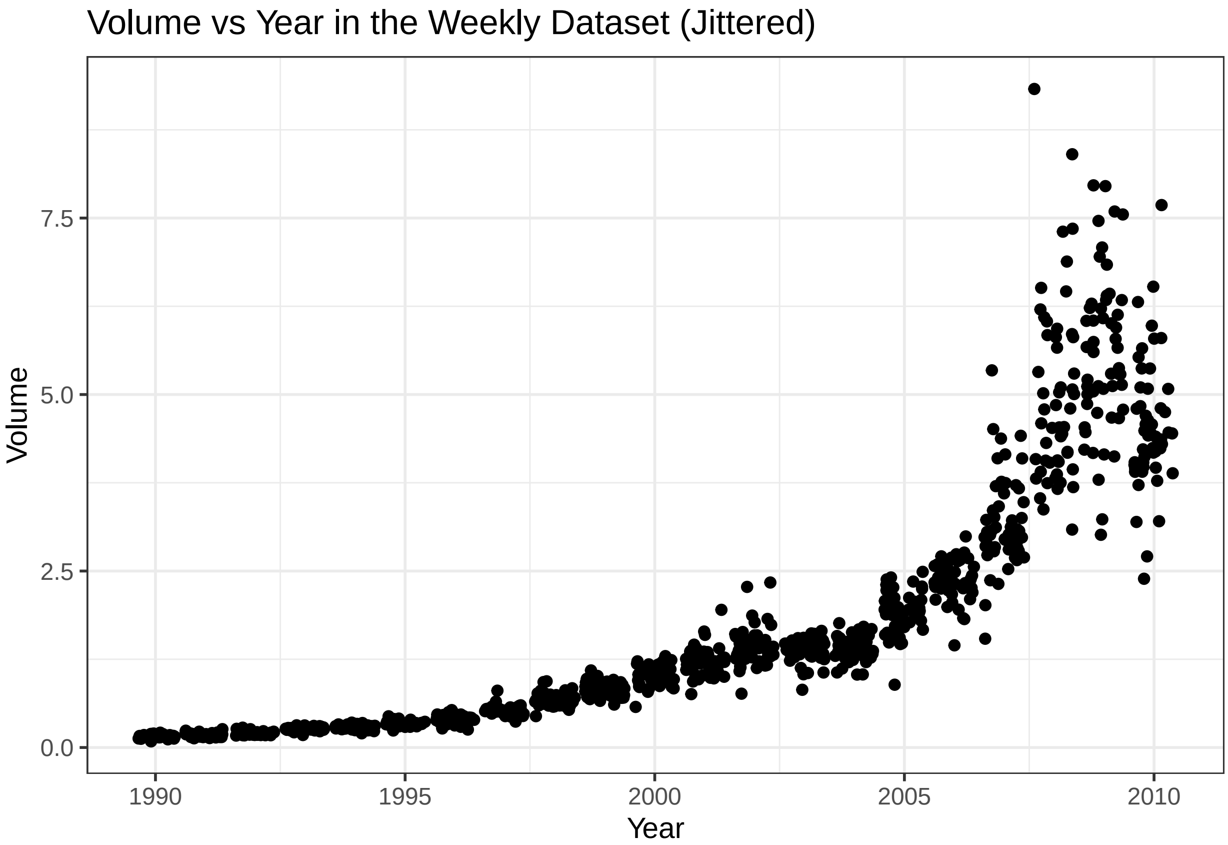

Figure 2: Volume vs Year Scatterplot

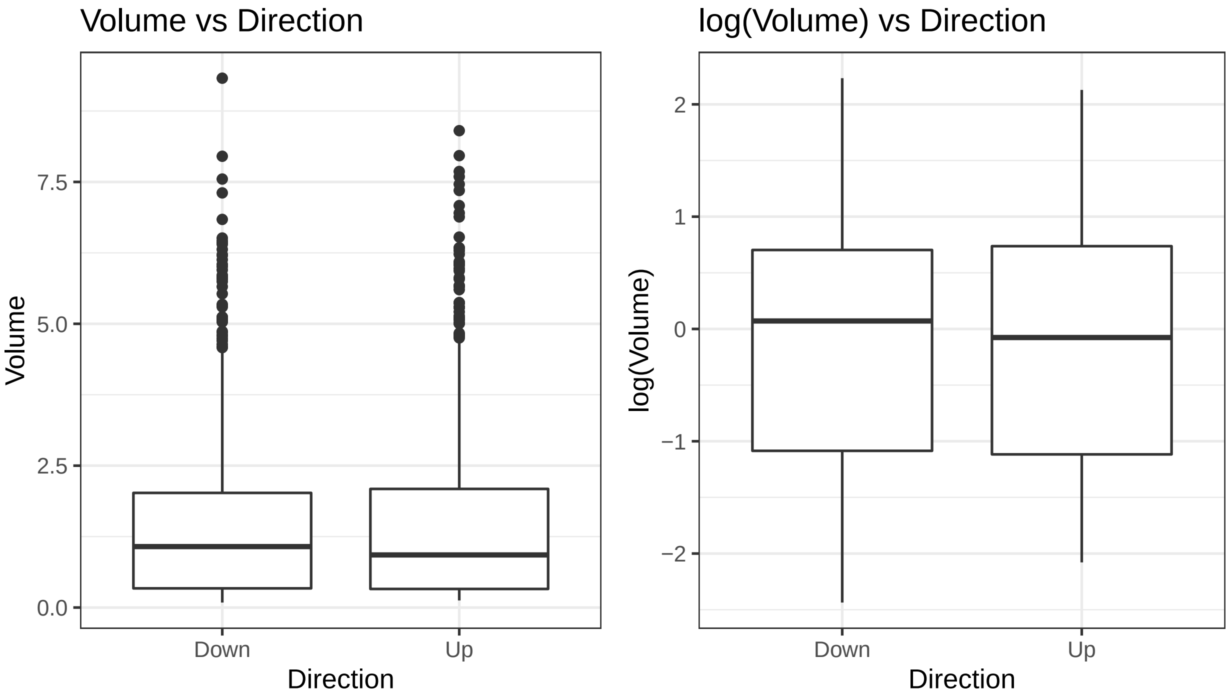

Figure 3: Volume vs Direction boxplots

Analysis

Training/Testing Sets

Just like in DA7, we will split the data into training and testing sets to observe our model's behavior on new data. The training set consists of the data from 1990-2008, while the testing set contains 2009-2010.

KNN

We begin by conducting a KNN with k=1 to predict the stock market direction with only `Lag2` as a predictor. We observe from Table 1 that the model almost exactly gets half of all true positives correct, and half of all true negatives. Thus, we see that it has an accuracy of 0.5, a sensitivity (true positive rate) of 0.525, and a specificity (true negative rate) of 0.488.

Table 1: KNN with K = 1 and Lag2 only

	True Down	True Up
Down	21	30
Up	22	31

Comparing to Logistic Regression, LDA, and QDA

Next, we obtain the results from the previous models conducted in DA7 and compare them to the KNN in Table 2. Immediately, we observe that the Logistic Regression and the LDA models were substantially better than the KNN in terms of accuracy and sensitivity. However, the KNN model had a sensitivity over double that of the other models, much more balanced the others.

Table 2: Model Comparison

	accuracy	sens	spec
LR	0.625	0.918	0.209
LDA	0.625	0.918	0.209
QDA	0.587	1.000	0.000
KNN	0.500	0.525	0.488

Other KNN Combinations

Lastly, we try a few more combinations of predictors and K values to see if any other KNN model would be more effective. To begin, we maintained K=1, but included the remaining `Lag` variables. From Table 3, we set that while this does have some improvements over the `Lag2` only model, we want to explore more options.

Table 3: Knn with K = 1 and all Lag variables

	Down	Up
Down	21	28
Up	22	33

Next, we experiment with different values of K for the KNN model. Using the caret package, we find that based on a 10-fold CV resampling and picking K based on accuracy (upwards of K=10), that K = 7 provides the best results. However, this model, as shown in Table 4, provides worse results than the one obtained previously, so we stop here. This model does seem to improve accuracy, and so we investigate further.

Table 4: KNN with K=7 and all Lag variables

	Down	Up
Down	20	22
Up	23	39

Lastly, we return to the `Lag2` only model, and again check for the optimal value of K based on 10-fold CV resampling and accuracy. Here, we find that the max value of K, K = 10, resulted in the best accuracy (note that the max value of 10 was to prevent overfitting). As shown in Table 5, there is a remarkable improvement in sensitivity, jumping up to 0.705 from the K = 1 model value of 0.525. It also presents a stronger accuracy than the larger model previously, and so we conclude with this model.

Table 5: Knn with K = 10 and Lag2 only

	Down	Up
Down	20	18
Up	23	43

To show this new model relative to the previous ones, we recreate Table 2 and include the new model. Again, we can note it's overall strong performance.

Table 6: Model Comparison with New

	accuracy	sens	spec
LR	0.625	0.918	0.209
LDA	0.625	0.918	0.209
QDA	0.587	1.000	0.000
KNN	0.500	0.525	0.488
KNN_adj	0.606	0.705	0.465

Conclusion

In conclusion, when predicting the stock market direction with just `Lag2` and k = 2, the KNN model performed substantially worse in accuracy and sensitivity compared to the Logistic Regression and LDA models, although the specificity was better. While experimenting with different values of K and total predictors, we found that with K = 10, the model with just `Lag2` vastly improved it's performance. With an accuracy of .606, a sensitivity of .705, and a specificity of 0.465, it seems to be the most well rounded model we have obtained so far. Given that we value specificity, we would choose this model over the Logistic Regression/LDA.