HW#4 Write-Up

**Question 2**

The goal for this problem is to analysis the Diabetes dataset. This dataset contains 145 observations and 6 variables. The 6 variables contain 5 numeric vectors of levels of measurement such as glucose, weight, and a categorized variable: group.

**Graphical user interface, text, application, chat or text message

Description automatically generated**

The first step is to set the work directory, load the Diabetes dataset, and preprocess the data. I preprocess the dataset by convert Diabetes to diabetes for convenience, then I check to see if there are any missing data, which there is none. Then, I take a closer look at our dataset such as summary(diabetes), head(diabetes), View(diabetes).

**A picture containing text

Description automatically generated**

**Calendar

Description automatically generated**

Base on the plot, we can see that different group of people are split up on the graph of glufast and glutest. And that is not surprising because as glucose level increase, the people will become more diabetic. Furthermore, I use the covariance matrices to take a closer look at each group of people in the dataset. It turns out that the glufast and glutest of an overt diabetic person is 22623, compare to chemical diabetic 319, and normal 86. Another variable that has major difference between different type of group is glutest and instest. Throughout this covariance matrices, we can identify difference in a more detail way.

**A picture containing text, newspaper

Description automatically generated**

**Graphical user interface, text, application

Description automatically generatedText

Description automatically generated**

In order to perform LDA and QDA, I first split our dataset into test and training set of 80% and 20%. Next, I am making sure some required packages are installed. I then perform the LDA and the result I came up is a 20.69%of error rate, whereas QDA has an error rate of 13.79%. This answer tells us that if we are given all the five other variables in the dataset and the error rate of predicting the group variables are 20.69% and 13.79%. Further demonstration is provided in the next section.

**Graphical user interface, text, application, email

Description automatically generated**

Finally, in part c, we are given one individual without the group classification. Our goal is to use LDA and QDA to predict the category of the given data. The first step is to insert the given data into a data.frame and line up each data with the correct variable. Then, to preform **LDA** use the lda.fit function previously defined above and the result is **Normal**. For **QDA** the result is **Overt\_Diabetic**. The result shows that the individual is predicted to be categorized as Normal under LDA and Overt\_diabetic under QDA.

**Question 3**

The goal of this problem is to check if there is any pattern in the given dataset, to perform generalize linear regression with “direction” column as the response and change around the predictor variables, use confusion matrix to determine which technique of regression is the best. The “Weekly” dataset in ISLR package, contains weekly percentage returns for the S&P 500 stock index between 1990 and 2010. There are 1089 observations on 9 different variables.

A picture containing text

Description automatically generated

My first step is to download some necessary package. ISLR is to load the Weekly dataset, and corrplot is for later usage. Secondly, I preprocess the data by convert Weekly into weekly for convenience, follow by check to see if there is any missing data which turns out to be none. Finally, I take a look at the dataset of weekly to get a brief understanding of the dataset.

Table

Description automatically generated

Chart, scatter chart

Description automatically generatedCalendar

Description automatically generated

Next, I am checking to see if there is any pattern between the variables of the weekly dataset except for direction variable. I first print out the summary of weekly, but it doesn’t really help me to observe any pattern. Furthermore, I use corrplot to plot graphic and numeric summaries to see if there is any pattern. Unfortunately, only volume index shows a pattern with the year index because as economics grow over the years, the volume will increase too. This is understandable because, stock is unpredictable in general, therefore it makes sense if we cannot observe any pattern other than a year increase the volume increase.

Table

Description automatically generated

I then perform a logistic regression using generalize linear model with Direction as the response and everything except for Year and Today as the predictors. After printing out the result, I observe that Lag2 is significant at alfa = 0.058. The result of this shows that the Direction variable has a relationship with Lag2 although it is not a strong significancy.

Graphical user interface, text

Description automatically generated

Next step is to compute the confusion matrix and overall fraction of correct predictions. Base on the confusion matrix, we can see that the column of down and up are the actual observed dataset and the rows of down and up are predicted result. As we can see from the matrix, when the direction is down, there are 54 predicted down and 430 predicted as up, therefore, the error rate is very high. However, when the actual direction is up, the predicted up is 557 and predicted down is 48, therefore after we take into account of all the observations, the rate of correct prediction is 56% where most mistake is made when we falsely predicted the trend as upward where it should’ve been downward.

Text

Description automatically generated

The problem is asking us to fit the logistic model after splitting the data into training and testing from 1990-2008 with Lag2 as the predictor. Therefore, my first step is to set the training data as data in prior of year 2009 and testing for those of after or equal to 2009. Then, in order to compute the confusion matrix and find the accuracy rate, I first fit a general linear model of direction and lag2. As the confusion matrix shown above this paragraph, we can see that when the training direction is down, the error rate is still pretty high, and when the training direction is up the error rate is lower. Therefore, after taking both predictions into account the accuracy rate for Lag2 as the only predictor is 62.5% which is higher than the previous model. This make sense because base on our generalized linear model, Lag2 is the only variable that shows a level of significancy with direction as the response. After we not include the noise of other predictors, the accuracy level would increase, but since the alfa of Lag2 is not very significant, the accuracy rate of prediction is not as high.

Graphical user interface, text, application

Description automatically generated

Next, I want to compute the confusion matrix and the correct fraction of predictions using Linear Discriminant Analysis. First, I check to see if I have MASS installed so I can use the LDA function. Then, I notice the confusion matrix I got back is give me similar result as the normal logistic regression model.

Graphical user interface, text, application

Description automatically generated

Repeat the same done for LDA and logistic regression model with KNN when k = 1, the result I got back is 50% accuracy rate. Therefore, it is not a good predictor as LDA and the logistic regression model. In summary, the best model is using LDA and logistic regression model when the predictor is Lag2 because they give the highest accuracy rate of 62.5%.