Name : Woosung Lee

Student ID: 50338213

Question 2.

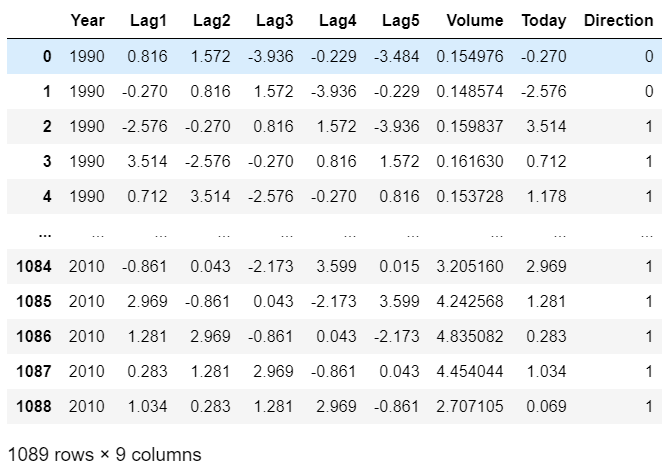
a)

Process:

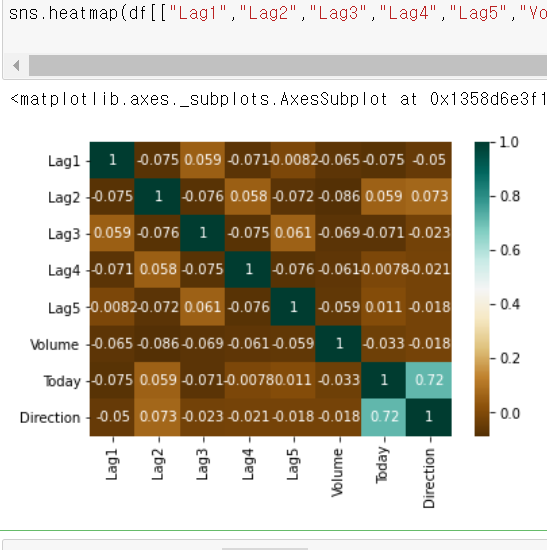
1. **Read data**: I read ‘Weekly’ dataset.
2. **change string Y(Direction) categorical data into num value**: To verify the Y(Direction) values, I used the ‘unique’ function and changed it into the number by using the ‘map’ function. **(Down=>0 , Up=>1)**
3. **Plot the correlation heatmap:** I plotted the heatmap to see correlations between factors.
4. **Plot the pairplot**: I plotted the pariplot with different colors representing the two different Direction values.
5. **Show the covariance matrix:** To check the covariance value, I made two covariance matrix by two Directions.

Output:

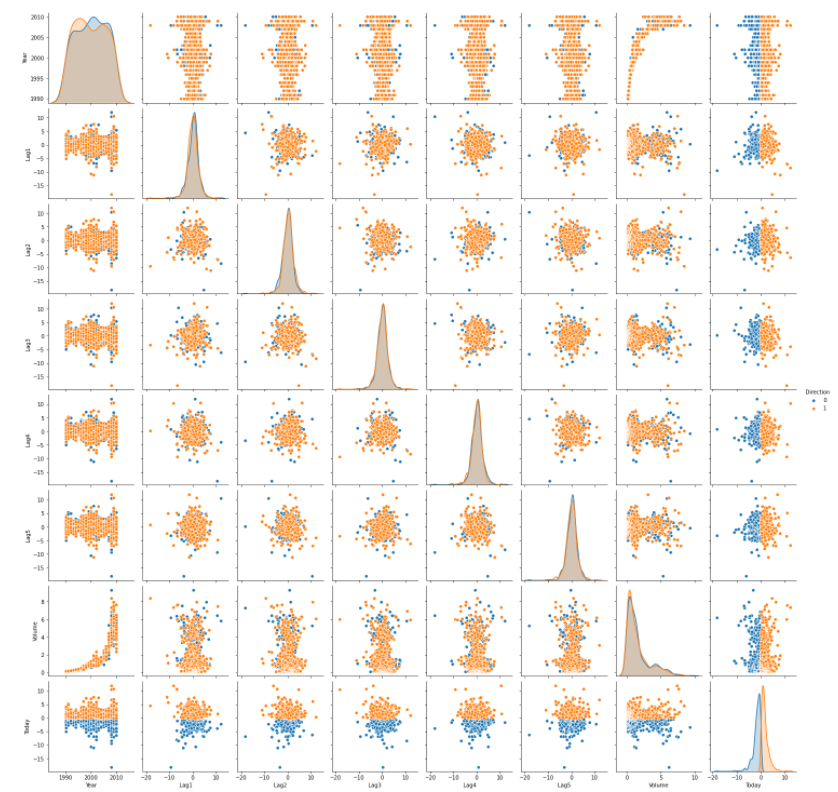
**DataFrame for the data**



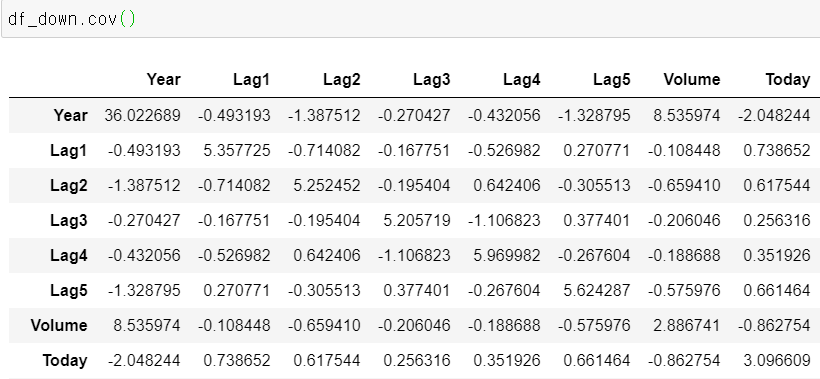
**Correlation Heatmap for the Data**



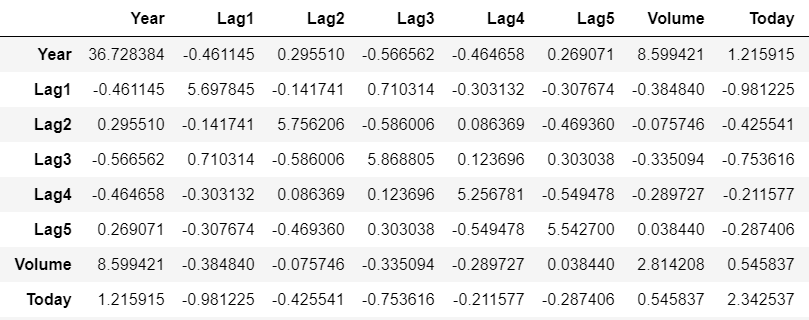
**Pairplot for the Data**



**Covariance matrix for the down direction**



**Covariance matrix for the Up direction**



Consideration:

**Correlation Heatmap for the Data**

* Most of the factor got **small** **correlation values** one another. However, ‘Direction’ and ‘Today’ have highly correlated with each other.

**Pariplot for the Data**

* When I see a distribution shape for the Direction up and down by the factors of the data, all the data have a similar distribution shape. Also, most of the Direction data **up** and **down** clusters are eclipsed. However, in the factor ‘today’ Direction data **up** and **down** seems to be easily distinguishable.

**The covariance matrix for the down direction**

* To confirm the covariance for the Direction data up and down by the factors in the data. I checked exact values with the covariance matrix. As we saw in pairplot, all the up and down data in factors have similar covariance.

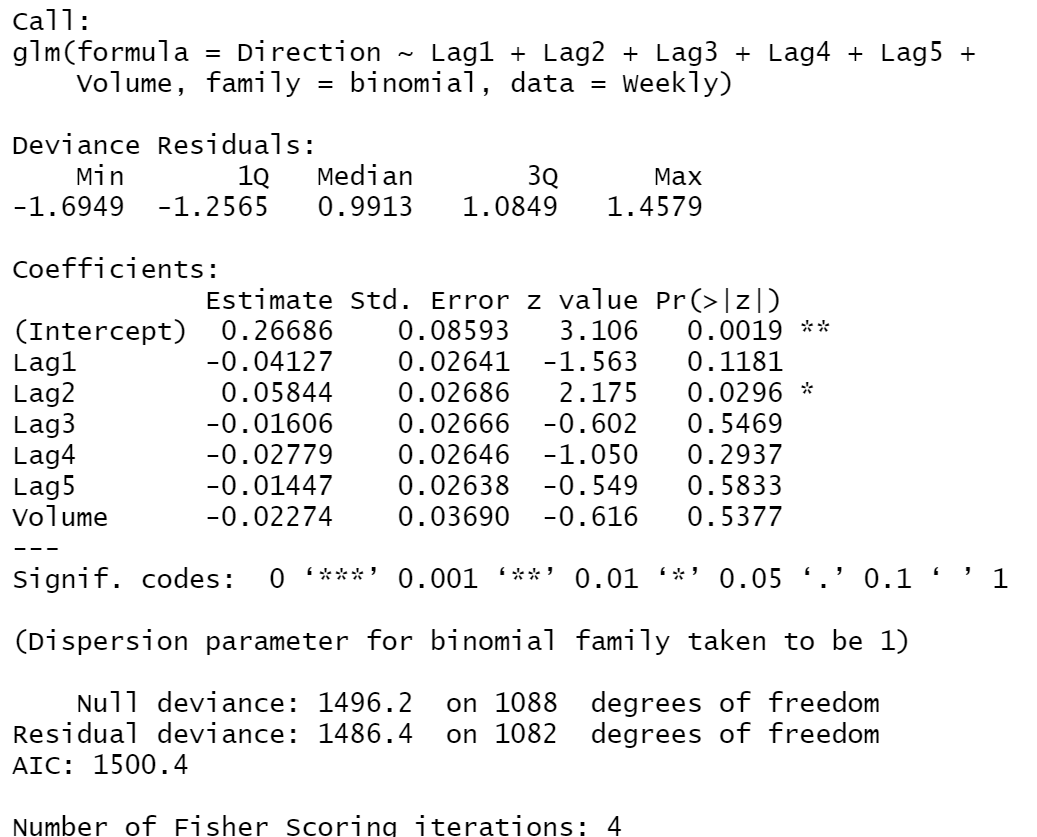
-

b)

Process:

1. **Read data**: I read ‘Weekly’ dataset.
2. **Fit the Logistic regression**: I fitted the data with “**Lag1+Lag2+Lag3+Lag4+Lag5+Volume**” factors with the ‘glm’ function and the option family as ‘binomial’ to get the Logistic regression model.
3. **Summary of the model**: I used the ‘summary’ function to see any significant factors in the model.

Output:



Consideration:

According to the ‘summary’ for the Logistic regression model, the factor ‘Lag2’ got the lowest p value which indicates that the factor ‘lag 2’ is the most significant factor in the model.

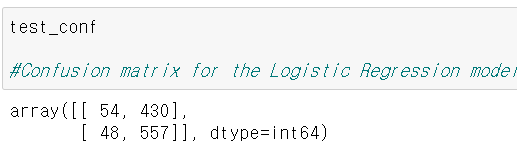
C)

Process:

1. **Continue to the Question b**:
2. **Show the confusion matrix**: With the Logistic Regression model from Question b, make predictions. After that, I made a confusion matrix to check the model accuracy.

Output

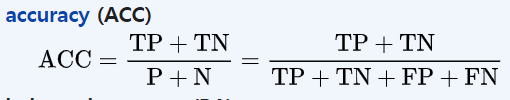
**The confusion matrix for the data from the Logistic Regression model.**



Consideration:

The ACCURACY = **correct predictions** **/ (correct predictions + incorrect predictions)**

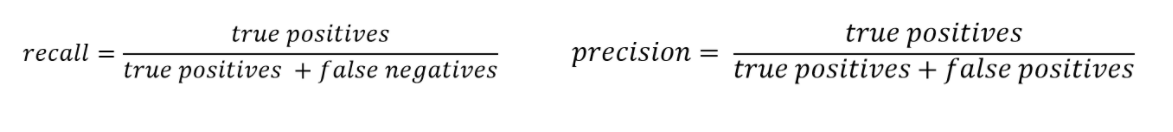
This accuracy also matches with the accuracy in the confusion matrix.



**(correct predictions + incorrect predictions)** indicate total rows which is same with (TP+TN) + (FP+FN)

**correct predictions** indicate TP + TN.

**Accuracy for the predicted data:** 557+54 / 1089 = 0.561..



**In the confusion matrix there are two values ‘Down’:0, 'Up':1**

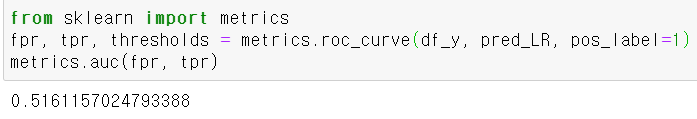
Precision (**Down**) = 54 / 54 + 430 = 0.11..

Precision (**Up**) = 557 / 48+ 557 = 0.92..

Recall(**Down**) = 54 / 54 + 48 = 0.529..

Recall(**Up**) = 557 / 430+ 557 = 0.56..

**The AUC scores.**

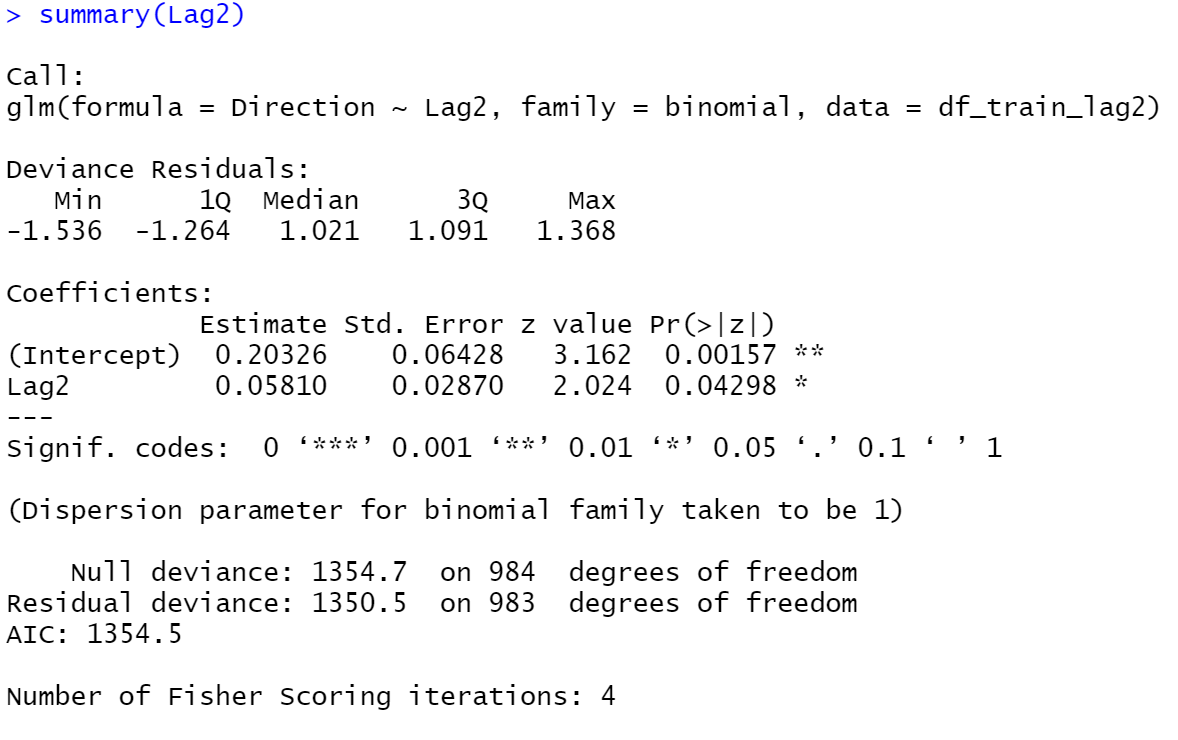


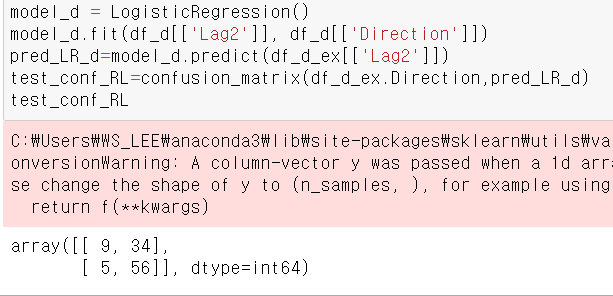
d)

Process:

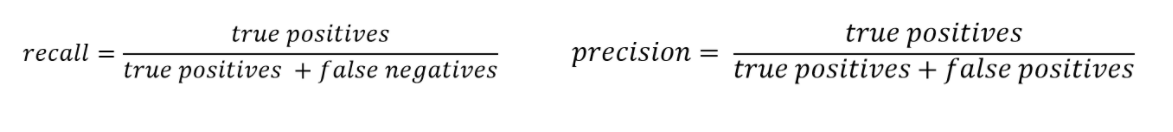
1. **Read data**: I read ‘Weekly’ dataset.
2. **change string Y(Direction) categorical data into num value**: To verify the Y(Direction) values, I used the ‘unique’ function and changed it into the number by using the ‘map’ function. **(Down=>0 , Up=>1)**
3. **Select train and test data:** By using the function ‘loc’, I select the train data which the factor ‘year’ is between 1990 and 2008. From the data that has the factor ‘year’ larger than 2008 becomes the test data.
4. **Fit the Logistic regression model with the train**: Using only one factor (Lag2) in the train data, I fitted the Logistic regression model.
5. **Show the confusion matrix:** With the fitted Logistic regression model, make a prediction list with the test dataset. Finally, I made the confusion matrix with the prediction list and Direction factor of the test data

.





**Accuracy for the predicted data:**  0.625



**In the confusion matrix there are two values ‘Down’:0, 'Up':1**

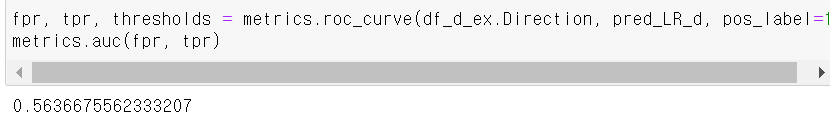
Precision (**Down**) = 0.209..

Precision (**Up**) = 0.91..

Recall(**Down**) = 0.642..

Recall(**Up**) = 0.6222…

**The AUC scores.**

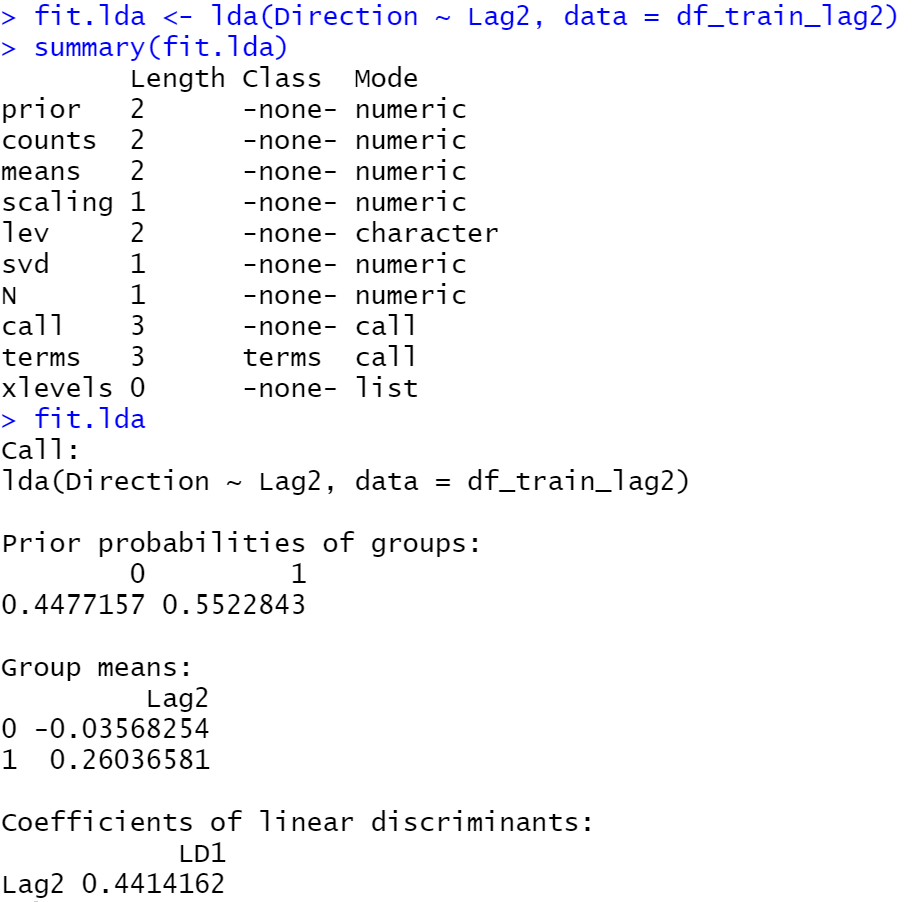


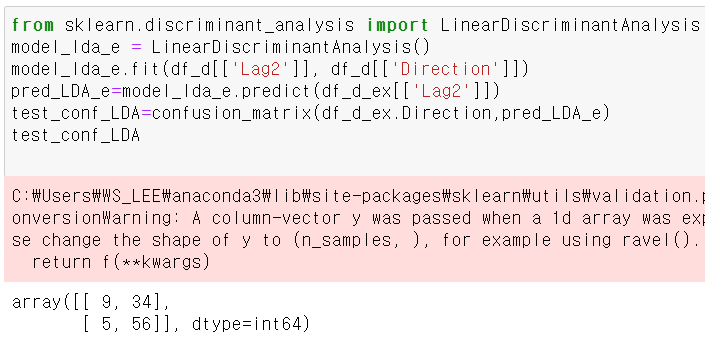
e)

Process:

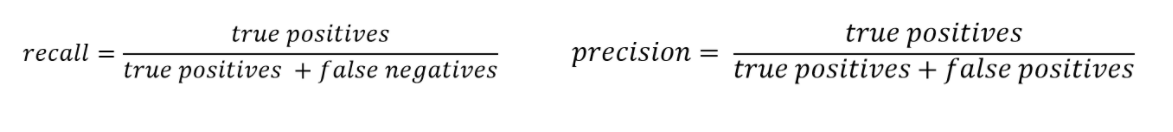
1. **Same Process with Question D**
2. **Fit the Linear Discriminant Analysis regression model with the train**: Using only one factor (Lag2) in the train data, I fitted the Linear Discriminant Analysis regression model.
3. **Show the confusion matrix:** With the Linear Discriminant Analysis regression model, make a prediction list with the test dataset. Finally, I made the confusion matrix with the prediction list and Direction factor of the test data

Outputs:





**Accuracy for the predicted data:**  0.625



**In the confusion matrix there are two values ‘Down’:0, 'Up':1**

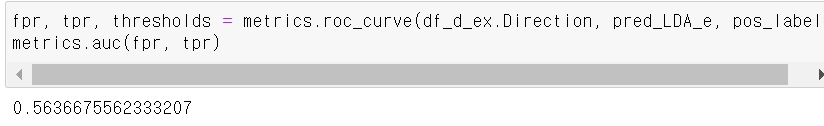
Precision (**Down**) = 0.209..

Precision (**Up**) = 0.91..

Recall(**Down**) = 0.642..

Recall(**Up**) = 0.6222…

**The AUC scores**

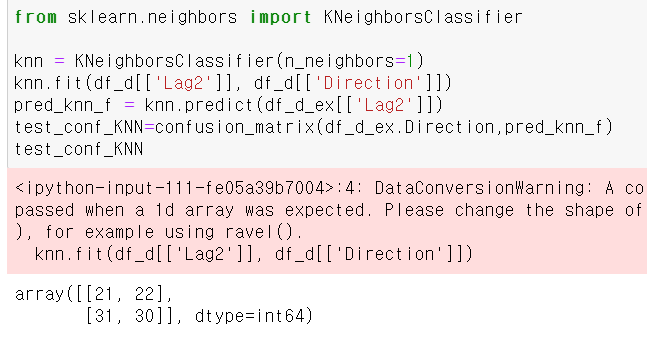


f)

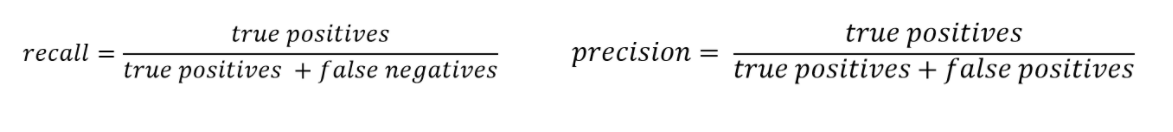
Process:

1. **Same Process with Question D and E**
2. **Fit the KNN model with the train**: Using only one factor (Lag2) in the train data, I fitted the KNN model.
3. **Show the confusion matrix:** With the KNN model, make a prediction list with the test dataset. Finally, I made the confusion matrix with the prediction list and Direction factor of the test data

Outputs:



**Accuracy for the predicted data:**  0.3923…



**In the confusion matrix there are two values ‘Down’:0, 'Up':1**

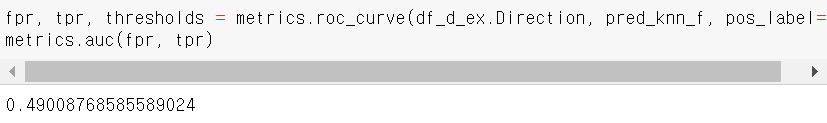
Precision (**Down**) = 0.4883…..

Precision (**Up**) = 0.49..

Recall(**Down**) = 0.403..

Recall(**Up**) = 0.576…

**The AUC scores**



g) Which method appear to provide the best results.

Consideration:

***THE LDA AND ROGISTIC MODEL with LAG2 PROVIDE THE BEST RESULTS.***

**Accuracy for the predicted data.**

1. The Logistic model (5Lag + volume): **0.561**
2. The Logistic model (2Lag): **0.625**
3. The LDA model (2Lag): **0.625**
4. The KNN model (2Lag, K=1): **0.3923**

**AUC for the predicted data.**

1. The Logistic model (5Lag + volume): **0.51611..**
2. The Logistic model (2Lag): **0.5636..**
3. The LDA model (2Lag): **0.5636..**
4. The KNN model (2Lag, K=1): **0.4900..**

According to the accuracy for all the four models, The Logistic model (2Lag) and The LDA model (2Lag) got the highest score of 0.625.

The Logistic model (5Lag + volume) **VS** The Logistic model (2Lag)

The Logistic model (5Lag + volume) got better performance than The Logistic model (2Lag). Because Lag 1,3,4,5 and the volume factor do not have information for classifying the group into ‘up’ and ‘down’. Even they may interfere with the prediction for the model.

Intuitively, it is easy to think that using more factor make a better model. In many cases, this assumption is true, however, this is only correct when the factor has fitful information for the model. We saw the correlation heat map for the data, pairplot for the data, and summary for the Logistic model(5Lag + volume). In the correlation heat map, we could not find any meaningful correlated factors with the Y(up and down). In the pariplot most of the data are eclipsed and it looks hard to make any decision boundary. In the summary for the Logistic model, the most important factor among the factors (5lags and volume) is 2Lag. In short, Lag 1,3,4,5 and the volume factor are not a useful factors for the Logistic model.

The Logistic model (2Lag) **VS** The LDA model (2Lag)

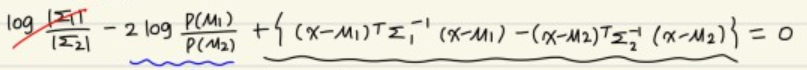
The Logistic model (2Lag) andthe LDA model (2Lag) got the same scores. We can find the reason for this from the pair plot. In the pair plot, we can see the ‘lag2-up’ and ‘lag2-down’ got exactly similar distribution. We can prove this with similarity to the two-covariance value. (5.2.. and 5.7..)

With the assumption that two groups of data have the same distribution, we can find out it’s **the division boundary became linear which is fitful to the LDA model.**

* The two group of data have same distribution (Σ1 = Σ2)

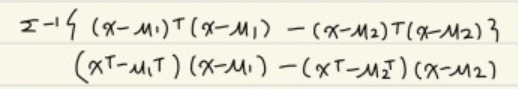
Σ1 = Σ2 = pooled variance = ((n1-1) Σ1 +(n2-1) Σ2)/ (n1 + n2 -2) = Σ

Apply it into discriminant function

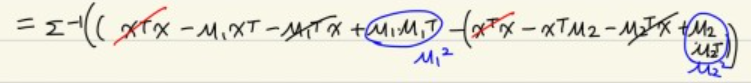


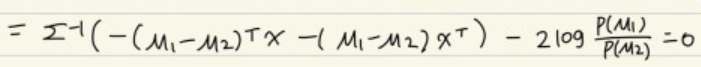
Blue line is a constant.

Rearrange the equation for the Black line with Σ-1 (notice that Σ1 = Σ**2**)



Multiply each equation and make it simpler.





The equation is become like the picture above.

This arranged equation became the linear function form as like **AX + B = 0.**

This is because the two groups have in pooled variance **(Σ1 = Σ2)**

**In this case we can see the division boundary became linear.**

**Precision and Recall for the predicted data group ‘up’.**

1. The Logistic model (5Lag + volume): **Precision: 0.92.. Recall: 0.56..**
2. The Logistic model (2Lag): **Precision:** **0.91.... Recall: 0.62..**
3. The LDA model (2Lag): **Precision:** **0.91.... Recall: 0.62..**
4. The KNN model (2Lag, K=1): **Precision: 0.49.. Recall: 0.57..**

We can see the best precision for the ‘up’ score is from The Logistic model (5Lag + volume).

The Logistic model (2Lag) and The LDA model (2Lag) got a similar score to the logistic model. However, The Logistic model (2Lag) and The LDA model (2Lag) got the exact better score in Recall than the LDA model (2Lag)

**Precision and Recall for the predicted data group ‘down’.**

1. The Logistic model (5Lag + volume): **Precision: 0.11.. Recall: 0.529..**
2. The Logistic model (2Lag): **Precision:** **0.209.... Recall: 0.64..**
3. The LDA model (2Lag): **Precision:** **0.209.... Recall: 0.64..**
4. The KNN model (2Lag, K=1): **Precision: 0.48.. Recall: 0.40..**

We can see the best precision for ‘down’ score is from The KNN model (2Lag, K=1).

The Logistic model (2Lag) and The LDA model (2Lag) got the highest score for recall.

Reference:

#https://blog.naver.com/goodgpt