This assignment should be answered using the Weekly data set, which is part of the ISLR package. This

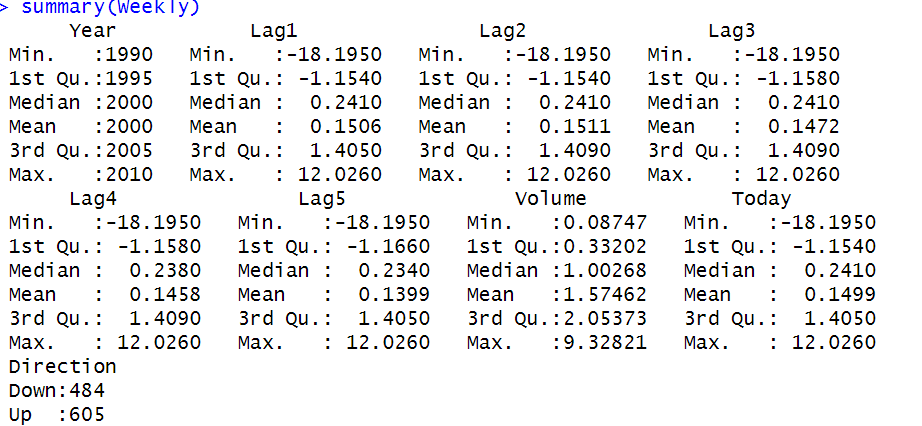
data is similar in nature to the Smarket data from this chapter’s lab, except that it contains 1,089 weekly

returns for 21 years, from the beginning of 1990 to the end of 2010. Use the first 16 years for training and

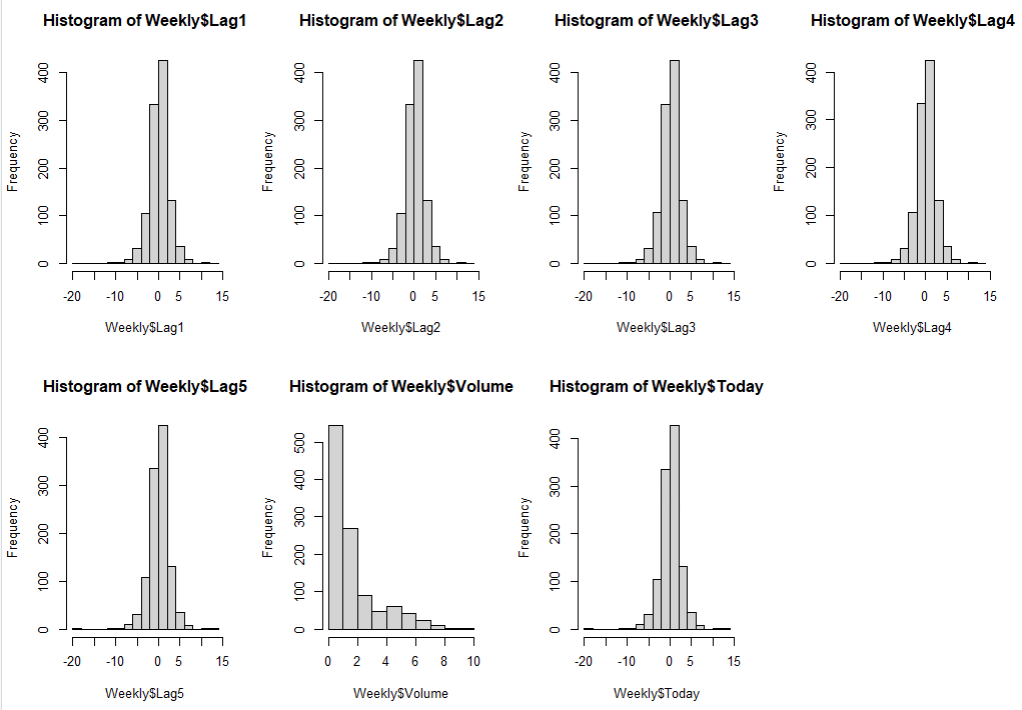
last 5 years for testing. Report results only on the test data.

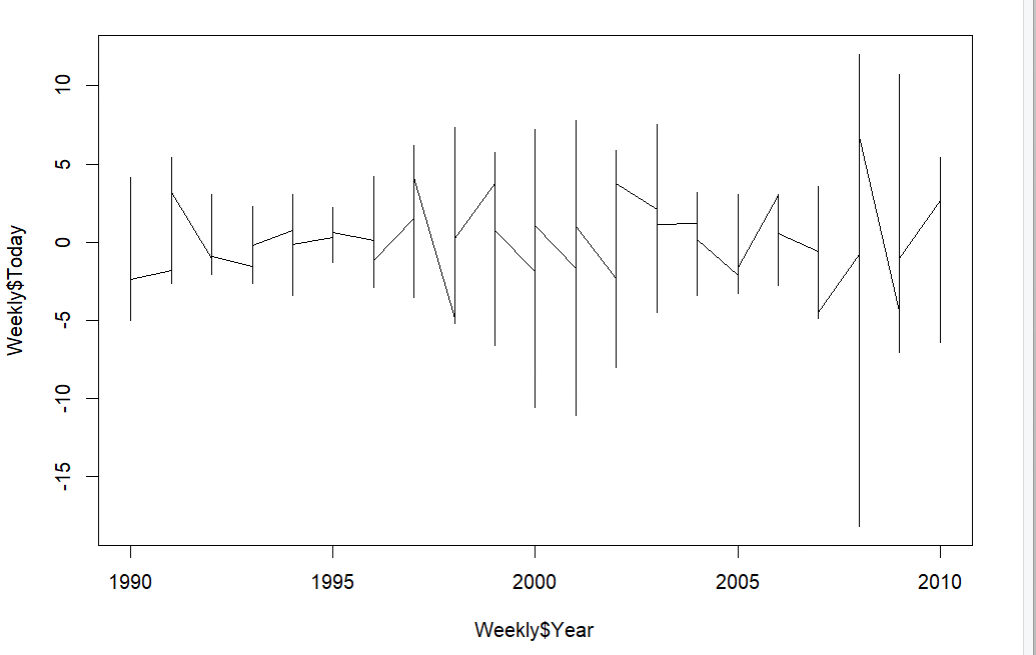
(a) Produce some numerical and graphical summaries of the Weekly data. Do there appear to be any patterns?

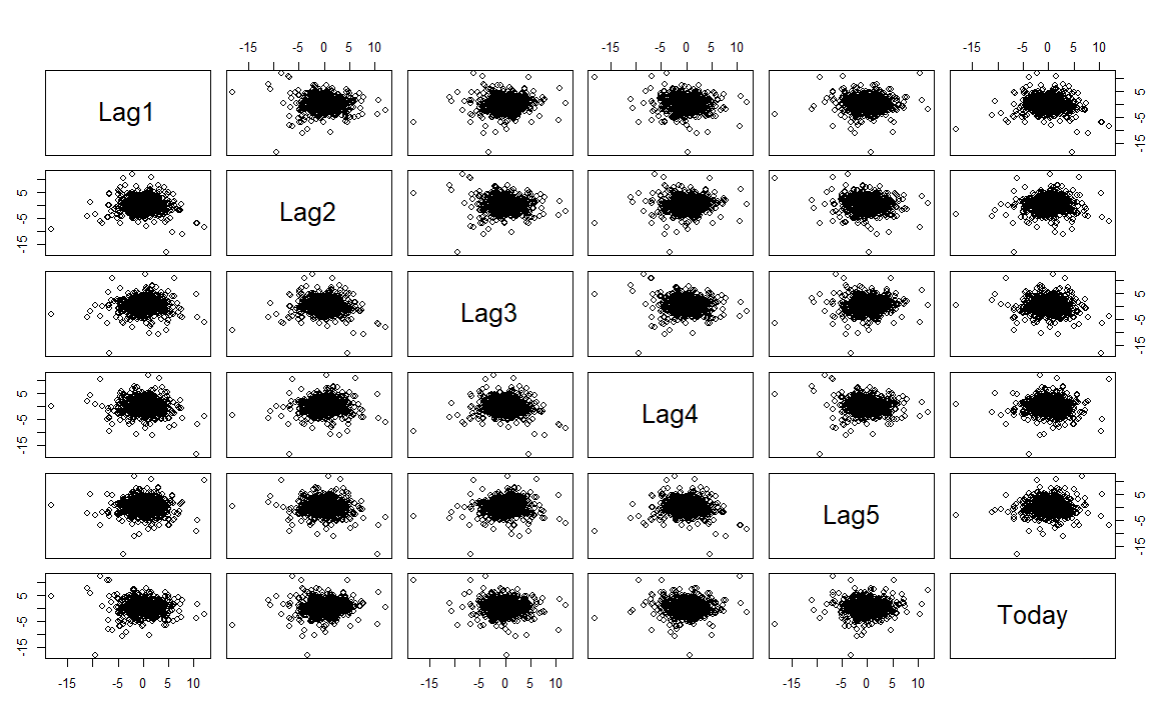
By summary the Weekly data set we can see:



By plotting histogram of each variable, we can see most of them follows normal distribution except weekly-volume is decreasing:



By plotting time series of the stock market, we can see that the stock market returns appear to have a cyclical pattern over time, with periods of high volatility followed by periods of relative stability : By plotting scatterplot matrix of the Lag1, Lag2, Lag3, Lag4, Lag5, and Volume variables we can see they are high-nonlinear relationship with each other.



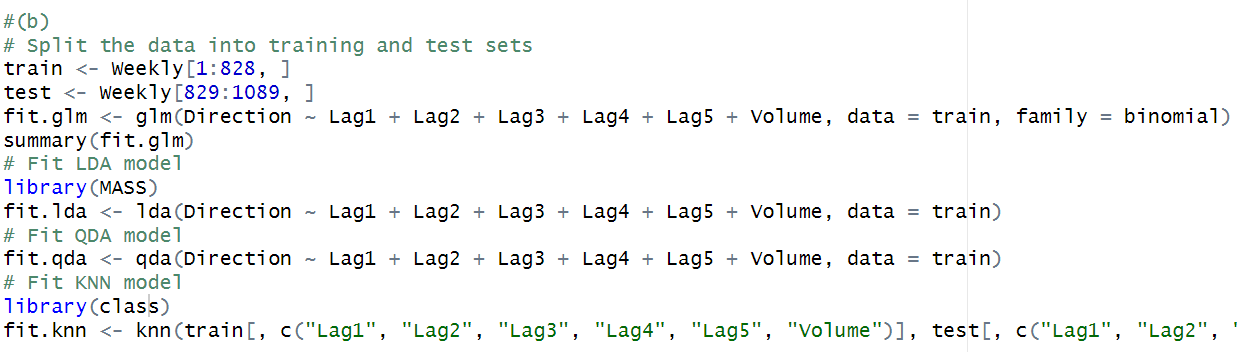
(b) Develop logistic regression, LDA, QDA, and KNN with Direction as the response and the five lag variables plus Volume as predictors.

Because for the first 16 years, data is from 1990 to 2005, the last 2005 data is in column 828, so the training data set will from 1-828, test data set will from 829 to the end (last 5 years).

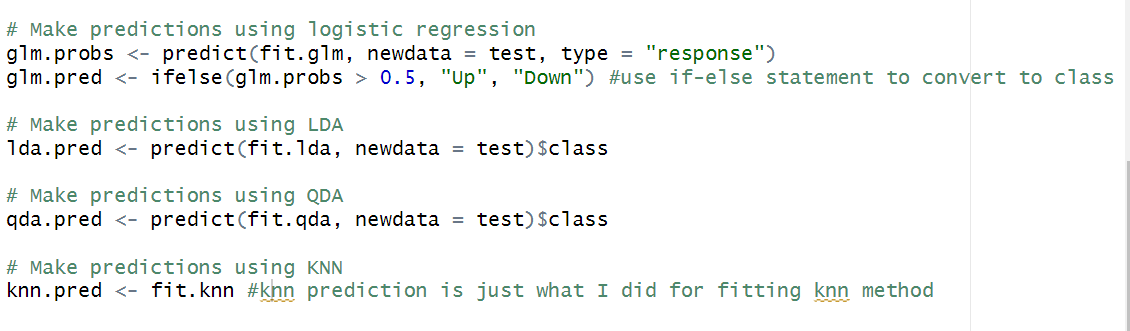
By summary the logistic regression we can see following



This splits weekly dataset into training and test sets. It then fits logistic regression, LDA, QDA, and KNN models to the training data, using Lag1, Lag2, Lag3, Lag4, Lag5, and Volume as predictors of the Direction variable.

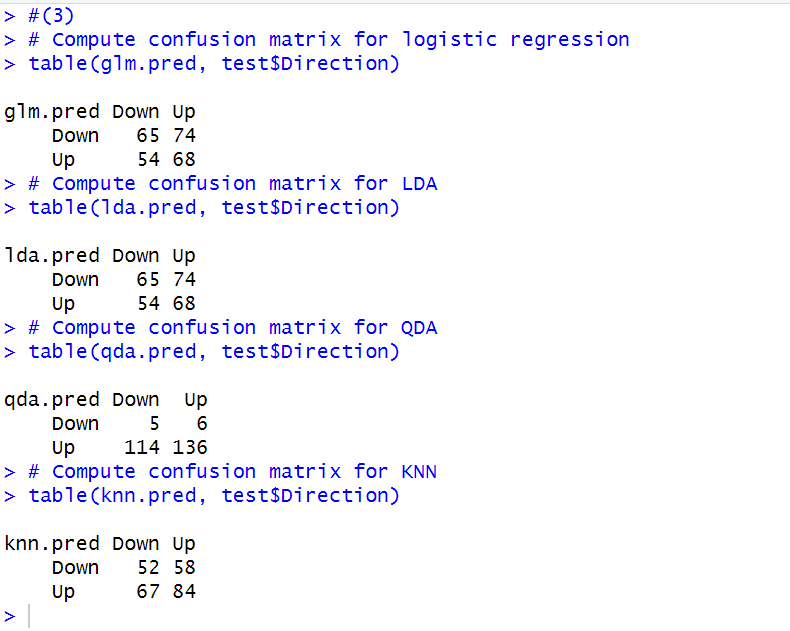


And then make predictions on the test data using each of the fitted models, and stores the predictions in glm.pred, lda.pred, qda.pred, and knn.pred, respectively.



(c) Compute the confusion matrix and interpret the results.

By using table function for each method:



Accuracy of logistic regression = (65+68)/ (65+74+54+68) = 133/261 = 0.51

Error rate of logistic regression = 1 – Accuracy = 0.49

Sensitivity of logistic regression = 68/ (54+68) = 0.5574

Specificity of logistic regression = 65/ (65+74) = 0.4676

Accuracy of LDA = (65+68)/ (65+74+54+68) = 133/261 = 0.51

Error rate of LDA = 1 – Accuracy = 0.49

Sensitivity of LDA = 68/ (54+68) = 0.5574

Specificity of LDA = 65/ (65+74) = 0.4676

Accuracy of QDA = (5+136)/ (5+6+114+136) = 141/261 = 0.4368

Error rate of QDA = 1 – Accuracy = 0.5632

Sensitivity of QDA = 136/ (114+136) = 0.544

Specificity of QDA = 5/ (5+6) = 0.4545

Accuracy of KNN = (52+84)/ (52+58+67+84) = 136/261 = 0.5211

Error rate of KNN = 1 – Accuracy = 0.4789

Sensitivity of KNN = 84/ (84+67) = 0.5563

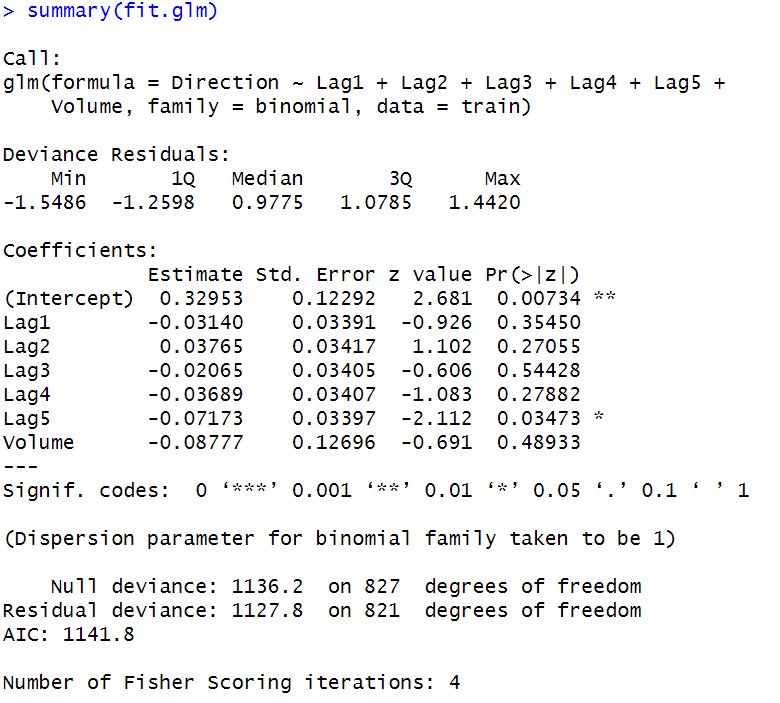
Specificity of KNN = 52/ (52+58) = 0.4727

The logistic regression, LDA, and KNN models all have similar accuracies, ranging from 0.51 to 0.5211. This means that all three models correctly predicted the direction of the market slightly more than half the time.

But the accuracy of QDA is less than 0.5, so half of the time, it didn’t predict the direction correctly.

And from above calculation, we can see that Logistic regression, LDA and KNNs’ predictions are very similar to each other. So, they are probably more effective in predicting the direction of the stock market using the given predictors.

(d) For what models it is possible to determine significant features? Determine the significant features based on at least one model.

We can determine significant features by logistic regression model. By using logistic regression model summary, we can see: 

So, let the significant level be 0.05. from these p-values, it appears that only the lag5 predictors has a p-value below 0.05, indicating that it may be a significant predictor of the response variable. Therefore, we can conclude that the significant feature based on the logistic regression model is lag5.

(e) Determine the sensitivity, specificity, accuracy, error rate for each of the models at probability thresholds of (i) 0.5, and (ii) 0.4. Present results as a single summary table covering all the models. Interpret what is happening when the thresholds are changed.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| P = 0.5 | Logistic | LDA | QDA | KNN(k = 5) |
| Accuracy | 0.51 | 0.51 | 0.4368 | 0.5211 |
| Error rate | 0.49 | 0.49 | 0.5632 | 0.4789 |
| Sensitivity | 0.5574 | 0.5574 | 0.544 | 0.5563 |
| Specificity | 0.4676 | 0.4674 | 0.4545 | 0.4727 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| P = 0.4 | Logistic | LDA | QDA | KNN |
| Accuracy | 0.5441 | 0.5441 | 0.5441 | 0.5172 |
| Error rate | 0.4559 | 0.4559 | 0.4559 | 0.4828 |
| Sensitivity | 0.5265 | 0.5265 | 0.5444 | 0.5526 |
| Specificity | 0.5 | 0.5 | 0.5 | 0.5321 |

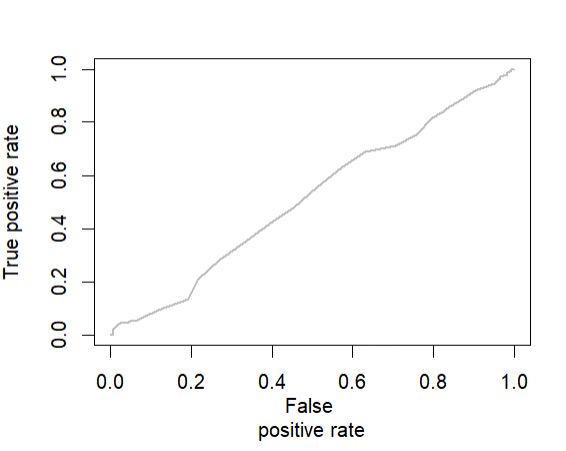
The knn function in R does not provide class probabilities, so we cannot set a probability threshold for KNN. I transfer it by calculating the predicted probabilities for the KNN model by converting the predicted class labels to numeric values using the as.numeric function.

On the other hand, if we change the probability threshold from 0.5 to 0.4, the sensitivity and specificity of each model changed, as the classification threshold for predicting the positive class (up) get lowered. This means that more instances will be classified as positive, decreasing the sensitivity but increasing the specificity.

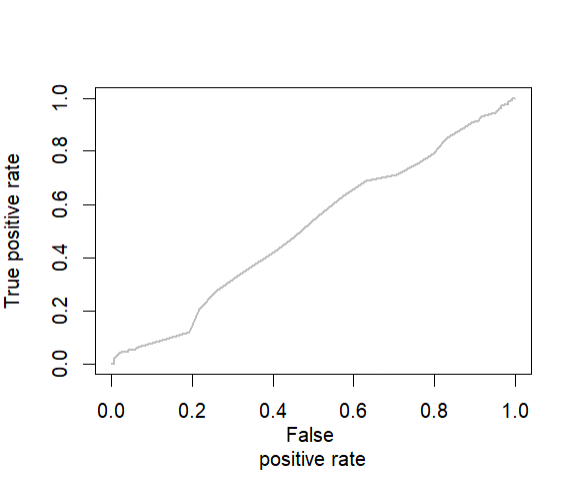
The accuracy and error rate of each model also changed, decreasing the error rate but increasing accuracy.

(f) Plot the ROC plot for each model. Based on the ROC which model has the best classification performance?

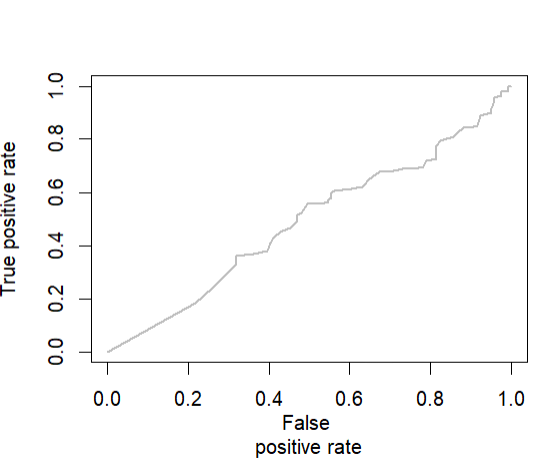
ROC for logistic regression:



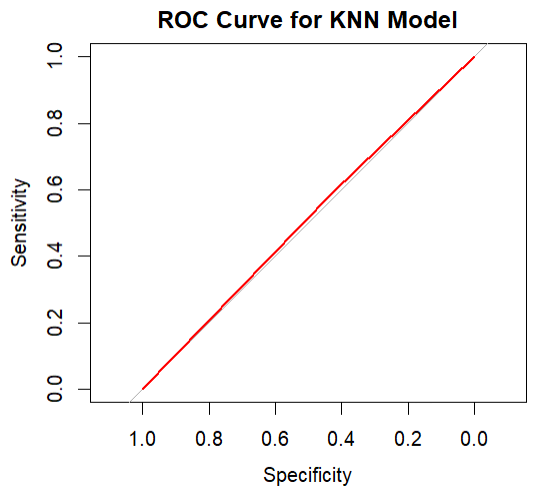
ROC for LDA:



ROC for QDA:

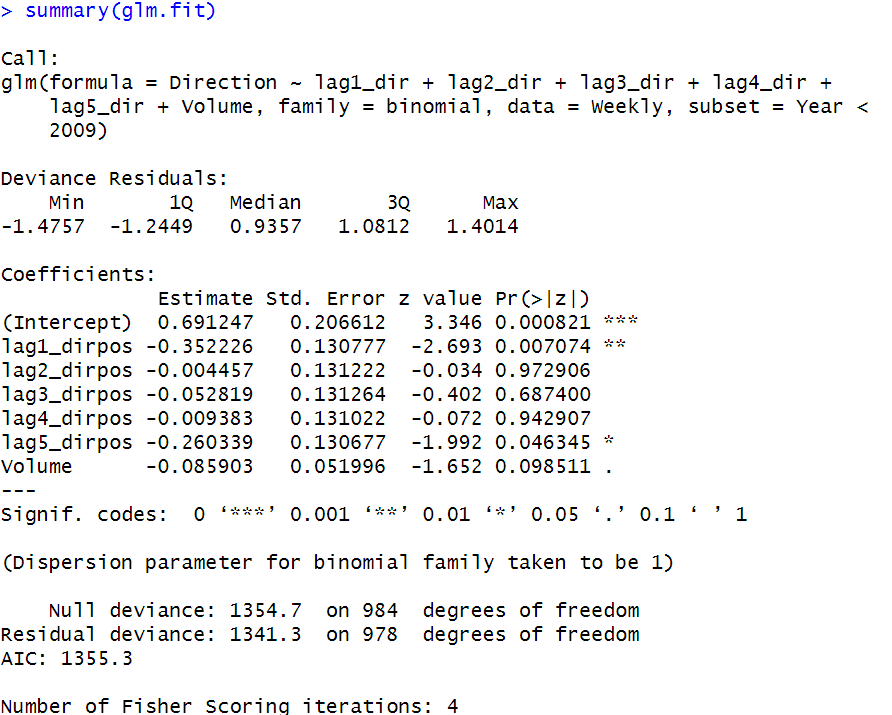


ROC for KNN:

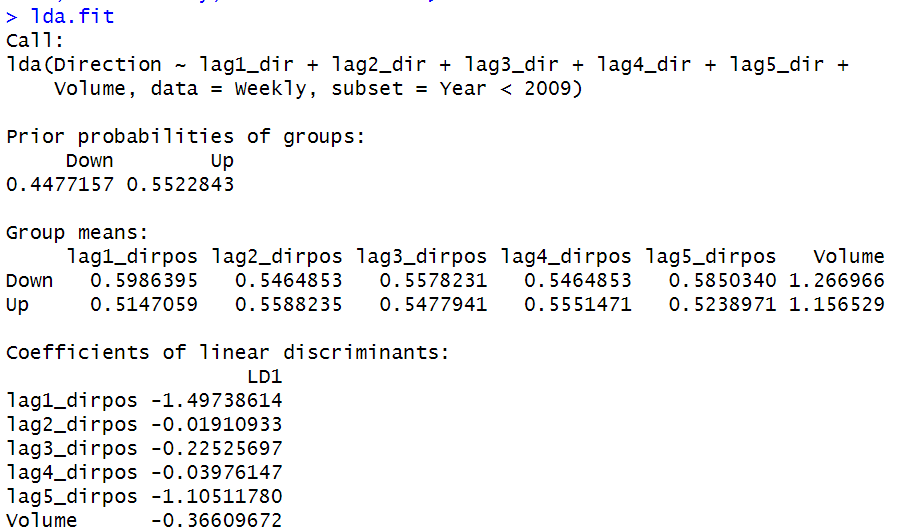


Based on the ROC curves, the logistic regression model appears to have the best classification performance, as it has the highest area under the curve value from rough view.

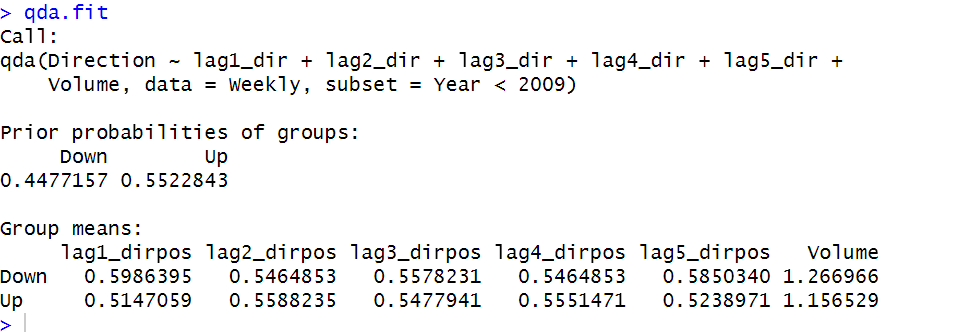
(g)Change all the lag variables to categorical based on whether they are positive or negative. Develop Logistic regression, LDA, QDA, and KNN models with the transformed lag variables and numerical Volume variable. Did the model performance improve?

Summary of logistic regression after transformed: 

Summary of LDA after transformed:



Summary of QDA after transformed:



So, we can see from the summary and output of the models that the transformed lag variables did not improve the model performance for Logistic regression, LDA, and QDA models. However, we can see that the KNN model accuracy improved from 0.5211 to 0.5709 when using the transformed lag variables. Therefore, we can conclude that the transformed lag variables did improve the performance of the KNN model.