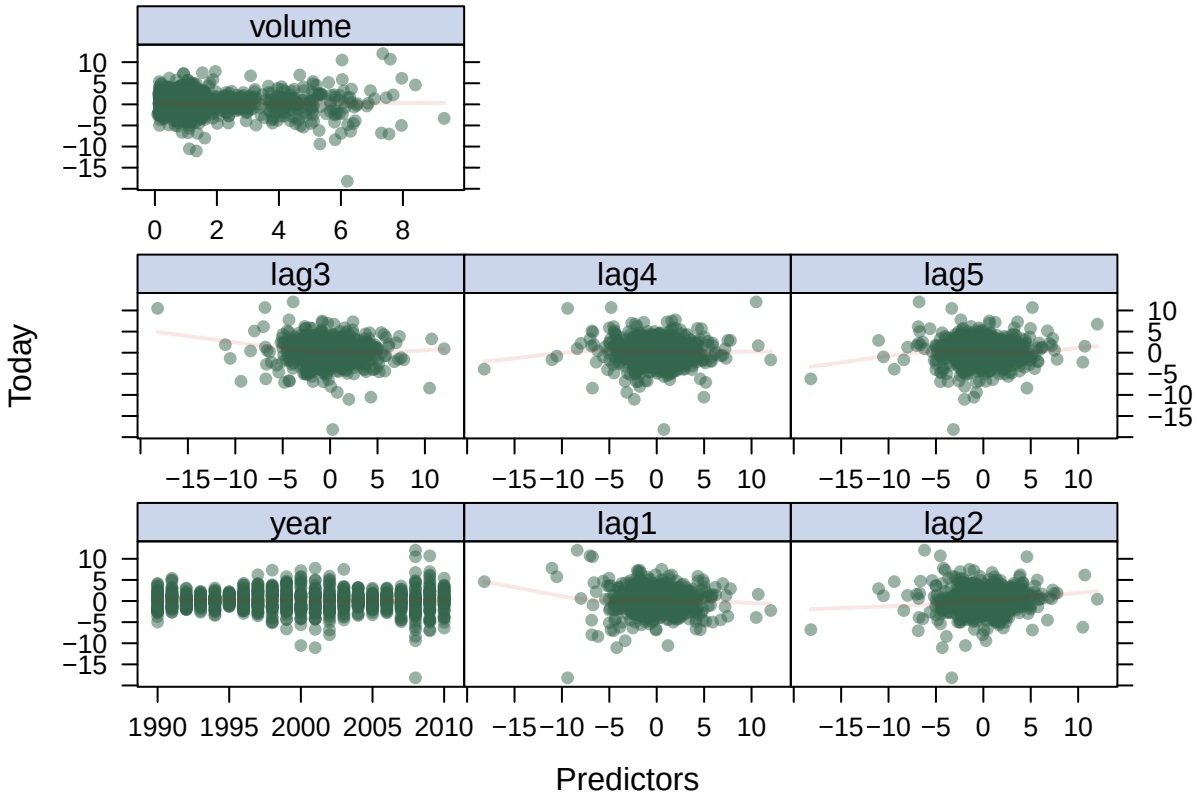
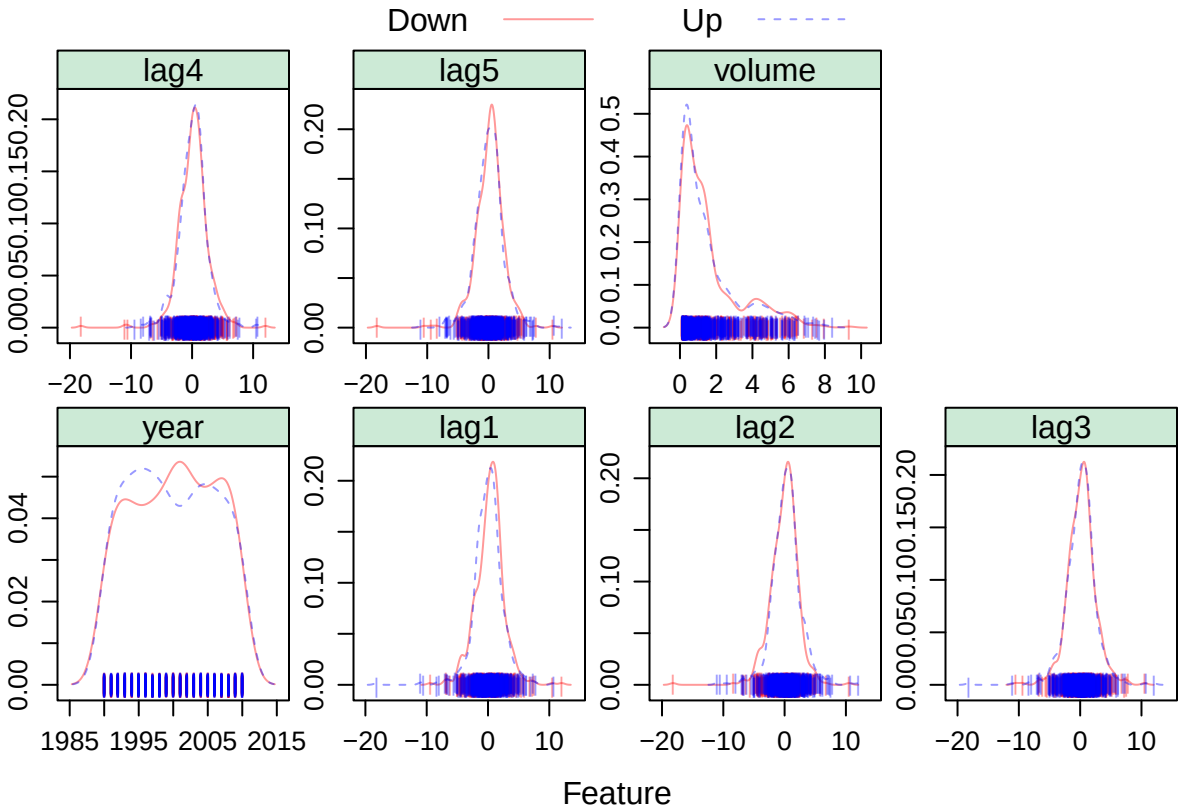
P8106 Homework 3

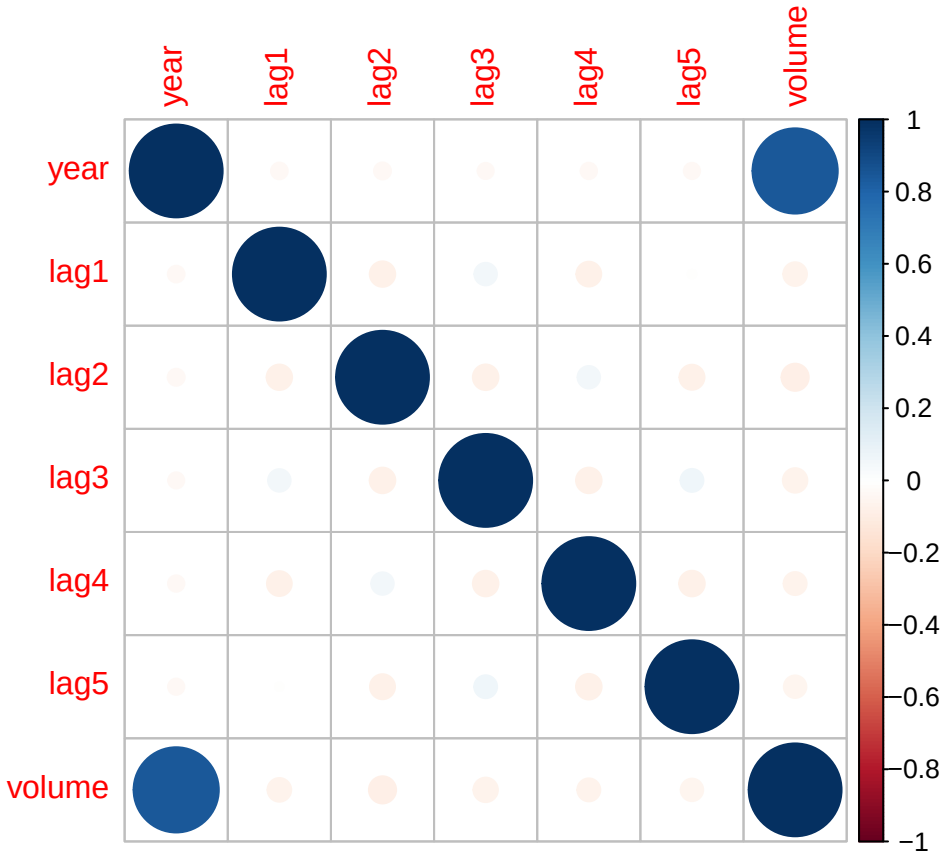
Ziyi Zhao

4/13/2020

Part A: Produce graphical summaries of weekly data







- the first plot show the scatter plots and density plots between variable 'today' and each predictors. In visual, we can hardly observe any linear or non-linear relationship.
- the second plot use the dichotomous variable 'direction' as outcome. We can see the lag1 to lag5 are normally distributed for both classes. The purple line indicate that direction goes up; the blue line indicate the direction goes down.
- the third plot is the correlation plot of those 7 features. We can easily observe the volume is positively correlated with year because the cell is shown in a dark blue. We don't observe any other strong correlation between other variables.

Part B: Use full dataset to perform logistic regression with outcome direction and predictor lag1 to lag5 and volume.

```
##      Up
## Down  0
## Up    1
```

```
##
## Call:
## glm(formula = direction ~ lag1 + lag2 + lag3 + lag4 + lag5 +
##     volume, family = binomial, data = dat)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -1.6949  -1.2565   0.9913   1.0849   1.4579
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.26686    0.08593   3.106   0.0019 **
## lag1        -0.04127    0.02641  -1.563   0.1181
## lag2         0.05844    0.02686   2.175   0.0296 *
## lag3        -0.01606    0.02666  -0.602   0.5469
## lag4        -0.02779    0.02646  -1.050   0.2937
## lag5        -0.01447    0.02638  -0.549   0.5833
## volume      -0.02274    0.03690  -0.616   0.5377
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1496.2  on 1088  degrees of freedom
## Residual deviance: 1486.4  on 1082  degrees of freedom
## AIC: 1500.4
##
## Number of Fisher Scoring iterations: 4
```

There is only one predictor appearing to be statistically significant. The p-value of predictor 'lag2' is smaller than 0.05, so we can conclude that the coefficient of lag2 is significantly different from 0.

Part C: compute a confusion matrix and overall fraction of correct predictions. Briefly explain what confution matrix tell us.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Down  Up
##       Down   54  48
##       Up    430 557
##
##                Accuracy : 0.5611
##                  95% CI : (0.531, 0.5908)
##     No Information Rate : 0.5556
##     P-Value [Acc > NIR] : 0.369
##
##                   Kappa : 0.035
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.9207
##             Specificity : 0.1116
##          Pos Pred Value : 0.5643
##          Neg Pred Value : 0.5294
##              Prevalence : 0.5556
##          Detection Rate : 0.5115
##    Detection Prevalence : 0.9063
```

```
##       Balanced Accuracy : 0.5161
##
##        'Positive' Class : Up
##
```

From the results of confusion matrix, we can find out that there are 557 true "Up" and 54 true "Down". The prevalence tell us that there are about 56% of "up" in the observed data. The overall fraction of correct prediction is 0.5611 with 95% CI from 0.531 to 0.5908.

The no information rate tell us that the fraction of "Up" class in both predicted and trained dataset is about 56%, which means the number of "up" and "down" approximate half and half. The p-value is larger than 0.05, which means that we failed to reject the null hypothesis that accuracy is equal to no information rate.

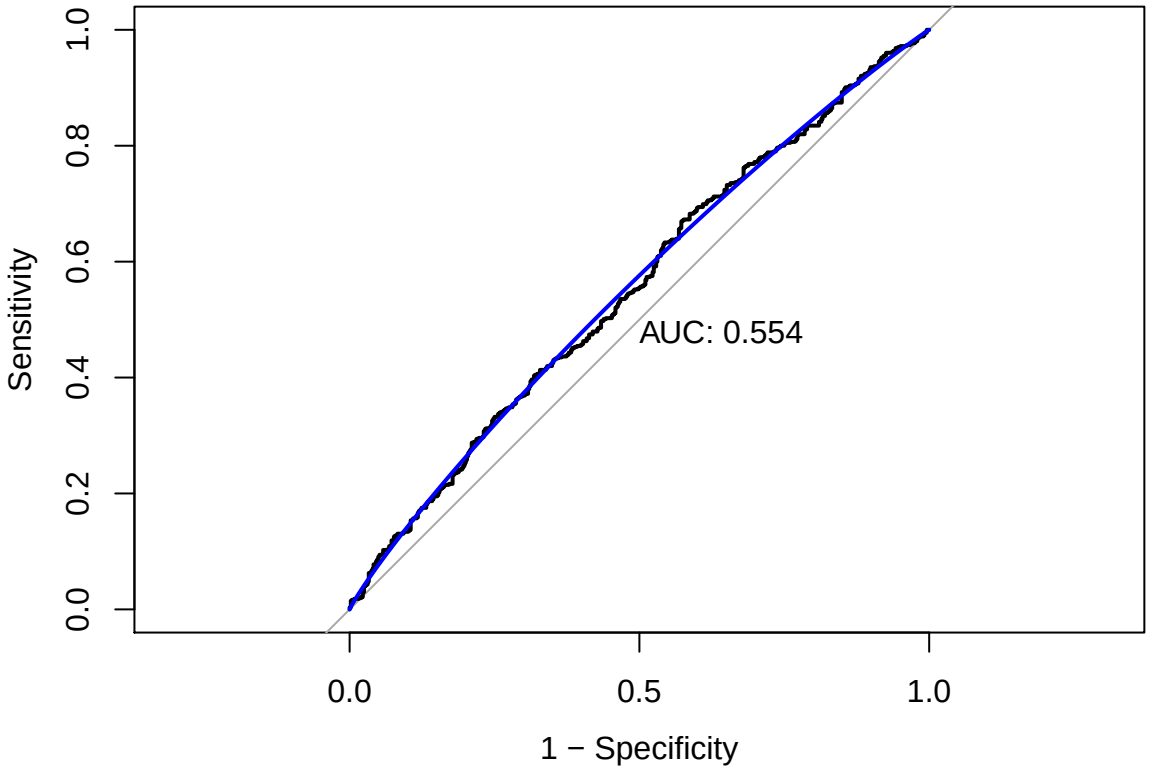
Kappa measure indicate how predicted values agreed with observed values. If the Kappa is close to 1, it means that the predicted values matched observed values perfectly. If not, vice versa. In this case, the Kappa is approaching to 0, which means that our predicted data agreed with the observed data by chance. This may not be a good predictive model.

We have 0.9207 for sensitivity and 0.1116 for specificity. The specificity is very low. It indicated that only 11.2% of observed "Down" is predicted correctly. Also, we have 56.4% for PPV and 52.9% for NPV, which means that true "Up" and "Down" are above half of the predicted "up" and "down", respectively.

Part D: Plot ROC curves using the predicted probability from the logistic regression and report AUC.

```
## Setting levels: control = Down, case = Up
```

```
## Setting direction: controls < cases
```

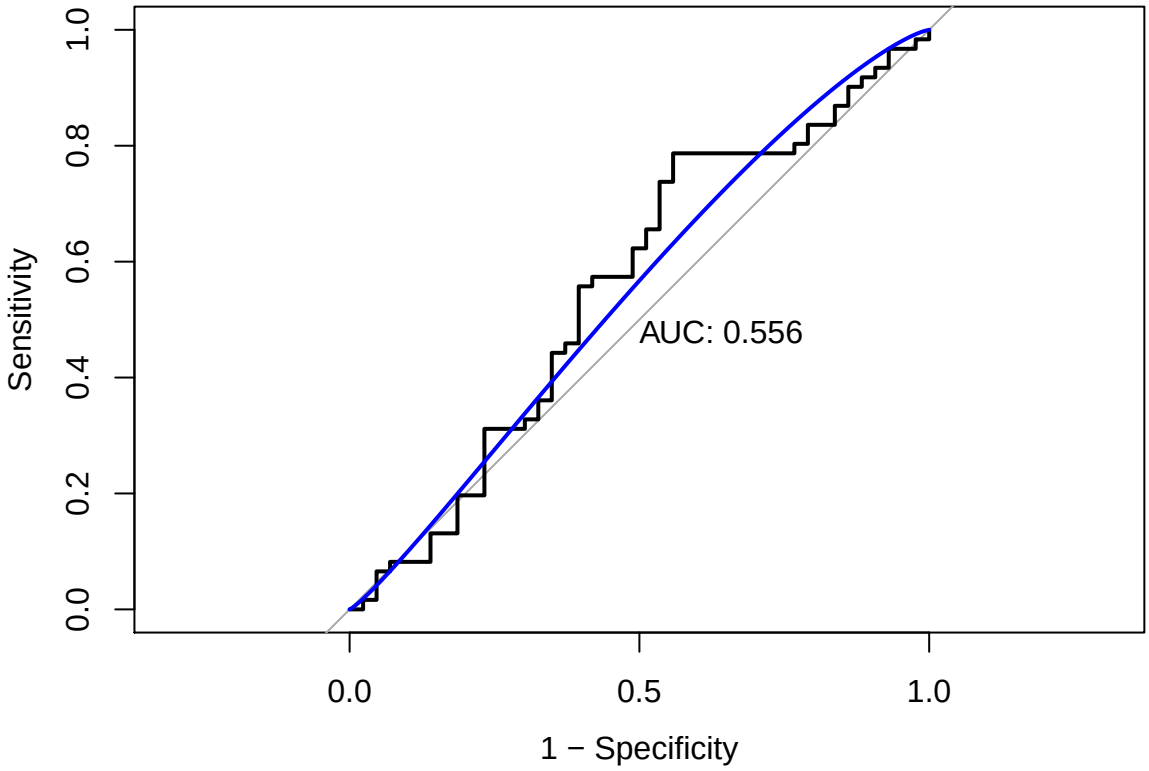


The AUC is 0.554.

Part E: Fit the logistic regression on the training data that selected from 1990 to 2008 with lag1 and lag2 as predictors; plot the ROC curve on the test data that selected from 2009 and 2010, and report the auc.

```
## Setting levels: control = Down, case = Up
```

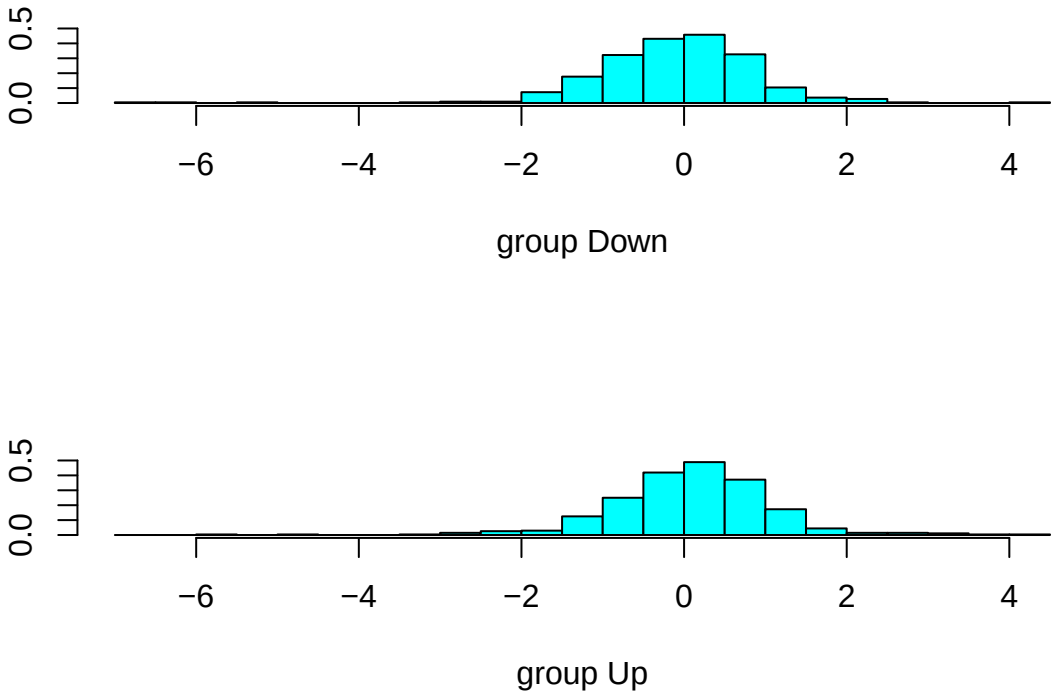
```
## Setting direction: controls < cases
```



The AUC is shown as 0.556.

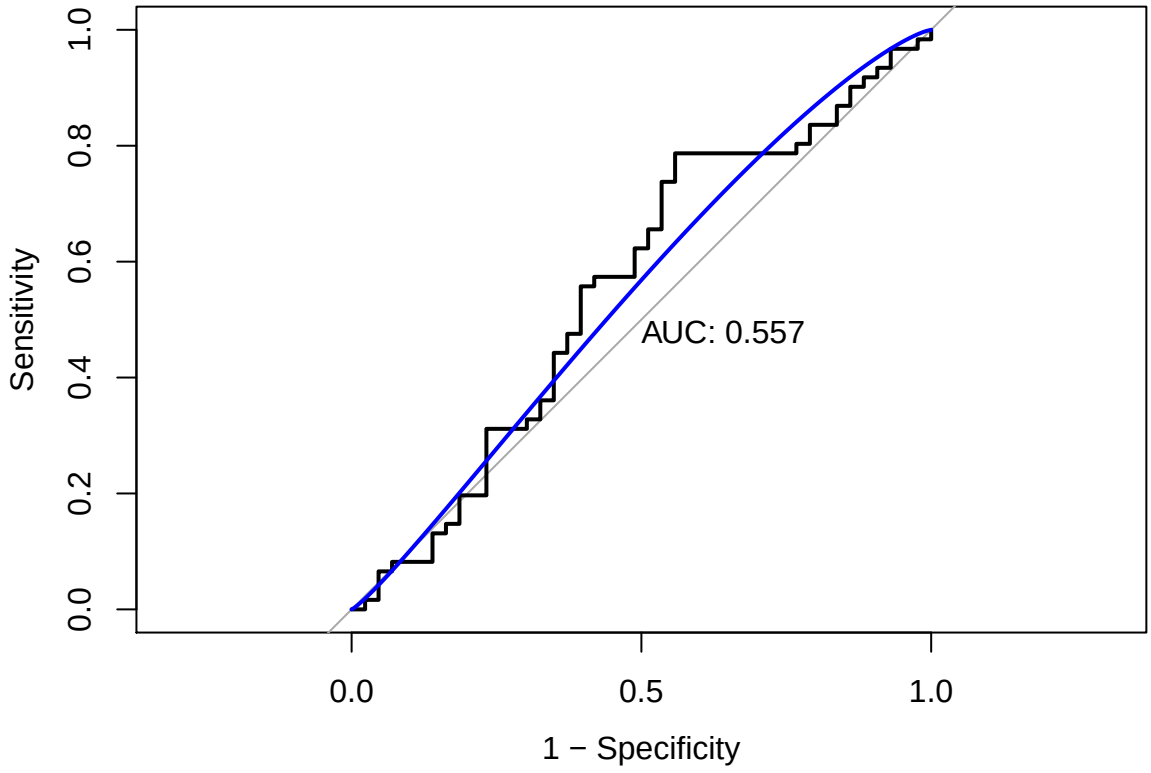
Part F: repeat the part E with LDA and QDA separately.

LDA



In this case, we have one discriminant variable with 2 classes. The plot above show the distribution of Z conditional on "Up" and "Down" class respectively. In visual, both of them are approximately normal distributed; however, the mean and model looks exactly equal, which means that two classes are not separated well by two predictors.

```
## Setting direction: controls < cases
```



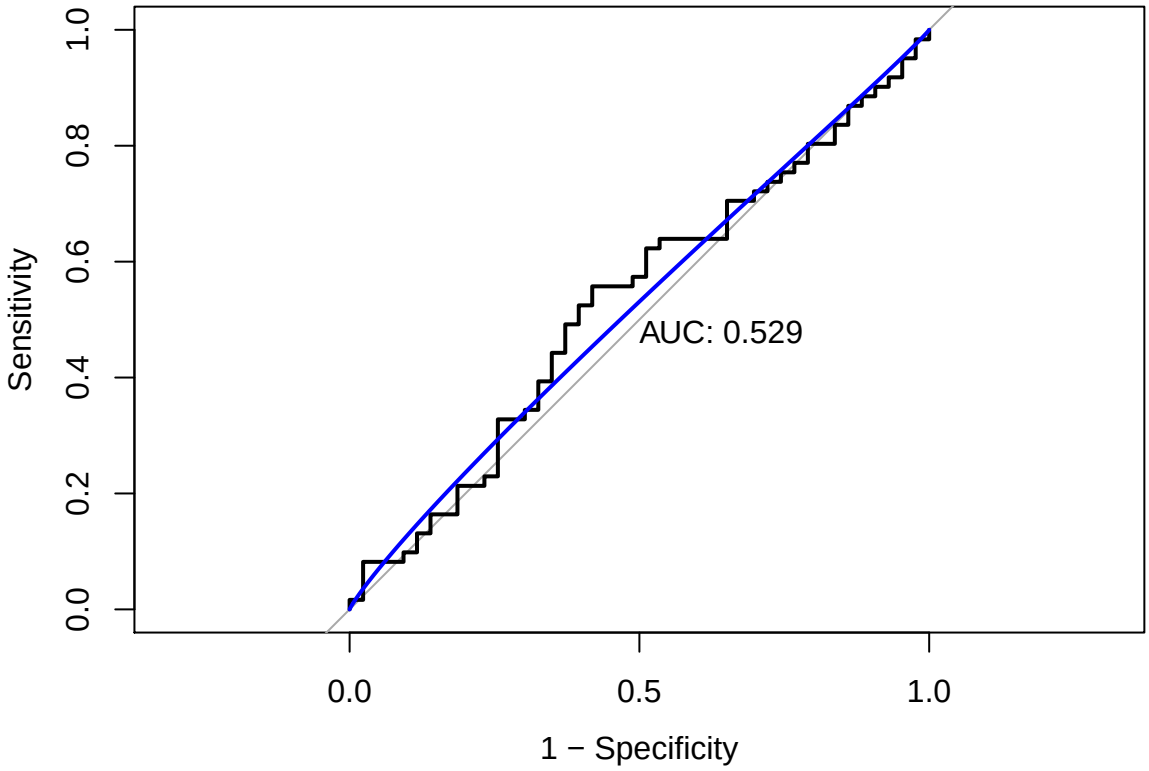
The AUC is shown as 0.557.

QDA

```
qda_fit <- qda(direction~.,dat_train)

qda_pred <- predict(qda_fit, newdata = dat_test)

roc_qda <- roc(dat_test$direction,qda_pred$posterior[,2],
               levels = c("Down","Up"))
```

```
## Setting direction: controls > cases
```

```
plot(roc_qda,legacy.axes = TRUE,print.auc = TRUE)
plot(smooth(roc_qda), col = 4, add = TRUE)
```

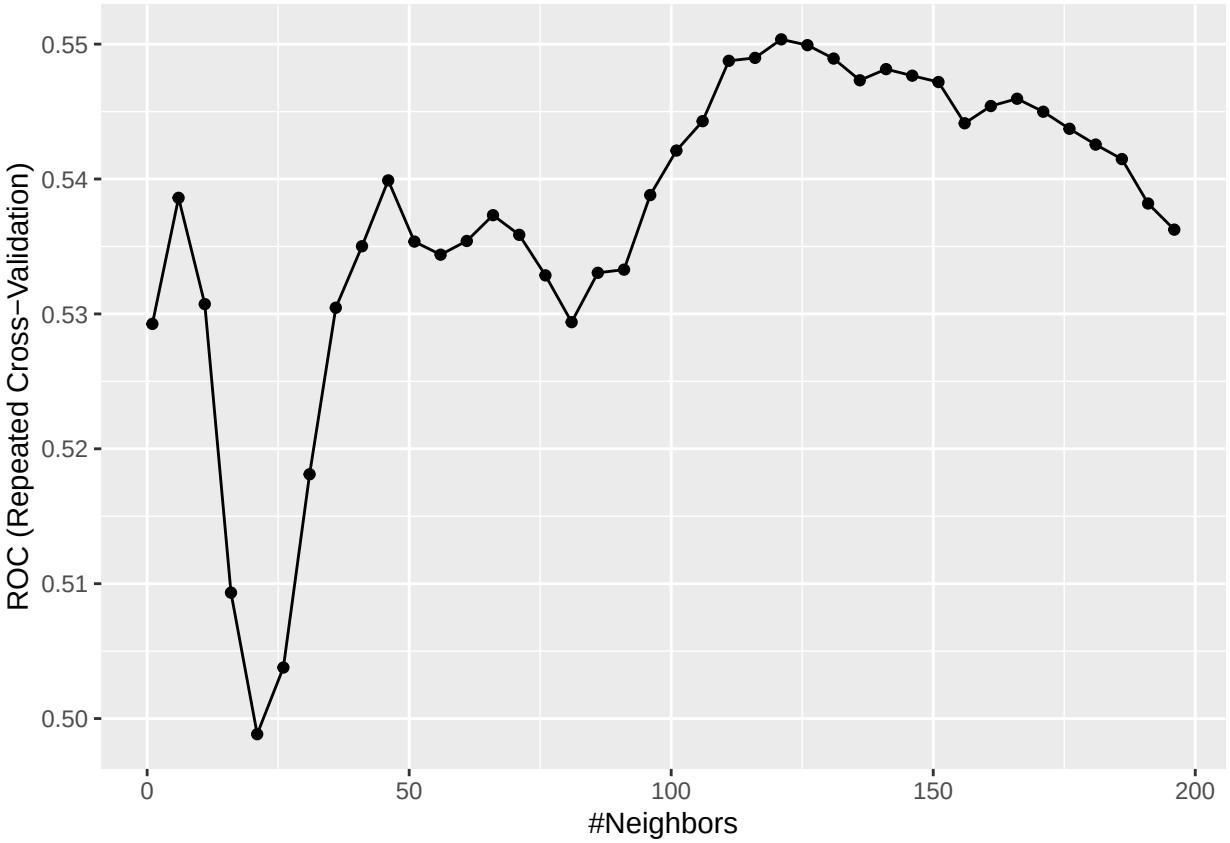


The AUC is shown as 0.529.

Part G: Repeat part E using KNN.

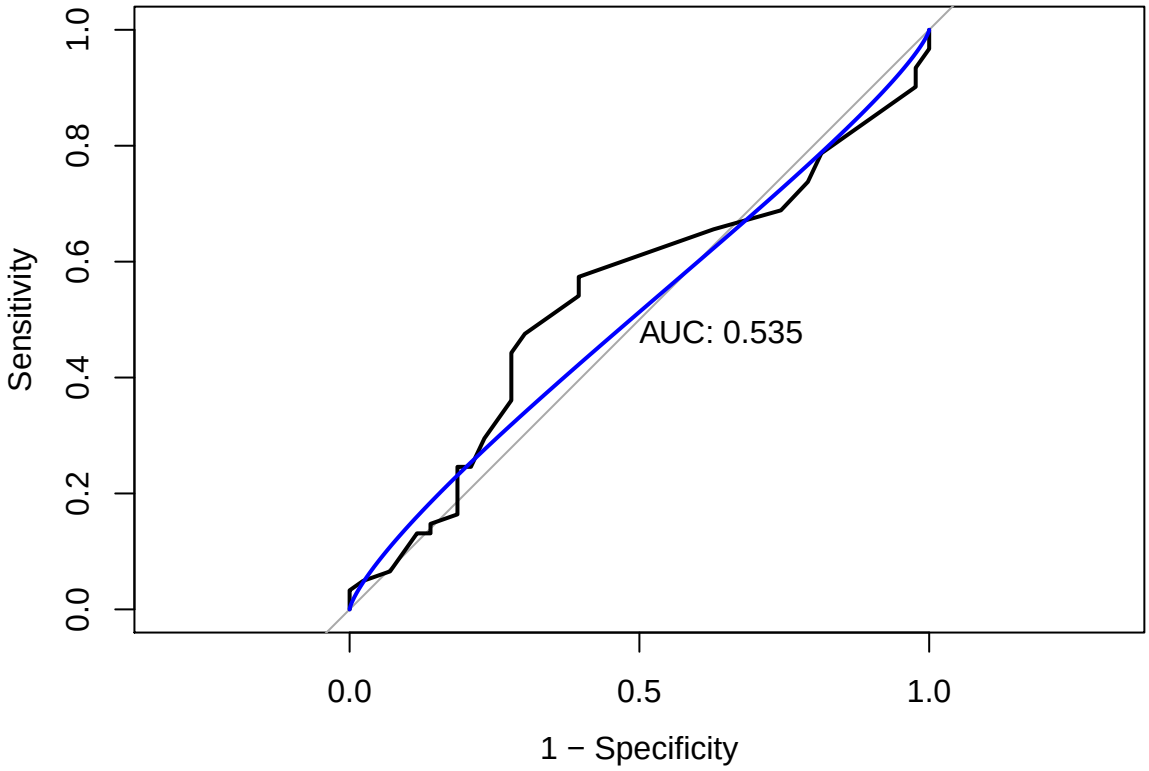
Fit KNN using caret

```
## Warning in train.default(x = dat_train[, 1:2], y = dat_train$direction, :
## The metric "Accuracy" was not in the result set. ROC will be used instead.
```




```
## Setting levels: control = Down, case = Up
```

```
## Setting direction: controls < cases
```



The AUC of KNN model was shown as 0.535 using caret.

Fit logistic, LDA, and QDA using caret

To compare each model, we use caret to fit other three models with repeated cross validation and compare the results.

```
set.seed(2)
model_glm <- train(x = dat_train[,c(1:2)],
                   y = dat_train$direction,
                   method = "glm",
                   metric = "ROC",
                   trControl = ctrl1)
glm_pred <- predict(model_glm,newdata = dat_test, type = "prob")[,2]
roc_glm2 <- roc(dat_test$direction,glm_pred)
```

```
## Setting levels: control = Down, case = Up
```

```
## Setting direction: controls < cases
```

```
set.seed(2)
model_lda <- train(x = dat_train[,c(1:2)],
                   y = dat_train$direction,
                   method = "lda",
```

```
                  metric = "ROC",
                  trControl = ctrl1)
lda_pred1 <- predict(model_lda,newdata = dat_test,type = "prob")[,2]
roc_lda1 <- roc(dat_test$direction,lda_pred1)
```

```
## Setting levels: control = Down, case = Up
## Setting direction: controls < cases
```

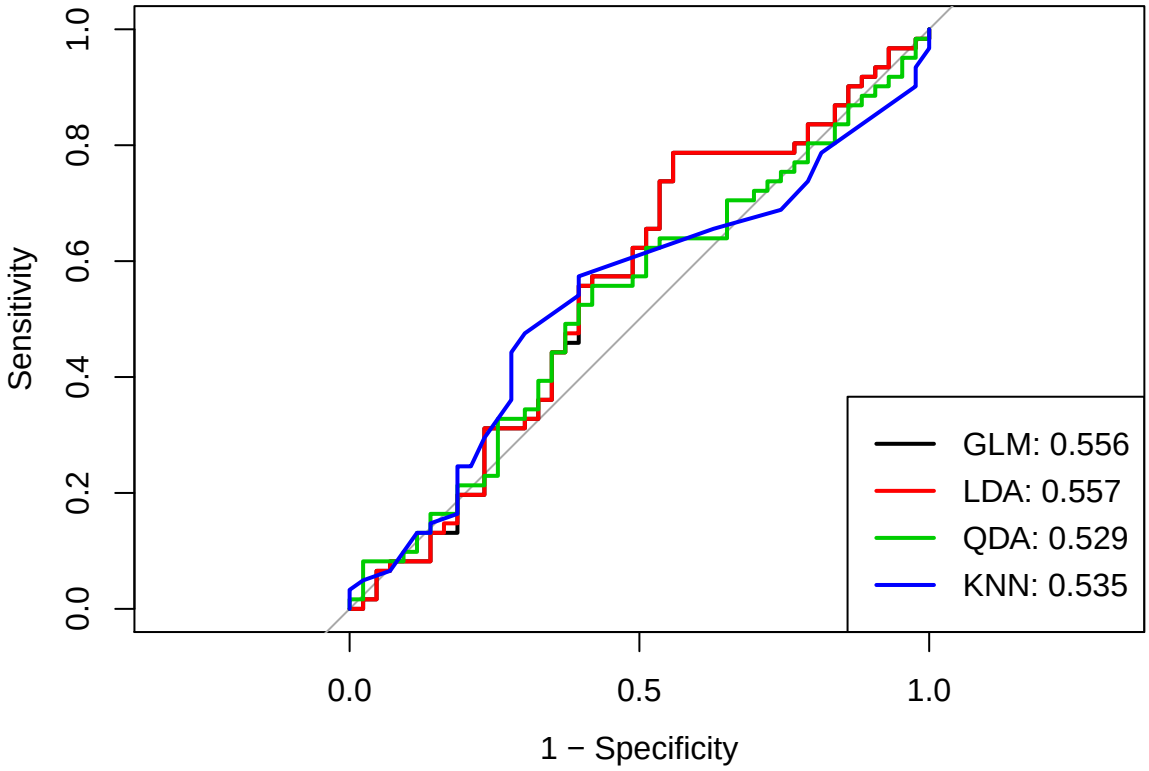
```
set.seed(2)
model_qda <- train(x=dat_train[,1:2],
                  y=dat_train$direction,
                  method = "qda",
                  metric = "ROC",
                  trControl = ctrl1)
qda_pred1 <- predict(model_qda,newdata = dat_test,type = "prob")[,2]
roc_qda1 <- roc(dat_test$direction,qda_pred1)
```

```
## Setting levels: control = Down, case = Up

## Setting direction: controls > cases
```

```
resamp <- resamples(list(GLM = model_glm,LDA = model_lda,
                         QDA = model_qda,KNN = model_knn))
summary(resamp)
```

```
##
## Call:
## summary.resamples(object = resamp)
##
## Models: GLM, LDA, QDA, KNN
## Number of resamples: 50
##
## ROC
##          Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## GLM 0.4191919 0.5089731 0.5494146 0.5460532 0.5857043 0.7623967    0
## LDA 0.4187710 0.5092761 0.5492080 0.5461869 0.5860199 0.7623967    0
## QDA 0.4023569 0.4867769 0.5324074 0.5249060 0.5519628 0.6396694    0
## KNN 0.4148148 0.5210438 0.5526389 0.5503547 0.5764941 0.7126033    0
##
## Sens
##           Min.    1st Qu.     Median       Mean   3rd Qu.      Max. NA's
## GLM 0.00000000 0.06818182 0.09090909 0.09749495 0.1136364 0.1818182    0
## LDA 0.00000000 0.06818182 0.09090909 0.09433333 0.1136364 0.1818182    0
## QDA 0.06818182 0.13636364 0.18181818 0.18496970 0.2272727 0.3863636    0
## KNN 0.11363636 0.24583333 0.27272727 0.28164646 0.3181818 0.5454545    0
##
## Spec
##          Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## GLM 0.8333333 0.8888889 0.9074074 0.9150303 0.9454545 1.0000000    0
## LDA 0.8363636 0.8893939 0.9175084 0.9172525 0.9454545 1.0000000    0
## QDA 0.7090909 0.7972222 0.8348485 0.8408081 0.8888889 0.9636364    0
## KNN 0.5925926 0.6758418 0.7272727 0.7352189 0.7818182 0.9636364    0
```

```r
auc <- c(roc_glm2$auc[1],roc_lda1$auc[1],roc_qda1$auc[1],
         roc_knn$auc[1])
modelNames <- c("GLM", "LDA", "QDA", "KNN")

plot(roc_glm2,legacy.axes = TRUE)
plot(roc_lda1, col = 2, add = TRUE)
plot(roc_qda1, col = 3, add = TRUE)
plot(roc_knn, col = 4, add = TRUE)
legend("bottomright",legend = paste0(modelNames, ": ", round(auc,3)),
       col = 1:4, lwd = 2)
```



We can find out GLM and LDA models have relatively higher mean specificity (over 0.9) than other two models; however, their mean sensitivity are less than 10%.

From the plot, the LDA have the best ROC curve and highest AUC values. However, the predictive performance on test data for those models still remain low because their AUC values are lower than 0.6.