# MF815 HW1

**Advanced Machine Learning Application for Finance: Classification**

**Shi Bo**

**2021/2/5**

*(a) Produce some numerical and graphical summaries of the Weekly data. Are there any apparent patterns?*

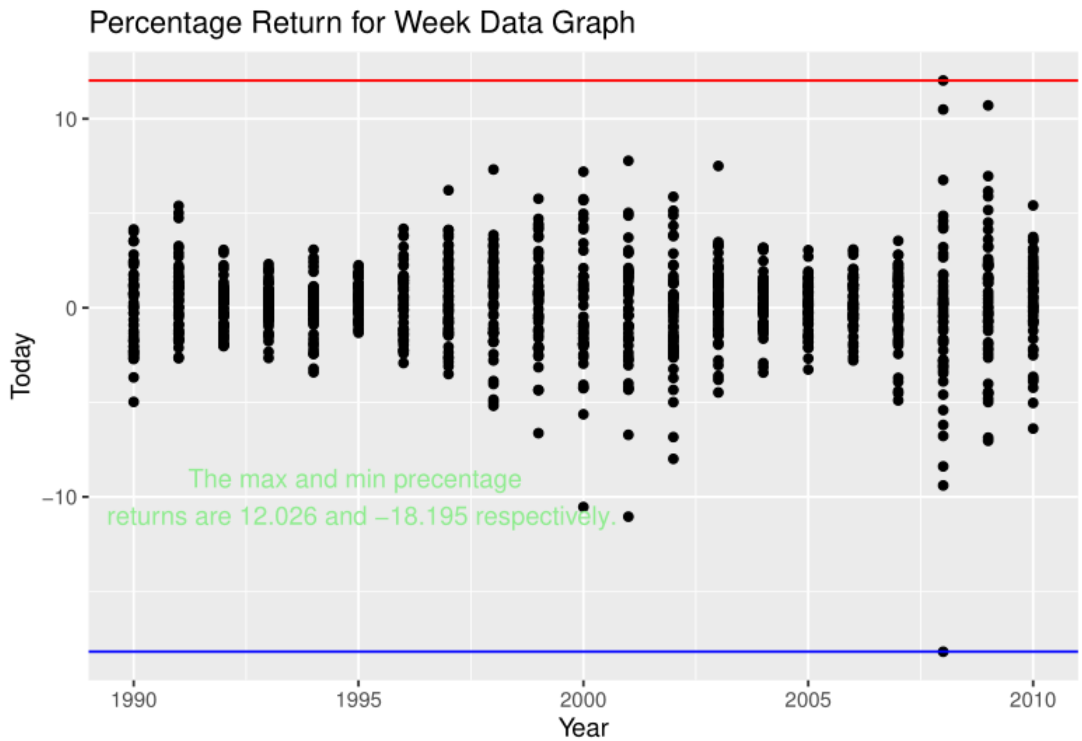
```
data(Weekly)
head(Weekly)
```

```
##   Year   Lag1   Lag2   Lag3   Lag4   Lag5    Volume  Today Direction
## 1 1990  0.816  1.572 -3.936 -0.229 -3.484 0.1549760 -0.270      Down
## 2 1990 -0.270  0.816  1.572 -3.936 -0.229 0.1485740 -2.576      Down
## 3 1990 -2.576 -0.270  0.816  1.572 -3.936 0.1598375  3.514        Up
## 4 1990  3.514 -2.576 -0.270  0.816  1.572 0.1616300  0.712        Up
## 5 1990  0.712  3.514 -2.576 -0.270  0.816 0.1537280  1.178        Up
## 6 1990  1.178  0.712  3.514 -2.576 -0.270 0.1544440 -1.372      Down
```

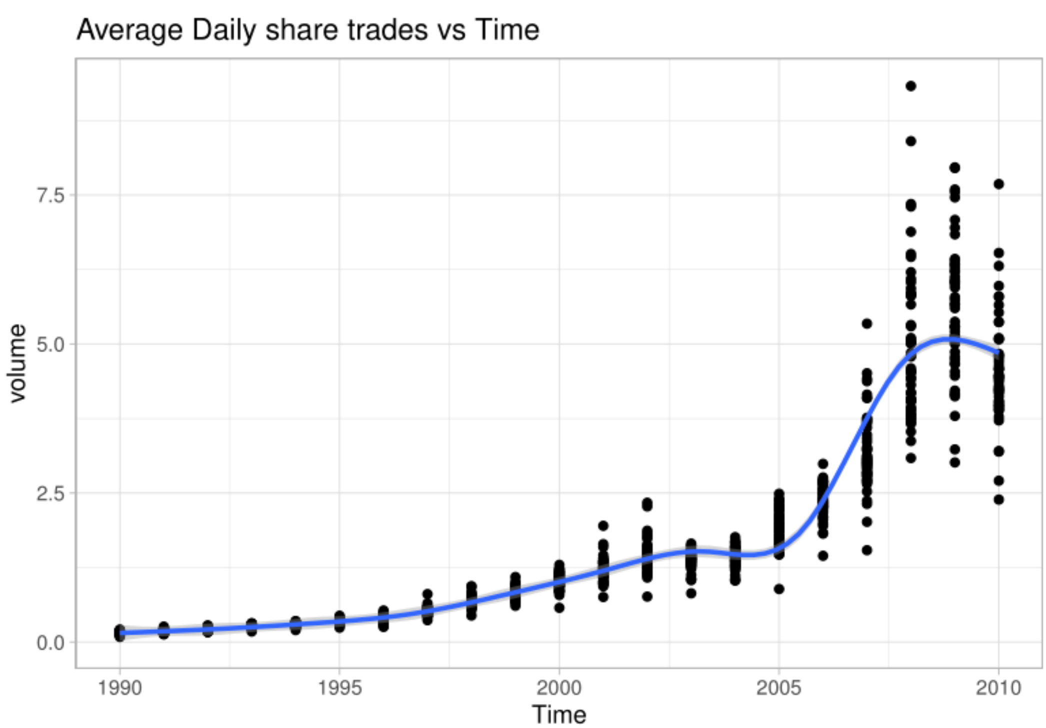
```
summary(Weekly)
```

```
##       Year           Lag1               Lag2               Lag3
##  Min.   :1990   Min.   :-18.1950   Min.   :-18.1950   Min.   :-18.1950
##  1st Qu.:1995   1st Qu.: -1.1540   1st Qu.: -1.1540   1st Qu.: -1.1580
##  Median :2000   Median :  0.2410   Median :  0.2410   Median :  0.2410
##  Mean   :2000   Mean   :  0.1506   Mean   :  0.1511   Mean   :  0.1472
##  3rd Qu.:2005   3rd Qu.:  1.4050   3rd Qu.:  1.4090   3rd Qu.:  1.4090
##  Max.   :2010   Max.   : 12.0260   Max.   : 12.0260   Max.   : 12.0260
##       Lag4               Lag5              Volume
##  Min.   :-18.1950   Min.   :-18.1950   Min.   :0.08747
##  1st Qu.: -1.1580   1st Qu.: -1.1660   1st Qu.:0.33202
##  Median :  0.2380   Median :  0.2340   Median :1.00268
##  Mean   :  0.1458   Mean   :  0.1399   Mean   :1.57462
##  3rd Qu.:  1.4090   3rd Qu.:  1.4050   3rd Qu.:2.05373
##  Max.   : 12.0260   Max.   : 12.0260   Max.   :9.32821
##      Today          Direction
##  Min.   :-18.1950   Down:484
##  1st Qu.: -1.1540   Up  :605
##  Median :  0.2410
##  Mean   :  0.1499
##  3rd Qu.:  1.4050
##  Max.   : 12.0260
```

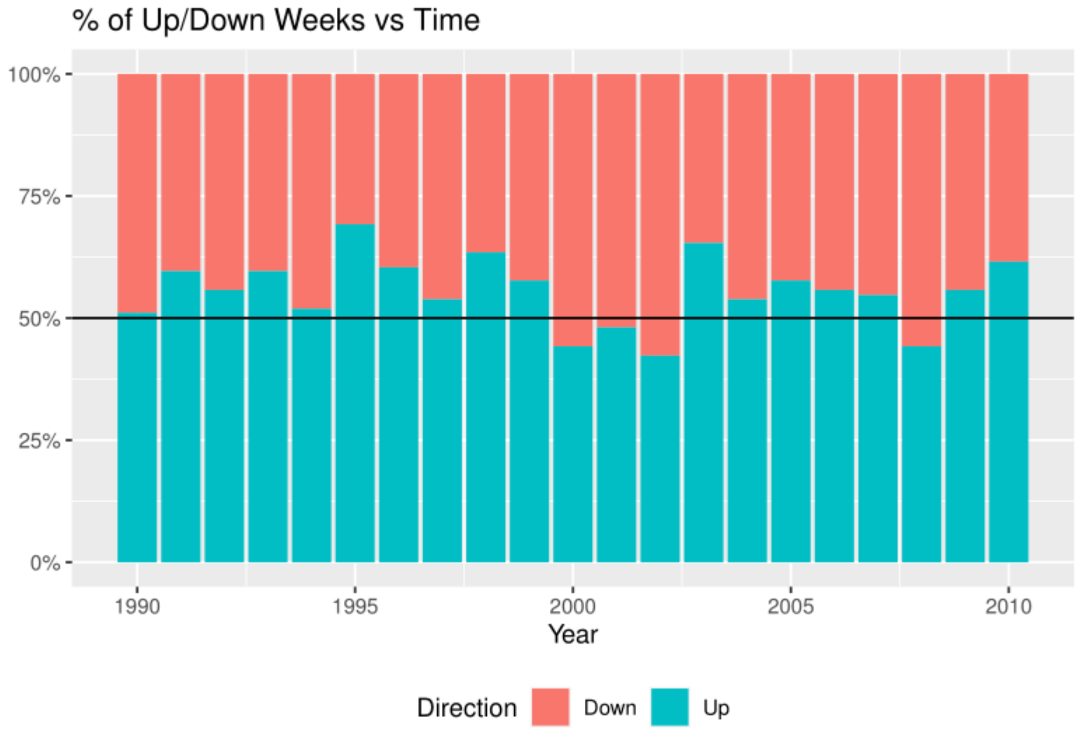
As we can see from the returns during whole period are very volatile, especially in 2008 in which there were many of negative returns.



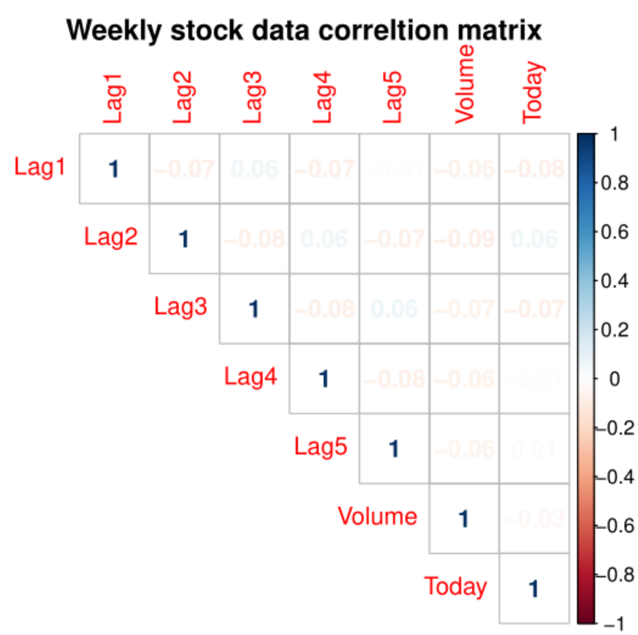

It is obvious that the trading volume was increasing as the time goes on.



The chart below shows that there are only four years (2000、2001、2002、2008) 50% of weeks that do not have positive return.




The correlation between these variables are almost uncorrelated with each other.



*(b) Use the full data to perform a logistic regression with Direction as the response variable and the five lags variables plus Volume as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so which ones?*

```
logit.fit <- glm(Direction ~ Lag1+Lag2+Lag3+Lag4+Lag5+Volume, data = Weekly, family="binomial")
summary(logit.fit)
```

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     Volume, family = "binomial", data = Weekly)
##
## 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
```

```
summary(logit.fit)$coef[,4] < 0.05
```

```
## (Intercept)        Lag1        Lag2        Lag3        Lag4        Lag5
##        TRUE       FALSE        TRUE       FALSE       FALSE       FALSE
##      Volume
##       FALSE
```

From this we know Lag2 is the most significant feature in prediction of class – direction of positive increase or negative increase in the weekly values.

*(c) Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes logistic regression is making.*

The false positive rate is 54/(54+430) = 11.16%. The false negative rate is 48/(48+557) = 7.93%.

```
logit.probs <- predict(logit.fit,type="response")
T.response <- Weekly$Direction %>% as.numeric() - 1
P.response <- logit.probs %>% round()
(log.confusion <- confusionMatrix(as.factor(P.response),as.factor(T.response)))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0  54  48
##          1 430 557
##
##                Accuracy : 0.5611
##                  95% CI : (0.531, 0.5908)
##     No Information Rate : 0.5556
##     P-Value [Acc > NIR] : 0.369
```

The confusion matrix tells us that there are 54 + 557 = 611 variables classified as correct (the accuracy is 56.11%) and rests are incorrect which accounted for 44.89% of proportion of the whole dataset.

*(d) Now fit the logistic regression using a training data period from 1990 to 2008 and Lag2 as the only predictor. Compute the confusion matrix and overall fraction of correct predictions for the hold out data, i.e., 2009 and 2010.*

```
log.train <- Weekly %>% filter(Year <= 2008)
log.test <- Weekly %>% filter(Year > 2008)
```

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0  9  5
         1 34 56

               Accuracy : 0.625
                 95% CI : (0.5247, 0.718)
    No Information Rate : 0.5865
    P-Value [Acc > NIR] : 0.2439

                  Kappa : 0.1414

 Mcnemar's Test P-Value : 7.34e-06
```

Let's split data into training – which includes data from 1990 to 2008 and testing – which includes data from 2009 to 2010. We can realize that the accuracy for testing data was same as LDA which has 62.5% accuracy as well.

*(e) Repeat (d) using the linear discriminant analysis (LDA).*

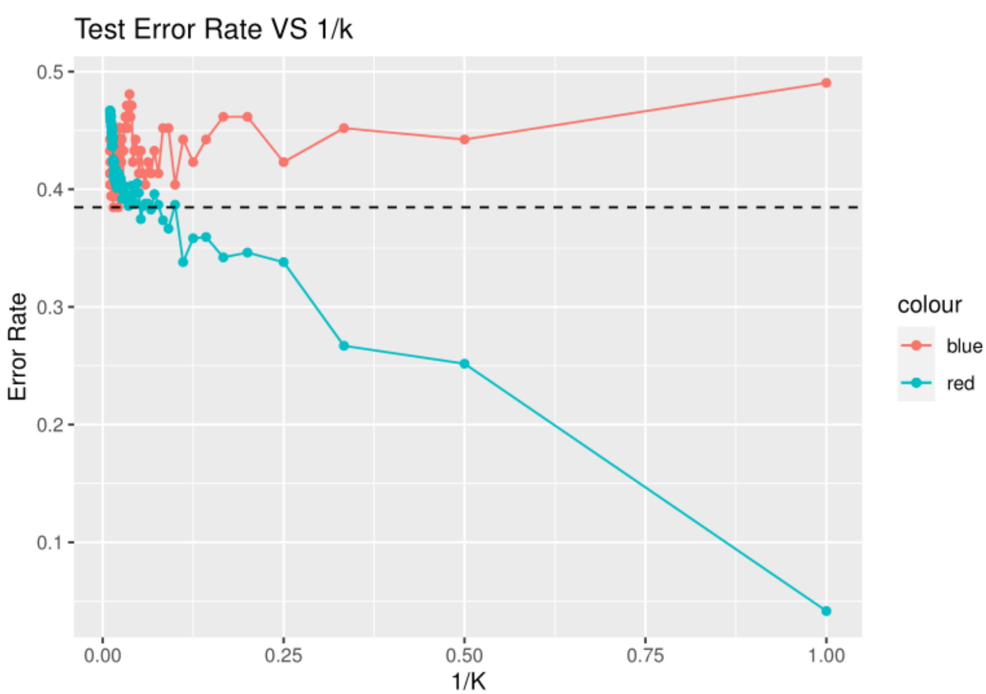
```r
LDA.pred <- predict(LDA.fit, log.test)

(confusion.lda <- confusionMatrix(as.factor(LDA.pred$class), as.factor(log.test$Direction)))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Down Up
##       Down    9  5
##       Up     34 56
##
##                Accuracy : 0.625
##                  95% CI : (0.5247, 0.718)
##     No Information Rate : 0.5865
##     P-Value [Acc > NIR] : 0.2439
```

By finding separation between classes using true decision boundary(LDA), the accuracy was better than logistic regression, which is 62.5%.

(f) For the test data using kNN, plot the misclassification error rate vs 1/k. What is the optimal k that minimizes the test misclassification error rate?




```r
print(paste('The optional K for minimum error is:', optimal_K))
```

```
## [1] "The optional K for minimum error is: 68"
## [2] "The optional K for minimum error is: 47"
```

```r
print(paste('The minimized error is:', min_error))
```

```
## [1] "The minimized error is: 0.384615384615385"
```

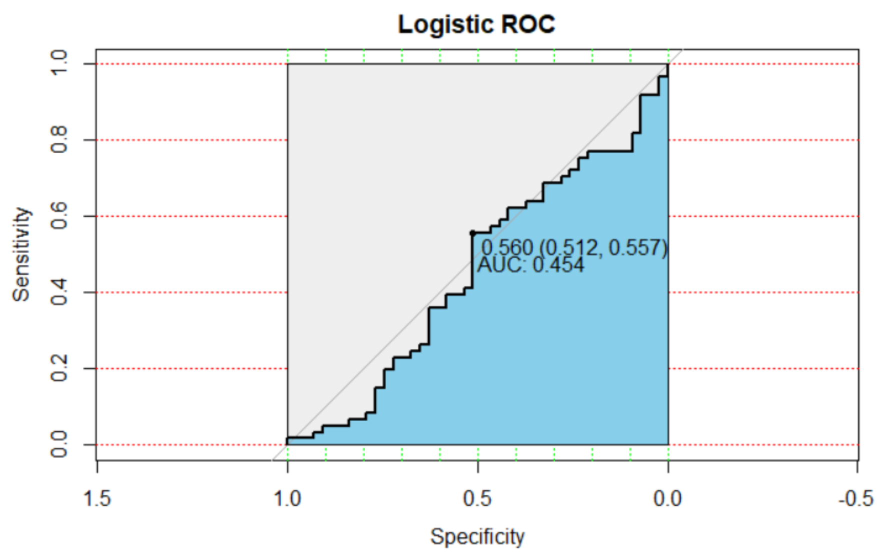
*(g) Which of these various methods appears to provide the best results on this data?*

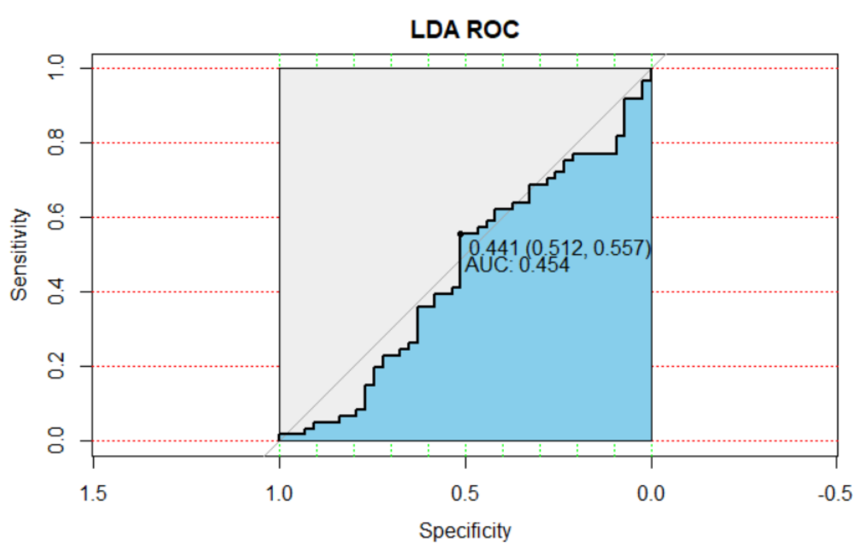
```r
log_acc <- log.confusion$overall[1] %>% as.numeric()
lda_acc <- confusion.lda$overall[1] %>% as.numeric()
knn_acc <- 1 - min_error
(df <- data.frame(Accuracy = c('Logistic','LDA','KNN'),
```

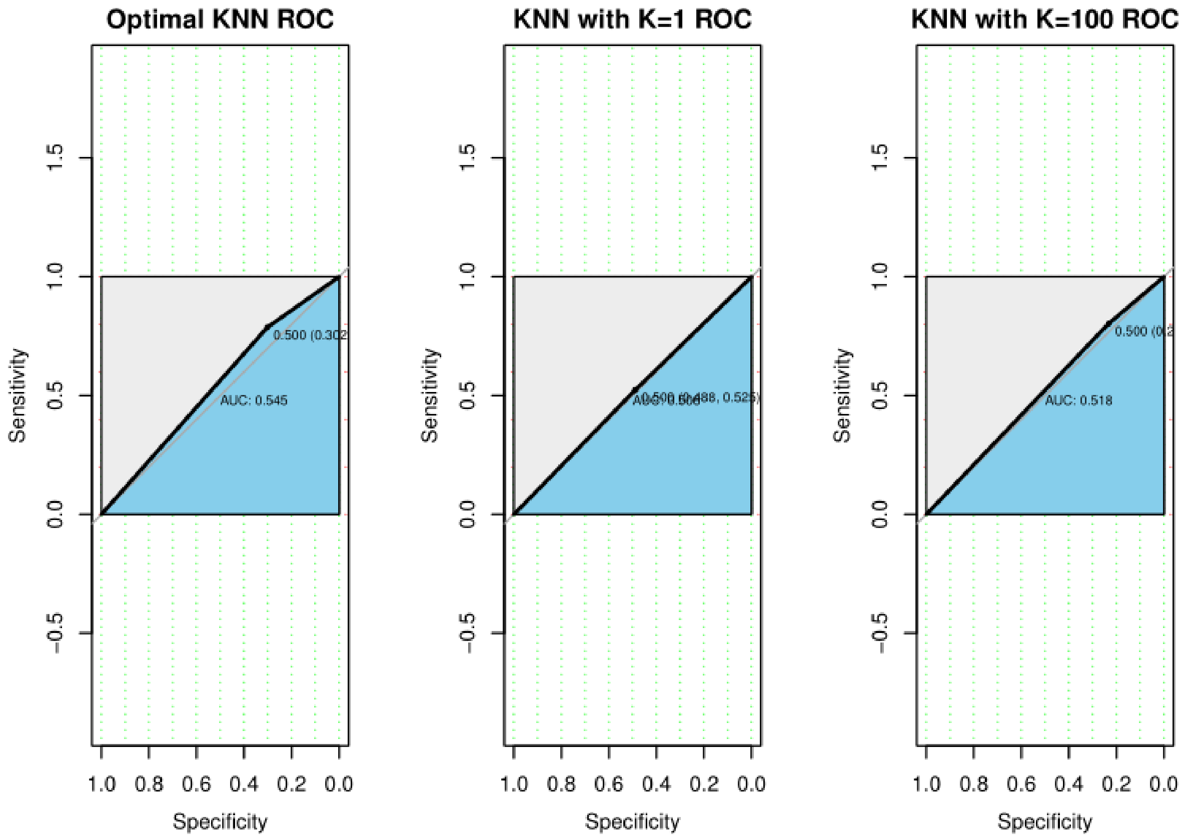
```
##    Accuracy       num
## 1 Logistic 0.6250000
## 2      LDA 0.6250000
## 3      KNN 0.6153846
```

The winner is LDA and logistic regression even though we chose the optimal K to implement KNN.

*(h) Plot the ROC curves for different classifiers, e.g. logistic regression, LDA, kNN with different k values and discuss the performance (the larger the area under the curve, the better the classifier). Optimal KNN performs best, second is logistic and lda.*





Optimal KNN ROC
KNN with K=1 ROC
KNN with K=100 ROC
Sensitivity
Specificity
0.500 (0.302
AUC: 0.545
AUC: 0.518
0.500 (0.2