STATISTICAL DATA MINING – I
HOMEWORK III

NAME: SIDDIQ SYED

UB Person # 50291566

Class # 50

1) Using the Boston data set (ISLR package), fit classification models in order to predict whether a given suburb has a crime rate above or below the median. Explore logistic regression, LDA and kNN models using various subsets of the predictors. Describe your findings.

Exploratory analysis of the data

Looking at the summary of the data for boston dataset and as the analysis of classification models to predict whether a given suburb has a crime rate above or below the median. The crime rate has many outliers which may led to quite the variation in outputs for a given subset of training and test datasets.

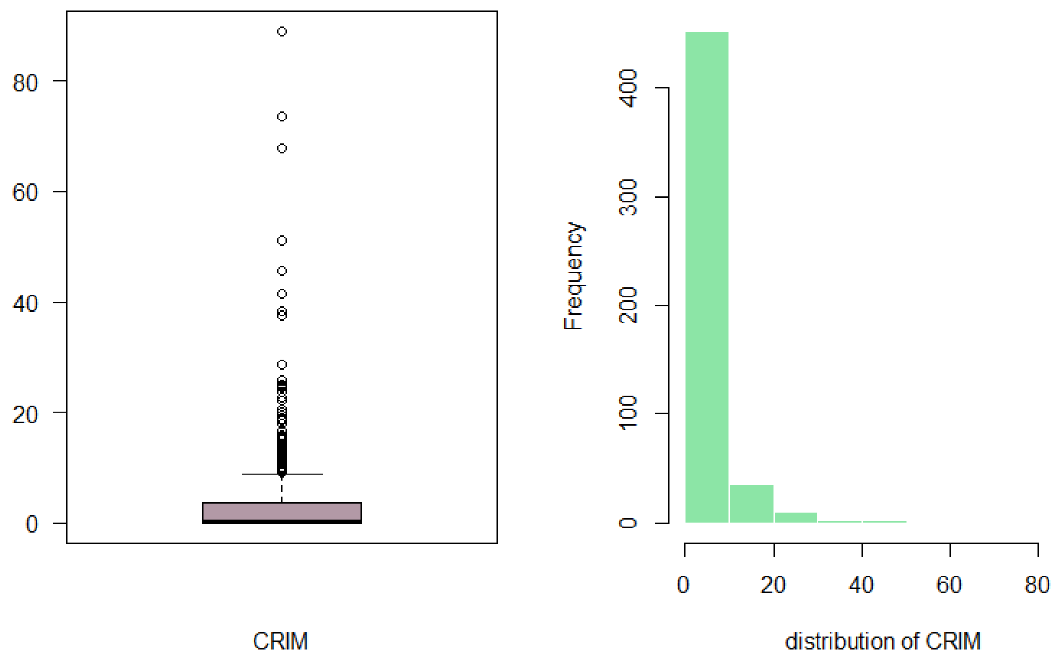


Figure 1: The above graph plots illustrates outliers details.

Let's analyse the dataset using Logistic regression, LDA and KNN. Considering the Logistic regression and fitting the model for the same we can see the summary of the model below.

```
Call:
glm(formula = CRIM ~ ., family = "binomial", data = boston_train)

Deviance Residuals:
    Min       1Q    Median       3Q      Max
-1.6224  -0.1039  -0.0009   0.0009   3.6232

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -22.550279   9.570026  -2.356 0.018456 *
ZN           -0.053746   0.039610  -1.357 0.174817
INDUS        -0.085119   0.062946  -1.352 0.176294
CHAS          1.043919   0.985295   1.059 0.289373
NOX          47.340357   9.700723   4.880 1.06e-06 ***
RM           -0.616574   0.992653  -0.621 0.534509
AGE           0.048388   0.016784   2.883 0.003940 **
DIS           0.626510   0.274648   2.281 0.022540 *
RAD           0.756329   0.198411   3.812 0.000138 ***
TAX          -0.008244   0.003417  -2.413 0.015835 *
PT            0.559887   0.185622   3.016 0.002559 **
B            -0.052026   0.017310  -3.005 0.002652 **
LSTAT         0.001320   0.062126   0.021 0.983042
MV            0.182293   0.086281   2.113 0.034620 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 490.70  on 353  degrees of freedom
Residual deviance: 128.55  on 340  degrees of freedom
AIC: 156.55

Number of Fisher Scoring iterations: 9
```

The various residual parameters and AIC value can be seen in above picture. The important parameters from the same are shown below in values.

```
> varImp(glm.fit)
         Overall
ZN     1.35688808
INDUS  1.35225529
CHAS   1.05949892
NOX    4.88008573
RM     0.62113813
AGE    2.88294256
DIS    2.28113697
RAD    3.81192267
TAX    2.41270360
PT     3.01627135
B      3.00547681
LSTAT  0.02125489
MV     2.11277738
```

Upon considering the values that of overall above one we can eliminate LSTAT and RM variables from the dataset and fit the model to predict the test and training errors.

Below is the confusion matrix for the Logistic Regression model fit.

```
conf$table
```

```
          Reference
Prediction  0  1
         0 75  8
         1  7 62
```

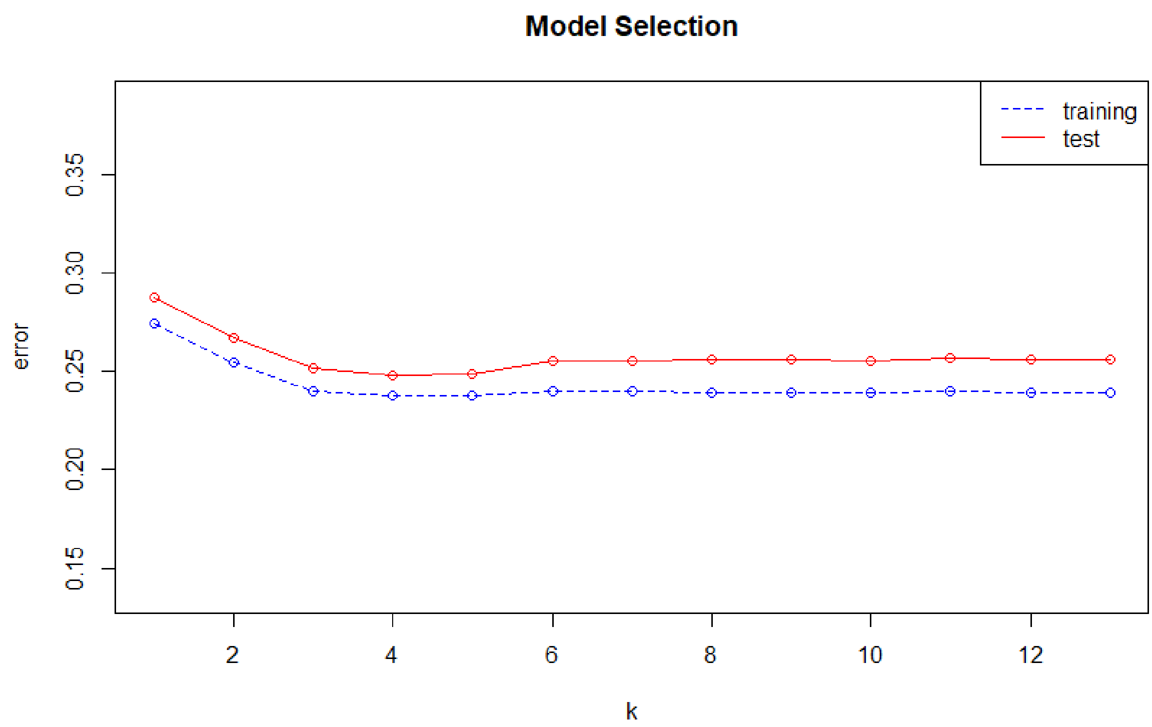
15 out of total are wrongly predicted.

Applying cross validation to the dataset of boston to check for the low error rates of the model taking 10 folds of the dataset boston.

```
Call:
NULL

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3946  -0.1585  -0.0004   0.0023   3.4239

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -34.103715   6.530015  -5.223 1.76e-07 ***
ZN           -0.079918   0.033731  -2.369  0.01782 *
INDUS        -0.059389   0.043722  -1.358  0.17436
CHAS          0.785327   0.728930   1.077  0.28132
NOX          48.523800   7.396499   6.560 5.37e-11 ***
RM           -0.425597   0.701104  -0.607  0.54383
AGE           0.022172   0.012221   1.814  0.06963 .
DIS           0.691400   0.218308   3.167  0.00154 **
RAD           0.656465   0.152452   4.306 1.66e-05 ***
TAX          -0.006412   0.002689  -2.385  0.01709 *
PT            0.368716   0.122136   3.019  0.00254 **
B            -0.013524   0.006536  -2.069  0.03853 *
LSTAT         0.043862   0.048981   0.895  0.37052
MV            0.167130   0.066940   2.497  0.01254 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 701.46  on 505  degrees of freedom
Residual deviance: 211.93  on 492  degrees of freedom
AIC: 239.93

Number of Fisher Scoring iterations: 9

> varImp(glm.cv.fit)
glm variable importance

      Overall
NOX   100.000
RAD    62.133
DIS    43.002
PT     40.513
MV     31.741
TAX    29.860
ZN     29.601
B      24.560
AGE    20.279
INDUS  12.620
CHAS    7.900
LSTAT   4.845
RM      0.000
```

Cross Validation results show us the greater approximation of what predictors to be included to fit the model for the best possible outputs.

Upon Model selection using hold out method and plotting the graph for the training and test errors in the below fig.



We can see that the error rates are constant from k=4.

Upon considering the first four variables from cross validation Important variables we can say that it is mostly dependent on NOX, RAD, DIS and PT.

Upon verification and sampling the various splits of the dataset and applying the Logistic regression model, LDA and KNN model for different sampling of the dataset. The model which is best can be defined by calculation the prediction error rates of the same.

The below are the results of the prediction error rates of the models.

```
> log_train_err
[1] 0.0990099
> log_test_err
[1] 0.06403941
> log_train_err_cv
[1] 0.09240924
> log_test_err_cv
[1] 0.07389163
> log_train_err_model
[1] 0.1254125
> log_test_err_model
[1] 0.1280788
> lda_train_error
[1] 0.1650165
> lda_test_error
[1] 0.1133005
> knn_train_error
[1] 0.09240924
> knn_test_error
[1] 0.09359606
```

Figure: For sampling 0.6 training and 0.4 testing

```
> log_train_err
[1] 0.07909605
> log_test_err
[1] 0.125
> log_train_err_cv
[1] 0.0819209
> log_test_err_cv
[1] 0.09210526
> log_train_err_model
[1] 0.1101695
> log_test_err_model
[1] 0.1052632
> lda_train_error
[1] 0.1468927
> lda_test_error
[1] 0.1381579
> knn_train_error
[1] 0.09039548
> knn_test_error
[1] 0.1118421
```

Figure: For sampling 0.7 training and 0.3 testing

```
> log_train_err
[1] 0.08415842
> log_test_err
[1] 0.09803922
> log_train_err_cv
[1] 0.08663366
> log_test_err_cv
[1] 0.07843137
> log_train_err_model
[1] 0.1237624
> log_test_err_model
[1] 0.127451
> lda_train_error
[1] 0.1559406
> lda_test_error
[1] 0.09803922
> knn_train_error
[1] 0.09405941
> knn_test_error
[1] 0.1078431
```

Figure: For sampling 0.8 training and 0.3 testing

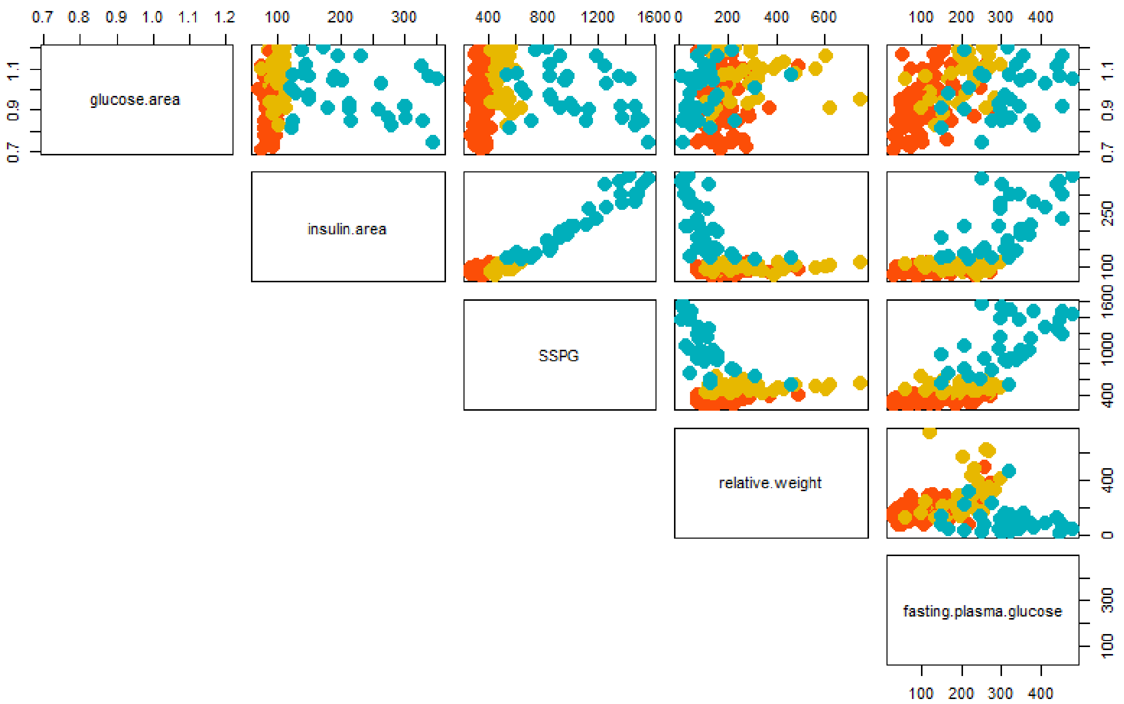
From the above results is clear that upon cross validation using K folds the test error for sampling 30 percentage testing with Logistic regression gives us the lowest error rate.

2) (10 points) Download the diabetes data set (http://astro.temple.edu/~alan/DiabetesAndrews36_1.txt). Disregard the first three columns. The fourth column is the observation number, and the next five columns are the variables (glucose.area,

insulin.area, SSPG, relative.weight, and fasting.plasma.glucose). The final column is the class number. Assume the population prior probabilities are estimated using the relative frequencies of the classes in the data. (Note: this data can also be found in the MMST library)

(a) Produce pairwise scatterplots for all five variables, with different symbols or colors representing the three different classes. Do you see any evidence that the classes may have difference covariance matrices? That they may not be multivariate normal?

Below is the scatterplot for the all the give five variables with different colors. Where Class Number 3 indicates Blue, class number 2 indicates Yellow and Class Number 1 indicates Red.



It is evident from the below matrices for Class Number 1 2 and 3 respectively that each class number will depend on different parameters to classify.

```
> cor(Dia_class1[,2:6])
                       glucose.area insulin.area       SSPG relative.weight fasting.plasma.glucose
glucose.area              1.0000000   -0.2795615 -0.3672603       0.3127337              0.2843730
insulin.area             -0.2795615    1.0000000  0.9550337      -0.6262652              0.5838766
SSPG                     -0.3672603    0.9550337  1.0000000      -0.6864899              0.5611345
relative.weight           0.3127337   -0.6262652 -0.6864899       1.0000000             -0.2004294
fasting.plasma.glucose    0.2843730    0.5838766  0.5611345      -0.2004294              1.0000000
> cor(Dia_class2[,2:6])
                       glucose.area insulin.area        SSPG relative.weight fasting.plasma.glucose
glucose.area             1.00000000  -0.07015646 -0.27200507      0.06688494             0.47727753
insulin.area            -0.07015646   1.00000000  0.60773933      0.10289976            -0.04121193
SSPG                    -0.27200507   0.60773933  1.00000000      0.12377645            -0.04268851
relative.weight          0.06688494   0.10289976  0.12377645      1.00000000             0.34471783
fasting.plasma.glucose   0.47727753  -0.04121193 -0.04268851      0.34471783             1.00000000
> cor(Dia_class3[,2:6])
                       glucose.area insulin.area      SSPG relative.weight fasting.plasma.glucose
glucose.area             1.00000000   0.29786966 0.2064439      0.05854811             0.44563087
insulin.area             0.29786966   1.00000000 0.2843786      0.03367841             0.09889239
SSPG                     0.20644385   0.28437858 1.0000000      0.22636071             0.21026412
relative.weight          0.05854811   0.03367841 0.2263607      1.00000000             0.49375038
fasting.plasma.glucose   0.44563087   0.09889239 0.2102641      0.49375038             1.00000000
```

Class 1 has highly corelated by parameter SSPG and Insulin.area also Class 2 has the same but Class 3 has fasting.plasma.glucose and Insulin.area are highly corelated.

(b) Apply linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA). How does the performance of QDA compare to that of LDA in this case?

The performance of QDA for the given dataset is better or with less error rate when compared with that of the LDA. Below are the error rates for the same when data is sample for 75% of training and remaining for test.

```
> lda_train_error
[1] 0.1203704
> lda_test_error
[1] 0.1081081
> qda_train_error
[1] 0.0462963
> qda_test_error
[1] 0.05405405
```

(c) Suppose an individual has (glucose area = 0.98, insulin area =122, SSPG = 544. Relative weight = 186, fasting plasma glucose = 184). To which class does LDA assign this individual? To which class does QDA?

```
> lda_pred
[1] 3
Levels: 1 2 3
> qda_pred
[1] 2
Levels: 1 2 3
```

The above is the prediction for the given data . LDA model predicts the class as 3 and QDA model predicts it as 2.

3)

a) Under the assumptions in the logistic regression model, the sum of posterior probabilities of classes is equal to one. Show that this holds for k=K.

We know that, the posterior probabilities are given by the following two equations:

$$Pr(G=k/X=x) = \frac{\exp(\beta_{k0}+\beta_k^T x)}{1+\sum_{l=1}^{K-1}\exp(\beta_{l0}+\beta_l^T x)} \text{ for } k=1,\dots,K-1$$

$$\& \quad Pr(G=K/X=x) = \frac{1}{1+\sum_{l=1}^{K-1}\exp(\beta_{l0}+\beta_l^T x)}$$

To find sum of posterior probabilities for $k=K$

$$\sum_{k=1}^{K} Pr(G=k/X=x) = \frac{\exp(\beta_{10}+\beta_1^T x)}{1+\sum_{l=1}^{K-1}\exp(\beta_{l0}+\beta_l^T x)} + \frac{\exp(\beta_{20}+\beta_2^T x)}{1+\sum_{l=1}^{K-1}\exp(\beta_{l0}+\beta_l^T x)}$$

$$\cdots + \frac{\exp(\beta_{(K-1)0}+\beta_{(K-1)}^T x)}{1+\sum_{l=1}^{K-1}\exp(\beta_{l0}+\beta_l^T x)} + Pr(G=K/X=x)$$

$$= \frac{\exp(\beta_{10}+\beta_1^T x)}{1+\sum_{l=1}^{K-1}\exp(\beta_{l0}+\beta_l^T x)} + \frac{\exp(\beta_{20}+\beta_2^T x)}{1+\sum_{l=1}^{K-1}\exp(\beta_{l0}+\beta_l^T x)} + \cdots$$

$$+ \frac{\exp(\beta_{(K-1)0}+\beta_{(K-1)}^T x)}{1+\sum_{l=1}^{K-1}\exp(\beta_{l0}+\beta_l^T x)} + \frac{1}{1+\sum_{l=1}^{K-1}\exp(\beta_{l0}+\beta_l^T x)}$$

all the denominators are equal add all the numerators.

$$= \frac{\exp(B_{10} + B_1^T x) + \exp(B_{20} + B_2^T x) + \cdots \exp(B_{(K-1)0} + B_{(K-1)}^T x) + 1}{1 + \sum_{l=1}^{K-1} \exp(B_{l0} + B_l^T x)}$$

$$= \frac{\sum_{l=1}^{K-1} \exp(B_{l0} + B_l^T x) + 1}{1 + \sum_{l=1}^{K-1} \exp(B_{l0} + B_l^T x)}$$

$$= 1$$

$$\therefore \sum_{K=1}^{K} P_r(G = k \mid X = x) = 1$$

b) Using a little bit of algebra, show that the logistic function representation and the logit representation for the logistic regression model are equivalent. In other words, show that the logistic function: �(�) = exp (�+ + �-�) 1 + exp (�+ + �-�) is equivalent to: �(�) 1 − �(�) = exp(�+ + �-�).

we have

$$p(x) = \frac{\exp(B_0 + B_1 x)}{1 + \exp(B_0 + B_1 x)}.$$

$$P(x)(1 + \exp(B_0 + B_1 x)) = \exp(B_0 + B_1 x)$$

$$P(x) + P(x) \times \exp(B_0 + B_1 x) = \exp(B_0 + B_1 x)$$

$$P(x) = \exp(B_0 + B_1 x) - P(x) \times \exp(B_0 + B_1 x)$$

$$= \exp(B_0 + B_1 x)(1 - P(x))$$

$$\frac{P(x)}{1 - P(x)} = \exp(B_0 + B_1 X)$$