CASE I

DATA MINING II

Jagarlapudi Krishna Teja
M10896584

CONTENTS

EXECUTIVE SUMMARY – BOSTON HOUSING DATA

Objective: Build multiple models – linear regression, regression tree, generalized additive model and neural network – and compare the performance of the models on training and testing data. Based on these, the best performing model needs to be chosen.

Data: The data provided consists of 13 different variables, for 506 localities in Boston. The response variable is the median value of house in the locality in $1000. All the 14 variables are numeric.

Approach:

- The data is split into testing and training sets in the ratio 75:25
- The four different models are built using appropriate R functions
- Mean Squared Error is used to compare the model performance.

Result: Below is a table showing the out-of-sample prediction performance of all the four models.

Table – Summary statistics

Model	**SSE**
Linear Regression	25.25
Regression Tree	26.43
GAM	13.28
Neural network	28.65

This shows that the generalized additive model is best performing among the four.

EXECUTIVE SUMMARY - GERMAN CREDIT SCORING DATA

Objective: Build multiple models – logistic regression, classification tree, generalized additive model and discriminant analysis – and compare the performance of the models on training and testing data. Based on these, the best performing model needs to be chosen.

Data: The data provided consists of 20 different variables, for 1000 previous customers and whether they were approved for a loan. Of the 20 variables, 13 are qualitative and the rest are numeric.

Approach:

- The data is split into testing and training sets in the ratio 75:25
- The four different models are built using appropriate R functions
- Asymmetric misclassification rate and area under ROC curve (AUC) are used to compare the model performance.

Result: Below is a table showing the out-of-sample prediction performance of all the four models.

Table – Summary statistics

Model	AUC	Asymm. error rate
Logistic Regression	0.78	0.58
Classification Tree	0.68	0.63
GAM	0.71	0.53
LDA	0.68	0.59

Based on AUC, the GLM model is preferred, and based on asymm. error rate, the GAM model is preferred. Since AUC is a more important parameter, the best model is the **classification tree.**

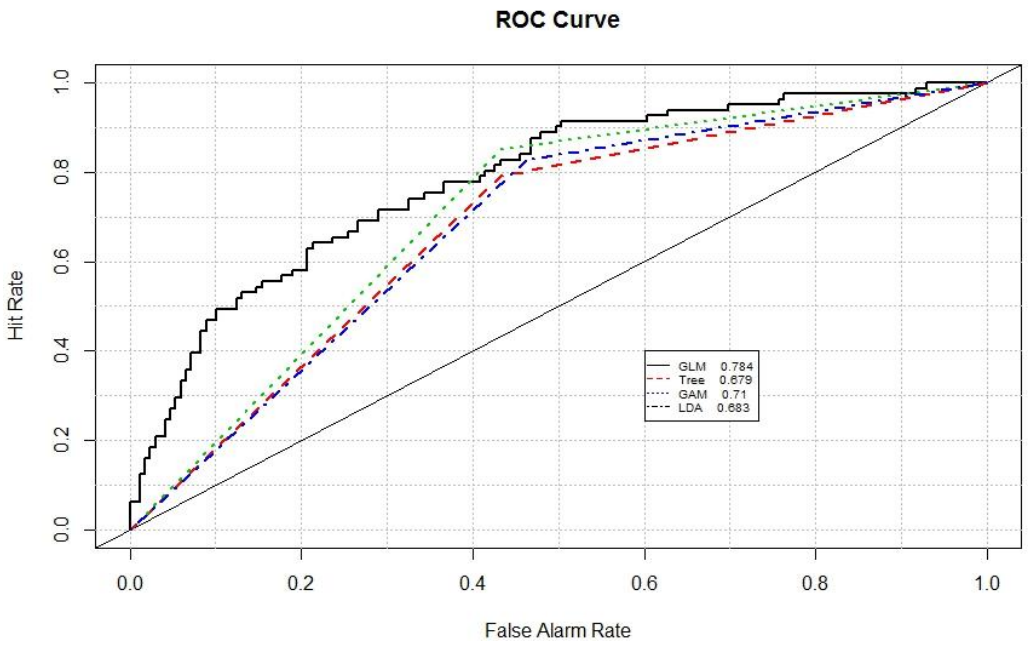


Figure – ROC curve and AUC for out-of-sample performance for all four models

Boston Housing Data

Linear regression

Linear regression model is built by first preparing a 'null model' with none of the variables, and a 'full model' which uses all the variables. Then, step-wise variable selection is used to prepare the final model. 11 of the 13 variables are significant in the final model, with 10 of the variables significant at 1% level.

The in-sample MSE is 21.01, and the out-of-sample MSE is 25.25.

The model details are shown below –

```
call:
lm(formula = medv ~ lstat + rm + ptratio + chas + dis + nox +
    black + zn + crim + rad + tax, data = Boston_train)

             Estimate
(Intercept)  33.594699
lstat        -0.471241
rm            3.951282
ptratio      -0.948892
chas          3.919259
dis          -1.388912
nox         -16.155682
black         0.009868
zn            0.041185
crim         -0.114169
rad           0.288143
tax          -0.011835
```

Figure – Linear regression results

Regression tree

A regression tree is created by first building a 'large tree', which has a very low complexity parameter. Then this tree can then be pruned and a final tree is obtained. The complexity parameter for the pruned tree is 0.017. Below is the output of `plotcp()` which is used to decide on 0.017 complexity parameter.

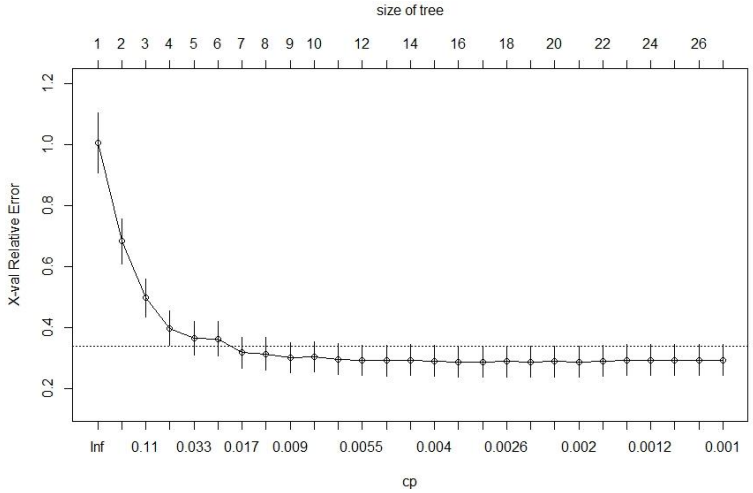


Figure – plotcp() output. The left-most point below the dotted line corresponds to cp = 0.017

The in-sample MSE is 18.04 and out-of-sample prediction MSE is 26.43. Below is the final tree.

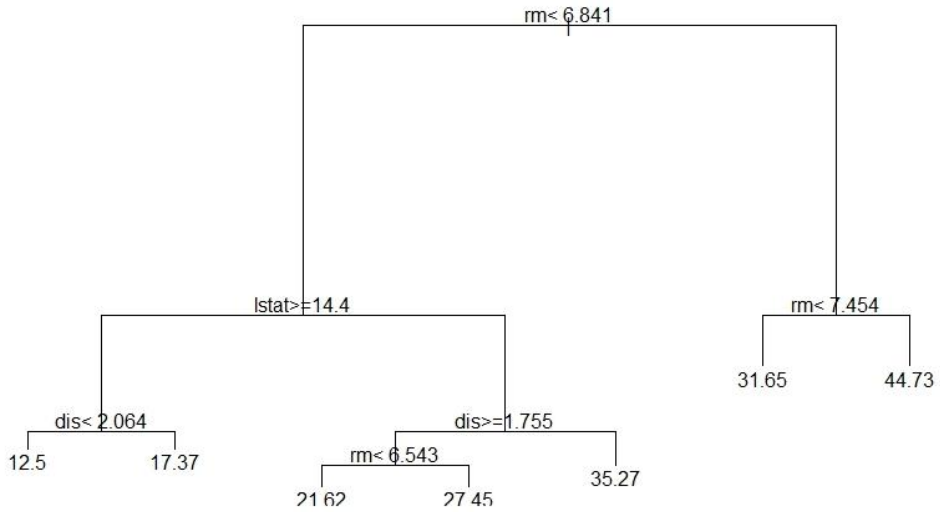


Figure – Regression tree

GENERALIZED ADDITIVE MODEL

GAM is built by summation smoothing or linear functions of predictor variables. While defining the formula for the `gam` function in R, the variables which are continuous and need smoothing splines are defined with `s()`. In the initial model, all continuous variables are smoothed. From this it is seen that smoothing is not significant for three of the continuous variables. Another model is built by removing smoothing for these variables. It is seen that these are not significant without smoothing also. They are removed from the model altogether and a final model is built.

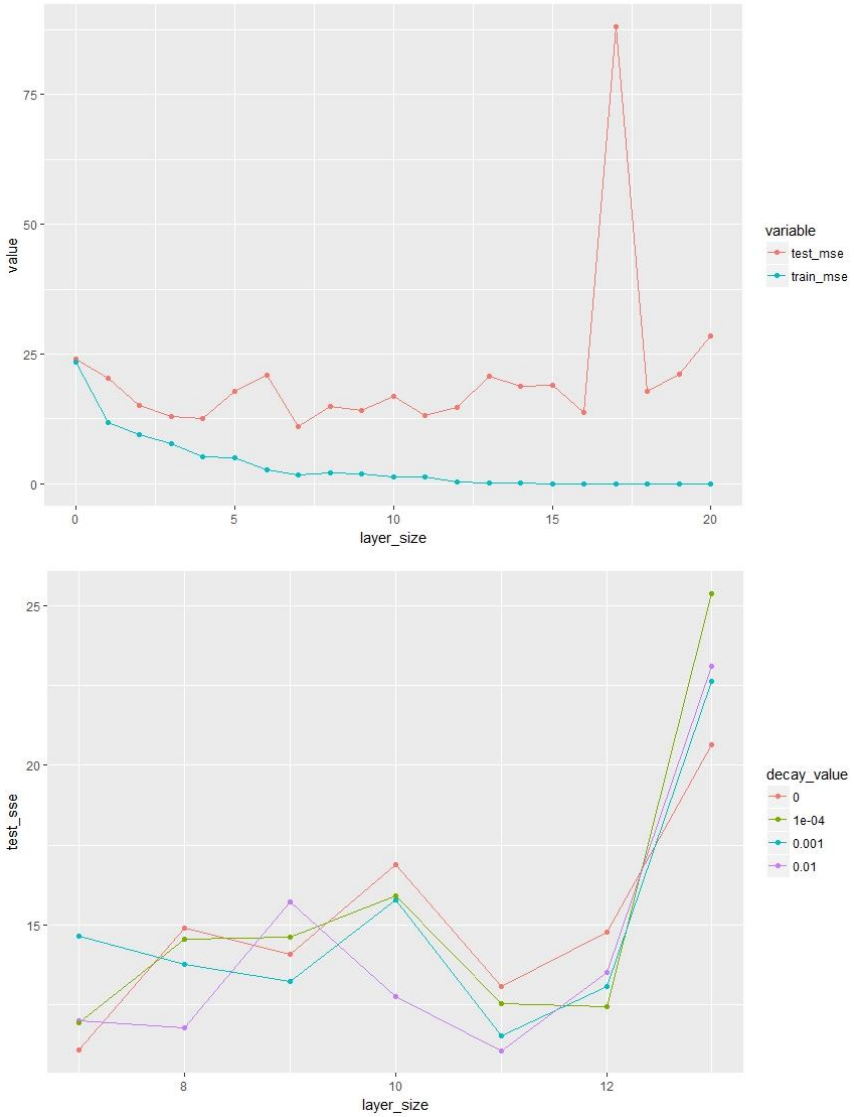
```
Formula:
medv ~ s(crim) + s(indus) + chas + s(nox) + s(rm) + s(dis) +
    rad + s(tax) + ptratio + s(lstat)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  35.1477     2.8433  12.361  < 2e-16 ***
chas          2.1994     0.7646   2.876  0.00429 **
rad           0.3883     0.1231   3.154  0.00176 **
ptratio      -0.8938     0.1329  -6.725 7.78e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
           edf Ref.df      F  p-value
s(crim)  3.782  4.650  9.262 1.57e-07 ***
s(indus) 7.210  8.147  3.736 0.000316 ***
s(nox)   8.994  9.000 13.621  < 2e-16 ***
s(rm)    8.160  8.788 17.383  < 2e-16 ***
s(dis)   8.682  8.967  8.115 5.86e-11 ***
s(tax)   3.655  4.406  8.908 6.56e-07 ***
s(lstat) 5.676  6.873 29.060  < 2e-16 ***
```

Figure – Final GAM

The in-sample prediction MSE is 8.13 and out-of-sample prediction MSE is 13.28.

Neural Network

A neural Network is a black-box model that uses two parameters – 'layer size' and 'decay' to find the final model. The dataset is broken in 75:25 ratio into training and validation sets. The training set is again broken down into training and testing sets. The final layer size is 11 with a decay value of 0.01.



Figure

Figure – Neural network parameters

The in-sample prediction MSE is 3.97 and the validation dataset prediction MSE is 28.65.

German Credit Scoring Data

Logistic Regression

In the first iteration of the model building, all the 20 variables are included. This led to 49 variables, including the intercept and dummy coded variables for the factor type ones, of which 17 are significant at the 5% level. The AIC value is 753.82 and BIC is 980.20.

Step-wise variable selection is done to improve the model. This gives 14 variables – leading to 38 variables including the dummy variables, with 16 of them significant. The AIC value is 743.12 BIC value is 918.67.

The asymmetric misclassification rate using a 5:1 cost is 0.452

The ROC curve with AUC value is shown below. The AUC value is 0.834

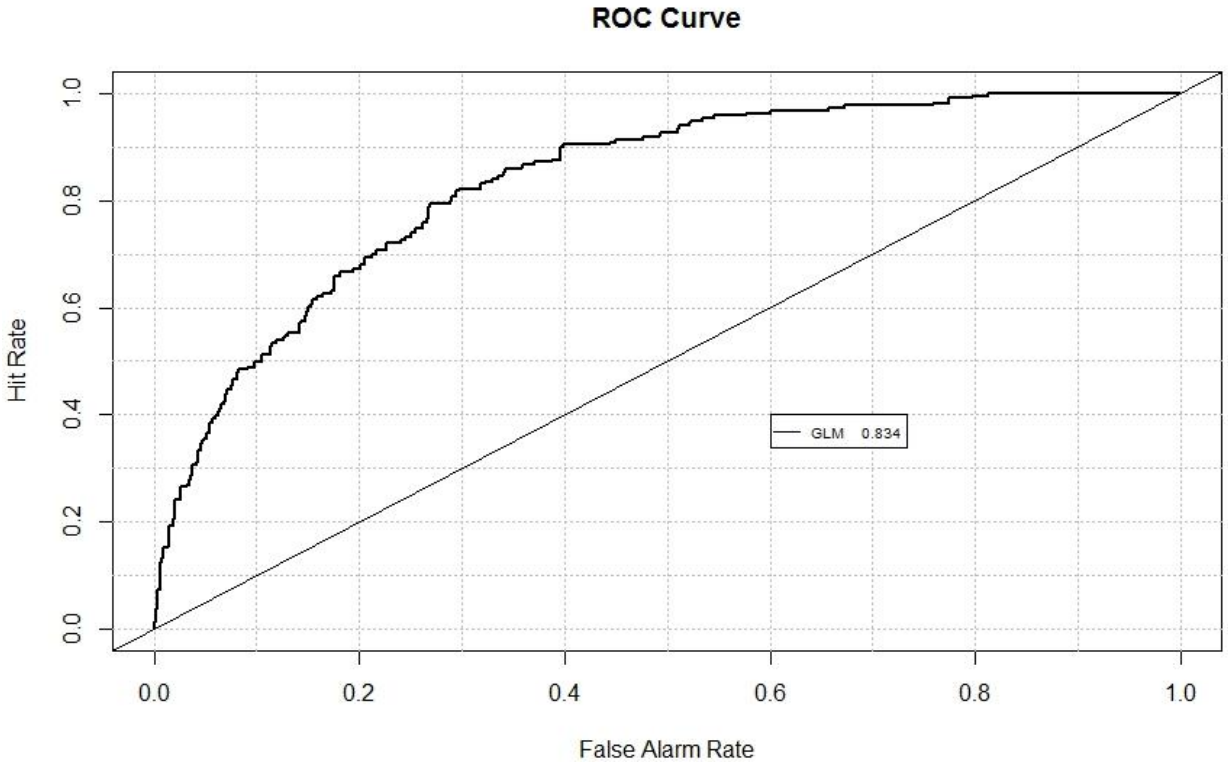


Figure – ROC curve and AUC value for in sample performance of GLM

The out of sample misclassification rate is 0.58. Similarly, ROC curve for out-of-sample performance is shown below. The AUC value is 0.784.

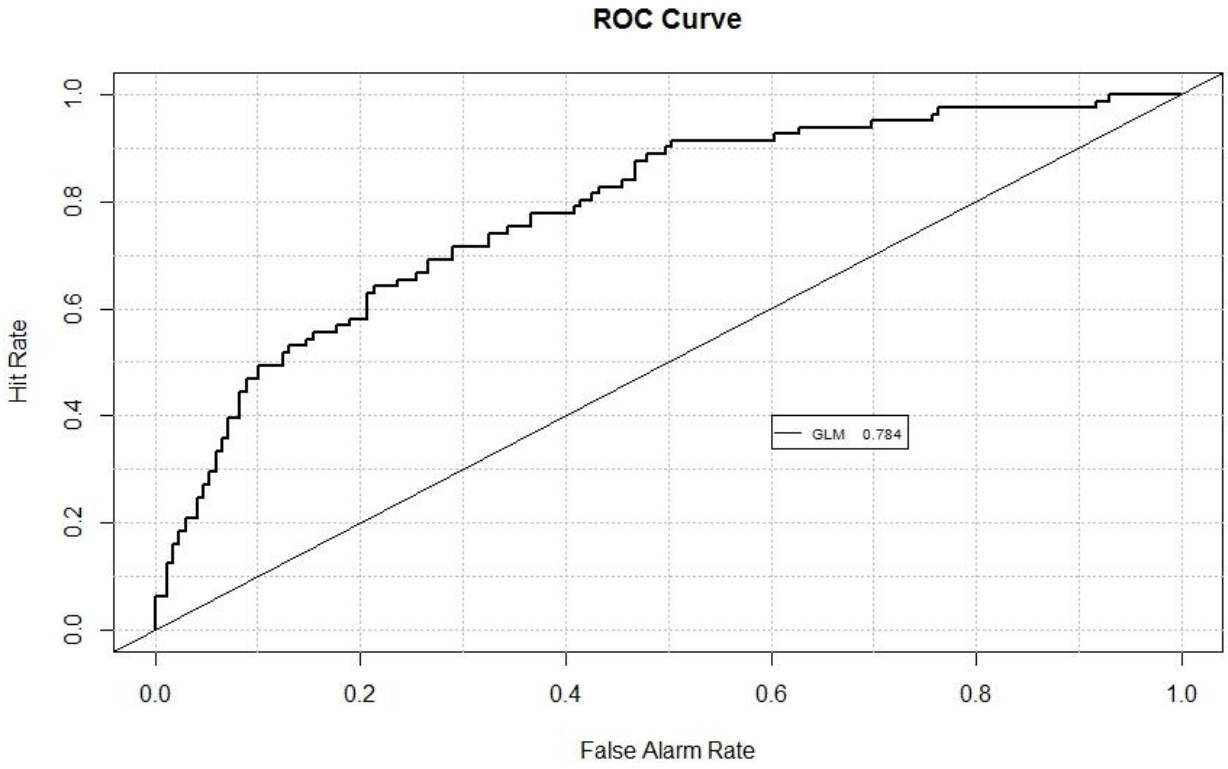


Figure – ROC curve and AUC value for out of sample performance of GLM

As can be expected, the out of sample performance of the model is lesser than the in-sample performance.

CLASSIFICATION TREE

Classification tree is shown below –

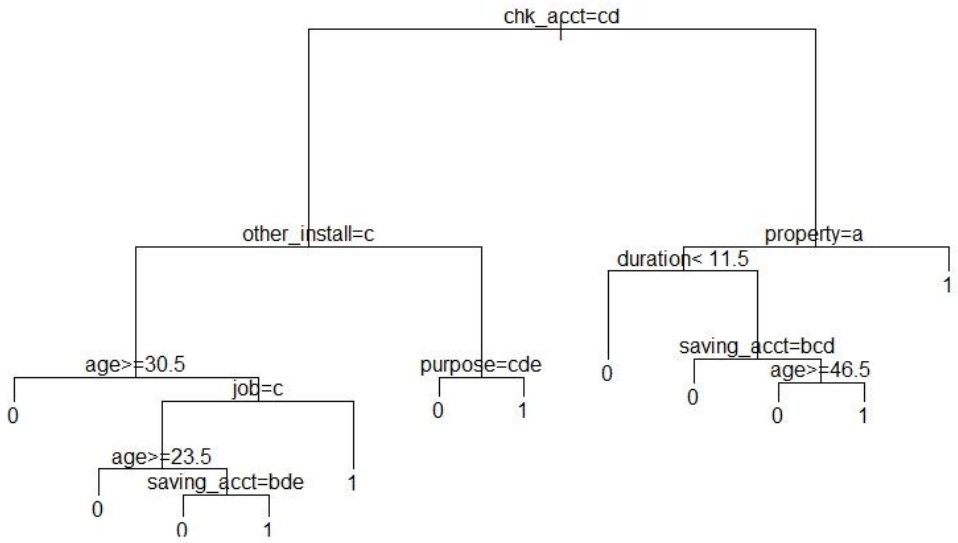


Figure – Classification tree

The misclassification rate for in sample prediction is 0.39. The ROC curve with AUC value is shown below. The AUC value is 0.76.

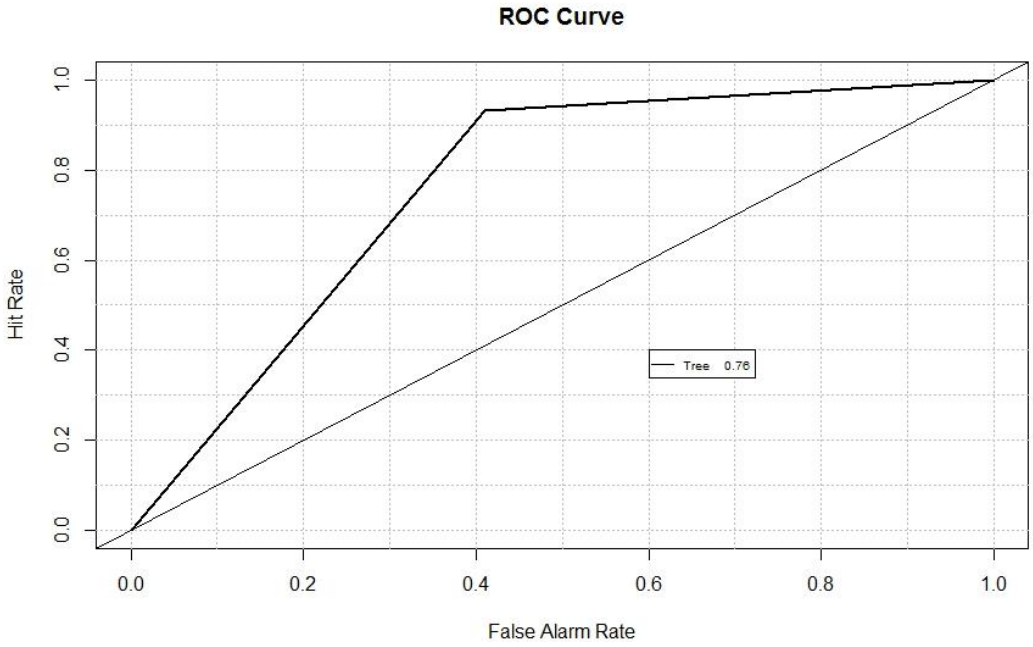


Figure – ROC curve and AUC value for in sample performance of classification tree

The same above metrics for out-of-sample performance are-

Asymmetric misclassification rate – 0.63

AUC – 0.68

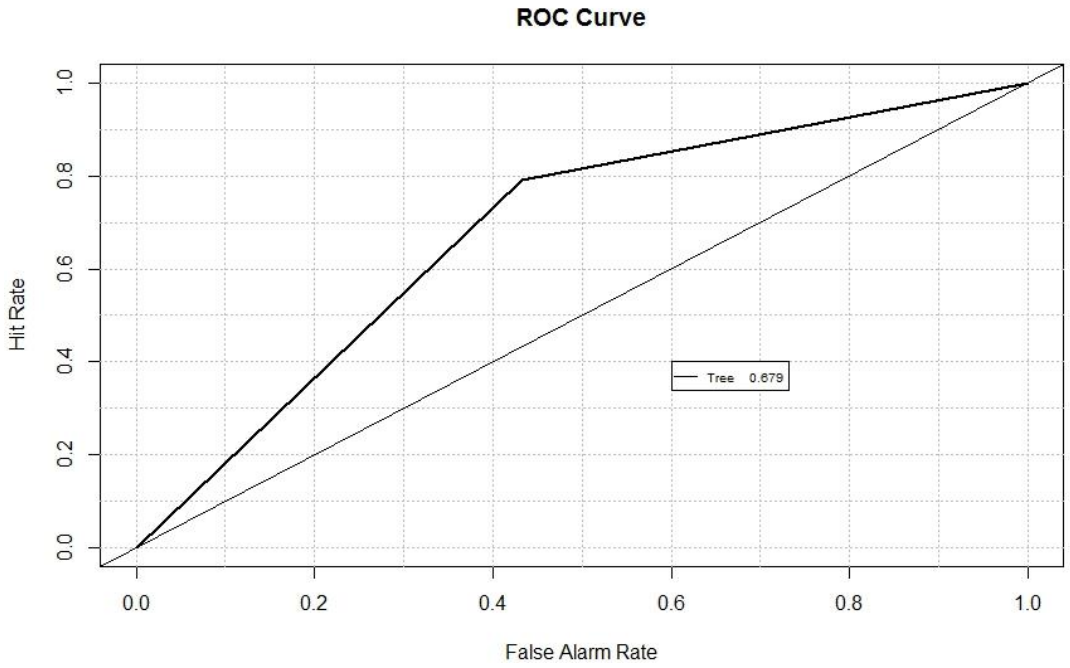


Figure – ROC curve and AUC value for out of sample performance of classification tree

The misclassification rate for out of sample performance (0.63) is much larger than that for in sample performance (0.39).

GENERALIZED ADDITIVE MODEL

The `gam` function is used to build the model. Three variables – duration, loan amount, and age – are defined using `s()` for non-parametric smoothing splines.

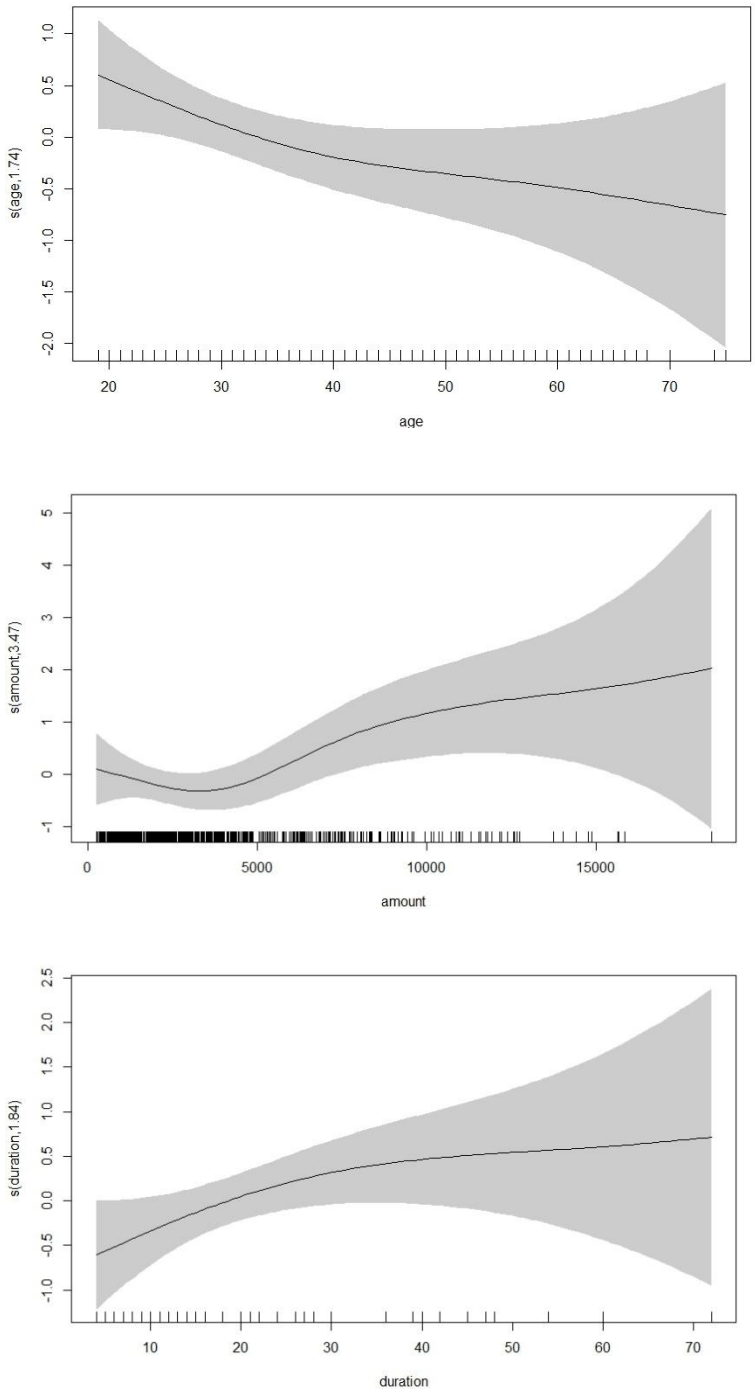


Figure – Splines for the three variables age, amount, and duration (top to bottom)

The in-sample prediction performance is shown below

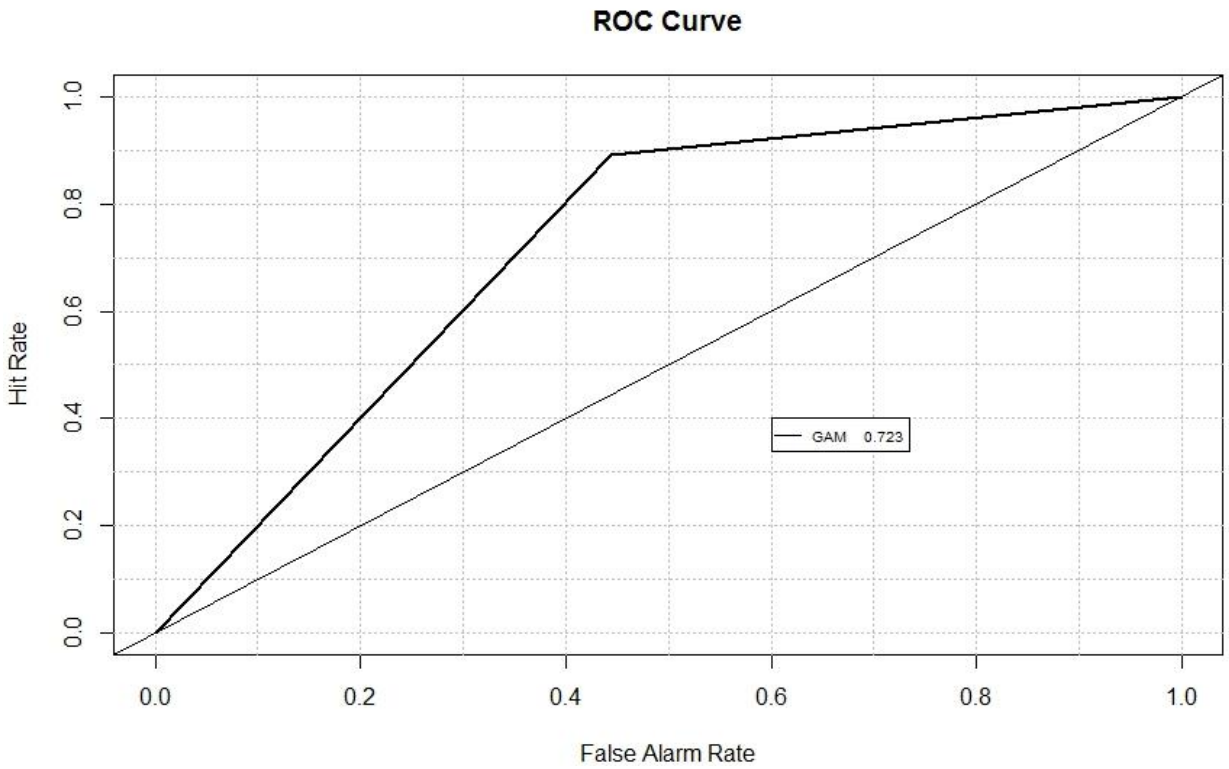


Figure – ROC curve and AUC value for in sample performance of GAM

Asymmetric misclassification rate – 0.47

AUC – 0.72

And the out of sample performance parameters are – asymmetric misclassification rate is 0.53 and AUC is 0.71.

Latent Discriminant Analysis

LDA model was built and the in-sample performance metrics are observed. The asymmetric misclassification rate is 0.47. The area under the ROC curve is 0.738. In contrast, the out-of-sample prediction performance parameters are – misclassification rate of 0.59 and area under the ROC curve is 0.68. The ROC curve for in-sample prediction and out-of-sample prediction are shown below.

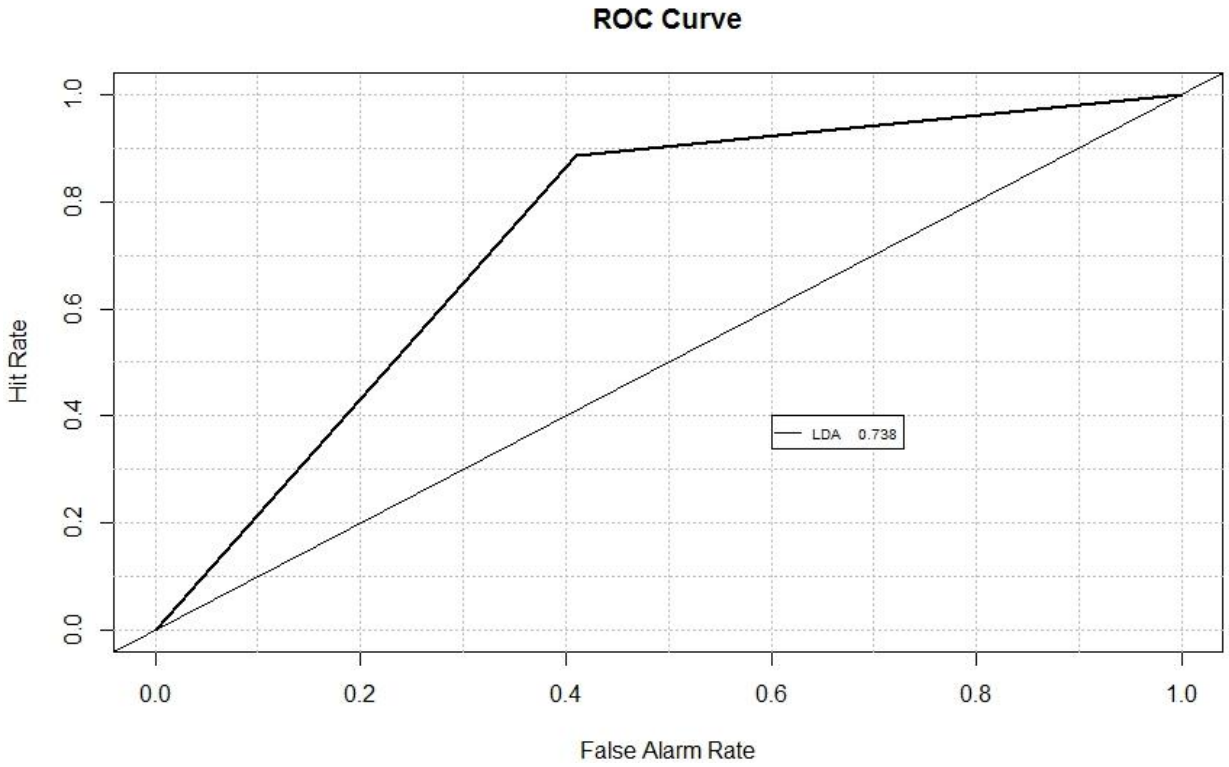


Figure – ROC curve and AUC value for in sample performance of LDA

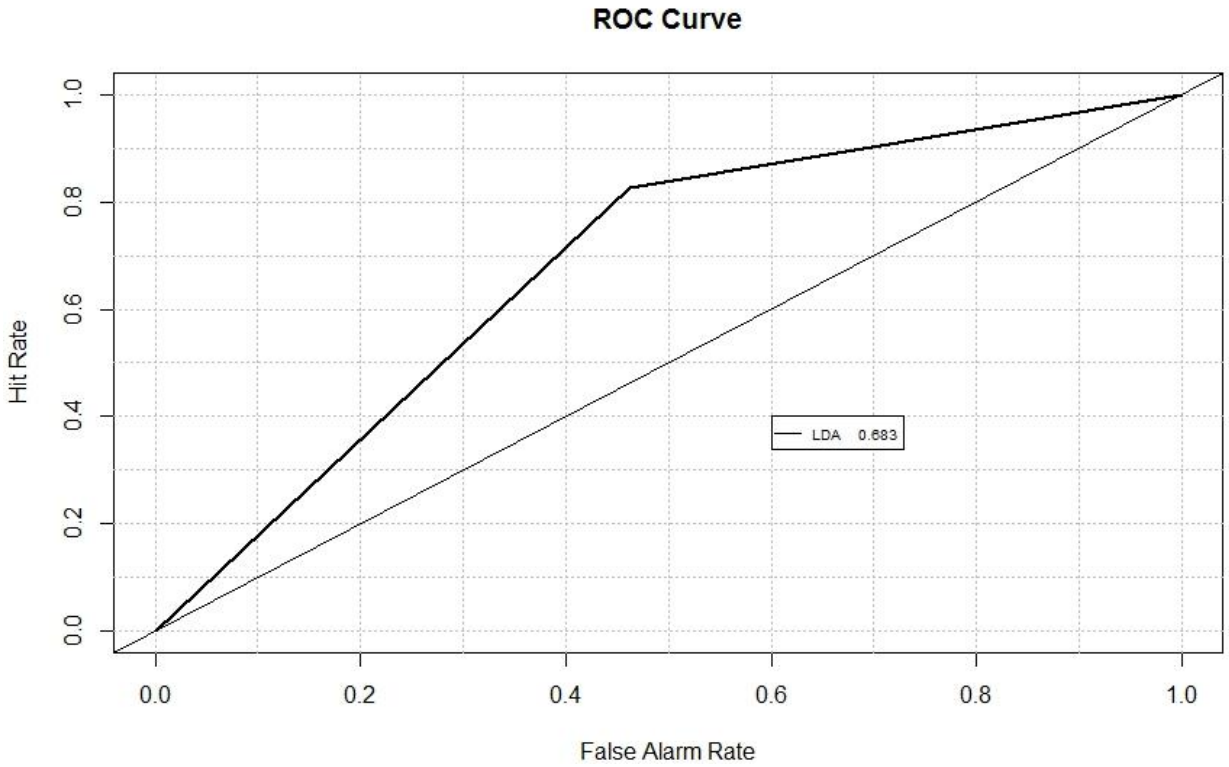


Figure – ROC curve and AUC value for out-of-sample sample performance of LDA