BANA 7047- INDIVIDUAL CASE I

Last Name: Sharma ***First Name: Swapnil***

MARCH 20, 2017

Table of Contents

Executive Summary: Boston Housing Data Analysis

Goal: The goal is to find the relationship on what parameters does the price of houses in Boston depend.

Approach: First the EDA is done to study and understand the variables. Four types of models are built. They are Linear Regression, Generalized Additive Model, Regression Tree, and Neural Network. Stepwise variable selection method is adopted after comparing BIC values with other variable selection methods. Regression tree is constructed after pruning and GAM and Neural Network models are also built. All the models are compared with each other using the mean square error value for in sample data. The model with the least MSE for training data set is chosen as the final model and that model is used to test the remaining 25% of the test data set.

Summary: Below table shows and compares the MSE across different statistical model

Table 1 In Sample MSE Comparison

	In Sample
	MSE
GLM	24.90
GAM	**9.01**
Regression Tree	12.29
NNET	82.42

GAM model gives the least In Sample error and is used to build the final model on which Out Sample data set is tested. Below table gives the Out sample MSE for GAM model.

Table 2 Out sample MSE for final model

	Out Sample
	MSE
GAM	10.09

Executive Summary: German Credit Scoring Data Analysis

Goal: Predictive Modelling has a major role to play in making the decision if a loan is to be given to an individual or not. The candidates are classified into good credit risk and bad credit risk candidates. The data set we have is having 20 variables for 1000 applicants. The model will help in making the decision if loan should be given to applicant or not.

Approach: The data set is divided into training set (75%) and test set (25%) randomly by setting the seed number as (10743959) my M-number. First EDA is done by finding summary of variables, plotting correlations, and looking for outliers. Then Variable selection is done using stepwise method after comparing backward, forward, and stepwise method based on the BIC criteria. Logistic regression model is built with 1/6 cut off probability and misclassification rate and AUC for Training data (In sample) and Testing (Out sample) is noted and ROC is plotted. In total, we build four different models which are Logistic regression, GAM, CART and LDA. They are compared based on AUC for in sample data and final model is chosen for which the out-sample data is tested.

Summary:

Table 3 In Sample AUC and Misclassification Rate Comparison

	In Sample	
	Misclassification Rate	**AUC**
GLM	34.40%	0.73
CART	32.67%	0.77
GAM	33.47%	0.73
LDA	33.50%	**0.85**

Based on the AUC we select LDA as final model. The misclassification rate and AUC for Out Sample data using LDA is shown in table below

Table 4 AUC and Misclassification rate for final LDA model

	Out Sample	
	Misclassification Rate	**AUC**
LDA	**34.40%**	**0.76**

Thus, LDA model outperforms other statistical models and should be used by Manager to decide if candidate should be given a loan or not

Supervised Learning: Questions

Q1) Which of the following model building approach should be used when you have Categorical Response Variable and categorical Predictor Variable?

a) ANOVA
b) Linear Regression
c) Logistic Regression
d) Analysis of Covariance

Q2) Which of the following model building approach should be used when you have Continuous Response Variable and both continuous and categorical Predictor Variable?

a) ANOVA
b) Linear Regression
c) Logistic Regression
d) Analysis of Covariance

Q1) Boston Housing Data

Linear Regression

Multivariate Linear regression model is built on the given data set with 506 observations and 13 variables. Different variable selections method is run and finally stepwise is selected based on the BIC criteria. The final model contains 10 variables

medv ~ lstat + rm + ptratio + dis + nox + rad + tax + zn + chas + black

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  39.100745   5.906146   6.620 1.27e-10 ***
lstat        -0.584761   0.055660 -10.506  < 2e-16 ***
rm            3.302782   0.468480   7.050 8.89e-12 ***
ptratio      -0.866189   0.155409  -5.574 4.83e-08 ***
dis          -1.549910   0.222323  -6.971 1.46e-11 ***
nox         -16.436108   4.171291  -3.940 9.74e-05 ***
chas          2.604710   1.007484   2.585  0.01011 *
zn            0.052583   0.016415   3.203  0.00148 **
black         0.009198   0.003276   2.807  0.00526 **
rad           0.294559   0.076894   3.831  0.00015 ***
tax          -0.013595   0.004209  -3.230  0.00135 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.99 on 368 degrees of freedom
Multiple R-squared:  0.7214,    Adjusted R-squared:  0.7138
F-statistic: 95.29 on 10 and 368 DF,  p-value: < 2.2e-16
```

Figure 1 Linear Regression Parameter estimates

In Sample

We get following In Sample model statistics

AIC	2306.811
BIC	2354.061
Model MSE	24.8989
Adj. R-squared	0.7138

Out Sample

Model MSE	18.11627

Regression Tree:

We have a continuous response variable and hence a regression tree is built. Pruning is done while building the tree and we obtain ideal nodes to be 7 with optimal cp value. It is observed that 9 variables are used in the tree.

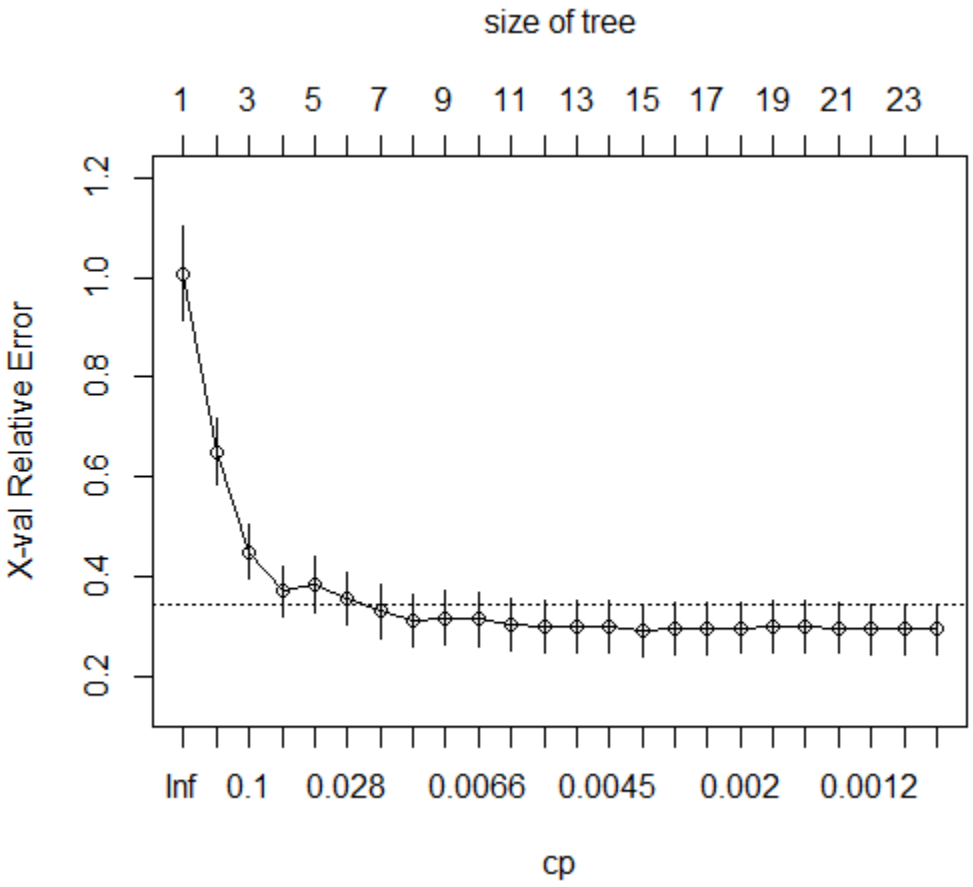

Figure 2 Plot of CP

In Sample statistics after Pruning

Model MSE	12.2872

Out Sample statistic after Pruning

Model MSE	10.95072

GAM

GAM is a generalized linear model in which we predict the behavior of the response variable by using a sum of smoothing and linear functions of predictor variable.

model.gam <- medv ~
s(crim)+zn+s(indus)+chas+s(nox)+s(rm)+age+s(dis)+rad+s(tax)+ptratio+s(black)+s(lstat)

In Sample statistics after Pruning

Model MSE	9.012159

Out Sample statistic after Pruning

Model MSE	10.08873

Neural Network

Neural Networks are black box models used for prediction. We use nnet package for fitting a neural network.

In Sample statistics

Model MSE	82.42

Out Sample statistic

Model MSE	90.40

Q 2) German Credit Scoring

Logistic Regression

- The first logistic regression model is built consisting of all the variables in the dataset. We check the summary analyses, parameters, and estimates. We also, check the AIC and BIC values. In this case 16 variables are statistically significant at a significance level of 5%. The AIC value is 733.9404 and BIC value is 960.324
- Next, we built a better model using stepwise variable selection using the AIC method. Here, only those variables are selected which are a best fit for the model. We check the parameters and estimates along with the AIC, BIC values. We also check the variables that get selected. Here we get an AIC value of 719.5545. This is much better than the previous case.
- We then make another logistic regression model using the BIC method. Here also we use stepwise variable selection. We check the parameters and estimates and the AIC and BIC values. The model selects the best variables and we get BIC as 814.6325.

Summary of stepwise model 1 (AIC)

1. AIC value 719.5545
2. BIC value 881.257

Summary of stepwise model 1 (BIC)

1. AIC value 791.5321
2. BIC value 814.6325

Final Model used:

german_credit_model = glm(response ~ chk_acct + duration + purpose + credit_his + saving_acct +installment_rate + amount + sex + telephone + foreign + other_install + other_debtor + present_resid)

In Sample:

The misclassification matrix for model 1 is as follows (with the cut off probability 1/6):

This is also known as an error matrix or confusion matrix which is a matrix of predicted value v/s true value.

	Predicted	
Truth	0	1
0	292	235
1	23	200

The correct rate: = 65.6%

The misclassification rate: = 34.4%

AUC: 0.7254703

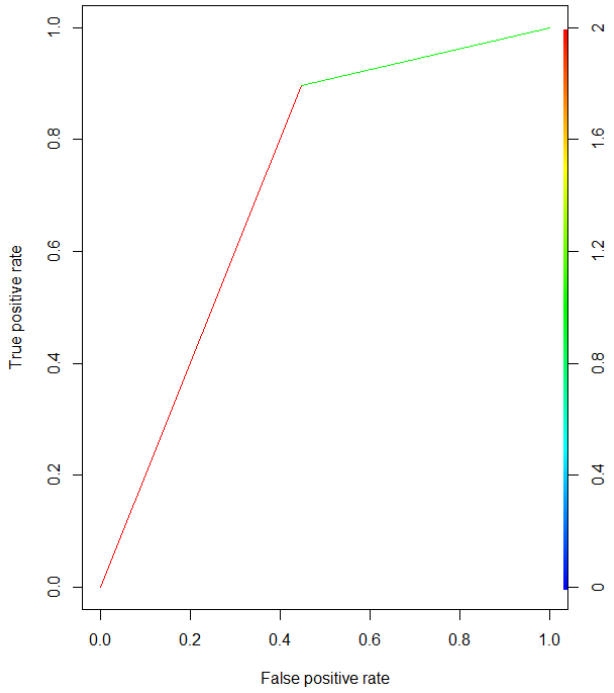

Figure 3 GLM: ROC In- Sample

Out Sample:

	Predicted	
Truth	0	1
0	102	71
1	17	60

The correct rate: = 64.8%

The misclassification rate: = 35.2%

AUC: 0.6844081

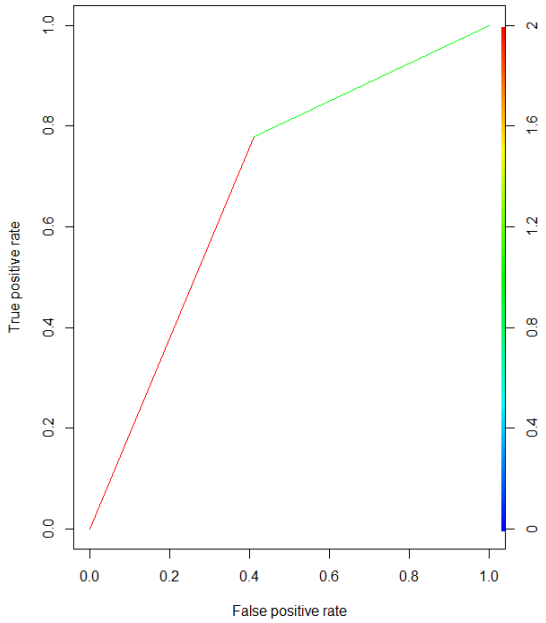

Figure 4 GLM: ROC Out- Sample

The model performance is compared with other statistical models based on misclassification rate and AUC

Classification Tree:

We use classification tree as the response variable is binary.

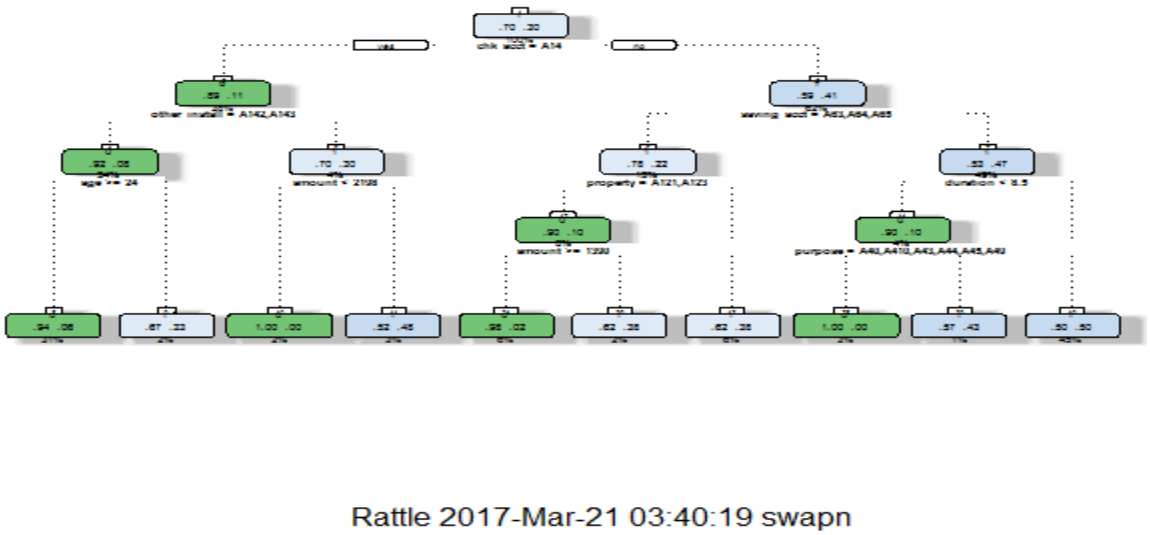

Figure 5 Classification Tree

Classification Tree

In Sample:

The misclassification rate: = 32.67%

AUC: 0.7668587

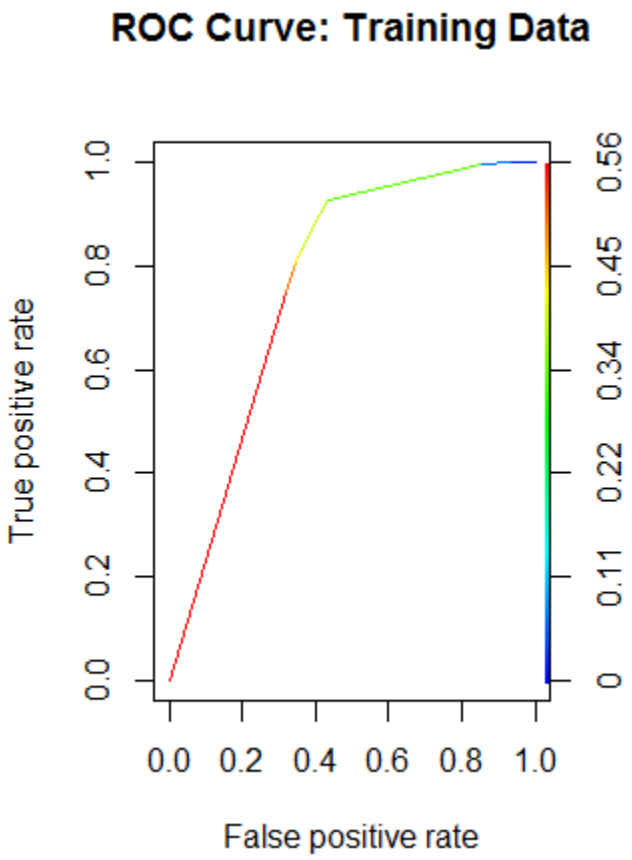

Figure 6 Classification Tree: In Sample ROC

Out Sample:

The misclassification rate: = 40 %

AUC: 0.6432325

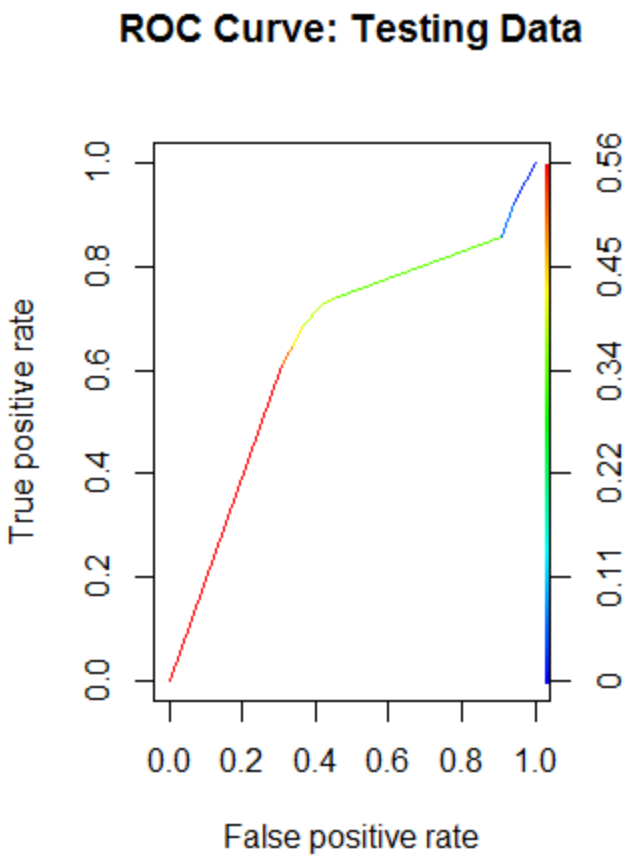

Figure 7 Classification Tree: Out-Sample ROC

GAM

Generalized additive model is built to compare with other models. It's a non-parametric extension of GLMs.

In Sample

Misclassification rate

	Predicted	
Truth	0	1
0	297	230
1	21	202

The misclassification rate: = 33.47 %

AUC: 0.7346985

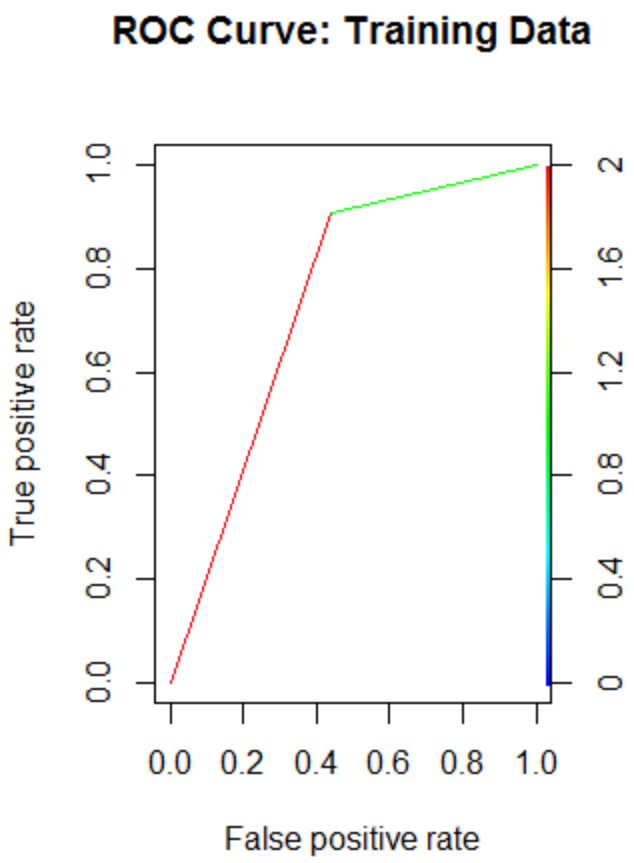

Figure 8 GAM: In-Sample ROC

Out Sample:

Misclassification rate

	Predicted	
Truth	0	1
0	111	62
1	19	58

The misclassification rate: = 32.4 %

AUC: 0.6974326

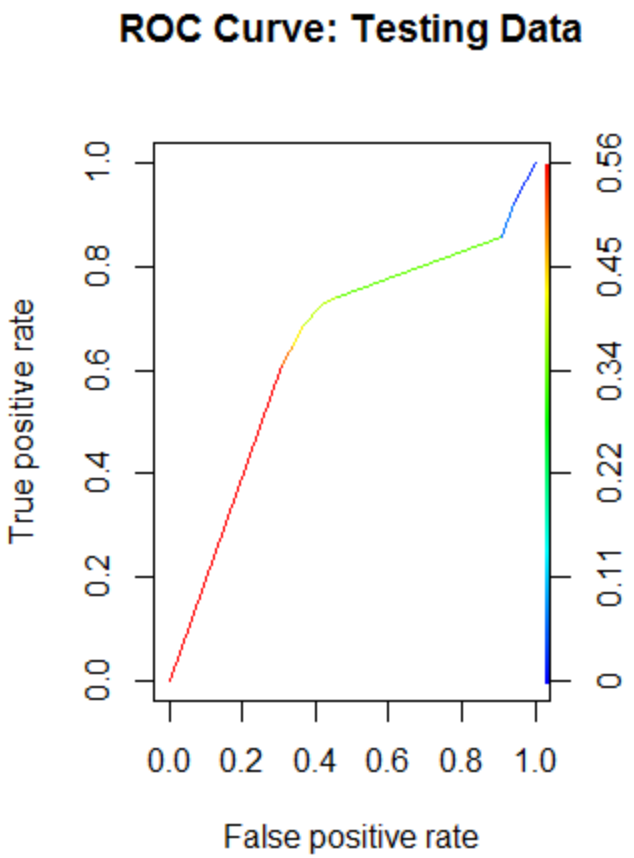

Figure 9 GAM: Out-Sample ROC

Linear Discriminant Analysis:

This is a method used in statistics which is generalization of Fisher's linear discriminant.

In Sample:

Misclassification rate

	Predicted	
Truth	0	1
0	303	224
1	27	196

The misclassification rate: = 33.5 %

AUC: 0.8482314

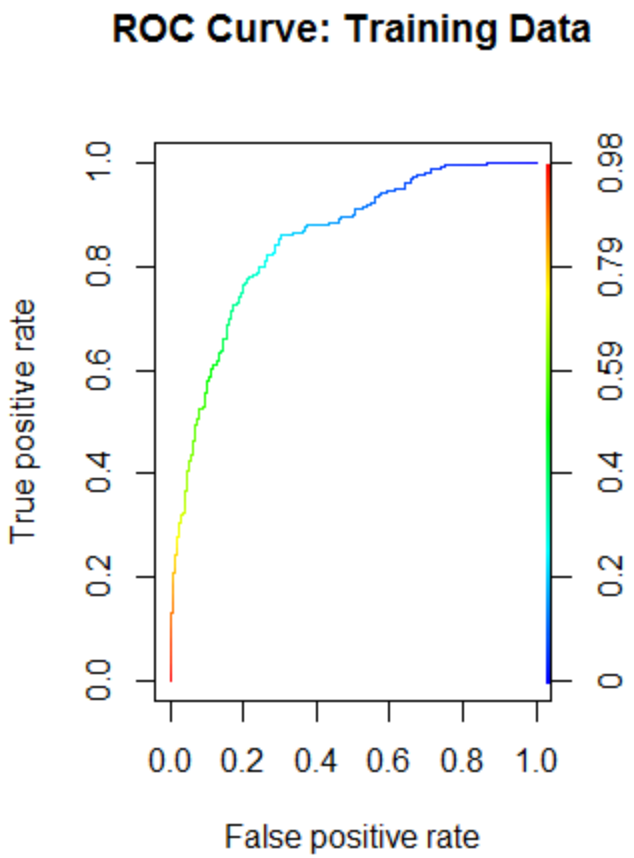

Figure 10 LDA: In-Sample ROC

Out Sample:

Misclassification rate

	Predicted	
Truth	0	1
0	101	72
1	14	63

The misclassification rate: = 34.4 %

AUC: 0.7612792

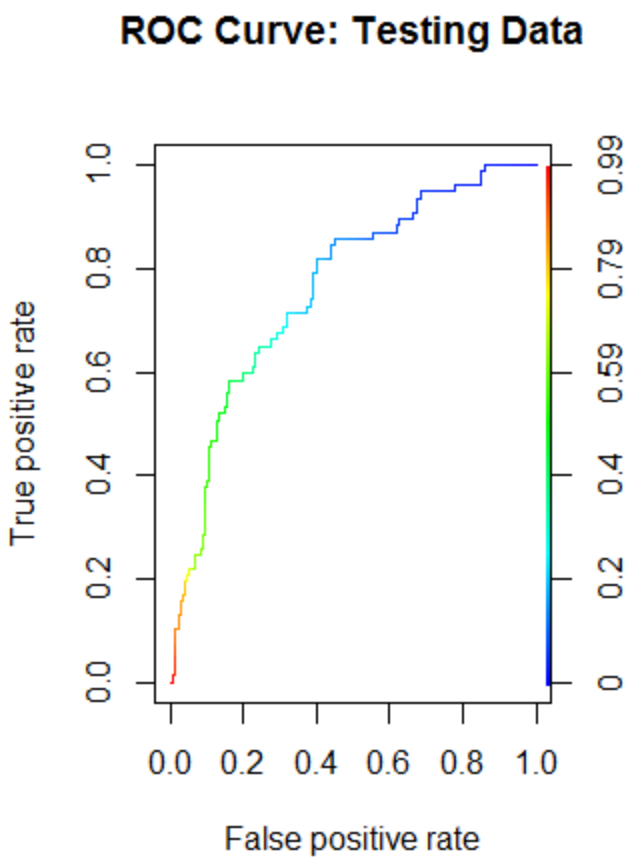

Figure 11 LDA: Out-Sample ROC