BANA 7047 – Prof. Yan Yu

Individual Case 1

Last Name: BADRE

First Name: SHASHANK

UCID: M12383328

Boston Housing Data

Summary

Boston data set contains housing values in suburbs of Boston. The Boston data set is a data frame and consists of 506 rows and 14 columns. I have used Linear Regression method, Regression Tree method, General Additive Model method and Neural Network method to see which method can give us the best predictive model to predict with lowest sum square error the median housing values in the suburbs of Boston. For Boston housing data set, I have looked at the structure of the data set and then summarized the data. As part of Exploratory Data Analysis on Boston housing data set, I have looked at the correlation of variables and it has been found that certain variables are strongly correlated.

Boston housing data set was split randomly in 80 % train data set and 20 % test data set. Predictive models were built using train data set and then measures were noted for IN sample data set. I have predicted values for test data set and then measured Out of sample measures. Following table gives the result obtained from Linear Regression method, Regression Tree method, General Additive Model method and Neural Network method.

Method	In Sample Average sum square error	Out of sample average sum square error
GLM	23.159	17.465
CART	14.987	26.725
GAM	8.89	9.17
neuralnet	4.279	8.934

Best model for predicting median housing values in the suburbs of Boston is found to be Neural Network Model. Final neural network model has 2 hidden layers with 5 and 3 nodes respectively. The In sample Average sum square error for this model is lowest among all models which is 4.279. Also, out of sample average sum square error is lowest for neural network model which is 8.934.

German Credit Score Data

Summary

German credit score data has records that are classified as good credit and bad credit. We will use this data set to train our model to predict whether a test record is a good credit or bad credit. German credit score data set consists of 1000 observations and 21 features. Variable "Response" is the dependent variable in this data set. Response variable 0 indicate "good credit risk" and 1 indicates "bad credit risk". For German credit score data, I have looked at the structure of the data set and then summarized the data. As part of Exploratory Data Analysis on German credit score data, I have looked at the correlation of variables and it has been found that variables amount and suration are strongly correlated.

German credit score data set was split randomly in 80 % train data set and 20 % test data set. Predictive models were built using train data set and then measures were noted for IN sample data set. I have predicted values for test data set and then measured Out of sample measures. Following table gives the result obtained from Logistic Regression method, Classification Tree method, General Additive Model method, Neural Network and Linear Discriminant Analysis method.

Method	Misclassification rate		Area under Curve	
	In Sample	Out of Sample	In Sample	Out of Sample
Logistic Regression	0.32	0.405	0.8477	0.7278
Classification Tree	0.2925	0.415	0.874	0.7014
GAM	0.3138	0.395	85.24	72.98
NNET	0.2988	0.38		
LDA	0.29	0.375	84.86	72.67

Best model for predicting good credit and bad credit for German Credit Score data is found to be Linear Discriminant Analysis Model. Final LDA model has misclassification rate of 0.29 for In sample data and 0.375 for Out of sample data, which is lowest as compared to all other models.

1. Boston Housing Data

Boston data set contains housing values in suburbs of Boston. The Boston data set is a data frame and consists of 506 rows and 14 columns. Let us look at the structure of this data frame.

Crim: This variable is per capita crime rate by town. It is numeric data type.

Zn: This variable is proportion of residential land zoned for lots over 25,000 sq.ft. It is numeric data type.

Indus: This variable is proportion of non-retail business acres per town. It is numeric data type.

Chas: This variable is Charles River dummy variable (= 1 if tract bounds river; 0 otherwise). It is integer data type

Nox: This variable is nitrogen oxides concentration (parts per 10 million). It is numeric data type.

Rm: This variable is average number of rooms per dwelling. It is numeric data type.

Age: This variable is proportion of owner-occupied units built prior to 1940. It is numeric data type.

Dis: This variable is weighted mean of distances to five Boston employment centres. It is numeric data type.

Rad: This variable is index of accessibility to radial highways. It is integer data type

Tax: This variable is full-value property-tax rate per \$10,000. It is numeric data type.

Ptratio: his variable is variable is pupil to teacher ratio. It is numeric data type.

Black: This variable is 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town. It is numeric data type.

Let us look at the summary of the Boston Housing data set.

	Crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	Medv
Min	0.01	0.00	0.46	0.00	0.39	3.56	2.90	1.13	1.00	187.00	12.60	0.32	1.73	5.00
First quartile	0.08	0.00	5.19	0.00	0.45	5.89	45.02	2.1	4.00	279	17.4	375.38	6.95	17.02
Median	0.26	0.00	9.69	0.00	0.53	6.21	77.50	3.21	5.00	330	19.05	391.44	11.36	21.2
Mean	3.61	11.36	11.14	0.07	0.55	6.29	68.57	3.80	9.55	408.2	18.46	356.67	12.65	22.53
Third quartile	3.68	12.50	18.10	0.00	0.62	6.62	94.08	5.19	24.00	666.00	20.20	396.23	16.95	25.00
Max	88.98	100.00	27.74	1.00	0.87	8.78	100.00	12.13	24.00	711.00	22.00	396.90	37.97	50.00

Table: Summary for Boston Data set

Let us look at the pairwise correlation between variables in the Boston data set.

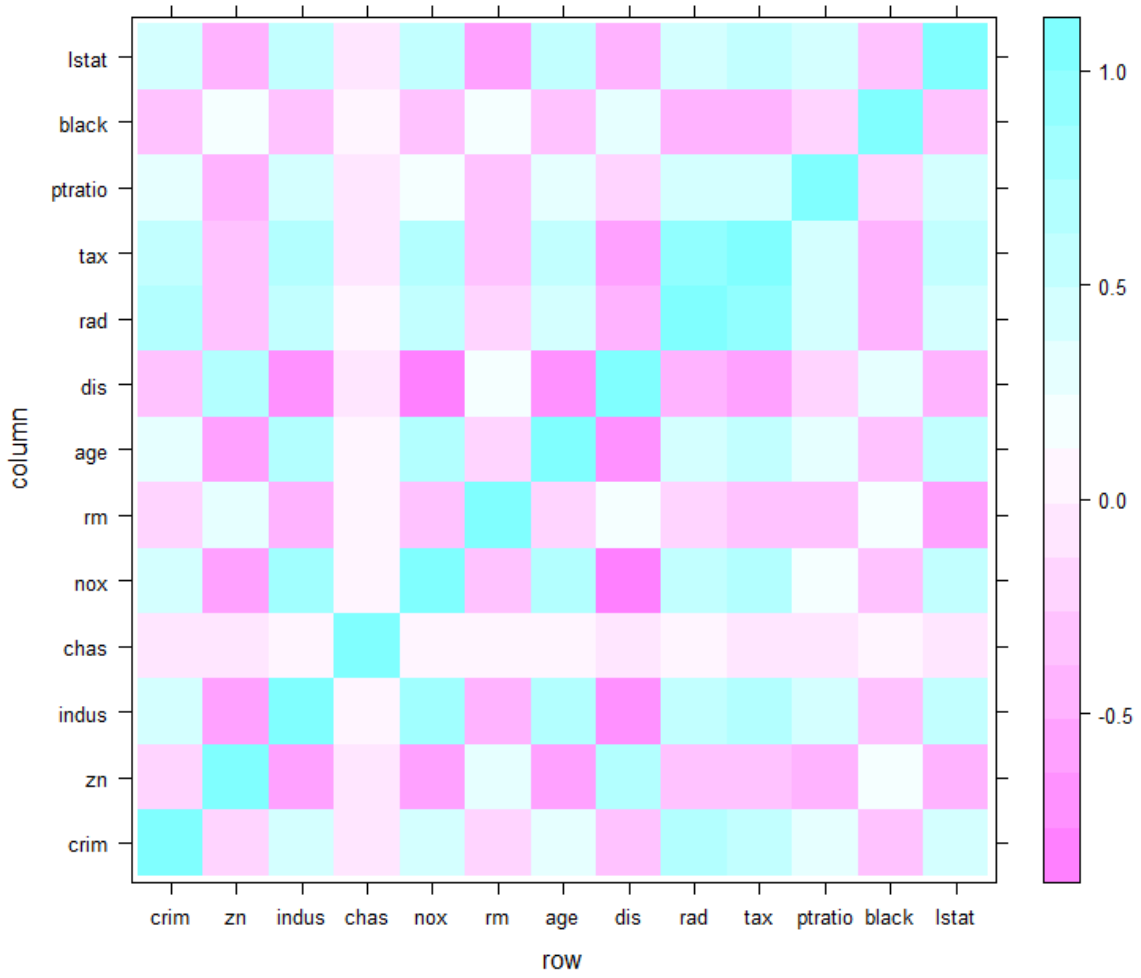


Fig: Correlation between variables in Boston data set

From the above correlation plot we can conclude below things.

- Dark blue color suggests strong positive correlation between variables
- Dark pink color suggests strong negative correlation between variables

To model Boston data set I have split the data into 80 percent train data and 20 percent test data by using seed 12383328

Linear Regression on Boston Data set

Using AIC in both direction the final regression model that I have got is given below

Medv = -0.12 * crom + 0.038 * zn + 3.28 * chas - 17.494 * nox + 3.98 * rm – 1.4 * dis + 0.36 * rad – 0.013 * tax – 0.992 * ptratio + 0.01 * black – 0.499 * lstat

Thus this regression model equation can be interpreted as if per capita crime rate by town i.e. Crim increased by 1 then median housing value in boston decreases by 0.12

In Sample: Average square error for this model is 23.159

AIC value for this model is 2448.018

Now we use this model to predict median housing values for test data.

Out of sample mean square prediction error is 17.465

Prediction housing values in Boston using CART

Above, we have used linear regression model to predict housing values in Boston. Now, I will use Classification and regression tree method to determine housing values in Boston. We would like to know if we can achieve better results using CART method as compared to Linear Regression method.

Using rpart in R on train data we get following best fitted regression tree.

```
n= 405

node), split, n, deviance, yval
      * denotes terminal node

 1) root 405 36191.0400 22.71506
   2) lstat>=9.725 234  5702.5870 17.22949
     4) crim>=5.84803 73  1051.2000 12.48356 *
     5) crim< 5.84803 161  2261.6240 19.38137
      10) lstat>=14.795 74   788.3659 17.22703 *
      11) lstat< 14.795 87   837.6834 21.21379 *
   3) lstat< 9.725 171 13811.4100 30.22164
     6) rm< 7.435 145  6069.5840 27.56138
      12) tax< 548 138  3458.5630 26.76739
        24) rm< 6.6385 81   792.5622 23.68148 *
        25) rm>=6.6385 57   798.5221 31.15263 *
      13) tax>=548 7   808.9286 43.21429 *
     7) rm>=7.435 26   992.8435 45.05769 *
```

Fig: Fitted regression tree on Boston values

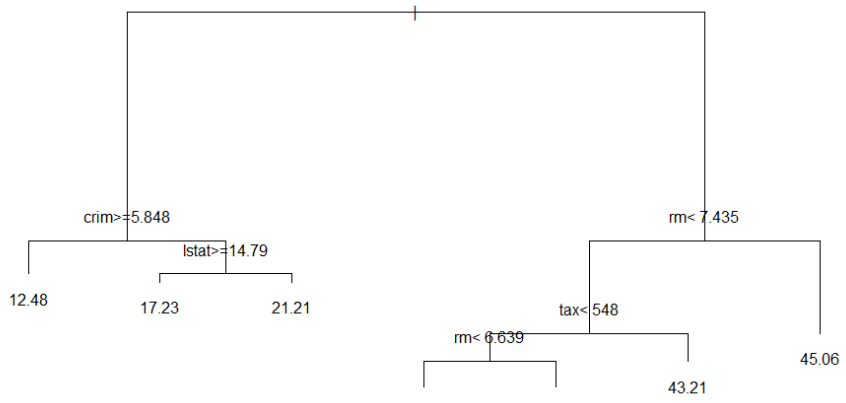


Fig: Regression tree plot for predicting Boston housing values

We prune the full grown tree to get the best tree.

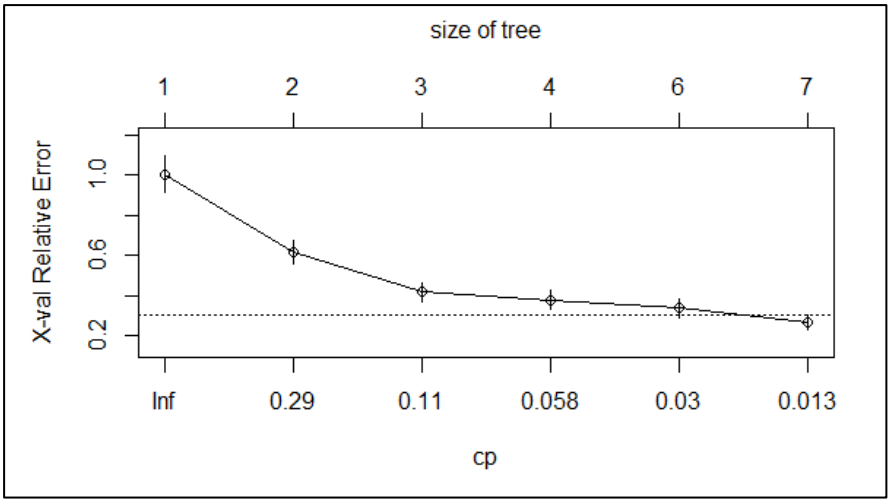


Fig: Pruning Tree

We can see that 0.013 is the best cp value as it is the leftmost value below the horizontal line.

After predicting values using the regression tree obtained from train data, I predicted values for train data. After comparing predicted values for train data and actual values of train data the average square error is found to be 14.987.

Thus In- Sample Average sum square error is 14.987

Now we predict values for test data by using regression tree obtained on train data and compare the predicted values for test data with actual values of test data.

Out of sample Mean square prediction error is 26.725

General Additive Model

The general additive model that we have consists of 11 variables and non parametric functions are applied on quantitative variables. Since effective degree of freedom is 1 for zn and age. Hence in the final general additive model we do not apply non parametric functions on zn and age variables.

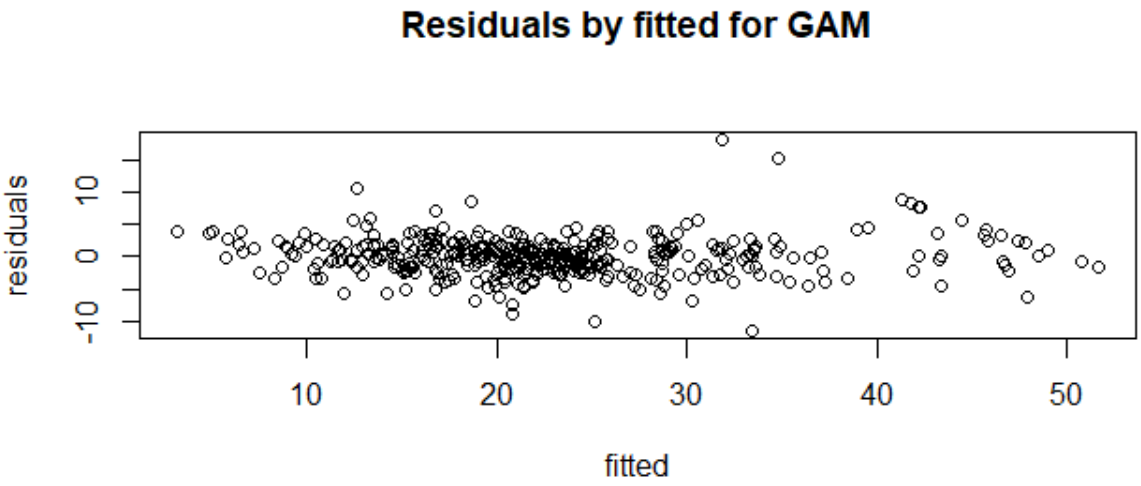
```
Family: gaussian
Link function: identity

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

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.819e+01  1.778e+00  10.228  < 2e-16 ***
zn          2.086e-02  1.543e-02   1.352  0.17713
chas        8.926e-01  7.078e-01   1.261  0.20816
age         5.798e-04  1.330e-02   0.044  0.96525
rad         4.010e-01  1.233e-01   3.253  0.00125 **
black       1.119e-03  2.312e-03   0.484  0.62864
---
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)    6.593  7.639  8.801 1.65e-10 ***
s(indus)   5.744  6.783  2.525   0.0174 *
s(nox)     8.924  8.993 13.092  < 2e-16 ***
s(rm)      7.681  8.554 24.162  < 2e-16 ***
s(dis)     8.755  8.977  6.792 5.33e-09 ***
s(tax)     3.083  3.730  9.807 5.91e-07 ***
s(ptratio) 1.037  1.071 29.544 6.68e-08 ***
s(lstat)   6.440  7.606 23.540  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.885   Deviance explained =   90%
GCV =  11.86  Scale est. = 10.271    n = 405
```



From the residual versus fitted values plot we can infer below points for General Additive Model.

- Residuals seems to be spread randomly
- Residuals seem to be normally distributed

We have used general additive model for train data. We predict values for train data and compare it with actual values of train data. The insample average mean square error we have is 8.89.

Now we used the gam model obtained from train data to predict values for test data. After that, we compare these predicted values with actual values of test data to calculate out of sample mean prediction square error.

Out of sample MSPE: 9.17

Neural network on Boston housing data set

We apply neural network method to determine housing prices in Boston. We would like to know if we can achieve better results by applying neural network as compare to Linear Regression, CART and Generalized additive models.

To apply neural network method, we first scale all our variables and then used scaled variables to build a neural network model. I have used neuralnet package to build a neural network model.

To build a neural network model, I have used 80 percent train data. We tried multiple combinations of hidden layers and number of nodes in a hidden layer. Final neural network model has 2 hidden layers with 5 and 3 nodes respectively.

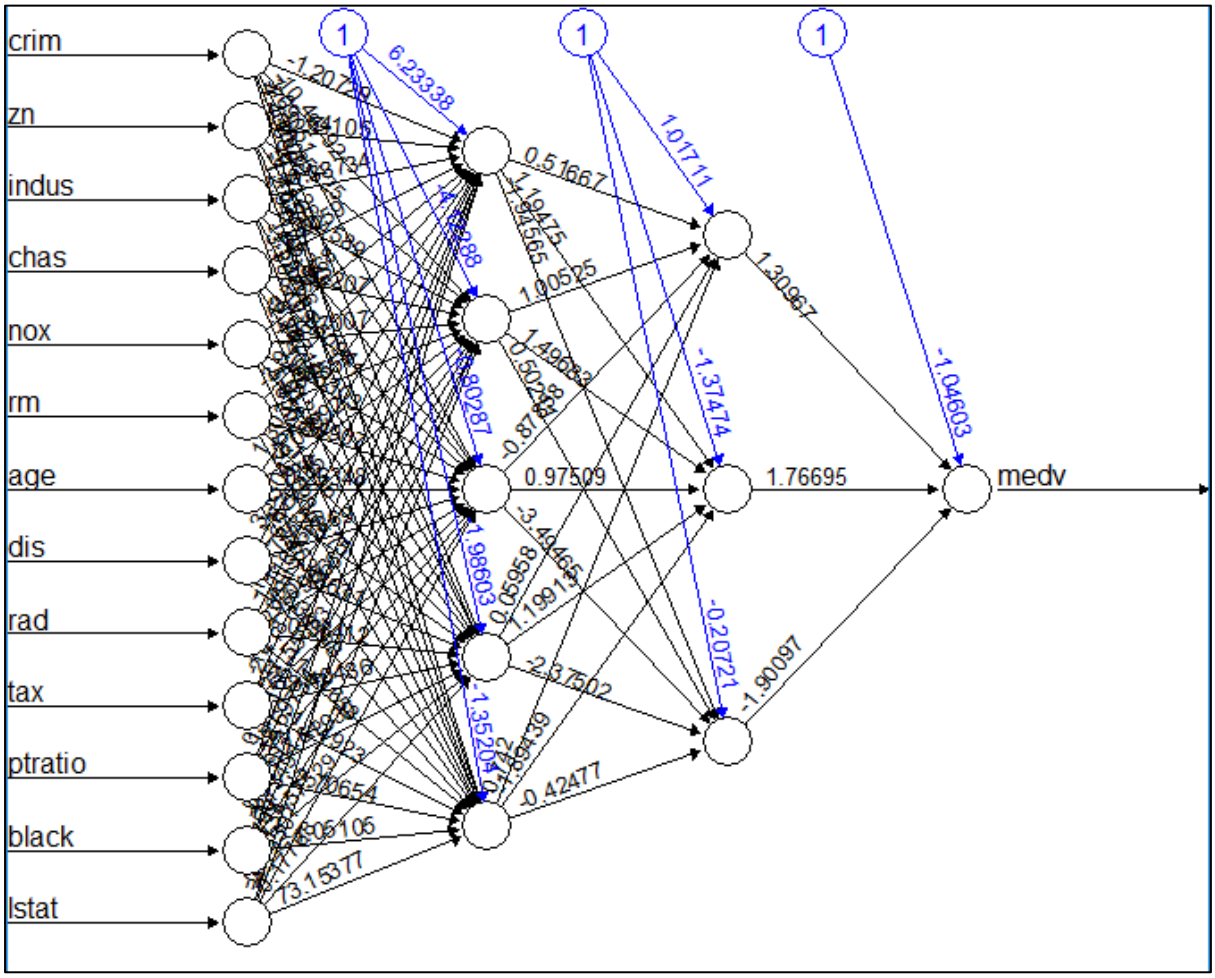


Figure: Neural network model for Boston housing data set

From neural network we have obtained following results.

Insample Average sum square of errors: 4.279

Out of sample Mean square prediction error: 8.934

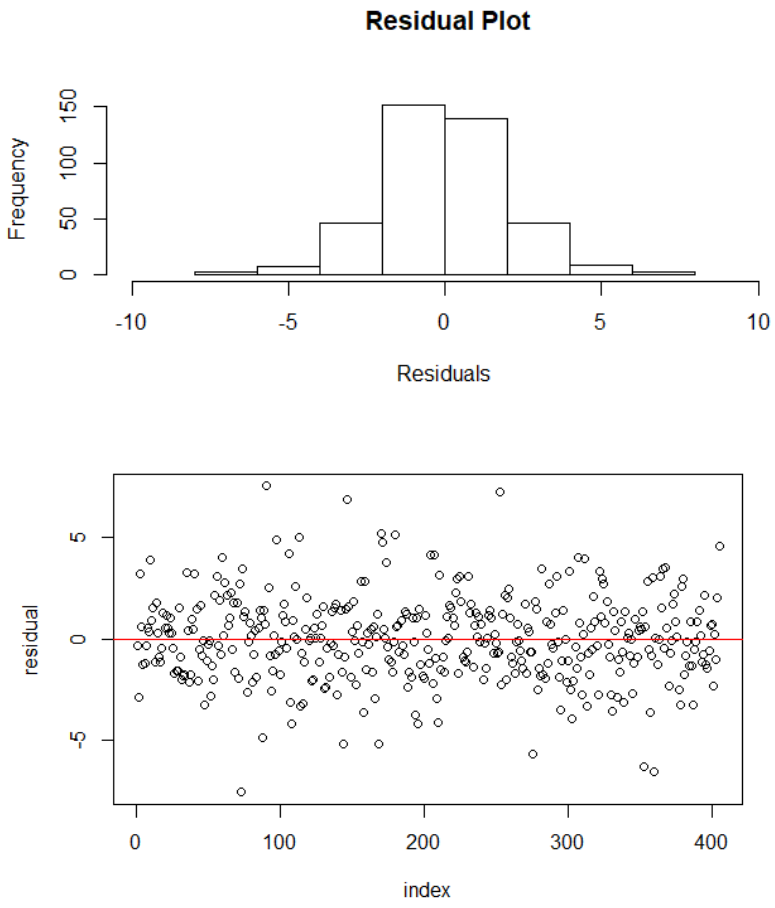


Figure: Residual Plot

Let us compare results of different models in the below table.

Method	In Sample Average sum square error	Out of sample average sum square error
GLM	23.159	17.465
CART	14.987	26.725
GAM	8.89	9.17
NNET	4.279	8.934

Best model for predicting median housing values in the suburbs of Boston is found to be **Neural Network Model.** Final neural network model has 2 hidden layers with 5 and 3 nodes respectively. The In sample Average sum square error for this model is lowest among all models which is 4.279. Also, out of sample average sum square error is lowest for neural network model which is 8.934.

2. German Credit Score Data

German credit score data has records that are classified as good credit and bad credit. We will use this data set to train our model to predict whether a test record is a good credit or bad credit. German credit score data set consists of 1000 observations and 21 features. Variable “Response” is the dependent variable in this data set. Response variable 0 indicate “good credit risk” and 1 indicates “bad credit risk”. Let us look at the structure of the data set below.

present_resid - This variable is integer in data type

property - This variable is factor in data type

Age - This variable is integer in data type

other_install - This variable is factor in data type

housing - This variable is factor in data type

n_credits - This variable is integer in data type

Job - This variable is factor in data type

n_people - This variable is integer in data type

telephone - This variable is factor in data type

foreign - This variable is factor in data type

response - This variable is integer in data type

chk_acct - This variable is factor in data type

duration - This variable is integer in data type

credit_his - This variable is factor in data type

purpose - This variable is factor in data type

amount - This variable is integer in data type

saving_acct - This variable is factor in data type

present_emp - This variable is factor in data type

installment_rate - This variable is integer in data type

sex - This variable is factor in data type

other_debtor - This variable is factor in data type

Let us look at the summary of the data.

```
 chk_acct     duration       credit_his    purpose        amount
 A11:274   Min.   : 4.000   A30: 40    A43    :280   Min.   :  250.000
 A12:269   1st Qu.:12.000   A31: 49    A40    :234   1st Qu.: 1365.500
 A13: 63   Median :18.000   A32:530    A42    :181   Median : 2319.500
 A14:394   Mean   :20.903   A33: 88    A41    :103   Mean   : 3271.258
           3rd Qu.:24.000   A34:293    A49    : 97   3rd Qu.: 3972.250
           Max.   :72.000              A46    : 50   Max.   :18424.000
                                       (Other): 55
```

```
saving_acct present_emp installment_rate  sex      other_debtor
A61:603     A71: 62     Min.   :1.000    A91: 50   A101:907
A62:103     A72:172     1st Qu.:2.000    A92:310   A102: 41
A63: 63     A73:339     Median :3.000    A93:548   A103: 52
A64: 48     A74:174     Mean   :2.973    A94: 92
A65:183     A75:253     3rd Qu.:4.000
                        Max.   :4.000
```

```
present_resid   property       age          other_install housing
Min.   :1.000   A121:282   Min.   :19.000   A141:139      A151:179
1st Qu.:2.000   A122:232   1st Qu.:27.000   A142: 47      A152:713
Median :3.000   A123:332   Median :33.000   A143:814      A153:108
Mean   :2.845   A124:154   Mean   :35.546
3rd Qu.:4.000              3rd Qu.:42.000
Max.   :4.000              Max.   :75.000

  n_credits       job         n_people     telephone  foreign
Min.   :1.000   A171: 22   Min.   :1.000   A191:596   A201:963
1st Qu.:1.000   A172:200   1st Qu.:1.000   A192:404   A202: 37
Median :1.000   A173:630   Median :1.000
Mean   :1.407   A174:148   Mean   :1.155
3rd Qu.:2.000              3rd Qu.:1.000
Max.   :4.000              Max.   :2.000

   response
Min.   :1.0
1st Qu.:1.0
Median :1.0
Mean   :1.3
3rd Qu.:2.0
Max.   :2.0
```

Let us at look at correlation between variables in german credit score data.

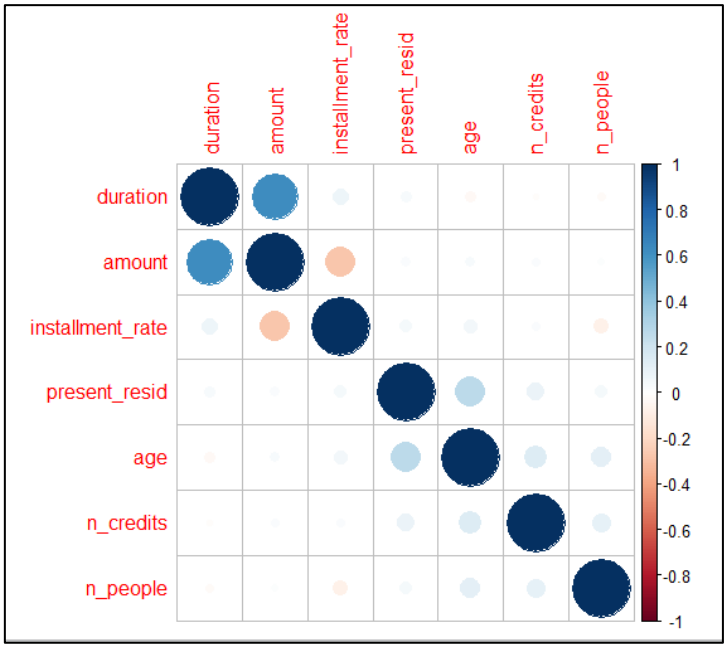


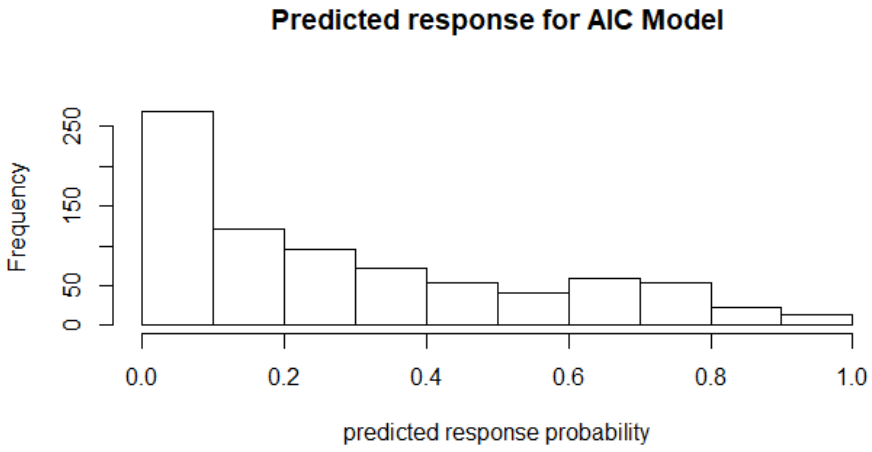
Figure: Correlation between variables

We can see that there is strong positive correlation between amount and duration variable.

Logistic regression on German Credit Score data

The final model that we have using Step AIC in both direction is given below .

response = chk_acct + duration + credit_his + purpose + amount + saving_acct + present_emp + installment_rate + sex + other_debtor + age + other_install + foreign



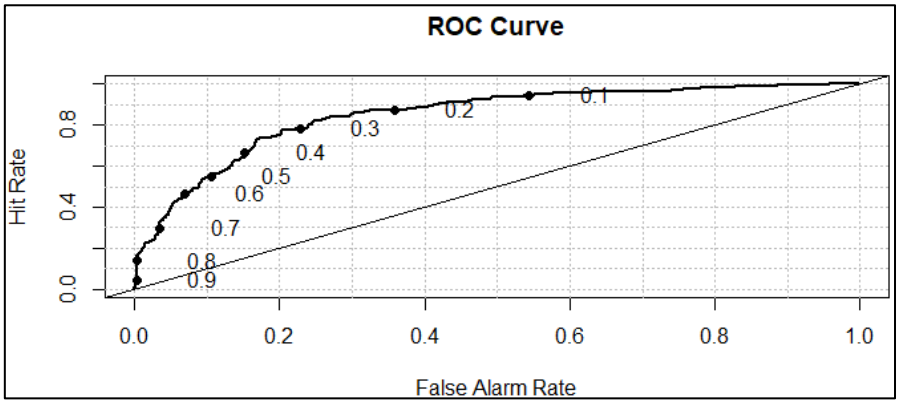
Let us look at In sample performance of Logistic regression model on German credit scoring data. The cut off probability is 1/6.

Residual Deviance: 689.517

Misclassification rate: 0.32

AUC: 0.8477

Let us look at ROC curve for Insample data.



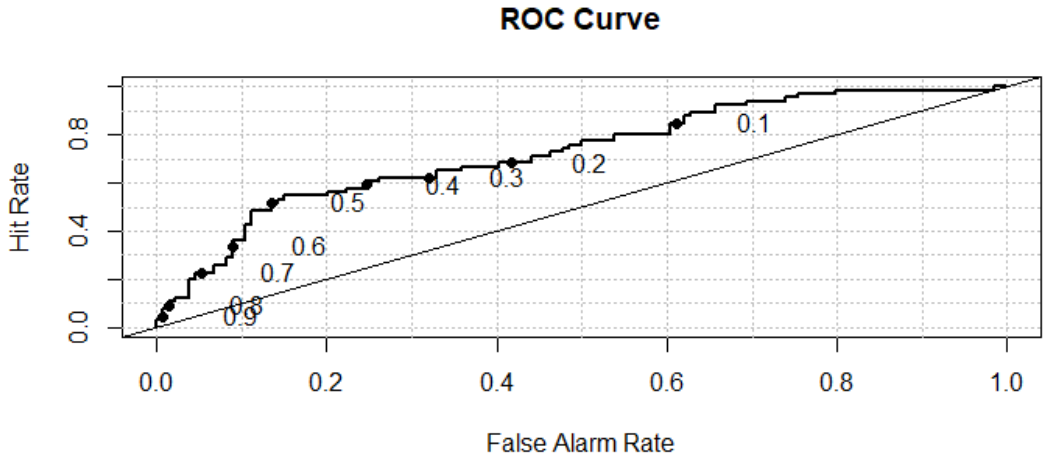
Using the model on test data I have measured the performance of the model on out of sample data.

Measures on Out of sample data

Misclassification rate: 0.405

AUC: 0.7278

Let us look at ROC curve for Out of sample data.



CART on German Credit Score data

I have applied classification tree on German Credit score data and splits for the final tree can be seen in the below tree plot.

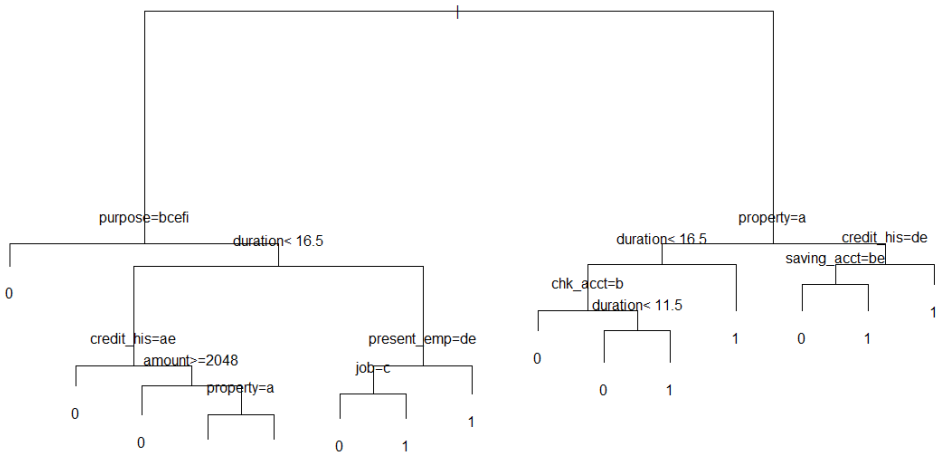


Figure: Classification Tree

I have built a classification tree model using training data and below are the measures of IN Sample performance of model.

Misclassification rate: 0.2925

AUC: 0.874

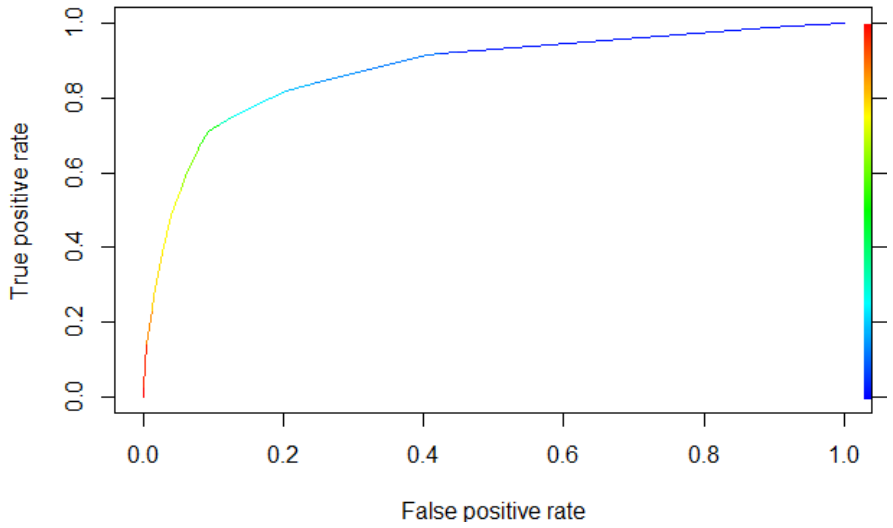


Figure: ROC curve for IN Sample data

Using classification tree model, I predicted classes for records in test data and measures of performance of model on Out of sample data is given below.

Misclassification rate: 0.415

AUC: 0.7014

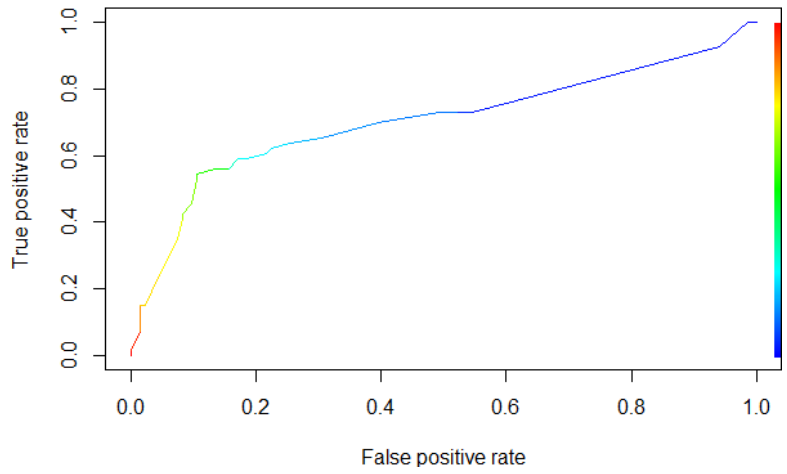
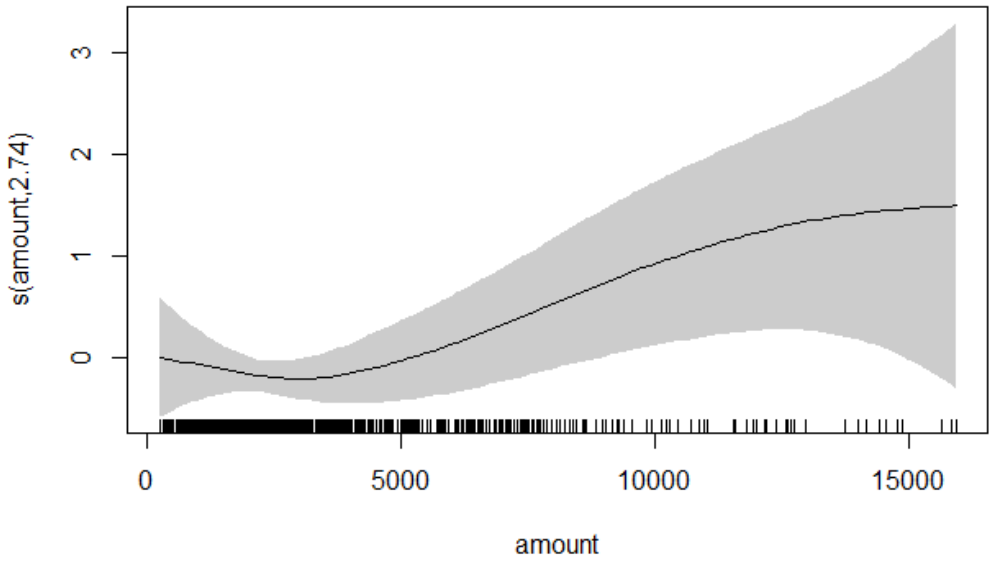


Figure: ROC curve for Out of Sample data

General Additive Model on German Credit Score data

I have applied general additive model on German Credit Score data. The Non-Parametric terms in the final model is only Amount as effective degree of freedom for Amount is 2.737.



In Sample performance measures for GAM Model on German Credit Score data

Misclassification rate: 0.3138

Deviance: 673.943

AUC: 85.24

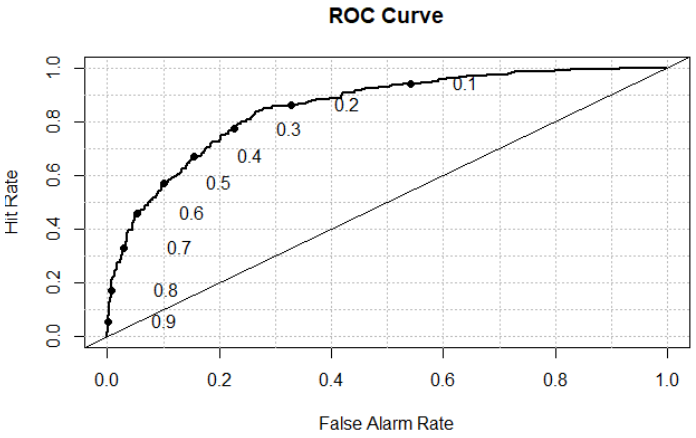


Figure: AUC for GAM In Sample

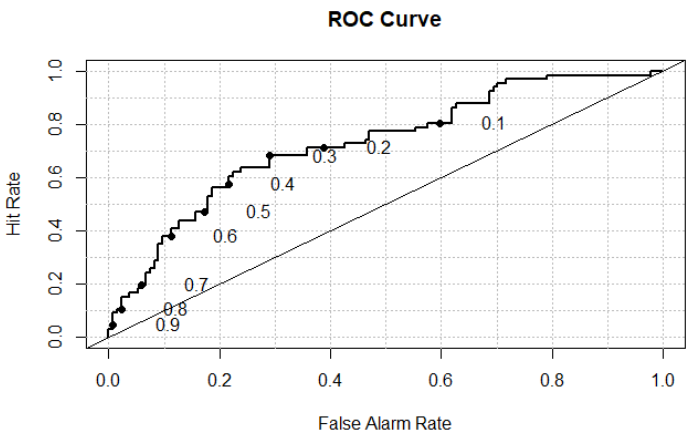


Figure: AUC for GAM Out of Sample

Out of Sample performance measures for GAM Model on German Credit Score data

Misclassification rate: 0.395

AUC: 72.98

Neural Network on German Credit Score data

I have applied neural network model in German Credit score data using nnet. The best neural network model that I have got is at size = 18 and decay at 0.

In Sample performance measures for Neural Network Model on German Credit Score data

Misclassification rate: 0.298

Out of Sample performance measures for Neural Network Model on German Credit Score data

Misclassification rate: 0.38

Discriminant Analysis on German Credit Score data

I have applied Linear discriminant analysis on German Credit Score data to measure performance of model on the data set.

In Sample performance measures for LDA Model on German Credit Score data

Misclassification rate: 0.29

AUC: 84.86

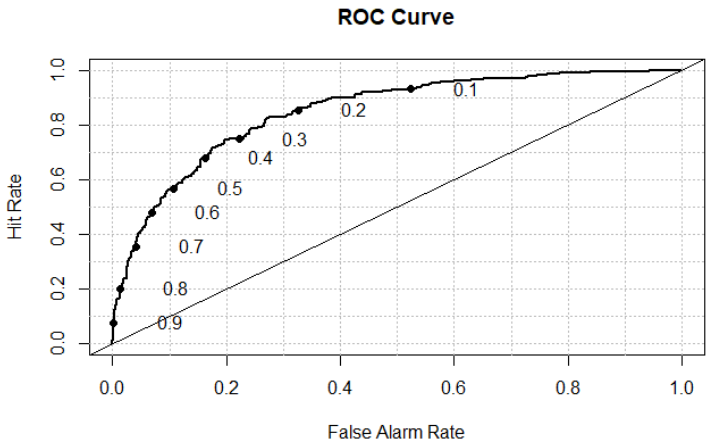


Figure: ROC curve for Insample LDA model

Out of Sample performance measures for Neural Network Model on German Credit Score data

Misclassification rate: 0.375

AUC: 72.67

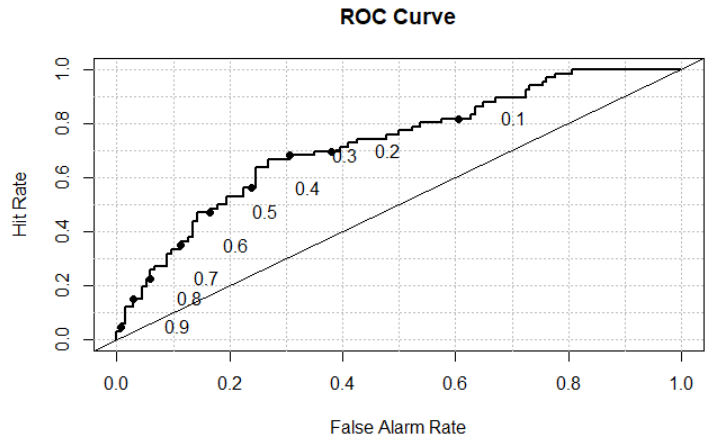


Figure: ROC curve for LDA Out of sample data

Method	Misclassification rate		Area under Curve	
	In Sample	Out of Sample	In Sample	Out of Sample
Logistic Regression	0.32	0.405	0.8477	0.7278
Classification Tree	0.2925	0.415	0.874	0.7014
GAM	0.3138	0.395	85.24	72.98
NNET	0.2988	0.38		
LDA	0.29	0.375	84.86	72.67

Best model for predicting good credit and bad credit for German Credit Score data is found to be Linear Discriminant Analysis Model. Final LDA model has misclassification rate of 0.29 for In sample data and 0.375 for Out of sample data, which is lowest as compared to all other models.