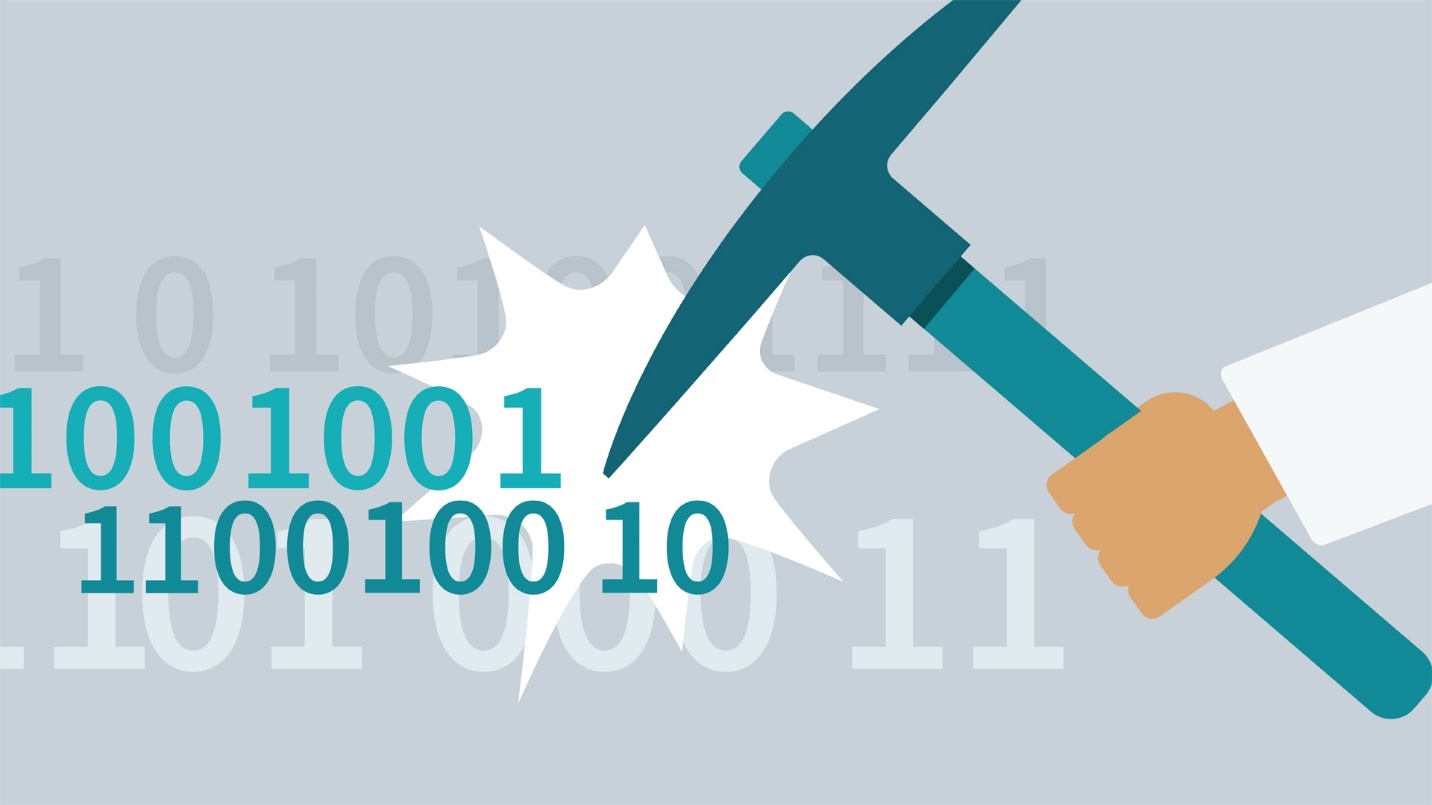
**DATA MINING - 2**

**INDIVIDUAL CASE -2**



Boston Housing and German Credit dataset

**Prof. Yan Yu**

**Executive Summary**

# Goal and Background:

The problem at hand is to fit generalized linear model, tree model (Random forest, Bagging, Boosting), Generalised Additive models and Neural Network for Boston Housing dataset and German Credit Scoring dataset. We need to do these for regression and classification problems and the response variable of Boston Housing dataset is quantitative and of German Credit Scoring dataset is qualitative.

# Approach:

For Boston Dataset:

* Divided the dataset into 75% train and 25% test dataset.
* Created a linear model using all the variables, with stepwise and best subset variable selection methods on training dataset.
* Calculated in sample and out of sample prediction error for all the models.
* Created tree model using all the variables, bagging, boosting and random forest.
* Calculated in sample and out of sample prediction error for all the tree models.
* Built GAM model using all the variables and revised the model using only non – linear variables.
* Calculated in sample and out of sample prediction error for GAM models.
* Designed neural network after scaling the data.
* Calculated in sample and out of sample prediction error for Neural Network model.

For Credit Scoring Dataset:

* Divided the dataset into 75% train and 25% test dataset.
* Created a logistic model using all the variables, with stepwise, LASSO and best subset variable selection methods on training dataset.
* Calculated optimal cutoff probability using 5:1 asymmetric cost function.
* Calculated in sample and out of sample confusion matrix.
* Created a classification tree to calculate in sample and out of sample confusion matrix and error.
* Built GAM model using all the variables to calculate in sample and out sample prediction error.
* Designed neural network and calculated in sample and out of sample prediction error.

# Major Findings

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Boston Data | | Credit Data | |
|  | MSE(In Sample) | MSE(Out of Sample) | MSE(In Sample) | MSE(Out of Sample) |
| GLM | 25.60242 | 13.92252 | 0.09899 | 0.1021007 |
| Tree | 2.478184 | 5.526306 | 0.34266 | 0.424 |
| GAM | 8.889896 | 14.09725 | 0.22 | 0.27 |
| Neural Network | 4.109685 | 10.57118 | 0.624 | 0.6373333 |

Table 1

We see that in Boston housing data, Tree and Neural Network models perform better than GAM and GLM. In Credit data, GLM and GAM perform better.

# Boston Housing Dataset

There are 14 attributes in each case of the dataset. They are:

1. CRIM - per capita crime rate by town
2. ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS - proportion of non-retail business acres per town.
4. CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
5. NOX - nitric oxides concentration (parts per 10 million)
6. RM - average number of rooms per dwelling
7. AGE - proportion of owner-occupied units built prior to 1940
8. DIS - weighted distances to five Boston employment centres
9. RAD - index of accessibility to radial highways
10. TAX - full-value property-tax rate per $10,000
11. PTRATIO - pupil-teacher ratio by town
12. B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13. LSTAT - % lower status of the population
14. MEDV - Median value of owner-occupied homes in $1000's

Our response variable is Medv.

## Linear Models

### Original model

This model contains all the variables in the dataset.

**Summary Statistics:**

|  |  |
| --- | --- |
| Adjusted R^2 | 62.09% |
| AIC | 2417.961 |
| BIC | 2469.149 |
| MSE (In-sample) | 33.29941 |
| MSE (Out of sample) | 13.31231 |

Table 2

**Plot:**

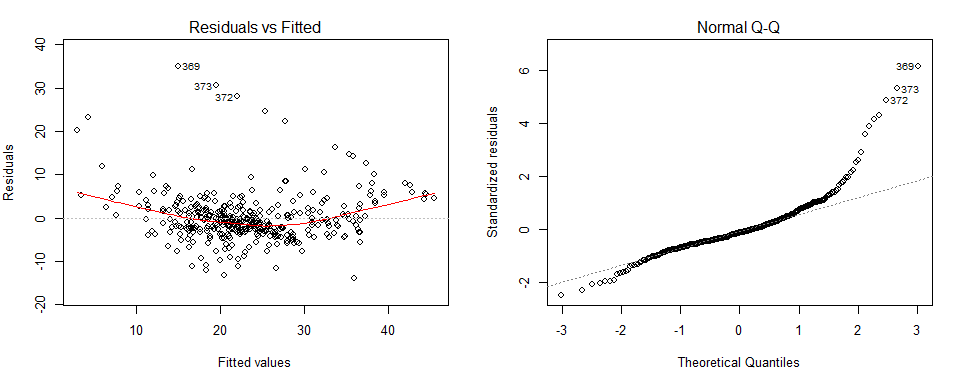


Figure 1

The plots suggest the relationship between response variable and other predictors is linear to some extent with some outliers. Residuals are not very independent of each other.

### Best Subset Regression

This model is built using these predictors: chas, nox, rm, dis, ptratio, black, lstat

**Summary Statistics:**

|  |  |
| --- | --- |
| Adjusted R^2 | 69.27% |
| AIC | 2334.463 |
| BIC | 2369.901 |
| MSE (In-sample) | 26.99079 |
| MSE (Out of sample) | 15.17247 |

Table 3

**Plot:**

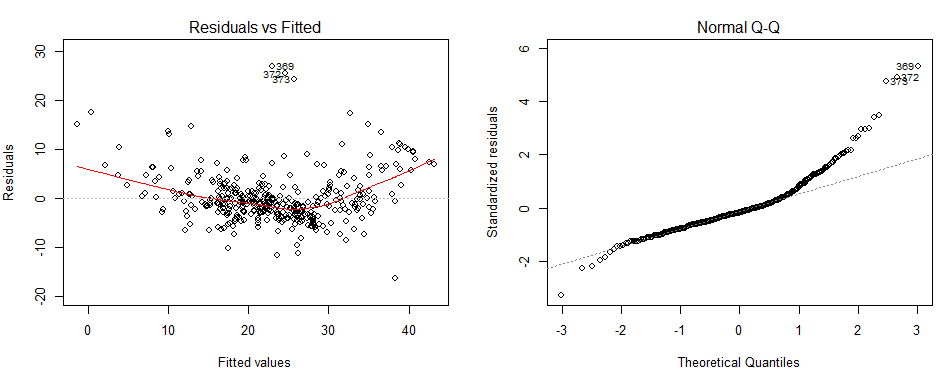


Figure 2

We do not see much of a difference as compared to previous plot. This data also has outliers and residuals are not independent but AIC,BIC and in sample prediction errors reduced.

### Stepwise Variable Selection

Variables used in making this model are : lstat, rm, ptratio, dis, nox, chas, zn, black, rad, tax, and crim.

**Summary Statistics:**

|  |  |
| --- | --- |
| Adjusted R^2 | 70.85% |
| AIC | 2318.34 |
| BIC | 2369.528 |
| MSE (In-sample) | 25.60242 |
| MSE (Out of sample) | 13.92252 |

Table 4

**Plot:**

The plot looks better than the previous two as it is less skewed and lightly tailed. Residuals are still not very independent but in a better shape. AIC and BIC are less as compared to other two. Same is the case with in sample and out of sample prediction error. We can conclude that this is the best linear model among three.

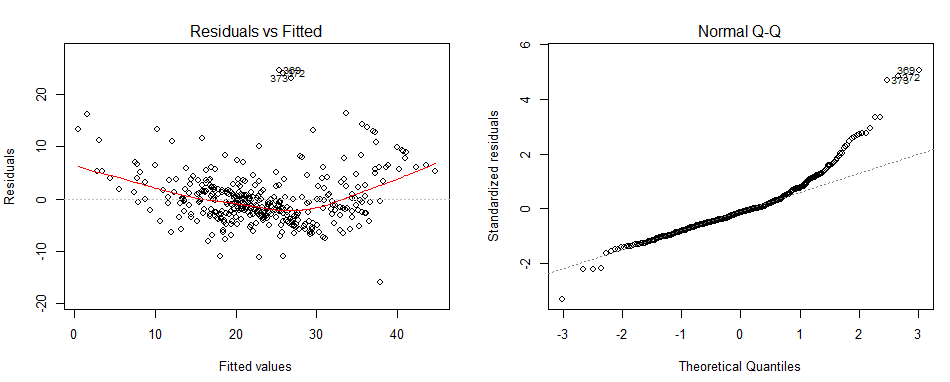


Figure 3

## Tree Models

### Original tree

This tree has been constructed using all the variables.

**Summary Statistics**

|  |  |  |
| --- | --- | --- |
|  | Original tree | Pruned tree |
| MSE (In sample) | 16.88527 | 20.62686 |
| MSE (Out of Sample) | 12.91423 | 17.96805 |

Table 5

**Plot:**

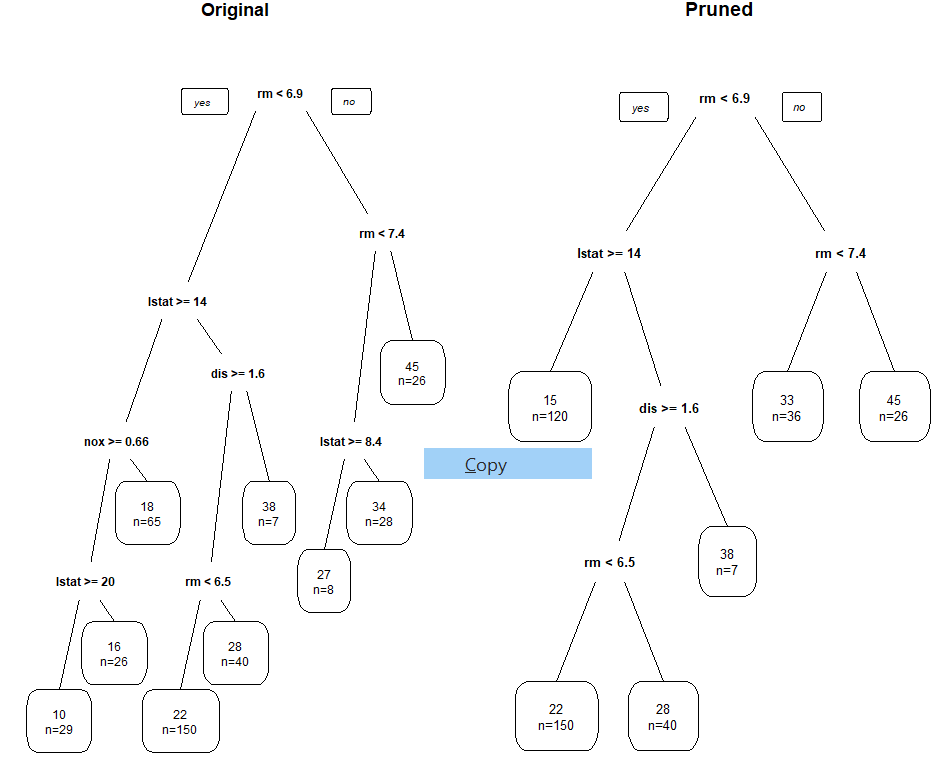


Figure 4

### Bagging, Boosting and Random Forest

|  |  |  |  |
| --- | --- | --- | --- |
|  | Bagging | Boosting | Random Forest |
| MSE (In sample) | 11.70027 | 6.163916 | 2.478184 |
| MSE (Out of Sample) | 10.21817 | 8.491944 | 5.526306 |

Table 6

The in and out sample errors in Random Forest are the least.

## Generalized Additive Model

GAMs are used when we need to model non- linearity in multiple predictors. GAM maintains additivity while allowing non-linear function of each variable. This model can be applied to both qualitative and quantitative data.

**Advantages**:

* There is no need to try out different transformations on each variable individually.
* Inference is easier. We can examine the effect of a particular ‘x’ on y while holding other variables fixed.

**Limitations:**

* With many variables, important interactions can be missed. This can be tackled using low dimension (2-D) interaction function such as local regression.

Although GAMs are flexible than linear models, they still can’t beat advanced tree algorithms such as random forest and boosting in flexibility.

### Original model

This model is built using all the variables using spline basis smoothing function. In the summary, I found out that zn, ptration and age are not non-linear because their estimated degree of freedom is 1.

We can see in the plot that ptratio, age and zn have pretty much straight lines and large confidence interval as we move further.

Hence, I build a new model using only non-linear and categorical terms.

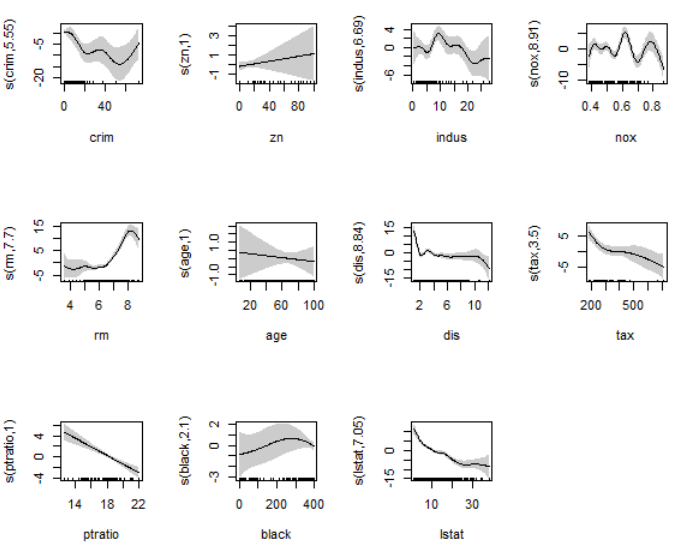


Figure 5

### Revised GAM using only non-linear terms

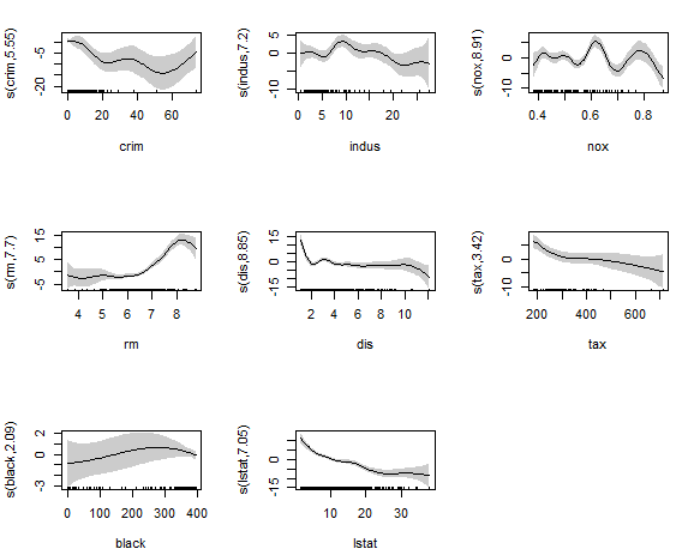


Figure 6

**Summary statistics**

|  |  |  |
| --- | --- | --- |
|  | **Original GAM** | **Revised GAM** |
| **AIC** | 2018.3 | 2019.048 |
| **BIC** | 2244.043 | 2246.482 |
| **Adjusted R^2** | 0.881 | 0.881 |
| **MSE (In sample)** | 8.889896 | 8.887286 |
| **MSE (Out of sample)** | 14.09725 | 14.1125 |

Table 7

We do not observe any serious changes in these statistics.

## Neural Network

The main agenda of neural network is to extract linear combinations of the input as derived features and then model the target as a non-linear function of these features. This model is highly non linear. This is also known as a blackbox method as it’s very difficult to interpret the middle nodes. The non linearity comes from sigmoidal activation function in multiple layers.

|  |  |
| --- | --- |
|  | Neural Network model |
| MSE (In sample) | 4.109685 |
| MSE (Out of sample) | 10.57118 |

Table 8

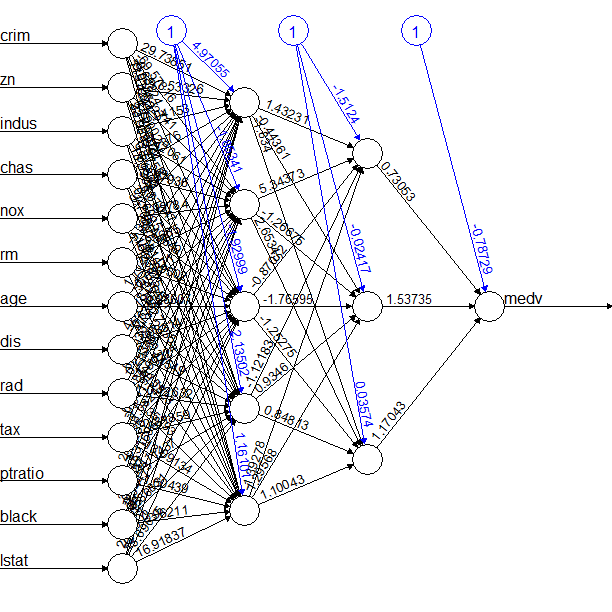


Figure 7

This tree has two hidden layers.

# German Credit Scoring data

## Generalized Linear model

### Original model

This model contains all the variables.

|  |  |
| --- | --- |
|  | Original Model |
| MSE (In sample) | 0.1425614 |
| MSE (Out of sample) | 0.2841683 |
| AIC | 767.4208 |

Table 9

### Stepwise GLM

This model contains all these variables : chk\_acct, duration, saving\_acct, housing, other\_debtor, credit\_his, installment\_rate, present\_emp, other\_install, sex, purpose, amount, and foreign.

|  |  |
| --- | --- |
|  | Stepwise Model |
| MSE (In sample) | 0.143833 |
| MSE (Out of sample) | 0.2814211 |
| AIC | 752.0808 |

Table 10

### LASSO variable selection

|  |  |
| --- | --- |
|  | LASSO Model |
| MSE (In sample) | 0.09899936 |
| MSE (Out of sample) | 0.1021007 |

Table 11

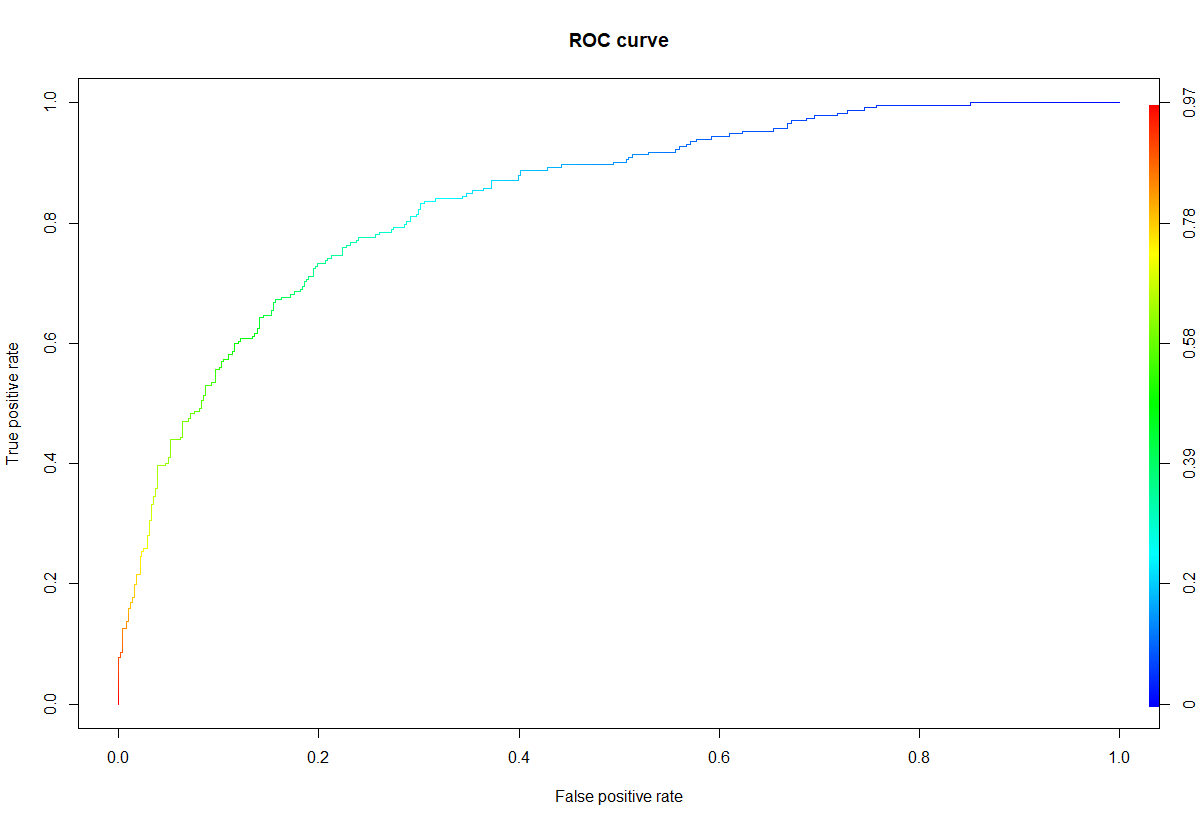
LASSO model has the least in and out sample error.

Figure 8

The performance of ROC curve is good – 97%.  
Cross Validation

In sample confusion matrix:

|  |  |  |
| --- | --- | --- |
|  | Predicted | |
| True | 0 | 1 |
| 0 | 352 | 166 |
| 1 | 34 | 198 |

Table 12

Out of sample confusion matrix:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted | | |
| True | 0 | | 1 |
| 0 | 36 | | 84 |
| 1 | 101 | | 29 |
|  | | Cross Validation | |
| MSE (In sample) | | 0.26 | |
| MSE (Out of Sample) | | 0.50 | |

Table 14,15

## Tree Model

Cut off probability on weight 6:1. We will use 0.27 as our cut off probability in all the models moving forward.

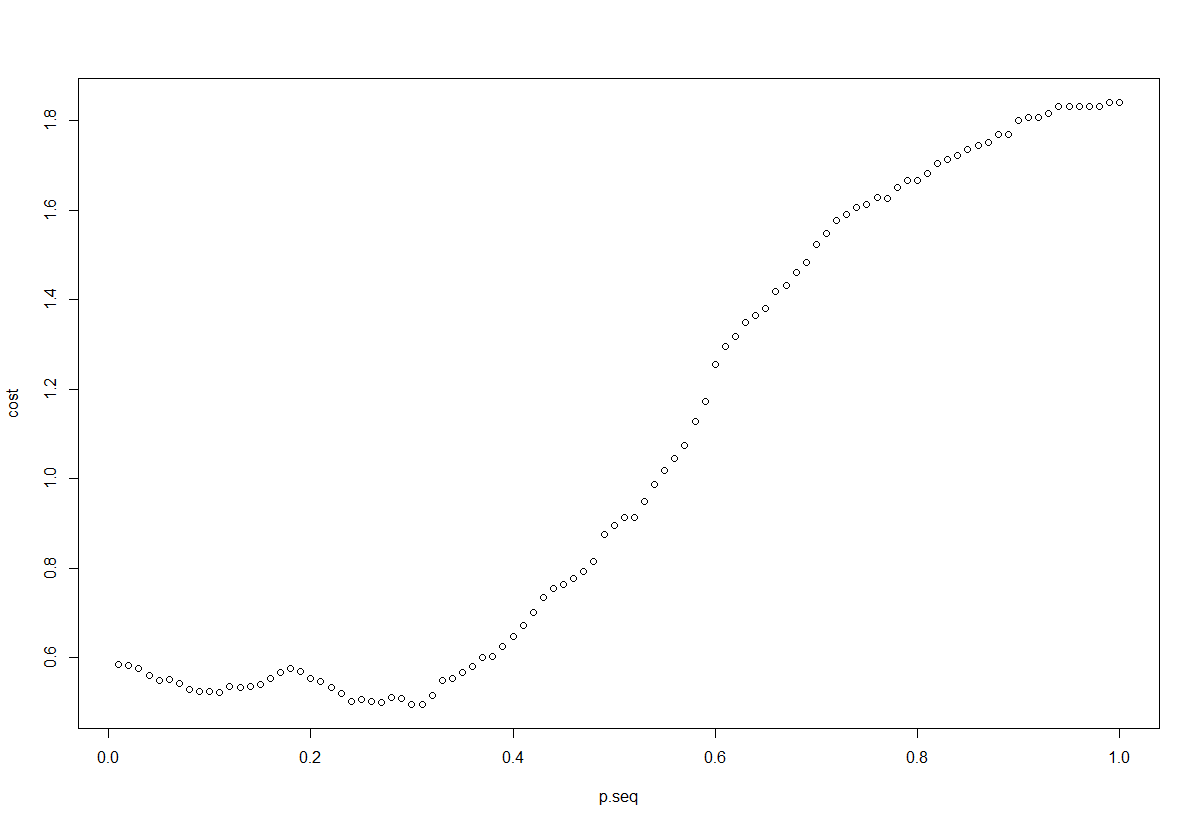


Figure 9

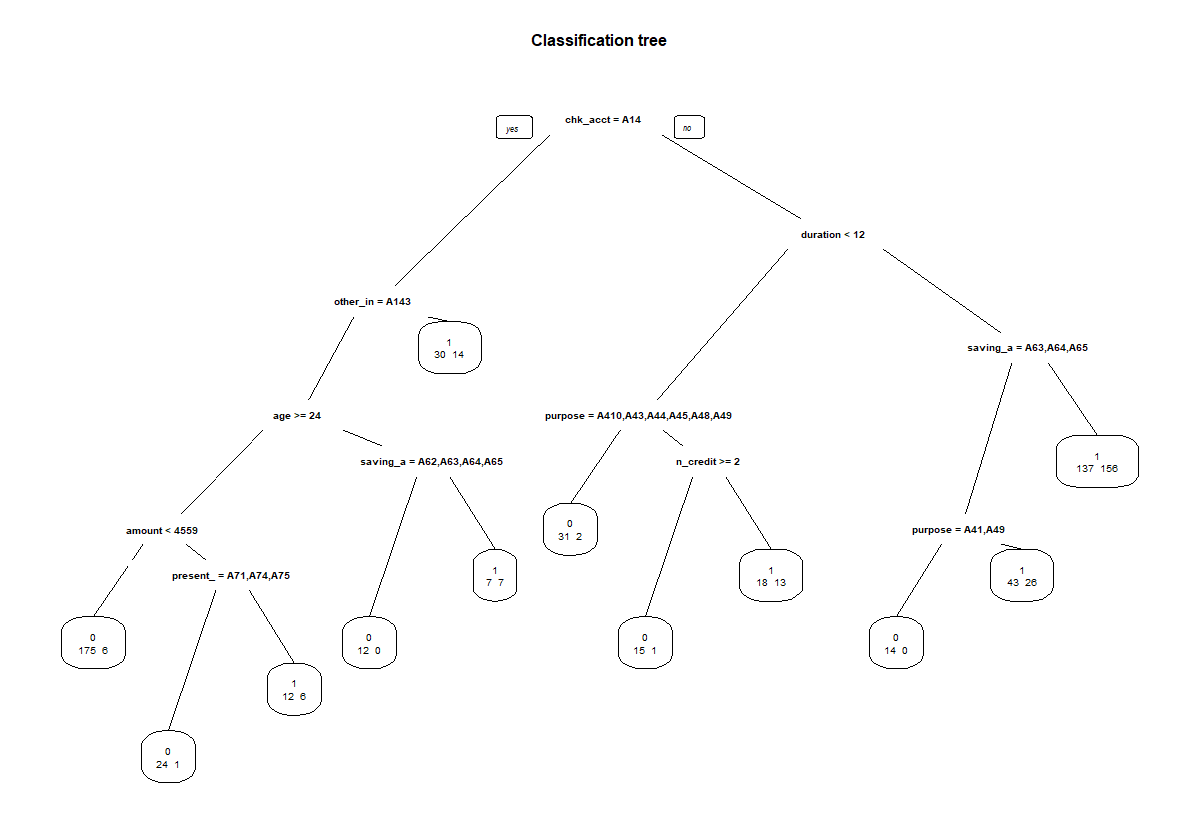


Figure 10

In sample confusion matrix:

|  |  |  |
| --- | --- | --- |
|  | Predicted | |
| True | 0 | 1 |
| 0 | 271 | 247 |
| 1 | 10 | 222 |

Table 16

Out of sample confusion matrix:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted | | |
| True | 0 | | 1 |
| 0 | 88 | | 94 |
| 1 | 12 | | 56 |
|  | | Tree model | |
| MSE (In sample) | | 0.3426667 | |
| MSE (Out of Sample) | | 0.424 | |

Table 17,18

## Generalized Additive Model

This model is generated using spline basis on amount and duration as these columns had the least unique values and are continuous. Their confidence interval is wide when we move forward.

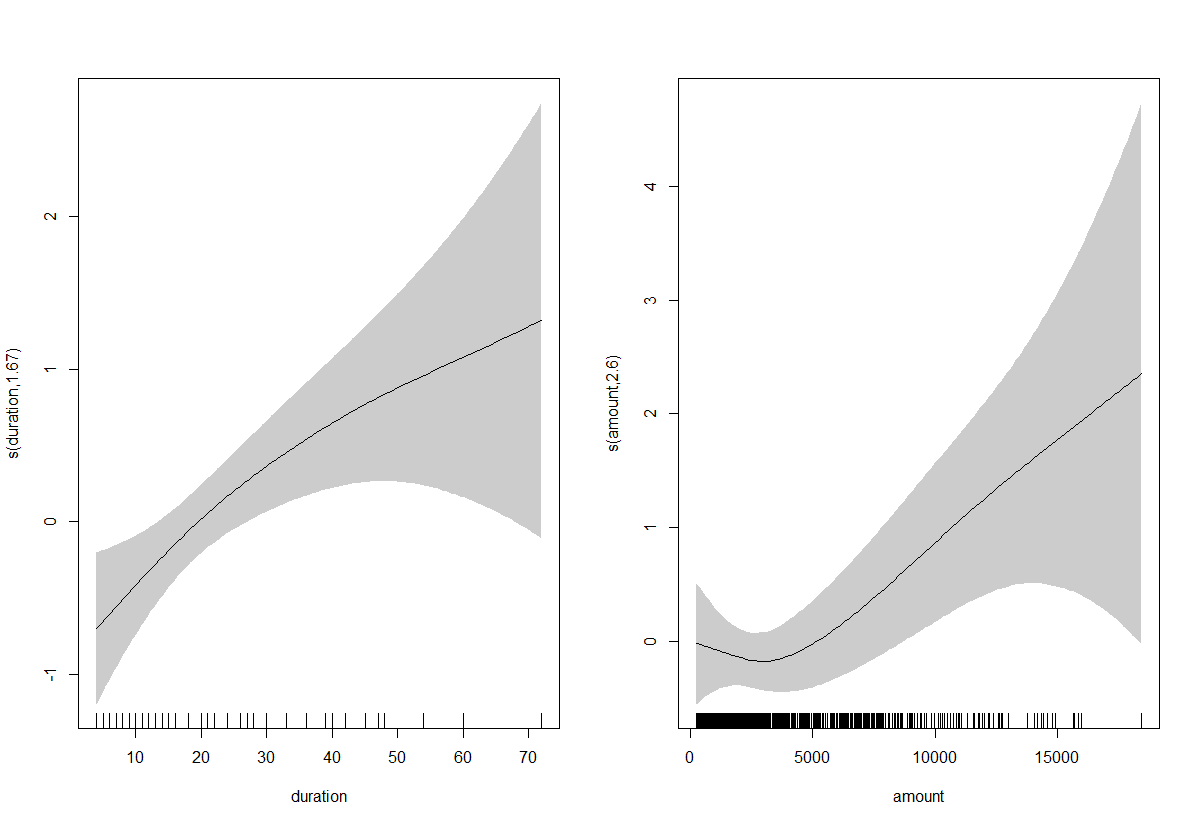


Figure 11

In sample confusion matrix:

|  |  |  |
| --- | --- | --- |
|  | Predicted | |
| True | 0 | 1 |
| 0 | 373 | 147 |
| 1 | 41 | 189 |

Table 19

Out of sample confusion matrix:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted | | |
| True | 0 | | 1 |
| 0 | 127 | | 53 |
| 1 | 15 | | 55 |
|  | | GAM model | |
| MSE (In sample) | | 0.22 | |
| MSE (Out of Sample) | | 0.27 | |

Table 20,21

## Neural Network

Designed a neural network with 13 middle layers.

In sample confusion matrix:

|  |  |  |
| --- | --- | --- |
|  | Predicted | |
| True | 0 | 1 |
| 0 | 55 | 471 |
| 1 | 7 | 217 |

Table 22

Out of sample confusion matrix:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted | | |
| True | 0 | | 1 |
| 0 | 19 | | 155 |
| 1 | 1 | | 75 |
|  | | Neural Network model | |
| MSE (In sample) | | 0.624 | |
| MSE (Out of Sample) | | 0.6373333 | |

Table 23,24