**Data Mining: Customer Loan Repayment Classification**

Priyanka Shah

Varun Kumar Chepuri

Amey Parab

**Purpose:**

The goal of the project is to find patterns in loan repayment of customers from various states of the U.S.A. The dataset used in this project pertains to historic loan data collected from various banks. The patterns obtained from the real world data aid to classify the customers into who may or may not be able to repay the loan.

**Result:**

Data mining techniques are applied to the data collected from various sources, to produce a classifier. This classifier can aid the bankers to identify customers who may not be able to repay the loan.

**Evaluation metrics and plan:**

During the initial stages of the project, preprocessing techniques have been used to clean and validate the customer’s loan information. Then, with the help of data visualization techniques, indicator variables are found. Finally, classification techniques have been implemented to fetch the patterns of loan repayment of customers from different states of the U.S.A.

**Week-1**: Understood the data mining process and gathered different ideas for the project.

**Week-2**: Familiarized with different technologies such as R, Hadoop and Hive that could be applicable for the project. Selected a suitable technology based on the requirements.

**Week-3**: Collected data from different sources that were suitable for the project. Studied data visualization and pre-processing techniques.

**Week-4**: Applied suitable data visualization and pre-processing techniques. Drew inferences based on the results obtained from the techniques.

**Week-5**: Studied various classification algorithms. Identified potential algorithms that can solve the problem.

**Week-6**: Recognized the boundaries of algorithms in the chosen platform. Implemented and tested an algorithm.

**Week-7**: Implemented and tested the remaining algorithms. Final documentation was drafted.

**Week-8**: Performed final testing and reviewed the documentation.

**Background research:**

Loan data was collected from various online resources. Out of which, three datasets were considered for the project.

All the finalized data sets were integrated into a single data set. It consists of more than a million tuples and about a hundred attributes. Then, using the data dictionary, all the attributes were thoroughly accessed. Various visualization algorithms have been applied to interpret the data. As a result, some of the attributes were found to be similar and the redundancies were eliminated. Correlation analysis was then applied to check the relation between different attributes. Finally, in order to reduce the size of the dataset, data reduction algorithms have been applied.

Since the goal of the project is to classify whether it is safe or risky to repay a loan to a customer, research was performed on how classification techniques have been used in various applications that relate to our project.

**Detailed Research methodology or approach taken in the project:**

After collecting all relevant information, and finalizing the dataset, various visualization techniques have been applied and each attribute was manually to understand the dataset in more detail. We then worked on various data preprocessing techniques to remove redundancy, inconsistencies and unnecessary information from the dataset. We used R language to visualize and preprocessing of dataset.

**Data Visualization:**

We have applied visualization techniques (i.e. bar plot) on interest rate, grades attributes over 9 years of historic data and observed below patterns.

1) Interest rate is increased based on grade but Net annual return is almost similar for all grades.

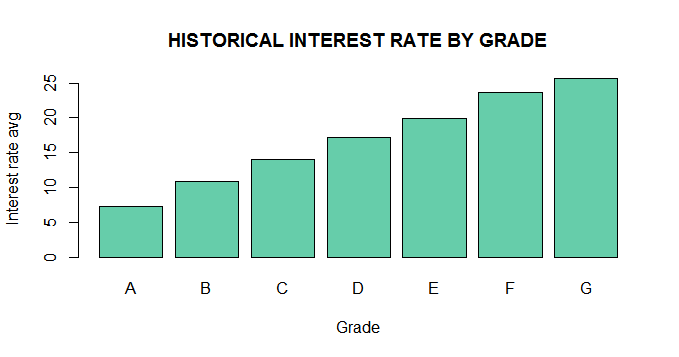
2) Average rate of interest is same in all years.

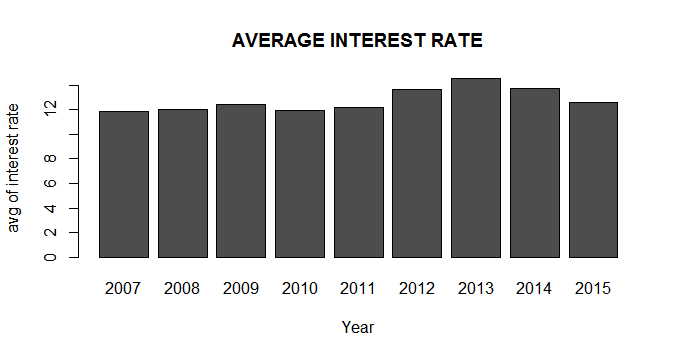
3) Percentages of all grades in each year.

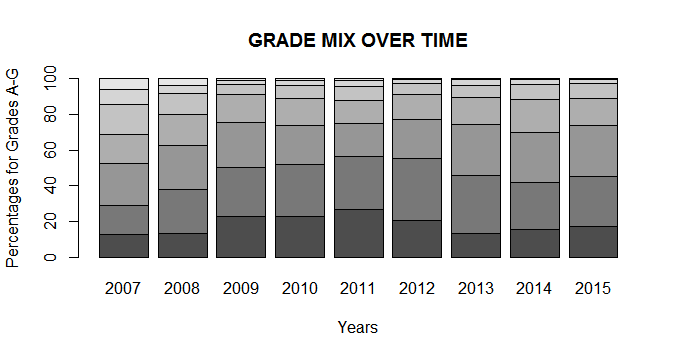
4) Debt to Income ratio of each state.

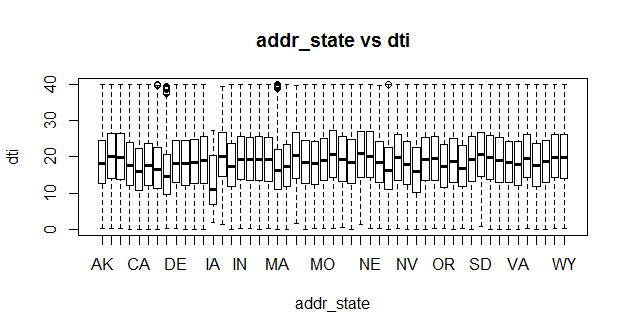
5) Annual\_income Vs Loan\_amount

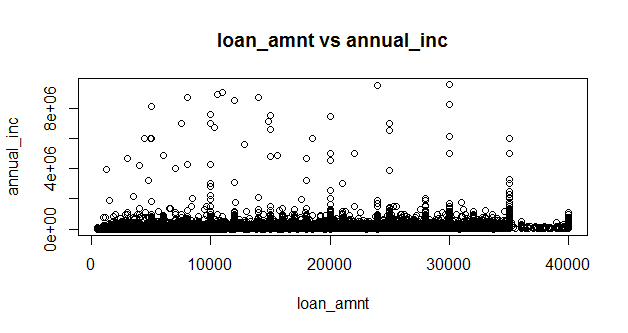
**Below are the few visualization diagrams:**

****

****

****





**Data Preprocessing and Implementation:**

Step 1: Checked for the uniqueness rule and null rule: loan\_id and customer\_id attributes were checked to ensure whether they are unique or not. Also checked for null rule, i.e. where there was missing data, we inserted “NA”.

Step 2: Used data dictionaries to check the meaning of each attribute and decided which attributes should be kept and which should be eliminated.

Step 3: We eliminated following attributes which are not relevant to our data mining process.

verification\_status\_joint, acc\_open\_past\_24mths, avg\_cur\_bal, bc\_open\_to\_buy, bc\_util, chargeoff\_within\_12\_mths, dti\_joint, acc\_now\_delinq, delinq\_amnt, mo\_sin\_old\_il\_acct, mo\_sin\_old\_rev\_tl\_op, mo\_sin\_rcnt\_rev\_tl\_op, mo\_sin\_rcnt\_t, mort\_acc, mths\_since\_recent\_bc, mths\_since\_recent\_bc\_dlq, mths\_since\_recent\_inq, mths\_since\_recent\_revol\_delinq, emp\_title, num\_accts\_ever\_120\_pd, num\_actv\_bc\_tl, num\_actv\_rev\_tl, num\_bc\_sats, num\_bc\_tl, num\_il\_tl, num\_op\_rev\_tl, num\_rev\_accts, num\_rev\_tl\_bal\_gt\_0, num\_sats, num\_tl\_120dpd\_2m, num\_tl\_30dpd, num\_tl\_90g\_dpd\_24m, num\_tl\_op\_past\_12m, pct\_tl\_nvr\_dlq, total\_bal\_ex\_mort, total\_bc\_limit, total\_il\_high\_credit\_limit, zip\_code.

Step 4: After observing data using visualization techniques, we concluded that interest rate increases based on grade (A - G). But, this did not impact on customer’s loan repayment. Even though interest rate is high, customers are able to repay full amount of the loan. So, grade attribute is not a valid attribute in our analysis. Hence, this attribute is considered unessential.

Step 5: An attribute “verification\_status” is used to show customer’s income has been verified or not. For the purpose of the project, annual income of the customers is an essential factor. It is necessary to ensure that it has been verified. All the tuples with verification\_status with “not verified” cannot be used. Hence, there is a need to eliminate data having “not verified” status. Later on, we eliminated entire attribute as all tuples contain “verified” or “source verified” status. This was done using SQL queries based on the sqldf package available in R.

Step 6: We applied correlation coefficient to identify the relationship between loan\_amount and funded\_amnt, loan\_amount and funded\_amnt\_inv, recoveries and collection\_recovery\_fee. We found that, all analysis got result near to 1, hence they are strongly correlated and we eliminated funded\_amnt, funded\_amnt\_inv and collection\_recovery\_fee attributes from the dataset.

Here are the outputs of our analysis:

Correlation – Coefficient (loan\_amount and funded\_amnt) = 0.9993684

Correlation – Coefficient (loan\_amount and funded\_amnt\_inv) = 0.9975263

Correlation – Coefficient (recoveries and collection\_recovery\_fee) = 0.8572363

Step 7: We applied object identification approach where we replaced member\_name to member\_title, as both have the same meaning and combined both attributes.

Step 8: We also checked for the derivable attributes. An attribute called installment is a derivable attribute, which is calculated using Simple Interest formula ((loan\_amount \* term \* interest\_rate) / 100). So we eliminated this attribute, as it can be calculated using the above formula.

Step 9: Data reduction techniques were applied such as, attribute ‘term’ had only two values “36 months” or “60 months”,so we replaced “36 months” as 3 and “60 months” as 5. Similarly, we have finite statuses attribute in loan\_status, so we classified them into a scale with ‘1’ as the customer most likely to repay the loan.

Step 11: We have calculated LTI based on formula loan\_amnt/ annual\_inc. LTI stands for loan to the income ratio. During our analysis, we have found that it is an important attribute for the classification.

Step 12: Identified key attributes to classify are term, dti, delinq\_2 yrs, loan status and interest\_rate.

Step 13: Identified loan status as label attribute to classify the data and scaled data into to 1 to 6.

1. Fully paid
2. Current
3. Grace Period
4. Late(16-30 days)
5. Late (30-120 days)
6. Default/Charged off

Step 14: We have eliminated all data with loan status as ‘Current’ as we are classifying who have paid or not.

Step 15: Converted loan status data to binary data (Yes – loan\_status as 1 and No – loan\_status as 3, 4, 5 and 6).

Step 16: We have used Neural Network and Decision Tree classification algorithm in r to classify data and identified the patterns of customers who have paid or not paid the amount.

Step 17: After applying both algorithms, we tested classifiers on Test Data and evaluate the results of Decision Tree and Neural Network.

**Status of Implementation:**

We applied two algorithms Neural Network and Decision Tree for finding classifiers of our project. A random sample of 2 lakh tuples from the processed dataset is chosen and both the algorithms are applied on that training set. Then, we applied this classification model to test the data and compared performance of both techniques. We then applied these algorithms on testing data of size 1 lakh which was taken from dataset randomly. Algorithms and their performance have been discussed as follows:

**Neural Network:** This algorithm was able to produce better results than the decision tree algorithm. A brief explanation of the usage and the results produced from this algorithm are given below.

1. Applying Neural Network algorithm on training data:

In R language, package neuralnet is used for Neural Network. The following command is used to execute the neural network.

nn <- neuralnet(formula = loan\_status\_scale ~ interest\_rate + dti + LTI + delinq\_2yrs + term, data = temp\_data, hidden = 3, stepmax = 1e6, linear.output = FALSE, algorithm = "backprop", learningrate = 0.01)

where,

formula is one of the parameters to be given to implement Neural Network. We gave loan\_status\_scale as a label for the classifiers which is 1 for safe and 0 for unsafe of loan repayment. Other attributes are the key attributes on which decision has been made. Step max is used to denote maximum steps to be taken by Neural Network to make decision. Linear output means if act.fct should not be applied to the output neurons set linear output to TRUE, otherwise to FALSE. Act.fct is a differentiable function that is used for smoothing the result of the cross product of the covariate or neurons and the weights. Learning rate is a numeric value specifying the learning rate used by traditional back propagation. Used only for traditional back propagation.

Variable nn will store the output produced after applying neural network algorithm on the dataset. Following outpur shows the error rate, and number of steps taken by Neural Network algorithm.

1 repetition was calculated.

Error Reached Threshold Steps

1 0.03017827207 0.009807199889 350664

Following is the Decision Matrix produced with the Neural Network which shows the weights assigned on different links.

|  |
| --- |
| 1  error 0.030178272067  reached.threshold 0.009807199889  steps 350664.000000000000  Intercept.to.1layhid1 1.119559115490  interest\_rate.to.1layhid1 1.214969113709  dti.to.1layhid1 -2.800579775599  LTI.to.1layhid1 -1.796644379001  delinq\_2yrs.to.1layhid1 -2.671824252053  term.to.1layhid1 -2.258077409628  Intercept.to.1layhid2 34.506721872063  interest\_rate.to.1layhid2 -2.682935089284  dti.to.1layhid2 0.047517998053  LTI.to.1layhid2 -6.871622275804  delinq\_2yrs.to.1layhid2 -4.137181026833  term.to.1layhid2 -0.099830805395  Intercept.to.1layhid3 -53.717273943383  interest\_rate.to.1layhid3 7.577741119390  dti.to.1layhid3 0.172656724824  LTI.to.1layhid3 4.658396981941  delinq\_2yrs.to.1layhid3 0.908152474328  term.to.1layhid3 -13.684070007310  Intercept.to.loan\_status\_scale 4.464383801183  1layhid.1.to.loan\_status\_scale -3.944591780427  1layhid.2.to.loan\_status\_scale -10.296653987983  1layhid.3.to.loan\_status\_scale 9.694053223478 |
|  |
| |  | | --- | |  | |

Following shows the output of the 25 tuples where 1st attribute is the row number and second attribute shows the predicted value of the row which can be computed as 0 or 1 after rounding up.

[,1]

256787 0.001306933596780

48261 0.980898108148450

112665 0.975315861413863

68244 0.005140151071020

192115 0.975243968591713

312040 0.995559115228409

343895 0.999998616327410

265073 0.002298431765039

64211 0.999939205424671

110660 0.091024995284347

348320 0.000075709972513

163923 0.000039847136137

306627 0.003115400826391

82581 0.997300620967078

62645 0.999938936229702

223559 0.999999994743025

195427 0.006391556980371

191184 0.999919853742878

186751 0.999808767628835

56386 0.996956297777231

160870 0.999934529834257

192772 0.999849715668213

32451 0.980932173387797

105643 0.980897554899414

309277 0.993550302900978

Following diagram shows the neural network generated:

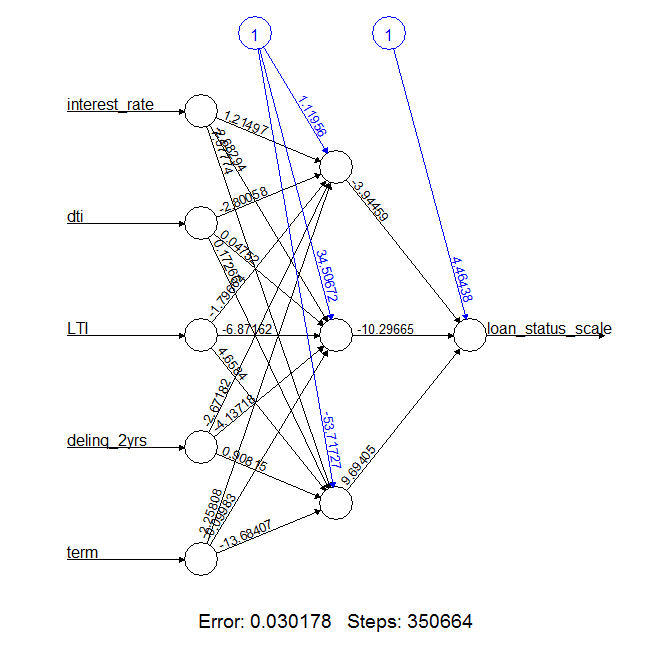


Figure 1 - Neural Network

1. Testing Neural Network on Test Data:

Following commands are used for the testing.

nn.results <- compute(nn, temp\_test)

results <- data.frame(actual = test\_data$loan\_status\_scale, prediction = nn.results$net.result)

Above output gave us the predicted and actual results. Following table shows the row number, actual and rounding of predicted output.

Actual Prediction

17440 1 1

302221 0 0

243882 1 1

171675 1 1

291419 1 1

118959 0 0

84923 1 1

251919 0 0

344686 1 1

71467 1 1

137188 1 1

172571 0 0

190563 1 1

232826 1 1

299634 1 1

138195 1 1

103788 0 0

184261 0 0

285218 0 0

185939 1 1

206992 1 1

132217 1 1

224441 0 0

133539 1 1

154780 0 0

94880 1 1

208227 0 0

7147 0 1

292724 1 1

**Decision Tree:** Another algorithm that we used for producing the classifiers is the decision tree algorithm. **Rpart, partykit and rattle** are the packages that we have used for implementing the decision tree. We used **Rattle**, GUI based tool, **simple** and **intuitive interface** that allows a user to quickly load data from a CSV file (or via ODBC). Following diagram shows the loaded csv file LDSPostPrecossing2 which was used for the Decision Tree. We can select Target Column, Input Columns using this tool.

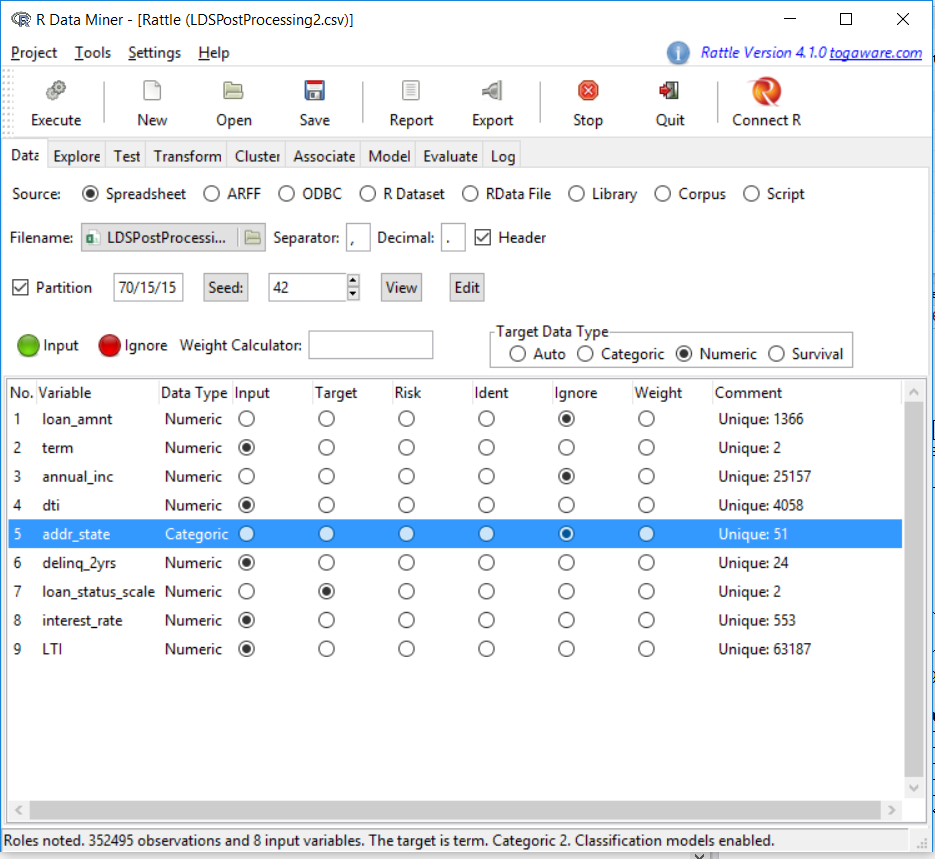


Figure 2 - Rattle Tool for Implementing Decision Tree

Then we applied Decision tree as follows:

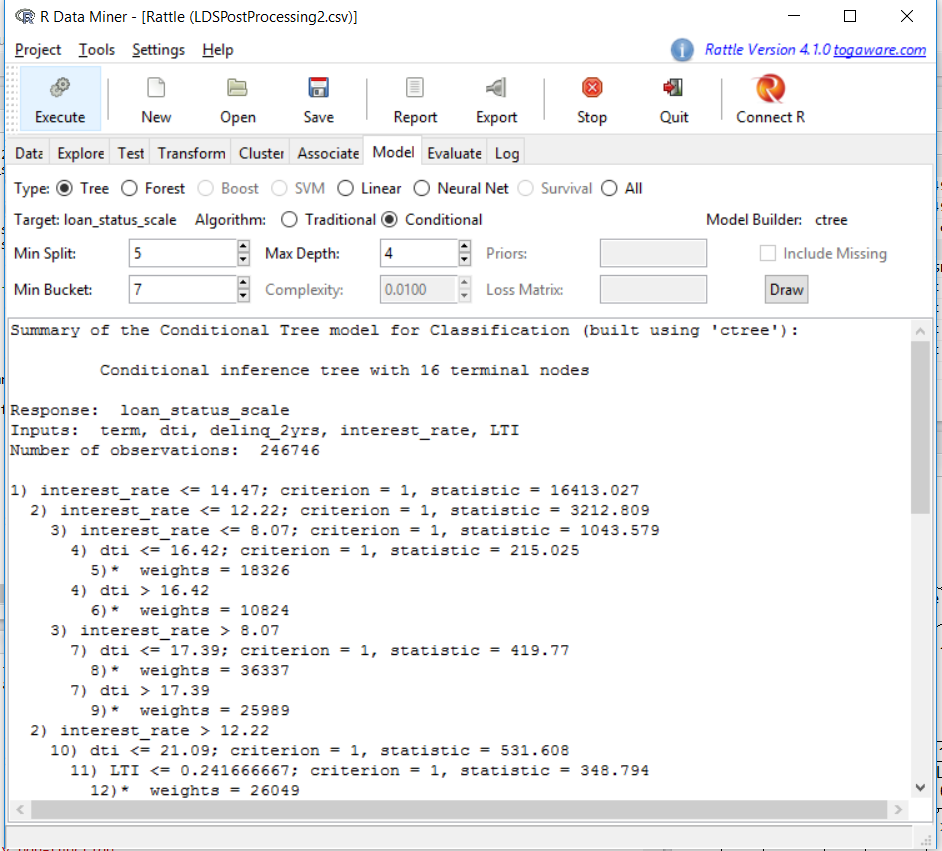


Figure 3 - Decision Tree result

**Parameters** usedin the algorithm:

1) Min Split – minimum observations in a node to split

2) Min Depth – minimum depth of the tree

3) Min Bucket – minimum observations in any leaf node

Following diagrams show how different Decision Tree generated by changing Min Depth Parameter from 4 to 3.

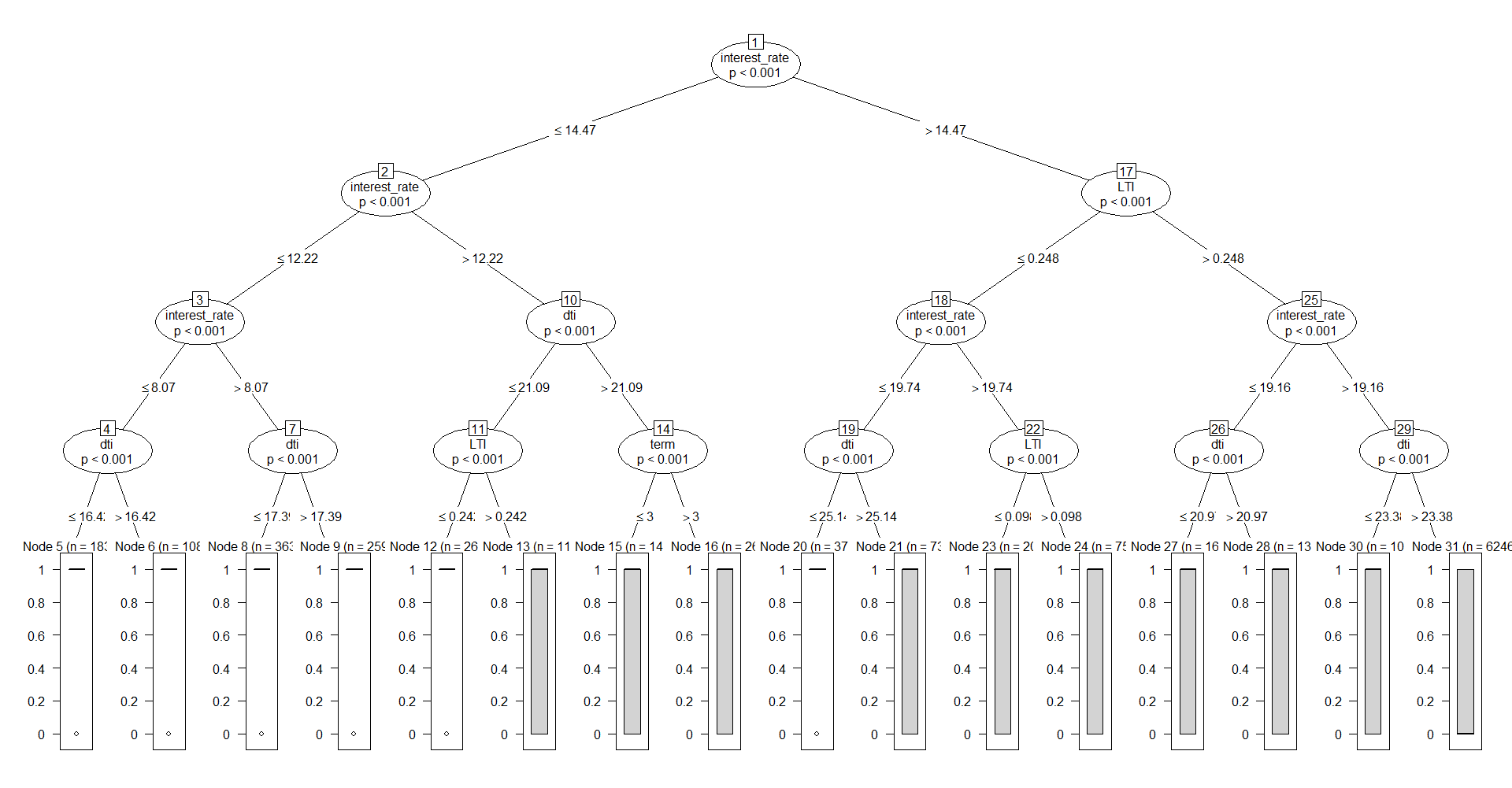


Figure 4 - Decision Tree with Min Depth = 4

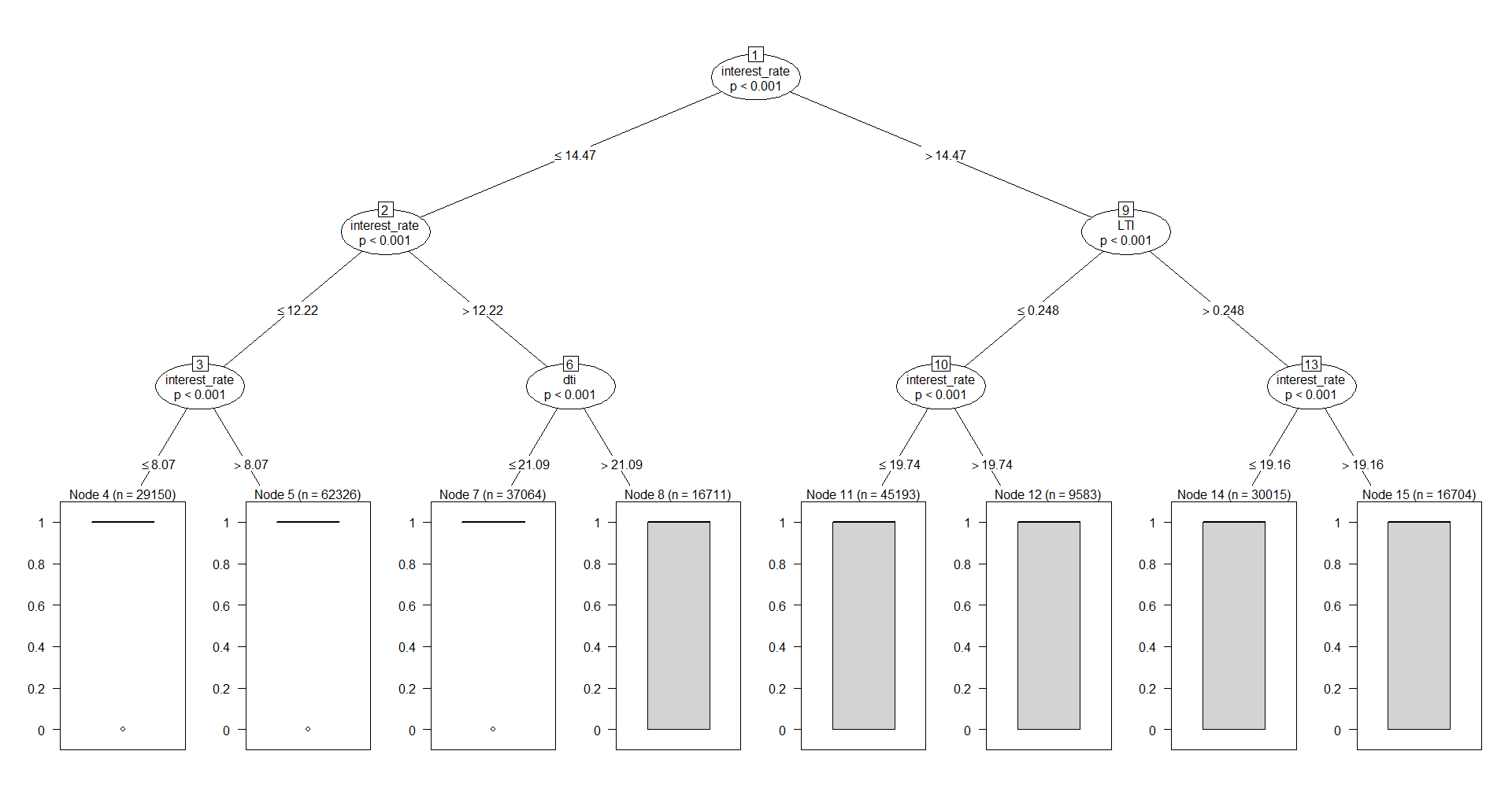


Figure 5 - Decision Tree with Min Depth = 3

Due to the drawbacks in decision tree, it was considered inappropriate for the project. The tendency of a decision tree to lead to drastic changes in the tree with minor changes in the data was one of the major reasons. Due to which, different decision trees were produced for different samples of the data. There was no consistency in the results produced from it. Since, there was no consistency on the results.

**Evaluation results:**

We evaluated our results of Neural Network using Confusion Matrix. Following shows outcomes of Classifier Evaluation.

Reference

Prediction 0 1

0 15 1

1 1 33

Accuracy : 0.96

95% CI : (0.8629, 0.9951)

No Information Rate : 0.68

P-Value [Acc > NIR] : 1.249e-06

Kappa : 0.9081

Mcnemar's Test P-Value : 1

Sensitivity : 0.9375

Specificity : 0.9706

Pos Pred Value : 0.9375

Neg Pred Value : 0.9706

Prevalence : 0.3200

Detection Rate : 0.3000

Detection Prevalence : 0.3200

Balanced Accuracy : 0.9540

'Positive' Class : 0

**Conclusion:**

This project is based on the mining performed on historical loan data collected from various banks. The results can aid bankers to identify customers who cannot repay the loan based on the classifiers produced. This was achieved by two algorithms i.e. Neural Network and Decision Tree. We compared performance of both techniques, out of which Neural Network works well with numeric as well as binary data and produce better and accurate results.

**Future Work:**

The rate of charged off loans has been a major issue in institutions like banks and private companies. It is necessary to interpret the attributes and the factors affecting the increasing rate. In the future, research can be extended by considering the standard of income pertaining to an area. Different classifiers can be produced based on the area.

Huge amount of information produced is produced from loan data. Hence, scalability issues need to be resolved by performing mining tasks better platforms such as Hadoop. The research can be extended by applying algorithms other than Decision Tree and Neural Network, and compare their performances.

**Responsibility of team members:**

1. Priyanka Shah: Collecting dataset and gathering project relevant information. Searching for any useful algorithm of finding patterns from dataset by reading various papers. Applying Data Reduction and Data Integration techniques. Applying Neural Network and Decision treealgorithm in R. Project Testing. Documentation.
2. Abilash Reddy Kommareddy: Collecting dataset and gathering project relevant information. Searching of any useful algorithm for finding patterns from dataset by reading various papers. Applying Visualization Techniques. Applying Neural Network and Decision tree algorithm in R. Project Testing. Documentation.
3. Varun Kumar Chepuri: Collecting dataset and gathering project relevant information. Searching of any useful algorithm for finding patterns from dataset by reading various papers. Applying Data Cleaning algorithms. Applying Neural Network and Decision tree algorithm in R. Project Testing. Documentation.

**Bibliography:**

Sridhar, V. **Predictive Mining on Loan Data using Rattle**. International Journal of Advanced Research in Computer Science, pp 1 – 10, Vol. 04/10, Sep – Oct 2013.

Edward, L., Gregory E. **A Comparative Analysis of Payday Loan Customers**. Contemporary Economic Policy, pp 299 – 316, Vol. 26/2, Apr 2008.

**Reby, D., Lek, S., Dimopoulos, I., Joachim, J., Lauga, J., &Aulagnier, S. (1997). Behavioural Processes. *Artificial Neural Networks as a Classification Method in the Behavioural Sciences,40*, 35-43.**

We referred website <https://www.lendingclub.com/> to understand how lending club company is dealing with Loan approval.

We referred Article “Using neural networks for credit scoring: a simple example” from the website <http://www.r-bloggers.com/using-neural-networks-for-credit-scoring-a-simple-example/>

We referred book “Data Mining Concepts and Techniques – Third Edition by Jiawei Han for understanding the working of Decision Tree and Neural Network algorithms.