

Decision Tree

Basic information **greatlearning**

- **Risk Analytics** – A wide domain in financial and banking industry , basically analyzing the risk of the customer.
 - Ex – 1) Whether the customer will be a transactor or revolver , etc..
 - 2) Whether a customer be delinquent or non-delinquent.
 - 3) Whether a customer be defaulter or non-defaulter.
- **Transactor** – A person who pays his due amount balance full and on time.
- **Revolver** – A person who pays the minimum due amount but keeps revolving his balance and does not pay the full amount.
- **Delinquent** - Delinquency means that you are behind on payments , a person who fails to pay even the minimum due amount.
- **Defaulter** – Once you are delinquent for a certain period of time your lender will declare you to be in default stage.

Data Summary

greatlearning

```
> str(pdata)
'data.frame':  11548 obs. of  7 variables:
 $ ID      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ delinquent : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ Sdelinquent: int  0 0 0 0 0 0 0 0 0 0 ...
 $ term     : Factor w/ 2 levels "36 months","60 months": 1 1 1 1 1 1 1 1 1 1 ...
 $ gender   : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 1 1 1 1 1 ...
 $ age      : Factor w/ 2 levels ">25","20-25": 2 2 2 2 2 2 2 2 2 2 ...
 $ FICO     : Factor w/ 2 levels ">500","300-500": 2 2 2 2 2 2 2 2 2 2 ...
```

- For further description of the variables please refer the attached data dictionary.



Data Dictionary

Exploratory Data Analysis



- The dataset consists of 7 variables and 11548 observations.
- We are trying to analyze attributes that leads to Delinquency.
- Dataset mainly contains Categorical variables.
- In the current dataset out of total 67% of the customers are delinquent, hence we need to analyze the attributes which can reduce this delinquency rate.
- Hence in our analysis **delinquent** would be the **target or the response variable** i.e the Dependent variable and other variables would be independent or the predictor variables.
- There are no missing values in the dataset , hence no treatment is required for the missing values.

Identifying Dependent & Independent Variables

Dependent Variable	Independent Variables
delinquent	term
	gender
	age
	FICO

CART(Classification and Regression Tree)

- The analysis to understand the major segments of delinquent customers has been implemented through CART.
- CART decision tree has been built in R and attached is the txt file that contains the code used for it.



R Code

- Attached is the pdf that shows the output of the tree , detailed output is given in the subsequent slides as well.



CART Output

CART(Classification and Regression Tree)

greatlearning

- The dataset has been split into training (70%) and test datasets (30%).
- Below are the libraries used in the code and the significance for using it.

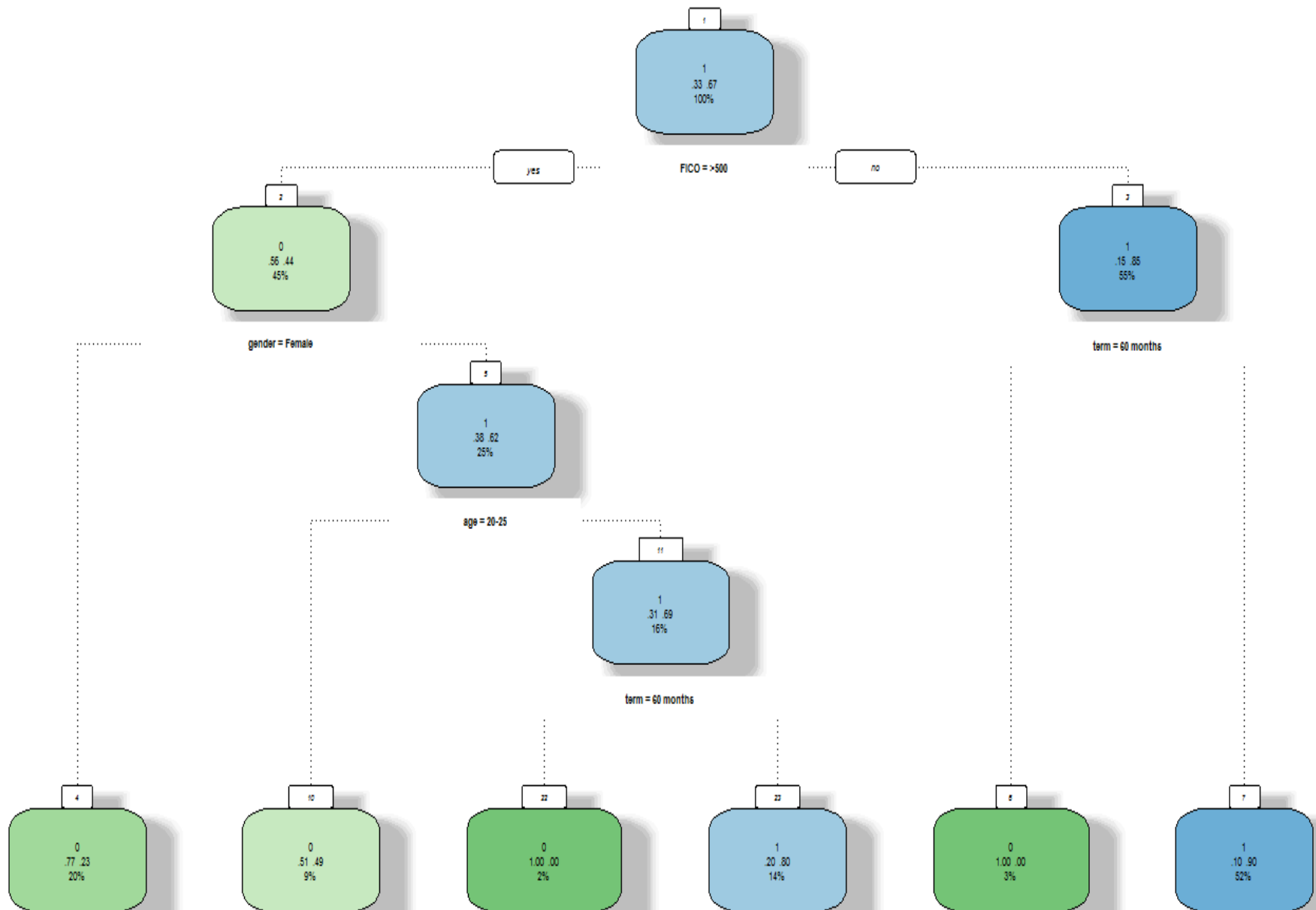
Library	Significance
caTools	Building the training and test dataset
Rpart	Built CART tree using rpart function and setting the control parameters using rpart.control function.
Rattle	Use the fancyRpartPlot function to display the tree
ROCR	Calculate KS , AUC , etc. statistics.
Ineq	Calculate GINI index.

CART(Classification and Regression Tree)

- Below are the control parameters that are used while building the decision tree:

Parameter	Value	Significance
minsplit	1000	If the node will have at least 1000 observations then only it will split.
minbucket	100	The terminal nodes should have at least 100 observations.
cp (Complexity Parameter)	0	Allowing the full tree to be grown.
xval(Cross Validation)	10	It will cross validate 10 times.

Decision Tree



Decision Tree output explanation & Insights

greatlearning

- First Node shows that there are 67% of delinquent customers and 33% non- delinquent customers.
- FICO is the first variable that is split in the decision tree hence is the most important variable as well while building any strategy.
- The highest risk segment is of 52% which means that 52% of the delinquent customers are being taken in that segment which is of customers having $FICO < 500$ and loan term of 36 months.
- The second risk segment is of 14% which means that 14% of the delinquent customers are being taken in that segment which is of customers having $FICO > 500$ and gender = Male and age > 25 and loan term = 36 months. Rest of the segments are non-risk segments.
- Next slide shows the detailed level decision tree as well.

Decision Tree

greatlearning

```
> m1 <- rpart(formula = sdelinquent~term+gender+FICO+age, data = p_train, method = "class", control = r.ctrl)
> m1
n= 7900

node), split, n, loss, yval, (yprob)
* denotes terminal node

1) root 7900 2615 1 (0.33101266 0.66898734)
 2) FICO=>500 3541 1568 0 (0.55718724 0.44281276)
   4) gender=Female 1591 367 0 (0.76932747 0.23067253) *
   5) gender=Male 1950 749 1 (0.38410256 0.61589744)
     10) age=20-25 701 343 0 (0.51069900 0.48930100) *
     11) age=>25 1249 391 1 (0.31305044 0.68694956)
       22) term=60 months 172 0 0 (1.00000000 0.00000000) *
       23) term=36 months 1077 219 1 (0.20334262 0.79665738) *
 3) FICO=300-500 4359 642 1 (0.14728149 0.85271851)
   6) term=60 months 234 0 0 (1.00000000 0.00000000) *
   7) term=36 months 4125 408 1 (0.09890909 0.90109091) *
```

- In the decision tree output 1 denotes delinquent and 0 represents non-delinquent.

Insights and implementation strategy

greatlearning

- Building and using this model the below insights are generated and the following strategies can be implemented:

Risk Segments	
Segment	Meaning
1) FICO < 500 and loan term = 36 months	1) Customers having low FICO score and having loan tenure of 36 months.
2) FICO > 500 and gender = Male and age > 25 and loan term = 36 months.	2) Male customers more than 25 years old having higher FICO score and having loan tenure of 36 months.

Strategy that can be implemented to decrease the delinquency rates

- 1) Reject the applications of those customers who have less FICO score and whose loan tenure is 36 months.
- 2) Proper verification needs to be done for male customers before approving the loan who have higher FICO score and have a loan tenure of 36 months.

Insights and implementation strategy

greatlearning

Non - Risk Segments

Segment	Meaning
1) FICO < 500 and loan term = 60 months	1) Customers having low FICO score and having loan tenure of 60 months.
2) FICO > 500 and gender = Female.	2) Female customers with higher FICO score.
3) FICO > 500 and gender = Female and age = 20-25.	3) Female customers more than 25 years old having higher FICO score and having loan tenure of 36 months.

Strategy that can be implemented to increase profits and having loyal customers

- 1) Increase the volume of accepting loans of female customers with higher FICO score and whose loan tenure is of 36 months.
- 2) Accept the applications of those customers whose loan tenure would be 60 months though having low FICO score.
- 3) Give some promotional offers as well to Female customers so that they can have a better and long term relationship with the organization.

Model performance and validation

- **Classification error** = $(627+710)/7900 = 16.9\%$

```
> with(p_train, table(sdelinquent, predict.class))
      predict.class
sdelinquent    0    1
0 1988   627
1   710 4575
> nrow(p_train)
[1] 7900
```

- Classification error gives a clear picture of the % of the observations that have been predicted wrongly and correctly by the model.
- The table of classification error gives the below insights:

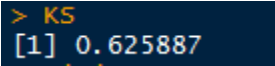

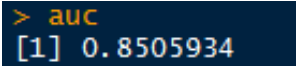

Classification	As per the above output
1) True Positive	1) 4575
2) True Negative	2) 1988
3) False Positive	3) 627
4) False Negative	4) 710

TPR = 86.5%

FPR = 23.9%

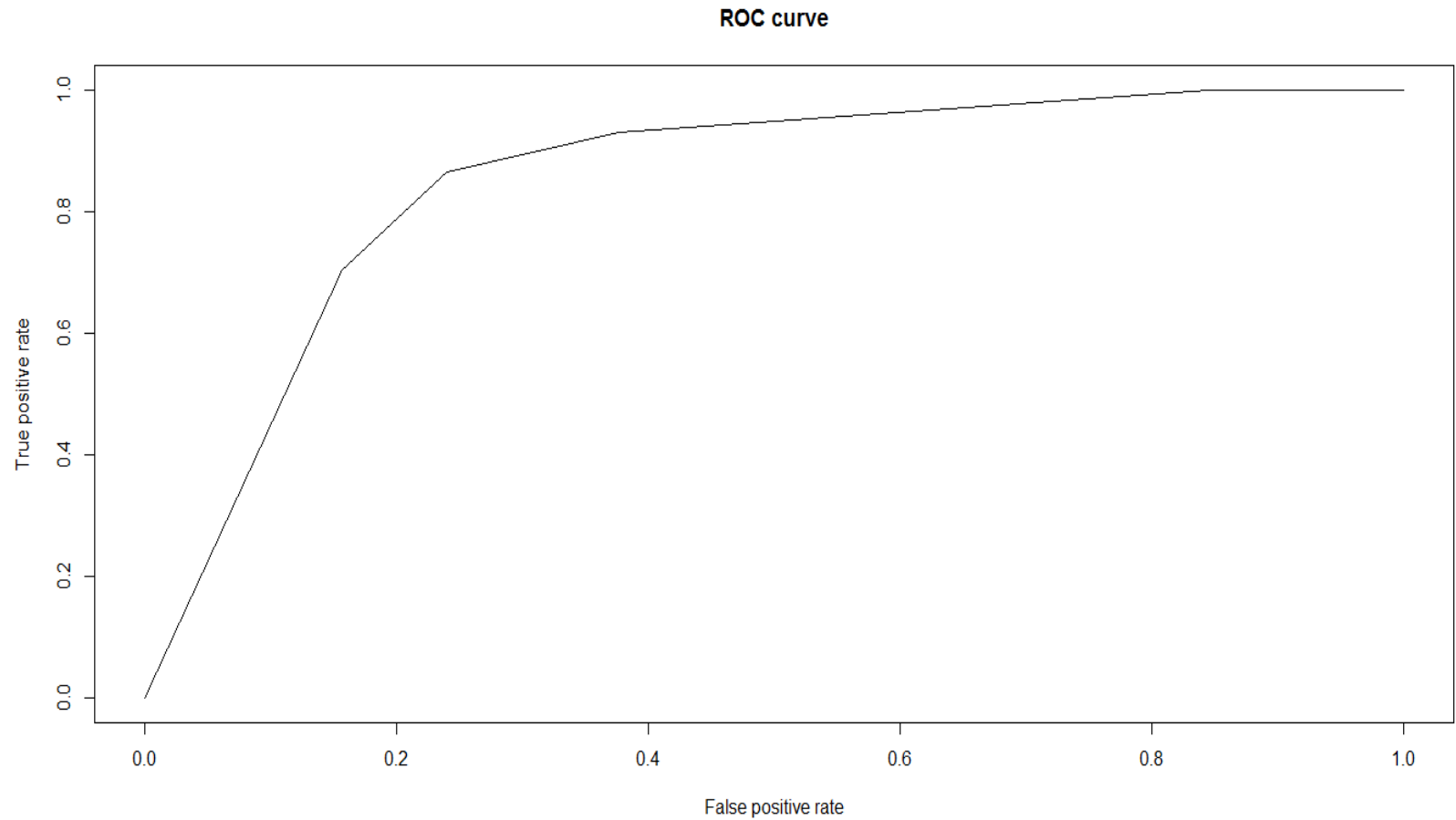
Model performance and validation

greatlearning

- **KS** = 62.5%   KS is higher which suggests the model holds good.
 - KS (Kolmogorov Statistic) is a **non-parametric** test that measures the degree of separation between the positive and negative distributions.
 - Higher the KS better the model is because higher KS means better the separation is by the model, its value is between 0 and 1 and has to be interpreted in %.
- **AUC** = 85%   AUC is higher which also suggests that the model also holds good.
 - AUC (Area Under Curve) is the % of the box that is under the ROC (Receiver Operating Characteristics) curve which is explained in the next slide.

Model performance and validation

greatlearning



Model performance and validation – Test Dataset

The model was tested on test dataset as well and below are the output of the validation measures:

```
> K5  
[1] 0.6236039  
> auc  
[1] 0.8490914
```



In line with the training dataset hence no issue with utilizing the model with the test dataset.

```
> with(p_test, table(Sdelinquent, predict.class))  
      predict.class  
Sdelinquent  0    1  
      0  919  293  
      1  328 2108  
> nrow(p_test)  
[1] 3648
```



Classification error
= $(293+328)/3648$
= 17% which is inline with the training dataset.

Thank you