

# Lending Club Loan Data

SUBMITTED BY –  
JSOM CRUNCHERS

**SAS GLOBAL FORUM-2018**  
**LENDING CLUB LOAN DATA**  
KRITI JAIN, GOPESH VIJAYVERGIYA  
UNIVERSITY OF TEXAS AT DALLAS

## 1. ABSTRACT

### 1.1 Mission Statement

The objective of this project to build two models:

- a model that will predict if a loan will be fully funded by investors or not
- a model that predicts if a loan will be defaulted or not
- determine if a borrower will receive a low or high interest rate

### 1.2 Introduction

Lending Club (LC) is an online peer to peer lending platform headquartered in San Francisco, California. It facilitates investors in searching and browsing the loan listings and helps them select loans that they want to invest in. Investors make money from interest. Lending Club makes money by charging borrowers an origination fee and investors a service fee.

For assessing the risk associated with their borrowers, Lending Club primarily relies on a grade and sub-grade system that it assigns them based on their credit history. This information is then made available to investors who fund the loan requests, so that the investors can decide which loan request and how much of that loan request they will fund. In addition to the grade information, Lending Club provides historical loan performance data to investors for more comprehensive analysis.

## 2. PROJECT MOTIVATION

We selected this dataset because our team was unfamiliar with the domain and it would add the challenge of learning about the data while trying to extract meaningful information from it. The dataset had many observations and attributes and it would give us the opportunity to explore different techniques in data pre-processing and data mining.

## 3. DATA DESCRIPTION

The data is second-hand data that was obtained from [www.kaggle.com](http://www.kaggle.com). The dataset contains the data of only approved loans by the LendingClub between 2007 and 2015.

The dataset has 75 variables and 880,000+ observations. The attributes include information about the borrowers such as their open accounts, months since last delinquency, amount of loan requested, amount of loan funded by investor, employment length, annual income, lending club assigned loan grade, interest rate, installments etc.

Following are our target variables:

- not\_fully\_funded (binary - 0 - No / 1 - Yes)
- will\_default (binary - 0 - No / 1 - Yes)

We have rejected variables having missing values greater than 50% and those that were not relevant for the analysis such as

- IDs
- URL
- Employee Title
- Description

## 4. BI Model: Target variable - Not\_fully\_funded

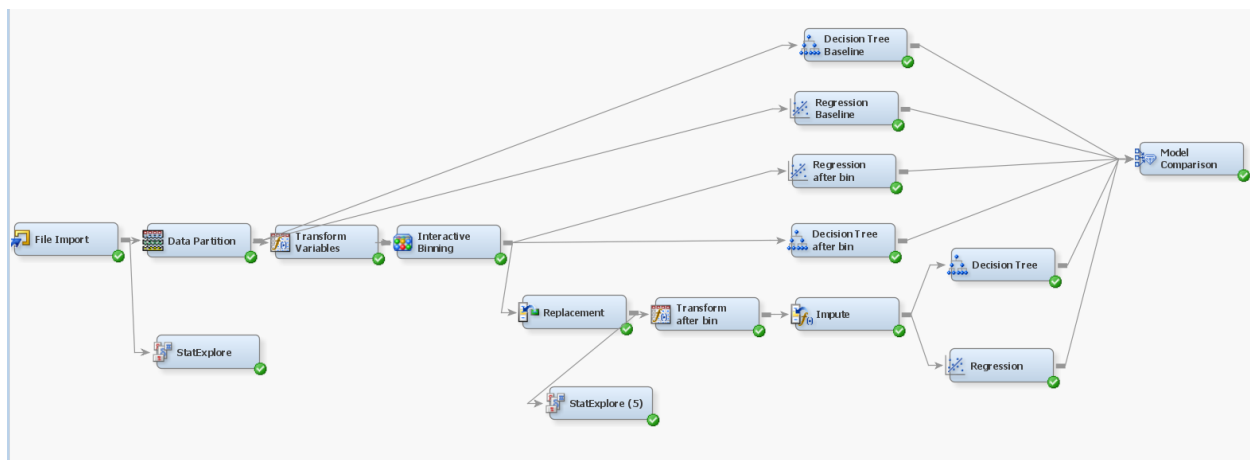


Fig: 4.1 – Process flow diagram for target variable not\_fully\_funded

We have created the target variable not\_fully\_funded that represents whether the loan is fully funded by an investor or not.

### 4.1 DATA EXPLORATION

The initial exploration was done using the File Import node by turning on the summarize option to 'Yes'. We can see from the below statistics table variables with high missing values:

Variable Name	Type	Number of Levels	Percent Missing ▼
dti joint	N		99.94264
verification status joint	C	3	99.94241
annual inc joint	N		99.94241
il util	N		97.90202
mths since rcnt il	N		97.65489
all util	N		97.59156
inq fi	N		97.59156
inq last 12m	N		97.59156
max bal bc	N		97.59156
open acc 6m	N		97.59156
open il 12m	N		97.59156
open il 24m	N		97.59156
open il 6m	N		97.59156
open rv 12m	N		97.59156
open rv 24m	N		97.59156
total bal il	N		97.59156
total cu tl	N		97.59156
mths since last record	N		84.5553
mths since last major derog	N		75.01597
mths since last delinq	N		51.19706
next pymnt d	N		28.50766

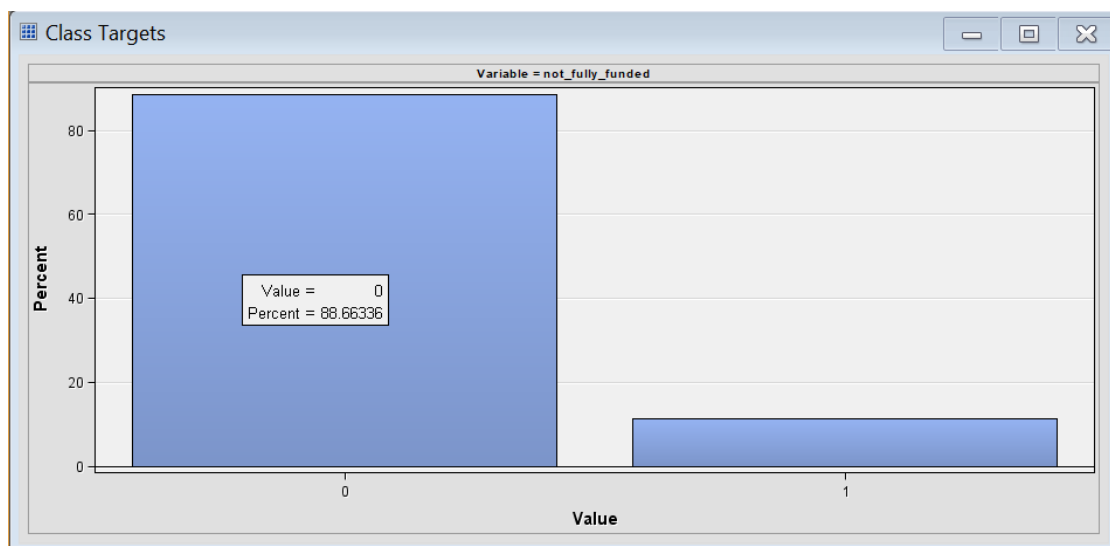
Fig: 4.2 – List of missing values

## Lending Club Loan Data



Fig: 4.4 – Classification of the target variable not\_fully\_funded

From the fig 4.4, we can see that the baseline misclassification rate is **11.33%**



## 4.2 DATA PARTITIONING

We have partitioned the data into training as 60% and validation as 40% of the initial dataset. The training data is used to create models to predict the outcome while the validation data helps to assess the model.

## 4.3 DATA PRE-PROCESSING

**Dropping variables:** Since we created the target variable Not\_fully\_funded from the two variables funded\_amt and funded\_amt\_inv, we have rejected funded\_amt\_inv as incorporating it in the analysis was resulting in near perfect model due to its direct relation with the target variable.

**Transform variables:** We have created 3 variables open\_rv, total\_open\_il, Total\_open\_acc by combining already existing columns in our dataset to simplify the computation. open\_rv was created combining open\_rv\_12m and open\_rv\_24m i.e. number of revolving trades opened in past 12 and 24 months. We did the same thing for total\_open\_il by combining open\_il\_6m, open\_il\_12m, open\_il\_24m i.e. the number of installment accounts opened in the past 6,12 and 24 months. And lastly, we created Total\_open\_acc by combining open\_acc and open\_acc\_6m which represents the number of open credit lines in the borrower's credit files.

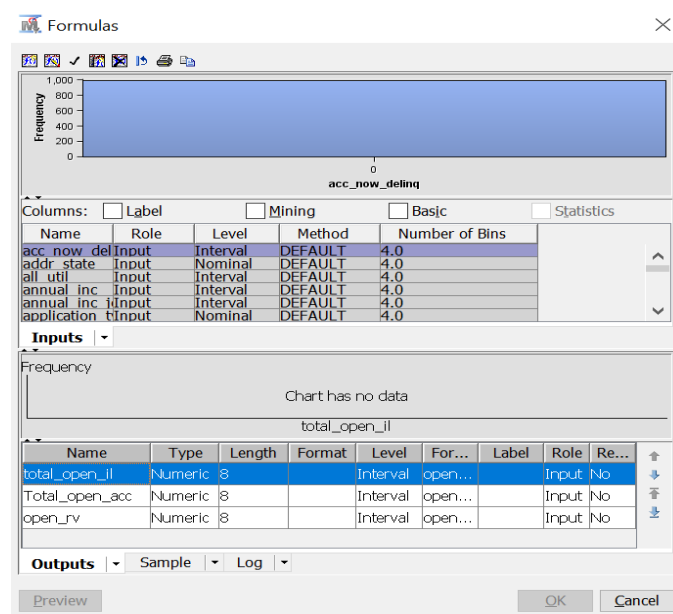


Fig: 4.5 – Creation of three new variables by combining existing variables

## 4.4 INTERACTIVE BINNING

**Transform Node:** We used transform node to transform variables such as:

- Months since last delinquency
- Months since last public record
- Months since last derogatory comment

These variables had approximately 97% missing values, but we have assumed that a missing value means that it has been a very long time since or the person has never had a record against them. For this reason, we have transformed these variables using transform and interactive binning to be used as whether a person has a record against them (irrespective of when) or not.

### Conversion of continuous variables to categorical:

To get more discernible results, we used interactive binning to convert continuous variables like interest rate, dti, last payment amount, total payment, employment length, delinquencies in the past 2 years into categorical variables.

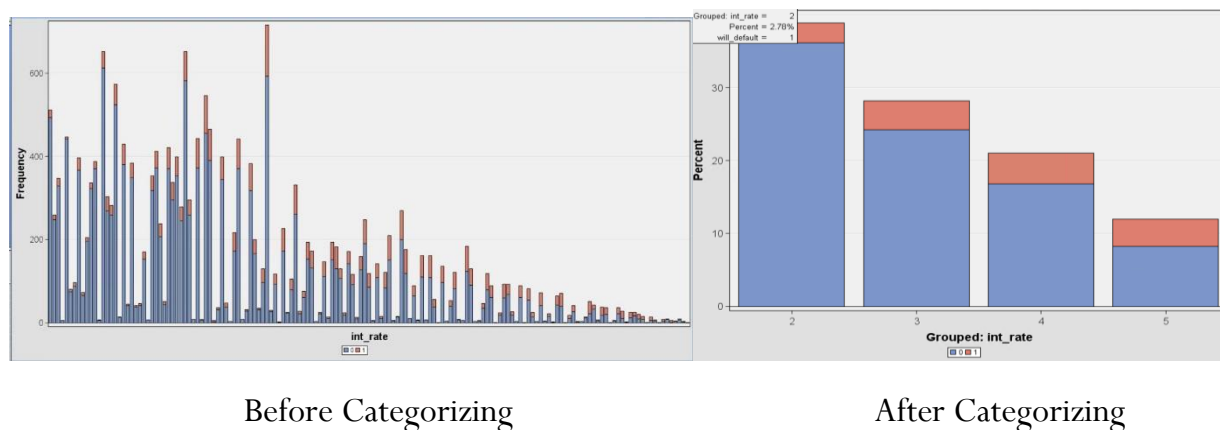
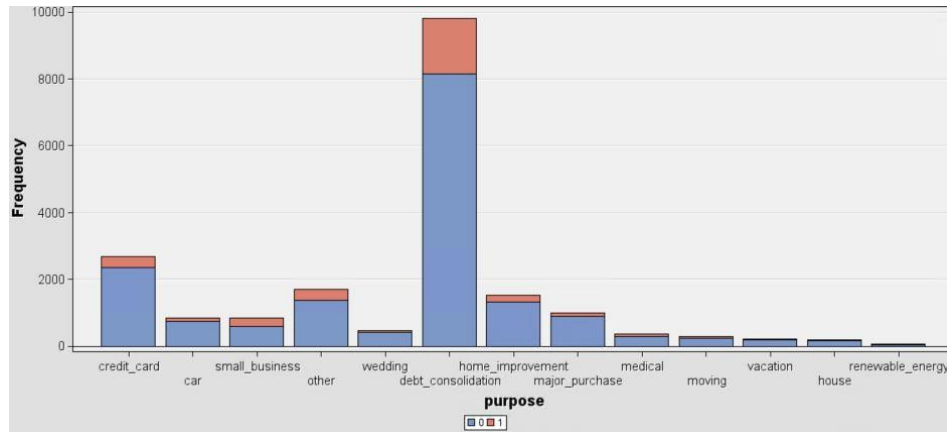


Fig: 4.6 – Categorizing continuous variable int\_rate

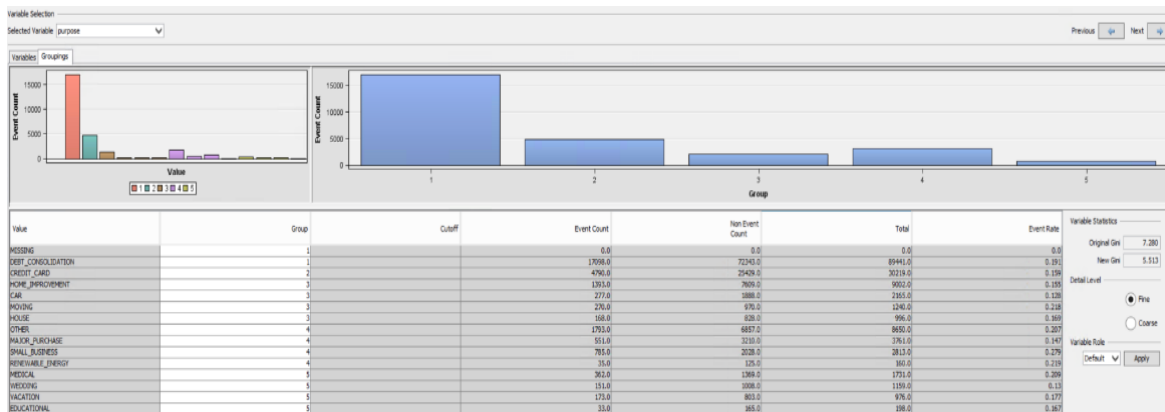
Fig: 4.6 shows that as the interest rate increases, the percent of defaulted loans (1) also increases.

### Category reduction of nominal variable - purpose:

The nominal variable purpose had 14 levels, we created 5 categories by combining similar purposes.



Before categorizing



After categorizing

Fig: 4.7 – Reducing categories of nominal variable purpose

**Removing outliers and reducing the skewness:** The replacement, transform and impute nodes were used in conjunction for removing the outliers and reducing the skewness of the data.

**Replacement node:** The replacement node was used to remove values that were outside of  $\pm 3$  standard deviations away from the mean.

**Transform Node:** We have used the transform node to reduce the skewness of variables using a logarithmic transformation for all the variables. This helped us in drastically reducing the skewness of highly skewed variables.



**Impute:** We have used the default impute method which uses the mean to impute the missing values for interval variables and count to impute for class variables.

Following are the results after removing outliers, transforming and imputing:

Variable Name	Impute Method	Imputed Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
REP annual inc	MEAN	IMP REP annual inc	73640.79	INPUT	INTERVAL	Replacement: annual inc	2
REP inq last 6mths	MEAN	IMP REP inq last 6mths	0.680345	INPUT	INTERVAL	Replacement: inq last 6mths	13
REP open acc	MEAN	IMP REP open acc	11.48789	INPUT	INTERVAL	Replacement: open acc	13
REP pub rec	MEAN	IMP REP pub rec	0.177163	INPUT	INTERVAL	Replacement: pub rec	13
REP revol util	MEAN	IMP REP revol util	55.06145	INPUT	INTERVAL	Replacement: revol util	303
REP tot coll amt	MEAN	IMP REP tot coll amt	204.8739	INPUT	INTERVAL	Replacement: tot coll amt	41921
REP tot cur bal	MEAN	IMP REP tot cur bal	136369.9	INPUT	INTERVAL	Replacement: tot cur bal	41921
REP total acc	MEAN	IMP REP total acc	25.20061	INPUT	INTERVAL	Replacement: total acc	13
REP total rev hi lim	MEAN	IMP REP total rev hi lim	31152.08	INPUT	INTERVAL	Replacement: total rev hi lim	41921

Fig: 4.8 – Results – removing outliers, reducing skewness and imputing

## 4.5 Models for target variable not\_fully\_funded

We used decision tree and logistics regression and tried 6 different models.

## 4.6 Model Comparison Results

Below are the results of the model comparison node. Using misclassification rate as the selection criterion, SAS EM has selected **“Tree7 (Decision tree after bin)”** as the best model.

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Selection Criterion : Valid: Misclassification Rate	Valid: Average Squared Error	Valid: Mean Square Error	Valid: Root Mean Square Error	Train: Misclassification Rate	Train: Average Squared Error	Train: Root Mean Squared Error
	Req3	Req3	Regression Baseline	not fully funded	0.11337	0.11337	0.11337	0.336705	0.113364	0.113364	0.336696
	Req7	Req7	Regression after bin	not fully funded	0.11339	0.100886	0.100886	0.317625	0.113342	0.100836	0.317573
	Tree3	Tree3	Decision Tree Baseline	not fully funded	0.1078...	0.086674	.	.	0.107506	0.08662	.
	Req8	Req8	Regression	not fully funded	0.1130...	0.086023	0.086023	0.293297	0.112662	0.085613	0.292886
	Tree8	Tree8	Decision Tree	not fully funded	0.10556	0.082651	.	.	0.104129	0.081827	.
Y	Tree7	Tree7	Decision Tree after bin	not fully funded	0.1052	0.081541	.	.	0.104738	0.081136	.

Fig: 4.9 – Fit statistics of models for target variable not\_fully\_funded

Below are the ROC curves for different models that we tried. The ROC curves show that the “**Tree7 (Decision tree after bin)**” is the best model.

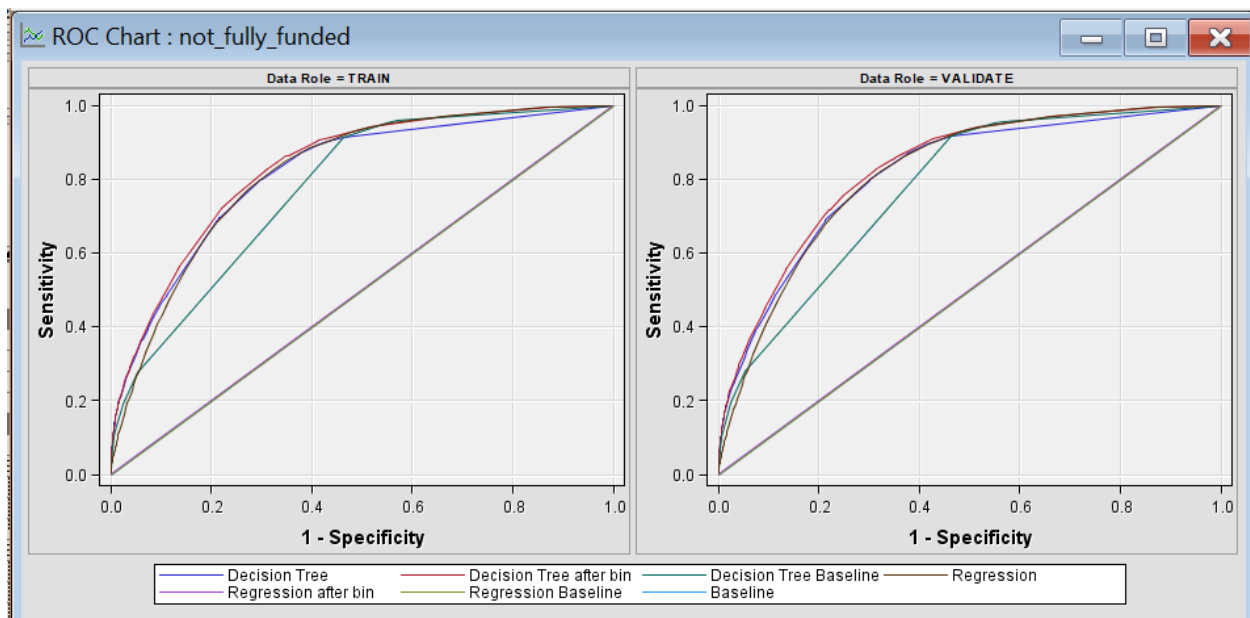


Fig: 4.10 – ROC curves of models for target variable not\_fully\_funded

Below is the cumulative lift chart for all the models:

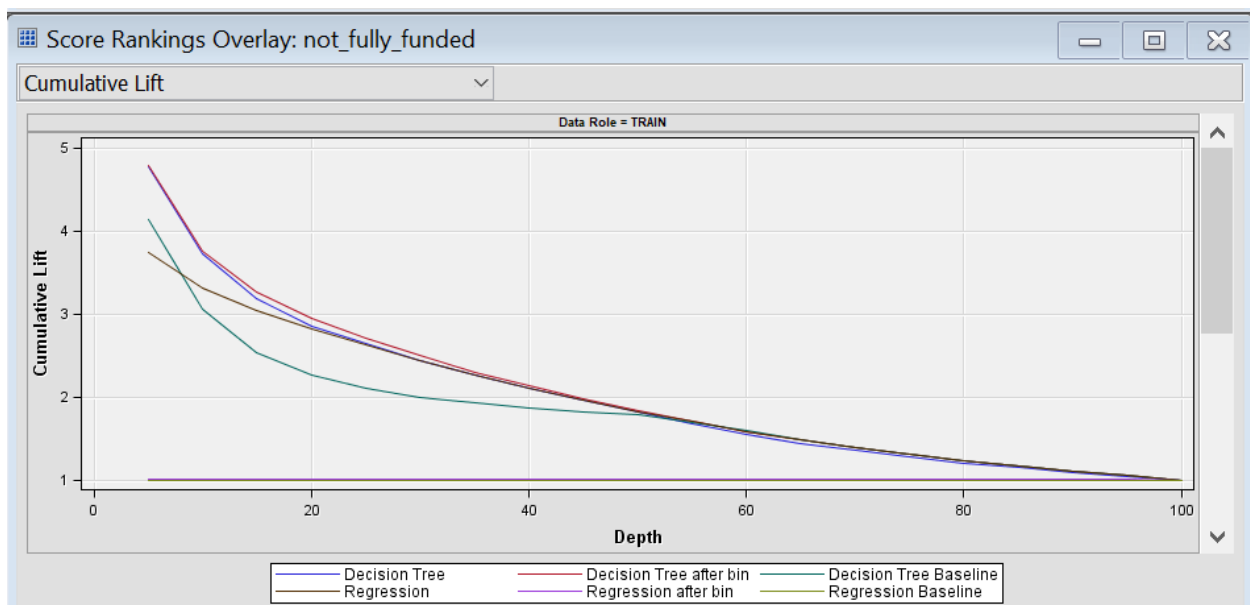


Fig: 4.11 – Cumulative lift of models for target variable not\_fully\_funded

Below are the results for our best model:

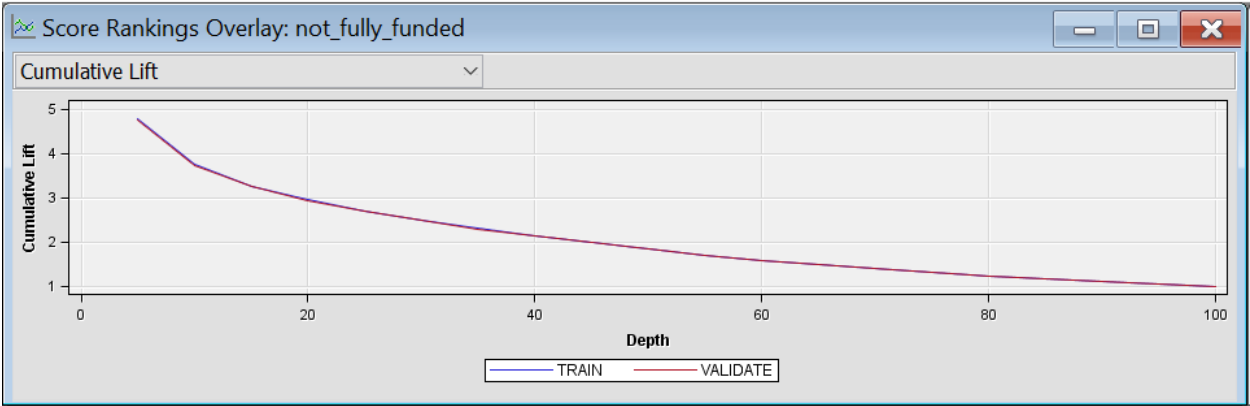


Fig: 4.12 – Cumulative lift of Tree7 for target variable not\_fully\_funded

From the Fig: 4.12, we can see that for 40% of the total approved loans, 2.11 times the loan will not be fully funded by the investors.

Fit Statistics						
Target	Target Label		Fit Statistics	Statistics Label	Train	Validation
not fully funded	not fully funded	not fully funded	NOBS	Sum of Frequencies	532426	354953
not fully funded	not fully funded	not fully funded	MISC	Misclassification	0.104738	0.1052
not fully funded	not fully funded	not fully funded	MAX	Maximum Absolute Error	0.995535	1
not fully funded	not fully funded	not fully funded	SSE	Sum of Squares	86397.85	57886.32
not fully funded	not fully funded	not fully funded	ASE	Average Squared Error	0.081136	0.081541
not fully funded	not fully funded	not fully funded	RASE	Root Average Squared Error	0.284844	0.285554
not fully funded	not fully funded	not fully funded	DIV	Divisor for ASE	1064852	709906
not fully funded	not fully funded	not fully funded	DFT	Total Degrees of Freedom	532426	

Fig: 4.13 – Fit statistics for Tree7 for target variable not\_fully\_funded

Confusion matrix:

		Actual	
		0	1
Predicted	0	311999 (True -)	34628 (False -)
	1	2713 (False +)	5613 (True +)

Fig: 4.14 – Confusion matrix for Tree7 for target variable not\_fully\_funded

Data Role=TRAIN Target=not_fully_funded Target Label=not_fully_funded			
False Negative	True Negative	False Positive	True Positive
51880	468183	3885	8478

Data Role=VALIDATE Target=not_fully_funded Target Label=not_fully_funded			
False Negative	True Negative	False Positive	True Positive
34628	311999	2713	5613

Fig: 4.15 – Event classification table for Tree7 for target variable not\_fully\_funded

## 4.7 Business implications

It is important to estimate the effect of misclassify the target variable. The errors can be classified into two types:

- False Positive
- False Negative

The more expensive error is the false negative error where the model predicts a 0 - meaning a loan is fully funded when it is actually not i.e. 1. The implications for this is that the lending club may not hold sufficient reserve funds to fund the loan thereby losing out on profit and customers.

## 4.8 Conclusion

This model was designed to help Lending club decide how much money they should keep in reserve, if they were to fund loans that were not fully funded by investors. Our best model returned a misclassification rate of 10.52% and we were unable to reduce it further. We suspect that the data is insufficient in the respect that we don't have the information about the credit history and credit score of the borrowers.

## 5. BI Model: Target Variable - Will Default

The following model will be used to predict the outcome of a loan whether it will be defaulted or not:

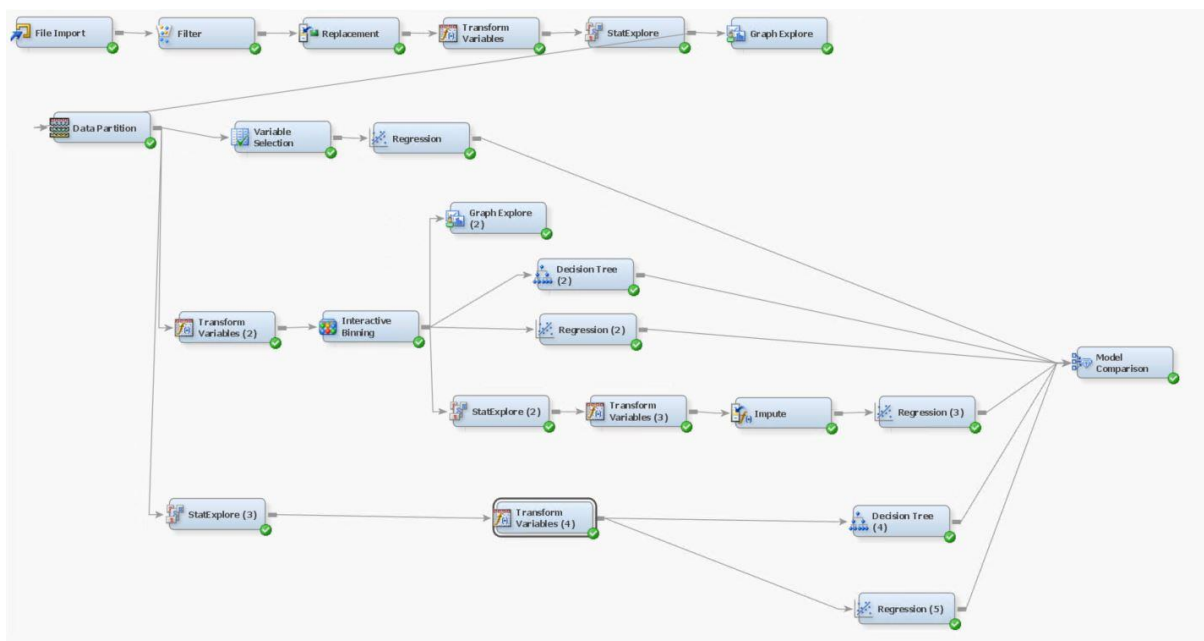


Fig: 5.1 – Process flow diagram for target variable will\_default

### 5.1 DATA EXPLORATION

From the diagram below, we can see that the baseline misclassification error is 19.4%.

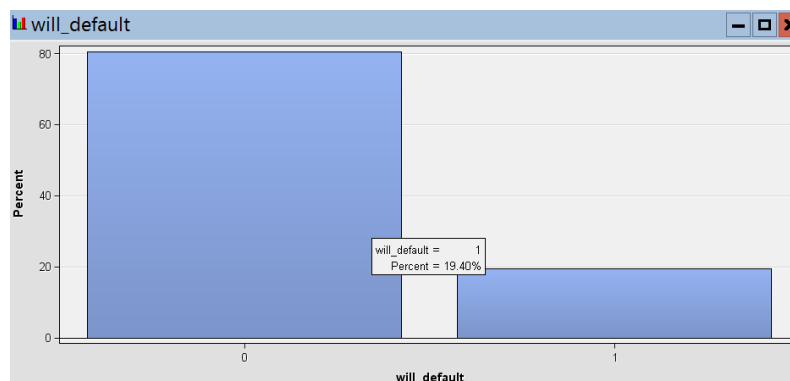


Fig: 5.2 – Classification for target variable will\_default

## 5.2 DATA PRE-PROCESSING

We have used a filter to remove observations that have a loan status of current, in grace period, late, and issued. These observations were removed since our target is to predict if a loan will be defaulted on or not. We were left with 250,000+ observations after filtering out the classes of target variables that were unnecessary to our model.

**Target Variable:** Replacement and Transform nodes were used to create a new binary target variable – Will\_Default where 0 – No (Loan Fully Repaid), 1 – Yes (Loan will be defaulted on).

Following is the picture showing the transformation for the target variable:

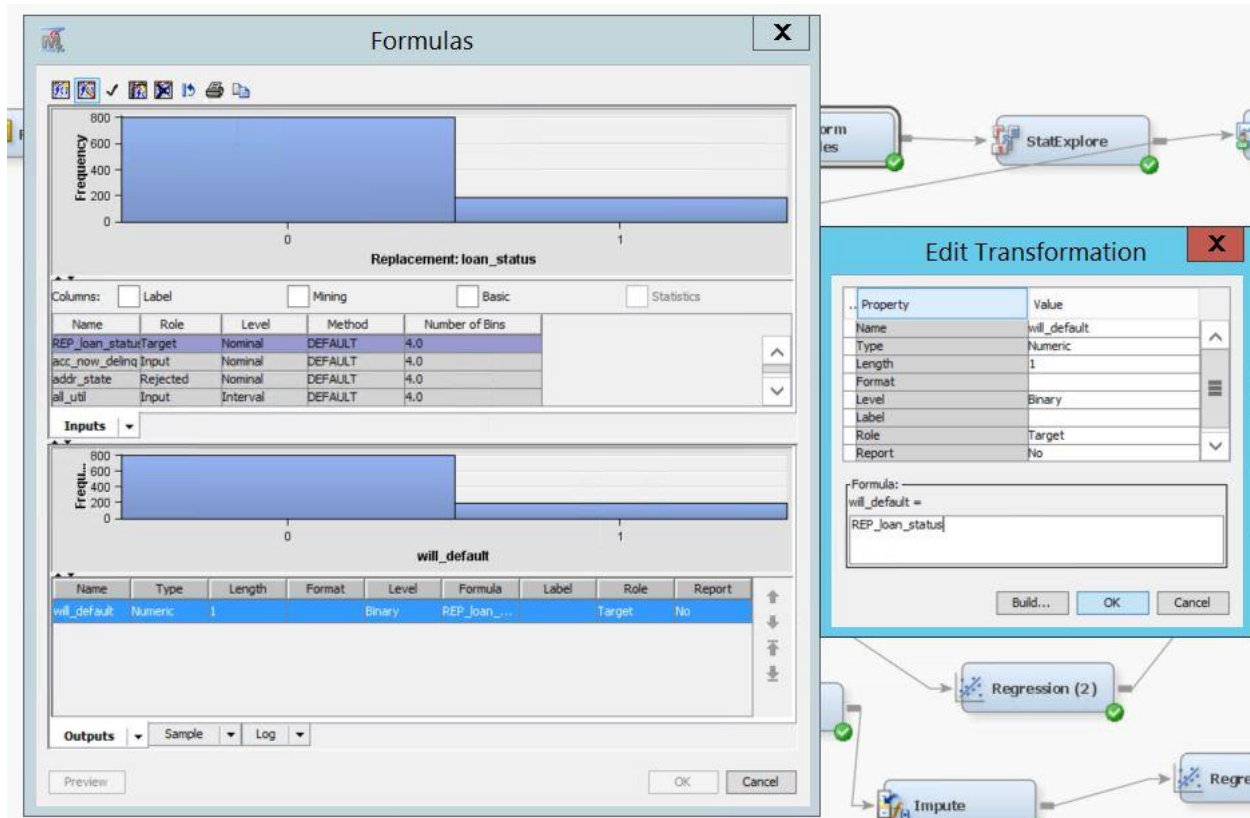


Fig: 5.3 – Creating target variable will\_default

**Data Partitioning:** The data was partitioned as Train – 60% and Validate – 40%.

**Dropping variables:** We rejected variables that had missing values above 50% and very less variable worth. We rejected variables like recoveries, coll\_rec\_fee (collection recovery fee), out\_prncp (outstanding principal) as incorporating it in the analysis was resulting in near perfect model due to its direct relation with the target variable. We also rejected variables like last\_pymnt (last payment), tot\_pymnt (total payment), tot\_received\_prncp (total received principal) and tot\_received\_int (total received interest) as this information would be unavailable at the time of deciding whether to fund a loan or not.

**Transform Node:** We used transform node to transform variables such as:

- Months since last delinquency
- Months since last public record
- Months since last derogatory comment

These variables had approximately 97% missing values, but we have assumed that a missing value means that it has been a very long time since or the person has never had a record against them. For this reason, we have transformed these variables using transform and interactive binning to be used has whether a person has a record against them (irrespective of when) or not.

### Conversion of continuous variables to categorical:

To get more discernible results, we used interactive binning to convert continuous variables like interest rate, dti, last payment amount, total payment, employment length, delinquencies in the past 2 years into categorical variables.

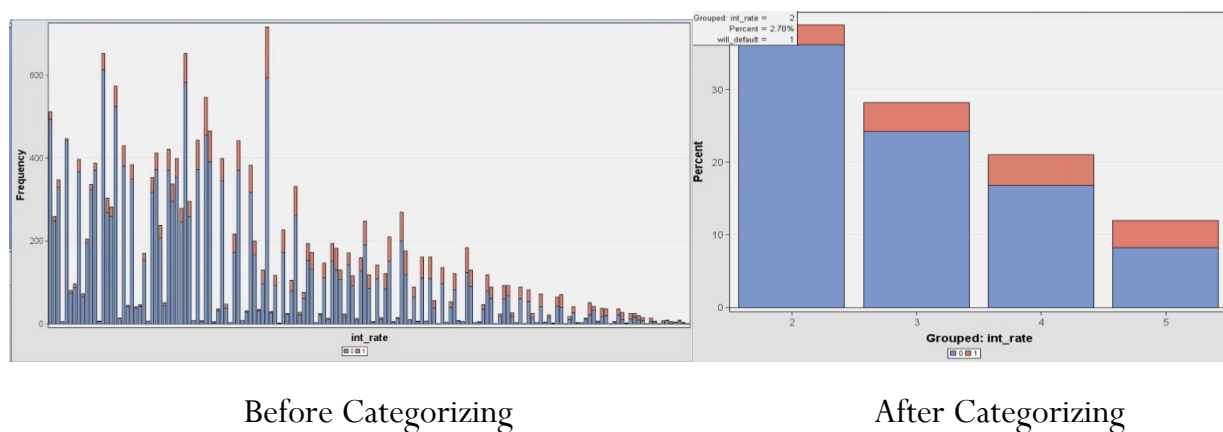
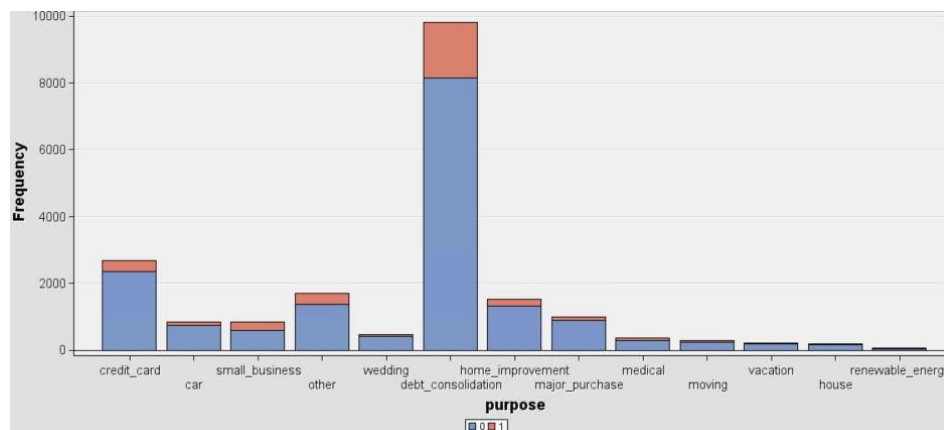


Fig: 5.4 – Categorizing continuous variable int\_rate

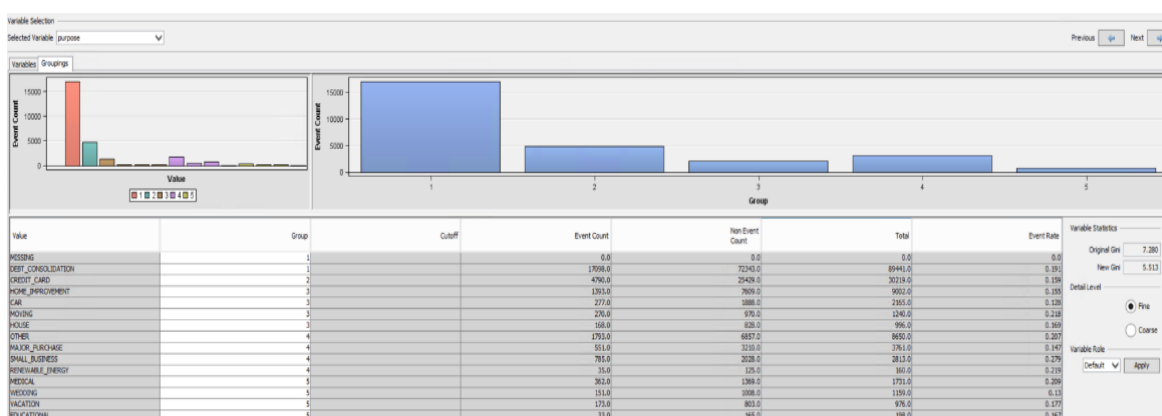
Fig: 5.4 shows that as the interest rate increases, the percent of defaulted loans (1) also increases.

### Category reduction of nominal variable - purpose:

The nominal variable purpose had 14 levels, we created 5 categories by combining similar purposes.



Before categorizing



After categorizing

Fig: 5.5 – Reducing categories of nominal variable purpose

To reduce the skewness of the following variables we used the Log 10 method in the Transform node:



Transformations Statistics				
Source ▲	Method	Variable Name	Formula	Skewness
Input	Original	annual inc		27.8173
Input	Original	dti		0.213386
Input	Original	revol bal		12.86989
Input	Original	tot coll amt		336.3033
Input	Original	total rev hi lim		7.793197
Output	Computed	LG10 annual inc	log10(annual inc ...	0.173422
Output	Computed	LG10 dti	log10(dti + 1)	-1.29266
Output	Computed	LG10 revol bal	log10(revol bal +...	-3.08175
Output	Computed	LG10 tot coll amt	log10(tot coll amt...	2.714481
Output	Computed	LG10 total rev h...	log10(total rev hi...	-1.02543

Fig: 5.6 – Transformation to reduce skewness

Two variables had missing values of 1% and 8% and they were imputed with the median value (replacing by mean did not change the misclassification rate of the model).

Variable Name	Impute Method	Impute Value
revol util	MEDIAN	55.81
tot cur bal	MEDIAN	801241

Fig: 5.7 – Imputing the missing values

### Variable selection using Variable Selection Node:

We used the variable selection node to check for the variables that SAS EM would select for the model based on their correlation to the target variable. We ran a regression model after running the variable selection node on default settings.

## 5.3 Models for target variable will\_default

We used decision tree and logistic regression to predict our target variable. Our models utilized different inputs:

- With/without reducing skewness
- With/without imputing for missing values
- Categorizing continuous and nominal variables

## 5.4 Model Comparison

We used the model comparison node to assess our models. Below is the output of the model comparison node:

Selected Model	Model Node	Model Description	Target Variable	Selection Criterion: Valid: Misclassification Rate	Train: Average Squared Error	Train: Average Error Function	Valid: Mean Square Error
Y	Req3	Regression (3)	will default	0.182134	0.139975	0.445361	0.140177
	Req5	Regression (5)	will default	0.182173	0.139919	0.445124	0.140059
	Req2	Regression (2)	will default	0.182232	0.140199	0.446147	0.140394
	Tree2	Decision Tree (2)	will default	0.182812	0.149384		
	Tree4	Decision Tree (4)	will default	0.182812	0.149384		
	Req	Regression	will default	0.183068	0.139568	0.442921	0.139673

Fig: 5.8 – Fit statistics for the models for target variable will\_default

Below are the ROC curves for the 6 models that we tried.

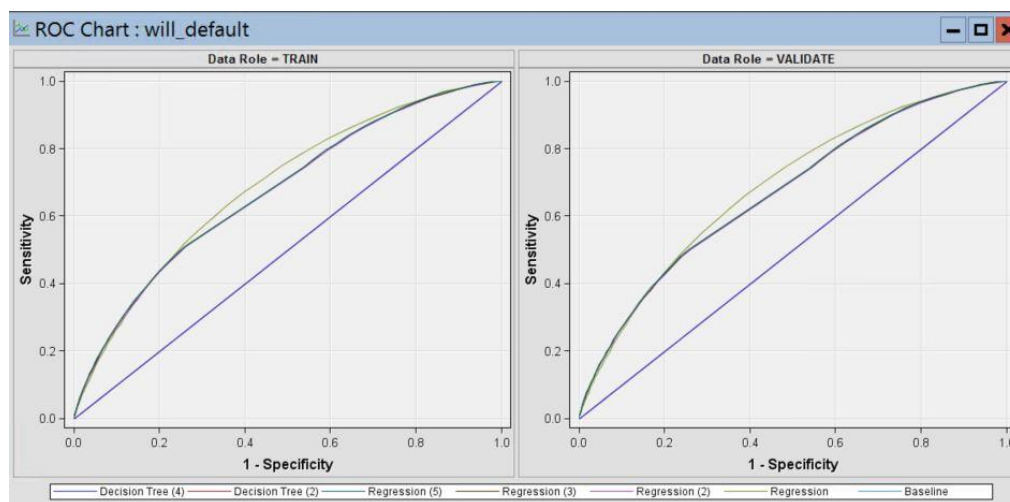


Fig: 5.9 – ROC curves for the models for target variable will\_default

The cumulative lift chart for the 6 models is below:

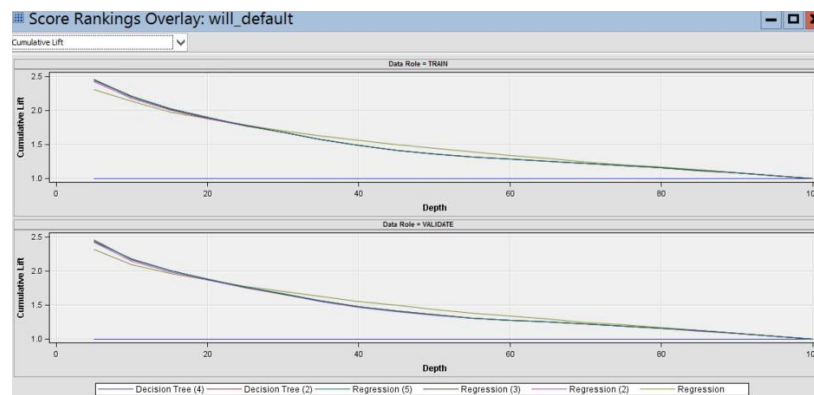


Fig: 5.10 – Cumulative lift chart for the models for target variable will\_default

### Best Model – Regression (3):

SAS EM chose Regression (3) as the best model for this data. This regression is the one that was used after categorizing some continuous and nominal variables, reducing skewness using Log 10 transformation, and imputing missing values as described in the preprocessing section.

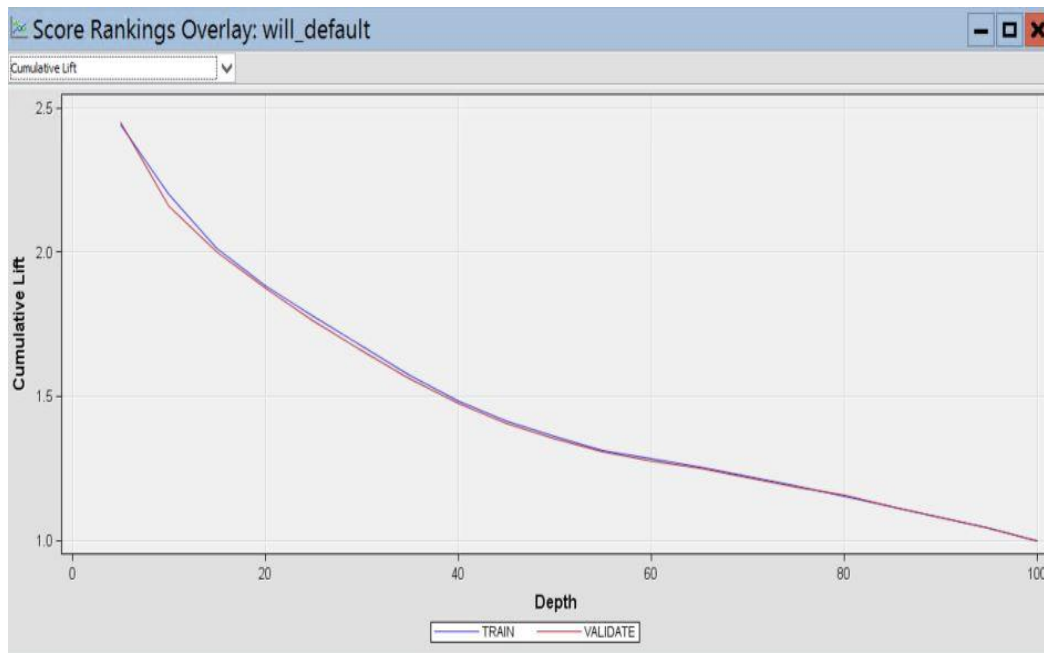


Fig: 5.11 – Cumulative lift chart for the regression (3) model for target variable will\_default

Target	Fit Statistics	Statistics Label	Train	Validation
will default	AIC	Akaike's Information Criterion	135952.9	
will default	ASE	Average Squared Error	0.139975	0.140177
will default	AVERR	Average Error Function	0.445361	0.445921
will default	DFE	Degrees of Freedom for Error	152457	
will default	DFM	Model Degrees of Freedom	54	
will default	DFT	Total Degrees of Freedom	152511	
will default	DIV	Divisor for ASE	305022	203351
will default	ERR	Error Function	135844.9	90680.6
will default	FPE	Final Prediction Error	0.140074	
will default	MAX	Maximum Absolute Error	0.988366	0.9931
will default	MSE	Mean Square Error	0.140024	0.140177
will default	NOBS	Sum of Frequencies	152511	101671
will default	NW	Number of Estimate Weights	54	
will default	RASE	Root Average Sum of Squares	0.374132	0.374401
will default	RFPE	Root Final Prediction Error	0.374264	
will default	RMSE	Root Mean Squared Error	0.374198	0.374401
will default	SBC	Schwarz's Bayesian Criterion	136489.4	
will default	SSE	Sum of Squared Errors	42695.31	28505.71
will default	SUMW	Sum of Case Weights Times Freq	305022	203351
will default	MISC	Misclassification Rate	0.182216	0.182131

Fig: 5.12 – Fit Statistics for the regression (3) model for target variable will\_default

## Confusion Matrix:

Actual - Will_Default			
Predicted - Will_Default		0 – [No]	1 – [Yes]
	0 – [No]	82819 (True -)	18248 (False -)
	1 – [Yes]	271 (False +)	340 (True +)

Fig: 5.13 – Confusion matrix for the regression (3) model for target variable will\_default

## 5.5 Business implication

When predicting whether the loan will be defaulted on or not, the more expensive error is the false negative. In this case, the model predicts a 0 i.e. loan will be fully paid when actually it will be defaulted on. This will lead to the investor losing both the principal and the profit (interest).

In the false positive error, the model predicts 1 i.e. loan will be defaulted on when it will be fully paid. In this case, the investors lose out on the profit that they would have made by funding the loan.

## 5.6 Conclusion

This model was designed to help investors decide which loans to invest in. Our best model returned a misclassification rate of 18.22%. This business model helps in the decision making of whether to invest in a loan or not.

## 6. BI Model – Will\_Default – Tracker

We tried to create a model that can be used as a tracker by investors to predict if their investment – a funded loan – will be fully repaid or defaulted on. For this reason, we have included previously rejected variables such as last\_pymnt (last payment), tot\_pymnt (total payment), tot\_received\_prncp (total received principal) and tot\_received\_int (total received interest). This model can be used by investors to track progress and outcome of their loan and plan their own finances accordingly. For instance, if they are relying on the profit from funding the loan and they can predict that the loan will be defaulted on, they can make alternative arrangements for the funds.

### 6.1 Model Comparison – Will\_Default Tracker:

We used the model comparison node to assess our models. Below is the output of the model comparison node:

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Selection Criterion: Valid: Misclassification Rate	Valid: Average Squared Error	Valid: Mean Square Error ▲	Valid: Root Mean Square Error	Train: Misclassification Rate	Train: Average Squared Error	Train: Root Mean Squared Error
Y	Tree2	Tree2	Decision Tree (2)	will default	0.031492	0.0251...			0.0303...	0.0241...	
	Tree4	Tree4	Decision Tree (4)	will default	0.053276	0.0425...			0.0515...	0.0415...	
	Req3	Req3	Regression (3)	will default	0.063288	0.0515...	0.0515...	0.2270...	0.0611...	0.0502...	0.2242...
	Req2	Req2	Regression (2)	will default	0.063347	0.0516...	0.0516...	0.2272...	0.06134	0.0503...	0.2243...
	Req	Req	Regression	will default	0.089803	0.0620...	0.0620...	0.24906	0.0880...	0.0609...	0.2468...
	Req5	Req5	Regression (5)	will default	0.106355	0.0721...	0.0721...	0.2686...	0.1034...	0.0711...	0.2667...

Fig: 6.1 – Fit statistics for the models for target variable will\_default tracker

SAS EM chose the decision tree (2) as the best model for this dataset based on the Misclassification rate of the Validation dataset. This decision tree was the model that used the inputs after categorizing the continuous and nominal variables as mentioned in the preprocessing step.

Below are the ROC curves for the 6 models that we tried. The ROC curves show that the decision tree (2) closely follows the left and top border and is the best model for this dataset.

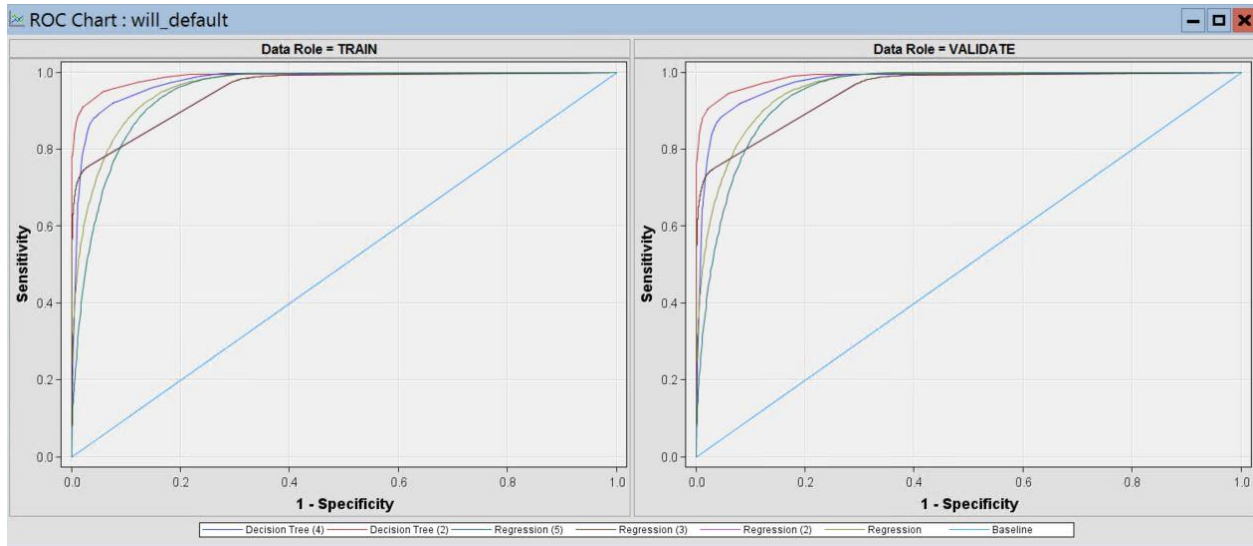


Fig: 6.2 – ROC curves for the models for target variable will\_default tracker

The cumulative lift chart for the 6 models that we tried is below:

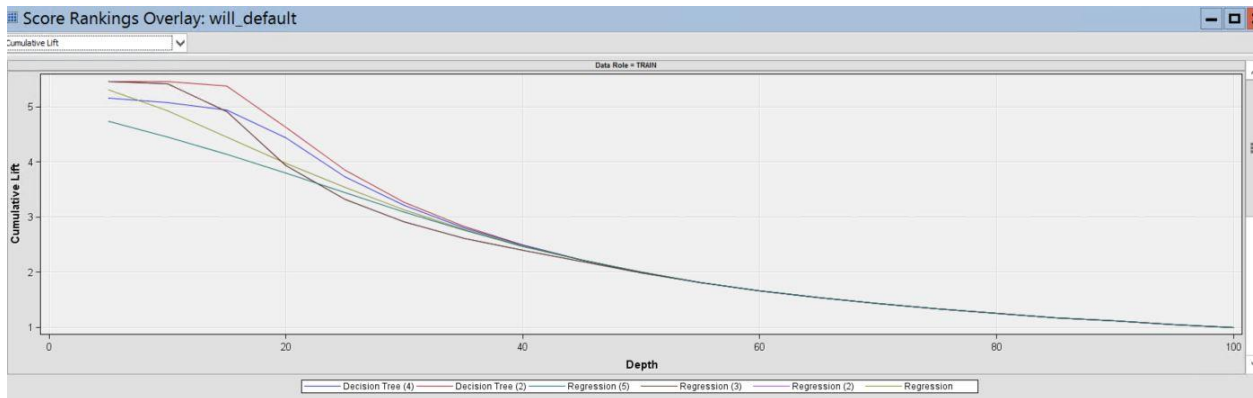


Fig: 6.3 – Cumulative lift chart for the models for target variable will\_default tracker

## Best Model – Decision Tree (2)

Cumulative Lift Chart and Fit Statistics:

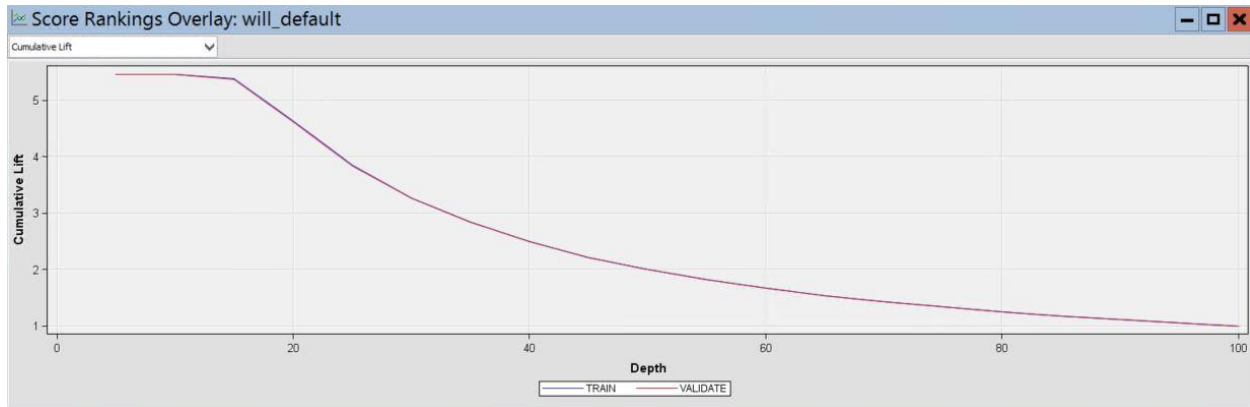


Fig: 6.4 – Cumulative lift chart for the decision tree (2) for target variable will\_default tracker

Target	Fit Statistics	Statistics Label	Train	Validation
will default	NOBS	Sum of Frequencies	152511	101678
will default	MISC	Misclassification Rate	0.030358	0.031492
will default	MAX	Maximum Absolute Error	0.999363	0.999363
will default	SSE	Sum of Squared Errors	7355.488	5111.47
will default	ASE	Average Squared Error	0.024115	0.025136
will default	RASE	Root Average Squared Error	0.155289	0.158542
will default	DIV	Divisor for ASE	305022	203356
will default	DFT	Total Degrees of Freedom	152511	

Fig: 6.5 – Fit Statistics for the decision tree (2) for target variable will\_default tracker

## Confusion Matrix:

		Actual	
		0 – [No]	1 – [Yes]
Predicted	0 – [No]	82137 (True -)	2249 (False -)
	1 – [Yes]	953 (False +)	16339 (True +)

Fig: 6.6 – Confusion matrix for the decision tree (2) for target variable will\_default tracker

## 6.2 Business implication

When using the tracker model, the more expensive error is the false negative. In the false negative error, the model predicts 0. loan will be fully paid when actually it will be defaulted on. This will lead to the investor losing both the principal and the profit (interest).

## 6.3 Conclusion

This model was designed to help investors track the progress and outcome of the loans to invest in. Our best model returned a misclassification rate of 3.03%.

## 7. BI Model: Target Variable –Loan Interest Rate Category

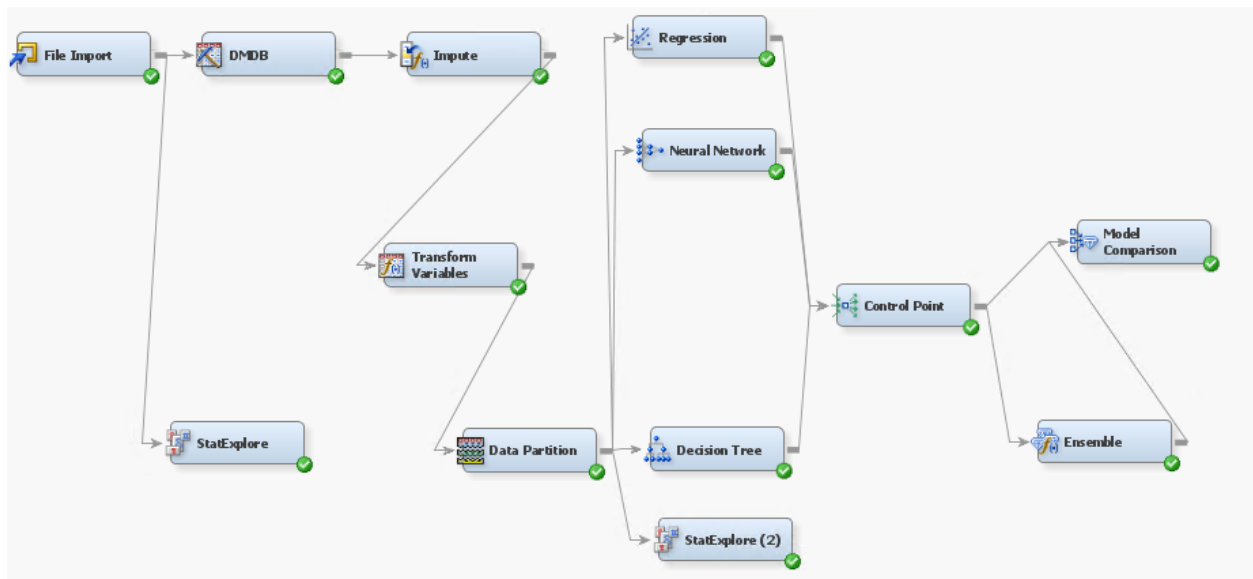


Fig: 7.1 – Process flow diagram for binary target variable Loan Interest Rate Category

The loan data is from 2007-2011 collected from lending club website having Fico score as one of the main variable having 42535 observations. We will be predicting if the person will get low or high interest rates.

### 7.1 DATA PRE-PROCESSING

We have removed the variables having high missing values. From the data we have found the median interest rate is 11.99%. The target variable is a binary variable where interest rate



above 11.99% is categorized as high: predicted output 0 and below 11.99% as low interest rate: predicted output =1.

We then imputed the missing values using “Impute” node and transformed the values for removing the skewness using “transform” node. We rejected variables like interest rate, subgrade, out\_prncp, installment, total\_rec\_int etc as these were the part of output and some are like other variables. Then the data was partitioned as training (70%) and validation (30%).

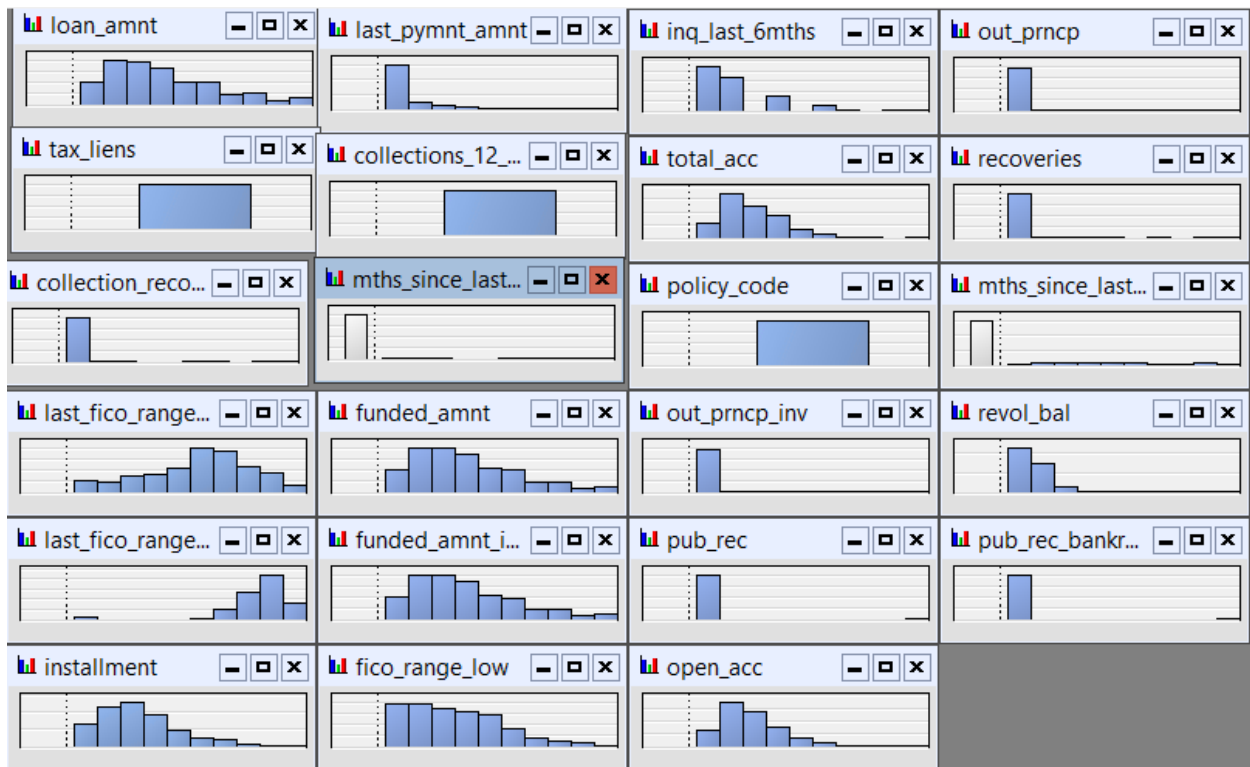


Fig: 7.2 – Variables before imputing for skewness

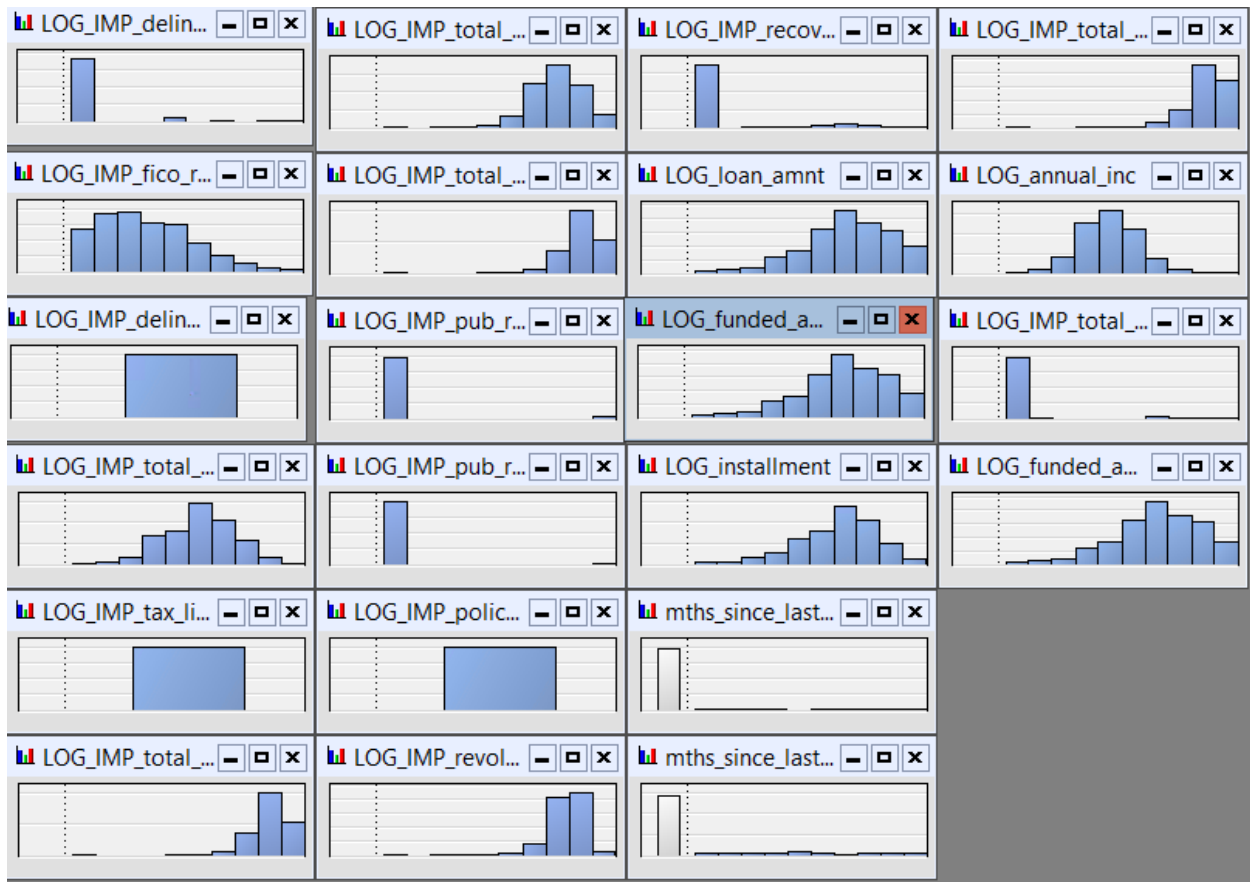


Fig: 7.3 – Variables after imputing for skewness

## 7.2 Models for target variable Loan Interest Rate Category

We have used Regression Model, Decision Trees, Neural network and Ensemble model

**Regression Model** – We have performed stepwise regression.

Fit Statistics	Statistics Label	Train	Validation
_AIC_	Akaike's Information Criterion	3378.543	.
_ASE_	Average Squared Error	0.025509	0.02672
_AVERR_	Average Error Function	0.09125	0.099119
_DFE_	Degrees of Freedom for Error	18393	.
_DFM_	Model Degrees of Freedom	10	.
_DFT_	Total Degrees of Freedom	18403	.
_DIV_	Divisor for ASE	36806	21040
_ERR_	Error Function	3358.543	2085.465
_FPE_	Final Prediction Error	0.025536	.
_MAX_	Maximum Absolute Error	0.999778	1
_MSE_	Mean Square Error	0.025523	0.02672
_NOBS_	Sum of Frequencies	18403	10520
_NW_	Number of Estimate Weights	10	.
_RASE_	Root Average Sum of Squares	0.159715	0.163463
_RFPE_	Root Final Prediction Error	0.159801	.
_RMSE_	Root Mean Squared Error	0.159758	0.163463
_SBC_	Schwarz's Bayesian Criterion	3456.746	.
_SSE_	Sum of Squared Errors	938.8748	562.1938
_SUMW_	Sum of Case Weights Times Freq	36806	21040
_MISC_	Misclassification Rate	0.031191	0.032985

Fig: 7.4 – Fit Statistics of Regression model for binary target variable Loan Interest Rate Category

Misclassification Rate of 3.3% is achieved for validation data.

**Decision Trees-** Assessment measure chosen for the stopping point of tree is misclassification rate with maximum depth of 10.

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
grade		2	1.0000	1.0000	1.0
LOG_IMP_fico_range_high	Transformed: Imputed fico_rang...	5	0.0991	0.1075	1.0
LOG_funded_amnt_inv	Transformed funded_amnt_inv	4	0.0781	0.0706	0.9
LOG_IMP_inq_last_6mths	Transformed: Imputed inq_last_...	2	0.0412	0.0360	0.8
term		1	0.0321	0.0261	0.8

Fig: 7.5 – Variable importance for binary target variable Loan Interest Rate Category

We found that important variables in predicting the interest category are grade, Loan amount invested, inquiry in last 4 months, Fico range, term.

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
Target		_NOBS_	Sum of Frequencies	18403	10520
Target		_MISC_	Misclassification Rate	0.028963	0.032224
Target		_MAX_	Maximum Absolute Error	0.999752	0.999752
Target		_SSE_	Sum of Squared Errors	904.8741	554.1258
Target		_ASE_	Average Squared Error	0.024585	0.026337
Target		_RASE_	Root Average Squared Error	0.156796	0.162286
Target		_DIV_	Divisor for ASE	36806	21040
Target		_DFT_	Total Degrees of Freedom	18403	.

Fig: 7.6 – Decision Tree Fit Statistics

Misclassification rate achieved by decision tree model is 3.2%

## 7.3 MODEL COMPARISON

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate
Y	Tree	Tree	Decision Tr...	Target		0.032224
	Ensmbl	Ensmbl	Ensemble	Target		0.034125
	Reg	Reg	Regression	Target		0.034125
	Neural	Neural	Neural Net...	Target		0.036407

Fig: 7.7 – Model Comparison Fit Statistics

Decision tree is chosen to be the best model.

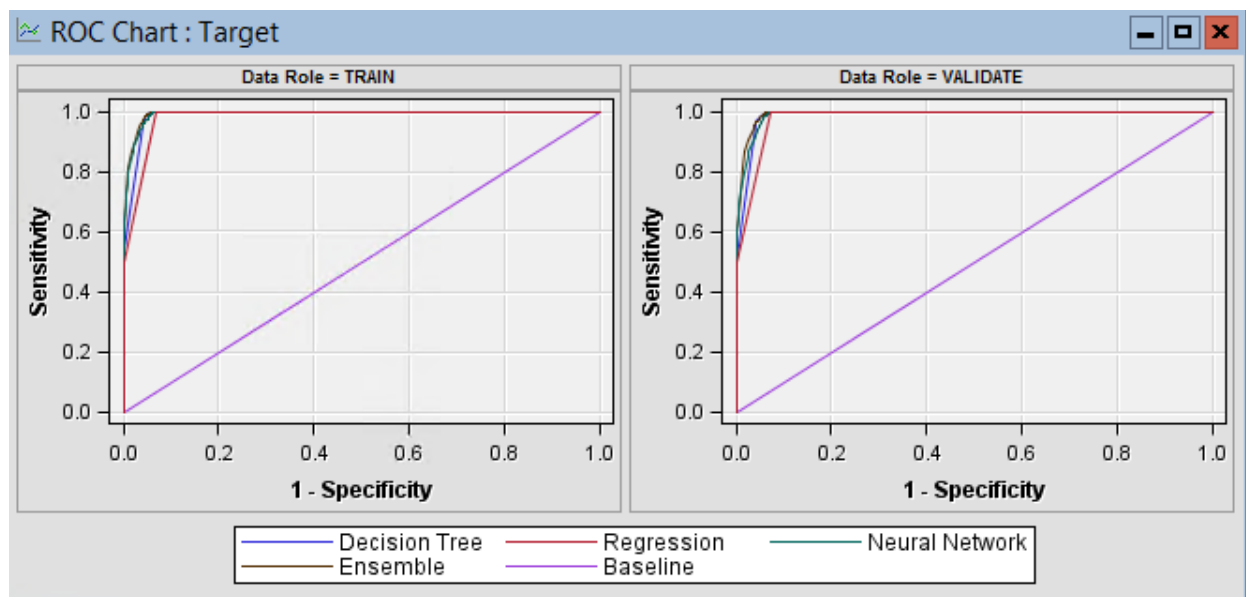


Fig: 7.8 – ROC curves for the models for target variable Loan Interest Rate Category

## Confusion Matrix:

		Actual	
Predicted		0 – [No]	1 – [Yes]
	0 – [No]	4672 (True -)	59 (False -)
	1 – [Yes]	280 (False +)	5509 (True +)

Fig: 7.9 – Confusion matrix for the decision tree for target variable Loan Interest Rate Category

Sensitivity =  $5509 / (5509 + 59) = 98.94\%$

Specificity =  $4672 / (4672 + 280) = 94.35\%$

This indicates that decision tree measures the proportion of positives that are correctly identified (98.94%) and the proportion of negatives that are correctly identified (94.35%)

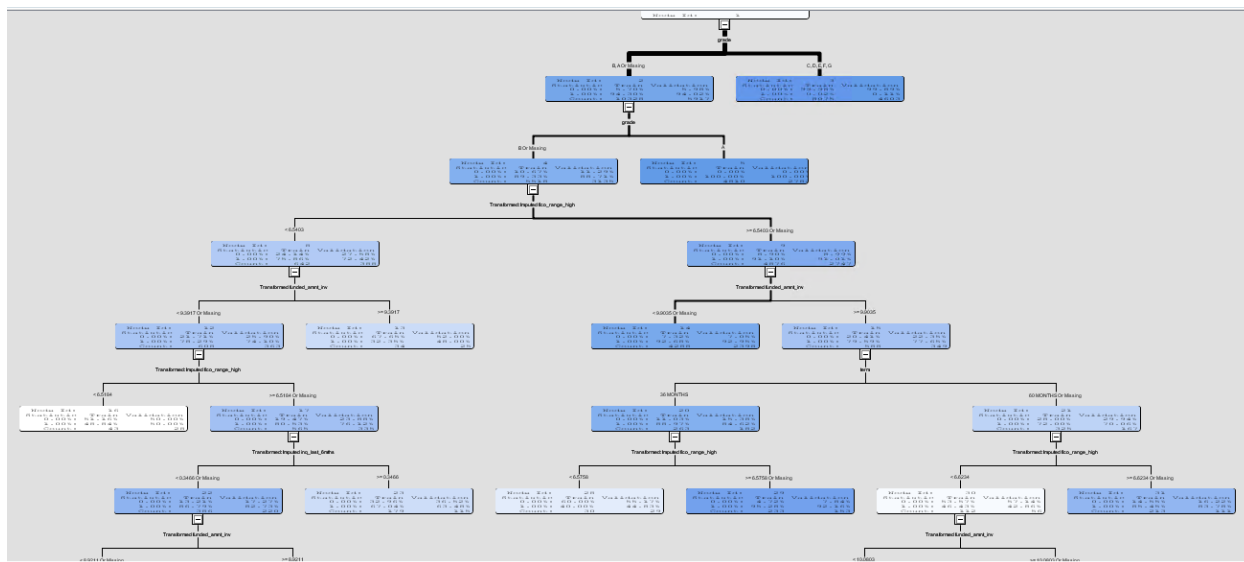


Fig: 7.9 – decision tree for target variable Loan Interest Rate Category

The first split occurs for the variable “grade” if it’s A, B or C, D,E,F. The second split occurs for the funded amount by investors. Next the model splits on basis Current fico score followed by tenure of the loan.

### 7.3 Conclusion

It is advisable for any borrower to maintain high upper credit score in the current credit check in order receive lower interest rates. The loans with grade A or B and with more tenure tend to have lower interest rate. The amount of loan doesn't play a significant part, but amount funded by investors does. Also, significant variable in predicting the high or low interest rate is enquires done in last 4 months.

### 7.4 Future Work

Can be enhanced bringing in macroeconomics like inflation rates, global liquidity, uncertainty for better analysis.

## 8. References

- [www.kaggle.com](http://www.kaggle.com)
- [www.lendingclub.com](http://www.lendingclub.com)
- <http://support.sas.com/kb/20/054.html>

## 9. Acknowledgment

We would like to express our thanks to Prof.Radha Mookherjee for advising and helping us in getting acquainted with various SAS Enterprise Miner concepts and data mining techniques. We thank our university and Prof. Mark Thouin for giving us this opportunity to participate in this global competition and show our analytics skills.