SESUG 2022 Paper 193

To Lend, or Not to Lend, That Is the Question! Evaluating Loan Application with Machine Learning

Rafae Abdullah, Oklahoma State University; Sucharitha Vallabhaneni, Oklahoma State University; Jaideep Muley, Walmart Global Tech

ABSTRACT

Have you ever lent money to someone? If yes, you must know how much headache lenders go through when they do not get their money back within the expected time. This is true for any financial institution. The goal of this paper is to apply various machine learning models by using SAS® Enterprise Miner in the loan lending process of a peer-to-peer lending company and result in the best approach that can be used by peer-to-peer lending companies to automate the loan acceptance process and by other financial institutions to build a more robust model for them since their amount of lend and amount of information on borrowers is huge. Algorithms such as Decision Tree, Random Forest, Neural Network, and so on were used to find the best alternative to ease the loan lending process. The data set used contains plenty of information about the accounts of the clients and the loans funded in the period from 2007 to 2018. We started our analysis with looking at statistical summary and visualizations of the variables, data cleaning and pre-processing, missing value imputation, and finally model building and evaluation. From our analysis, we have found that the Random Forest model outperforms the others. Based on the outcome, we hope that financial institutions can apply Machine Learning to automate their loan approval procedure. Also, to handle, transform, and keep track of around 100 inputs, a tool like SAS® Enterprise Miner is very much useful.

INTRODUCTION

Besides banks, there are some unorthodox types of institutions which gives loan for various purposes. One of such types of institutions is peer to peer lending. Lending Club, situated in San Francisco, is pioneer of this concept. Lending Club was established in 2006. The following picture summarizes this type of business model.

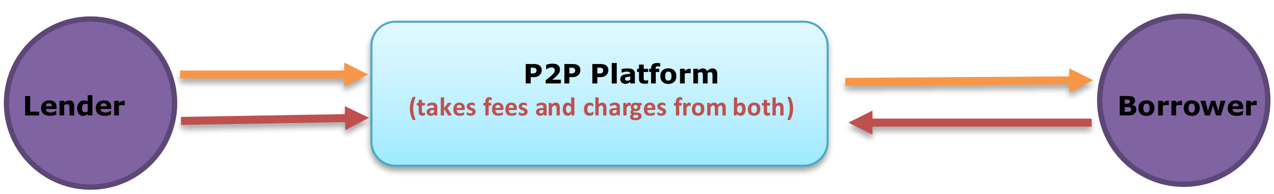


Figure 1. How P2P platforms work

Lenders usually come to know various information about loan applicants through lending club website. We will try to build a predictive model that can predict whether the loanee will be 'good' or 'bad' based they paid fully at least within the grace period.

LITERATURE REVIEW

Researchers took different approaches to find best model to automate loan approval process. Dong, L B Thong (2020)[1] took a different approach. He classifies the loans as "fully paid" and "charged-off" and tried to classify loans according to the amount of profit. He wrote "empirical evidences indicate that investors should prioritize predicting profitable loans instead of just worrying about whether a loan will default since even a fully paid loan can give negative return and vice-versa"; however, we will try to classify the loans as good (fully paid at least within grace period), and bad (not fully paid or paid in delay). We could go for profit/loss calculation, but we have found that profit/loss could not be calculated directly from this data and data dictionary that is available online without making a lot of assumptions. Dong, T had to assume many things to calculate profit generated per account to create target variable as well.

Different researchers drop the records of loans paid fully but in delay, for instance, Teply and Polena (2020)[2]; but we have classified them as good or bad based on length of delay.

DATA OVERVIEW

Source of our dataset is Kaggle[3]. The original dataset contains 2,260,701 records. The file was in csv format and file size was 1.55 GB. Timeframe of the dataset is 2007 to 2018. However, "loan status", which is our target variable, originally contains 9 levels of data including "Current". Here, we cannot create model based on "Current" status as it is neither good nor bad, thus we removed all the records of current loan. We also removed 13 records where "loan status" column has missing values, resulting in a new dataset of 1,382,351 records. The rest of the discussion will be based on this trimmed dataset.

METHODOLOGY

The dataset contains 151 columns. However, as per our understanding, it is not possible to know some of the columns' information at the time of lending. We deleted such 47 columns from the dataset. Also, we deleted two columns which will not be useful for analysis: member_id and url.

Also, columns 'purpose' and 'title' refer to the same thing: purpose of taking the loan. Since the column purpose is clean while title is messy, we dropped title.

Here is an important decision we made. The loan status contain following levels and we converted them to Good loan and Bad loan as shown in the following table:

Old Level	New Level
'Charged Off'	'Bad'
'Default'	'Bad'
'Does not meet the credit policy. Status:Charged Off'	'Bad'
'Late (31-120 days)'	'Bad'
'Late (16-30 days)'	'Bad'
'In Grace Period'	'Good'
'Fully Paid'	'Good'
'Does not meet the credit policy. Status:Fully Paid'	'Good'

Table 1. Table showing the conversion of levels of loan status

It is a business decision which should be treated as 'Good' or 'Bad'. It might happen that a borrower paid fully but made late, but since the p2p platform can earn money, they are still interested in those categories of borrowers. On the other hand, some lenders might want to be risk averse. Whatever decision they make, based on that decision, the model outcome might be changed.

It was a bit challenging to understand the meanings of each variable since we could not find a proper data dictionary. Assuming 'last_fico_range' high/low refers to the value when the data was generated and not when the loan was disbursed, we removed it from the model building process, and kept only 'fico_range' high/low.

Also, we did not consider 'desc', 'earliest credit line' and 'issue_date'. Although earliest credit line might be a good predictor, since FICO score takes that factor into consideration, we did not move forward with it.

EXPLORATORY ANALYSIS

At first, we can look at distribution of good loan and bad loan in dataset. 78% are good loan, making it a slightly unbalanced dataset. Here is the graph from JMP:

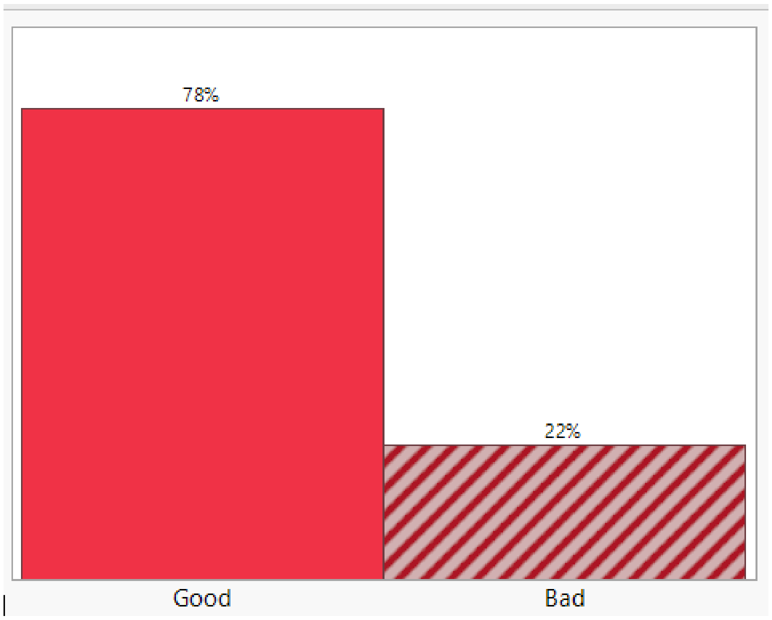


Figure 2. Distribution of good and bad loan in dataset

We looked at the factors which might play influential role in final prediction. Since there are too many variables, we will try to show few of them. These variables can be classified into different categories such as borrower characteristics.

1. Borrower Characteristics

'FICO Score low' indicates an interesting distribution of FICO scores of loanee. While the mean and median are both between 690 and 697, there are hardly any datapoint below 660. From 660, it shows a right distribution.

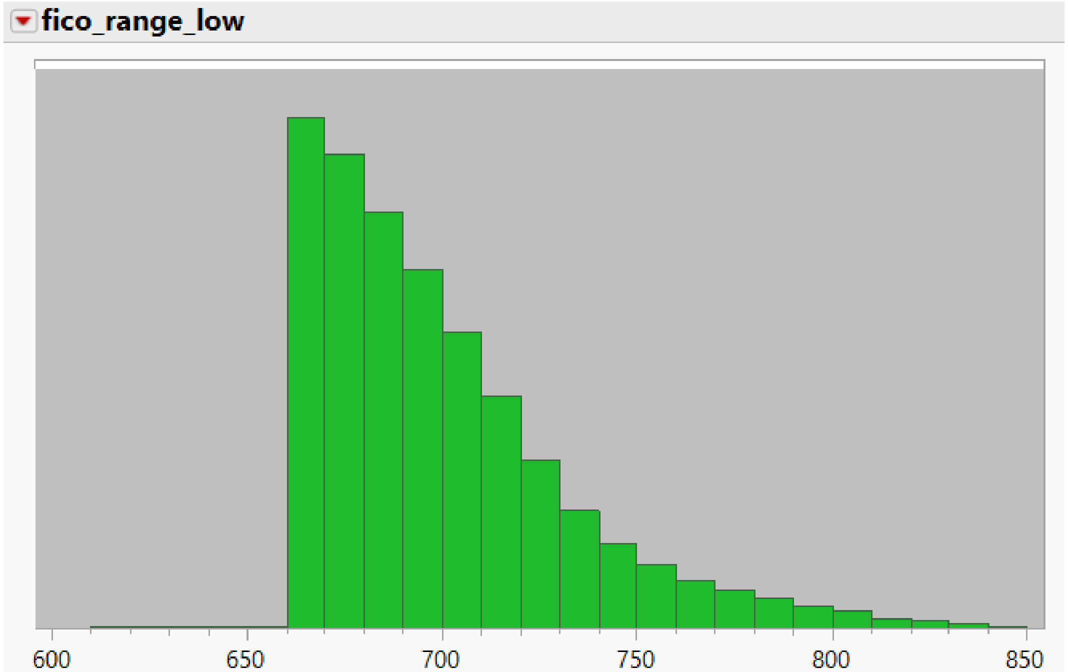


Figure 3. Histogram of FICO range low

Classifying the income of loanees into shows another interesting outcome. While most of the loanee's annual income is between 30,000 and 100,000 USD, people who earn more than this level constitute a surprising 21% of records.

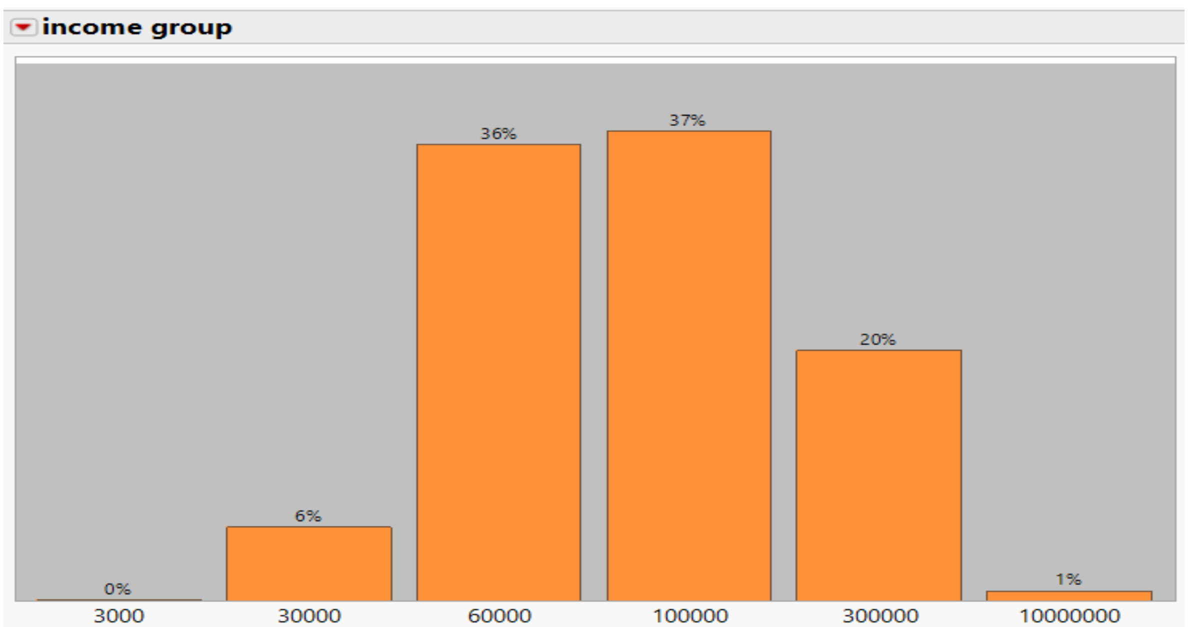


Figure 4. Histogram of different income groups

Relation between annual income and loan status is important, with the increase in annual income, the possibility of being defaulter drops, holding others constant.

Employment tenure of the loanees column was transformed as follows:

For employment tenure of 'more than 10 years' was categorized as 11 and 'less than 1 year' was categorized as 0 to make the datatype numeric. We can see an even (almost uniform) distribution of loanee based on their career timeframe.

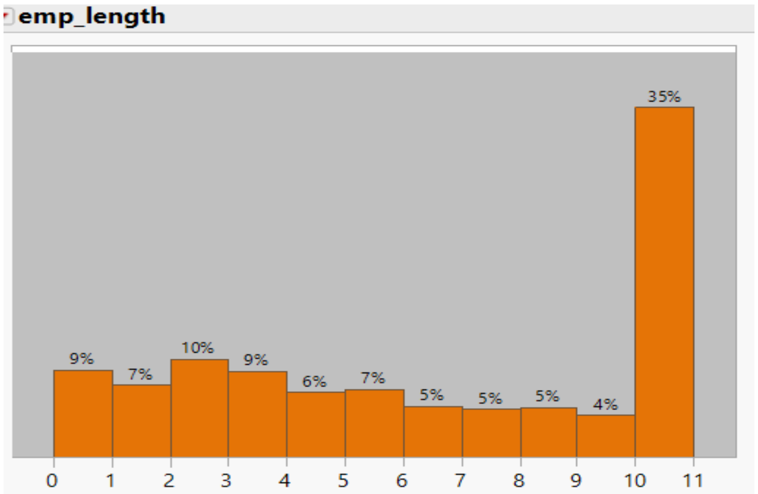


Figure 5. Distribution of length of employment or work experience

It is deceptive that most of them are in service for more than 10 years while if we imagine a career of 30 years, 35% represents 20 years, that means 1.7% per year. We can infer that while the distribution looks even, it skewed to the right very slowly and steadily.

Most of the loanee have mortgaged house. Further investigation might reveal whether they took lean to repay their mortgage. While income group shows that high income people are also taking loan via this platform, home ownership type shows very few (11%) are homeowner. From the following diagram, we can also understand that mortgage, 'own' (i.e. home owner), and 'rent' are 3 main levels of 'home ownership' variable. The others, thus, can be consolidated into one of them.

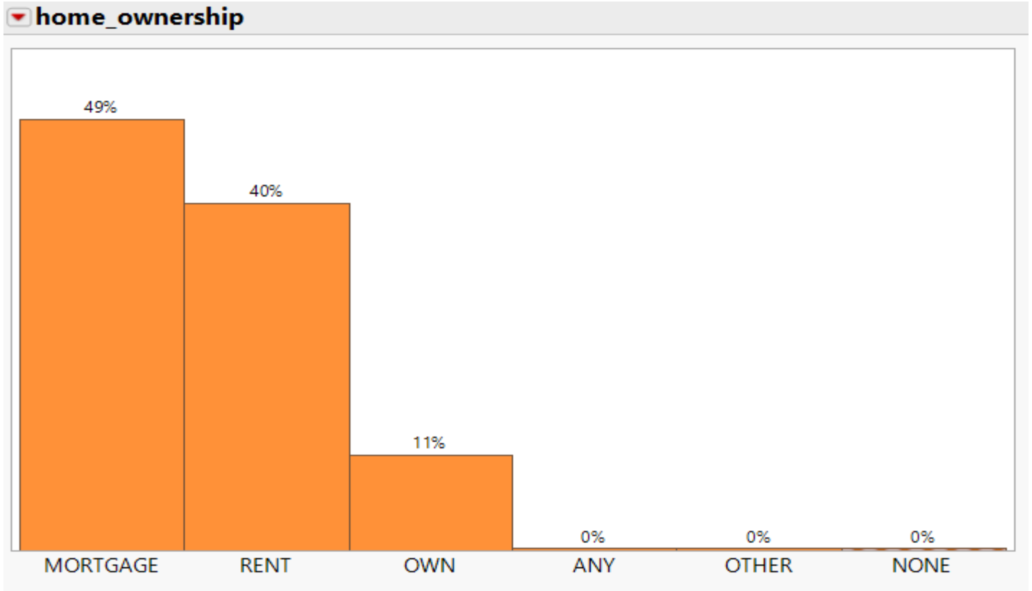


Figure 6. Distribution of types of home ownership

The dataset shows California, New York, Texas, and Florida, these 4 states are address of almost 40% of the loanee. However, we did not consider this variable to build model.

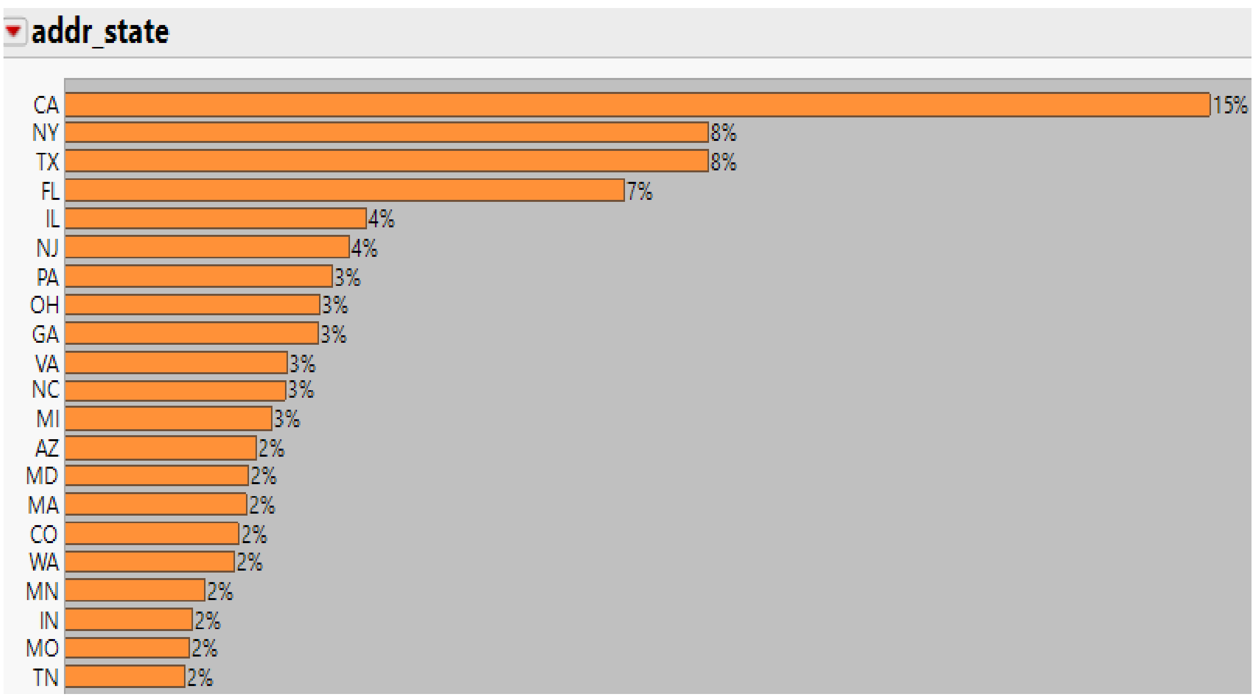


Figure 7. Distribution of client's location

2. Loan Characteristics

While the distribution of purpose shows debt_consolidation and credit_card constitutes 50% of the reasons behind taking loan, digging in 'desc' column, we have found that debt consolidation, in many of the cases, are to consolidate credit card debt.

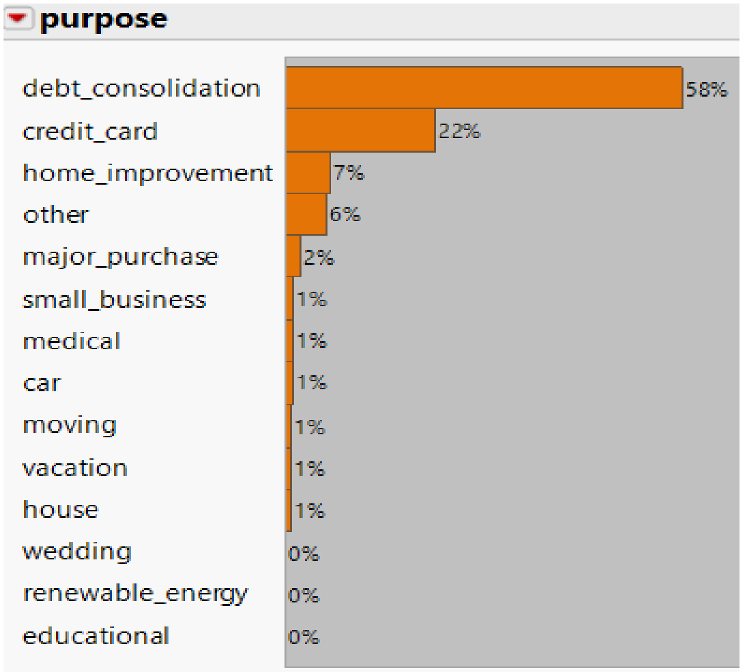


Figure 8. Distribution of purpose of taking loan

Interestingly, percentage of bad loanee among homeowners is slightly higher than that of loanee who had mortgage.

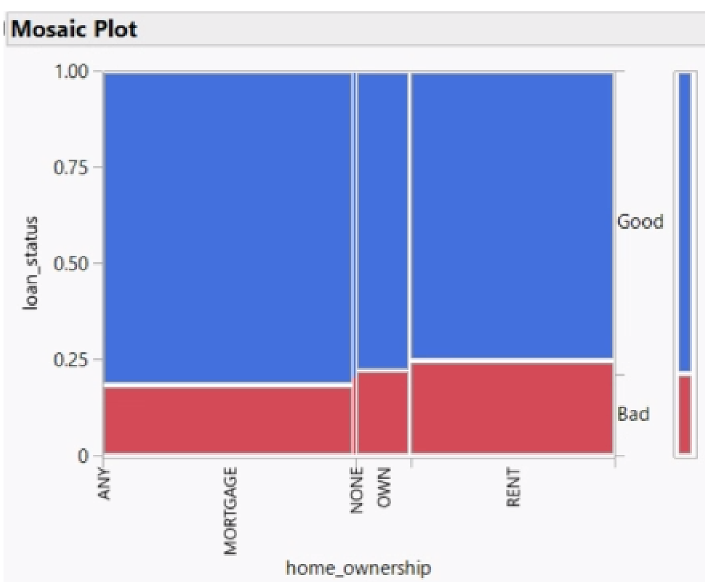


Figure 9. Mosaic plot showing relation between home ownership and loan status

The funded amount by investors shows a slightly right skewed data. While 12,000 USD is the median amount funded, the highest amount is USD 40,000. It should be noted that the minimum value is zero here, which should not be the case, so it requires replacement before creating logistic models.

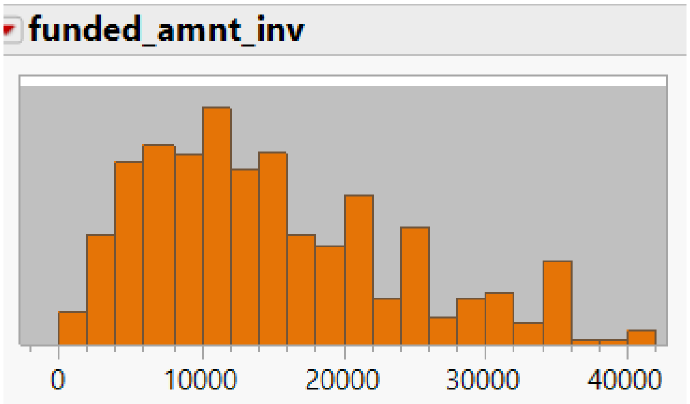


Figure 10. Distribution of the total amount committed by investors for that loan at that point in time.

Interest rates in P2P platforms are usually high. It can be seen from graph that the data is right skewed with median of 12.79%, and maximum of 30.99%

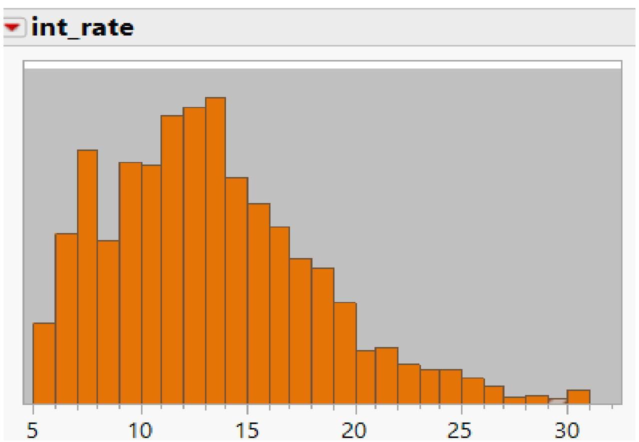


Figure 11. Distribution of interest rate

There are only two available options of term for loanees in Lending Club. 75% choose 3 years option, while the others went for 5 years.

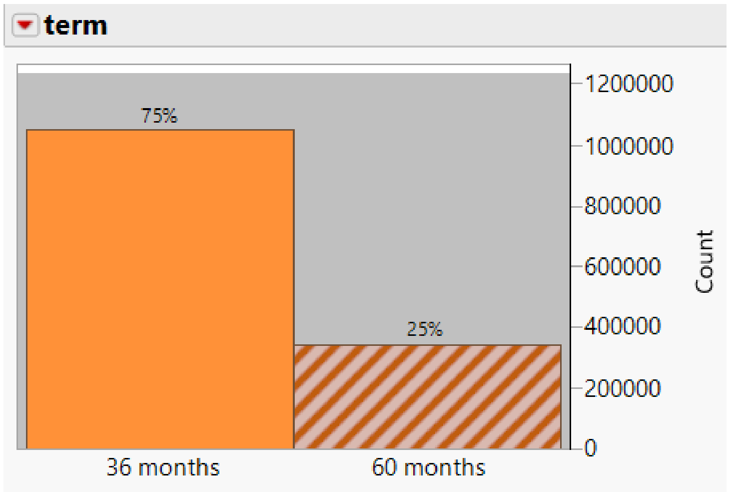


Figure 12. Distribution of term or duration of loan

Both relation between term and loan status, and between interest and loan status shows an expected outcome. For 36 months term, the percentage of bad loan is 17.6%, on the other hand, for 60 months it is almost double (35.2%). Here is mosaic plot to depict the relationship between term and loan status:

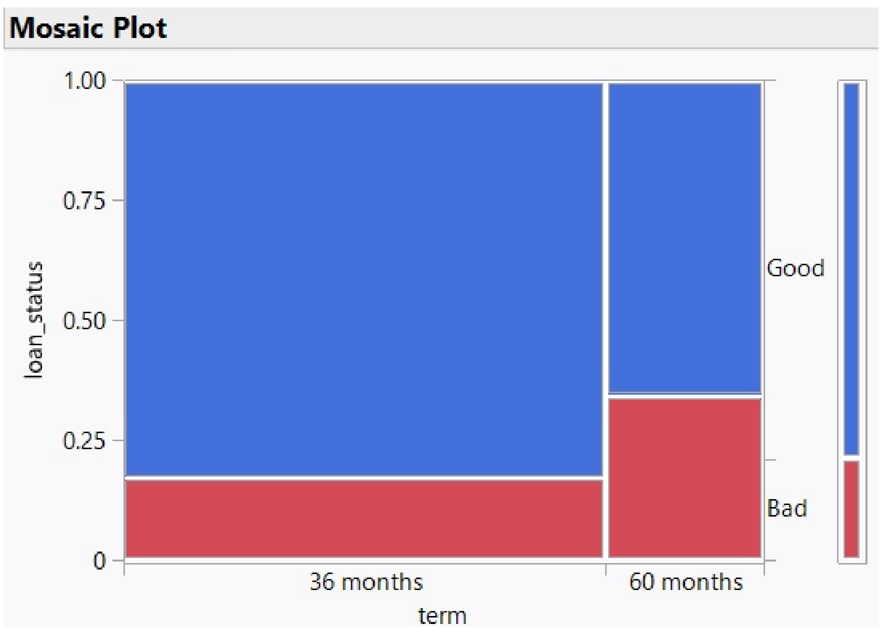


Figure 13. Mosaic plot showing relation between term and loan status

3. Lender's Assessment

Lending Club classifies their loan applicant in grades and subgrades. We can see loanee with grade F, and G also received loan.

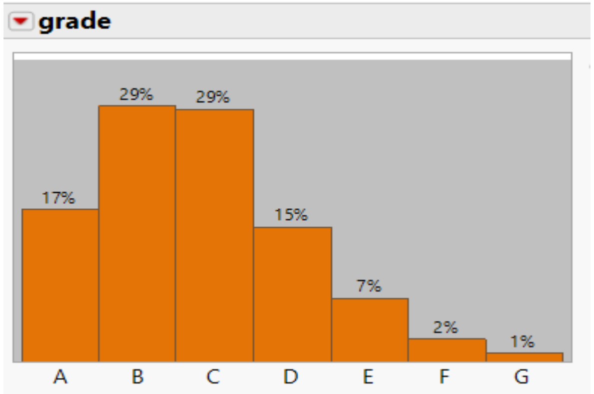


Figure 14. Histogram of grade

Illustration of 35 'subgrade' levels further proofs that B and C are the most used classification.

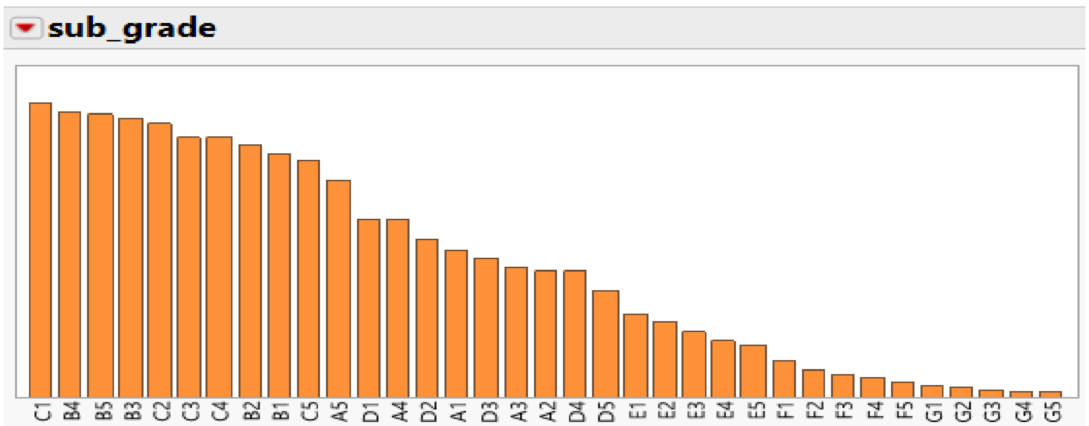


Figure 15. Histogram of subgrade

Grades deserve attention since it is a good predictor of good and bad loan as the following picture depicts:

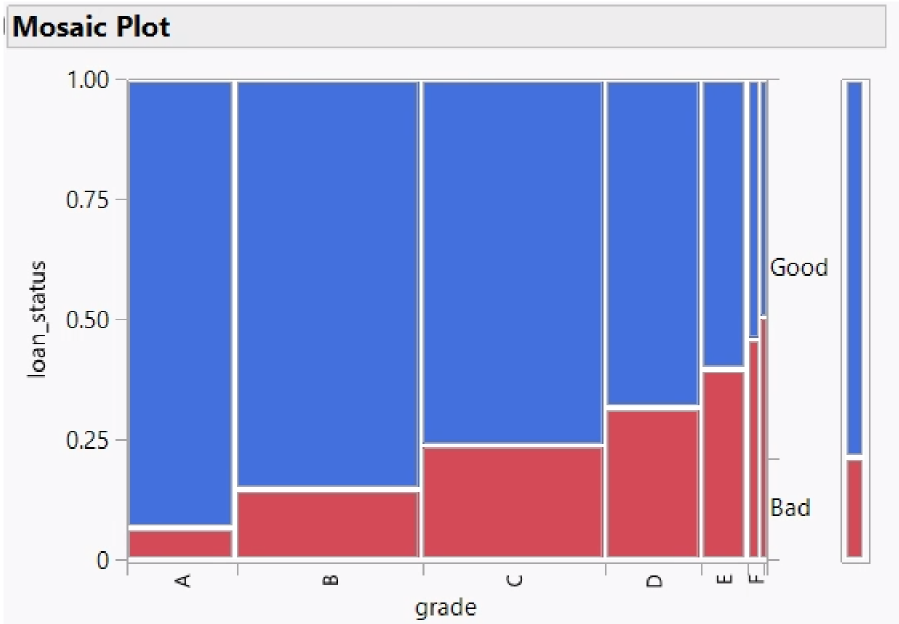


Figure 16. Mosaic plot showing relation between grade and loan status

4. Timeline

The dataset contains loan records which were disbursed between 2007 and 2018. We get fewer number of records for the first and last few years. Since, running loans were filtered out and most of them were disbursed in the last few years, it is normal to get such distribution.

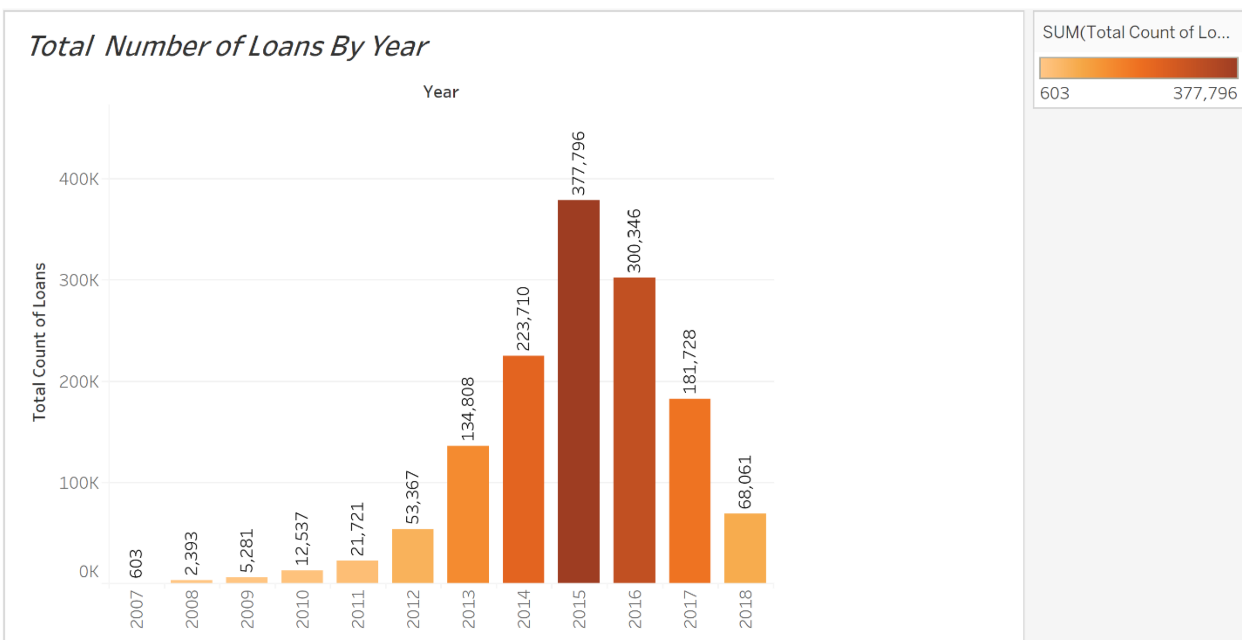


Figure 17. Histogram depicting total number of loans per year

The distribution of percentage of bad loans issued each year is as follows:

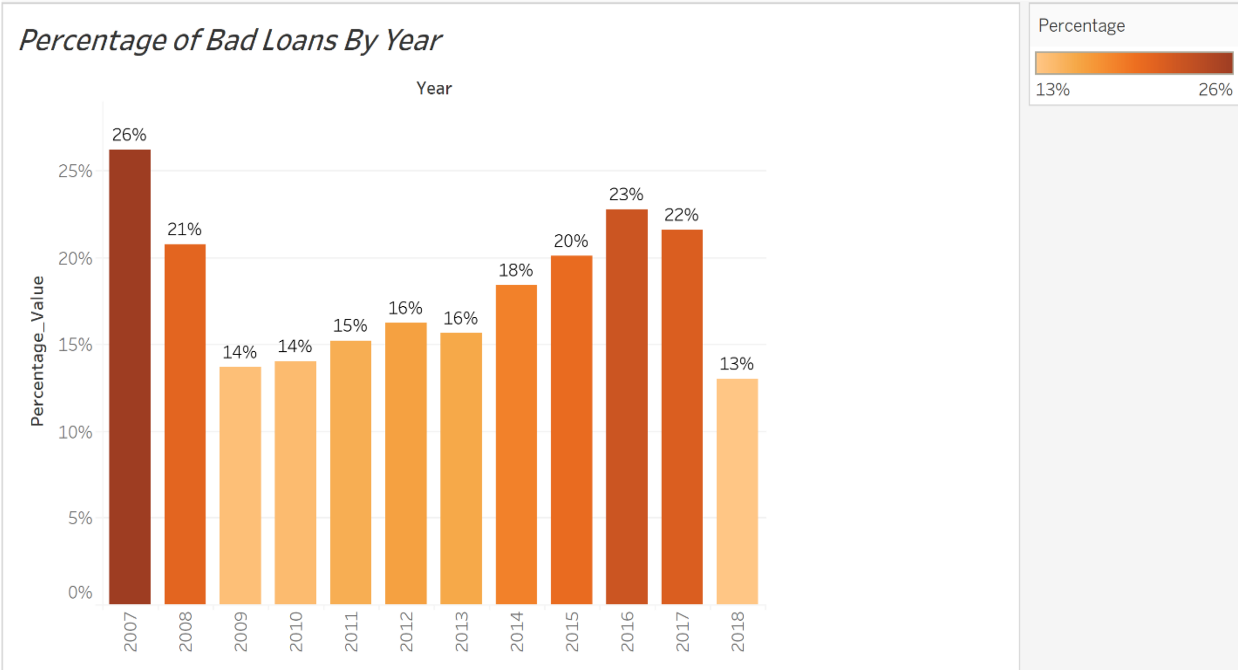


Figure 18. Histogram depicting percentage of bad loans per year

DATA REPLACEMENT:

SAS Enterprise Mines provides an easy way of replacing values of variables via its replacement node. By using it, we have made the following changes.

Since in the home ownership variable, 'any', 'other' and 'none' were rare incident, we replaced them with 'Mortgage', which is the most common phenomenon for this variable.

In case of 'purpose' of taking loan variable, we have replaced house with home improvement since both are similar and there are very few records for house. Also, we replaced moving, vacation, wedding, renewable energy, and educational with another level: other. Here, instead of putting all purposes with low frequency with the level with high frequency, we wanted to understand the purpose of taking loan and bucket them accordingly.

For 'Employee length (emp_length)', we classified the available records into 3 categories. 0-3 years, 4-6 years, and 7-9 years. 81,437 missing values were replaced with unknown.

There were few variables which are related to second applicants. Since such cases are not very common, most of them are null values. Although we did not think that those variables will be good predictors, we tried to keep them to see whether they become important, so we replaced nulls with NA. In case of 'sec_app_charge_off_within_12_mths' column we replaced 0,1,2,3,4,5,6,7,8,9 with 0-9 because 1-9 makes up only around 800 records. We considered number of records for each level during the decision-making process. We took the same strategy for 'sec_app_collections_12_mths_ex_m'. For 'sec_app_inq_last_6_mths', we classified available records into 0-2 and 2+ categories. 'sec_app_mort_acc' variable's available records were classified into 0, 1, and 2-9 categories. For 'sec_app_mths_since_last_major_de', available records were classified into 0-3 and 4-9. 'sec_app_open_acc', 'sec_app_open_acc_il', 'sec_app_num_rev_acc' and 'sec_app_num_rev_acc_il' columns' available records were classified into 0-2 and 3-9.

In case of 'subgrade', G1,G2,G3,G4,G5 were replaced with G since the frequency of these levels were very low. F1, F2 were replaced with F1-F2, and the rest with F3-F5.

For interval variables, few modifications were needed to deal with unrealistic values. For ‘dti’, upper limit was set to 60, and ‘dti_joint’ upper limit was set to 40, since lending club does not provide loan to people whose debt-to-income is more than 60%, and joint-dti is more than 40%.[4] Although the limit might have been changed over years, since the data is right skewed, it will also help us normalize the data.

DATA IMPUTE:

Impute node has been used to deal with the missing values. Tree surrogate method were used to replace the missing values. Here is the summary outcome from impute node.

Imputation Summary

Variable Name	Impute Method	Imputed Variable	Indicator Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
REP dti	TREESURR	IMP REP dti	M REP dti		.INPUT	INTERVAL	Replacement: dti	146
acc now delinq	TREESURR	IMP acc now delinq	M acc now delinq		.INPUT	INTERVAL		7
acc open past 24mths	TREESURR	IMP acc open past 2...	M acc open past 24...		.INPUT	INTERVAL		15387
all util	TREESURR	IMP all util	M all util		.INPUT	INTERVAL		264591
annual inc	TREESURR	IMP annual inc	M annual inc		.INPUT	INTERVAL		1
avg cur bal	TREESURR	IMP avg cur bal	M avg cur bal		.INPUT	INTERVAL		21757
bc open to buy	TREESURR	IMP bc open to buy	M bc open to buy		.INPUT	INTERVAL		20399
bc util	TREESURR	IMP bc util	M bc util		.INPUT	INTERVAL		20658
cr length	TREESURR	IMP cr length	M cr length		.INPUT	INTERVAL		7
delinq 2yrs	TREESURR	IMP delinq 2yrs	M delinq 2yrs		.INPUT	INTERVAL		7
delinq amnt	TREESURR	IMP delinq amnt	M delinq amnt		.INPUT	INTERVAL		7
inq fi	TREESURR	IMP inq fi	M inq fi		.INPUT	INTERVAL		264569
inq last 12m	TREESURR	IMP inq last 12m	M inq last 12m		.INPUT	INTERVAL		264569
inq last 6mths	TREESURR	IMP inq last 6mths	M inq last 6mths		.INPUT	INTERVAL		7
max bal bc	TREESURR	IMP max bal bc	M max bal bc		.INPUT	INTERVAL		264569
mo sin old il acct	TREESURR	IMP mo sin old il acct	M mo sin old il acct		.INPUT	INTERVAL		35661
mo sin old rev tl op	TREESURR	IMP mo sin old rev tl...	M mo sin old rev tl op		.INPUT	INTERVAL		21745
mo sin rcnt rev tl op	TREESURR	IMP mo sin rcnt rev t...	M mo sin rcnt rev tl ...		.INPUT	INTERVAL		21745
mo sin rcnt tl	TREESURR	IMP mo sin rcnt tl	M mo sin rcnt tl		.INPUT	INTERVAL		21744
mort acc	TREESURR	IMP mort acc	M mort acc		.INPUT	INTERVAL		15387
mths since last delinq	TREESURR	IMP mths since last ...	M mths since last del...		.INPUT	INTERVAL		234764
mths since rcnt il	TREESURR	IMP mths since rcnt il	M mths since rcnt il		.INPUT	INTERVAL		270619
mths since recent bc	TREESURR	IMP mths since recen...	M mths since recent ...		.INPUT	INTERVAL		20046
mths since recent inq	TREESURR	IMP mths since recen...	M mths since recent ...		.INPUT	INTERVAL		55976
mths since recent rev...	TREESURR	IMP mths since recen...	M mths since recent ...		.INPUT	INTERVAL		312322
num accts ever 120 pd	TREESURR	IMP num accts ever ...	M num accts ever 12...		.INPUT	INTERVAL		21744
num actv bc tl	TREESURR	IMP num actv bc tl	M num actv bc tl		.INPUT	INTERVAL		21744
num actv rev tl	TREESURR	IMP num actv rev tl	M num actv rev tl		.INPUT	INTERVAL		21744
num bc sats	TREESURR	IMP num bc sats	M num bc sats		.INPUT	INTERVAL		18055
num bc tl	TREESURR	IMP num bc tl	M num bc tl		.INPUT	INTERVAL		21744
num il tl	TREESURR	IMP num il tl	M num il tl		.INPUT	INTERVAL		21744
num op rev tl	TREESURR	IMP num op rev tl	M num op rev tl		.INPUT	INTERVAL		21744
num rev accts	TREESURR	IMP num rev accts	M num rev accts		.INPUT	INTERVAL		21744
num rev tl bal gt 0	TREESURR	IMP num rev tl bal gt...	M num rev tl bal gt 0		.INPUT	INTERVAL		21744
num sats	TREESURR	IMP num sats	M num sats		.INPUT	INTERVAL		18055
open acc	TREESURR	IMP open acc	M open acc		.INPUT	INTERVAL		7
open acc 6m	TREESURR	IMP open acc 6m	M open acc 6m		.INPUT	INTERVAL		264569
open act il	TREESURR	IMP open act il	M open act il		.INPUT	INTERVAL		264569
open il 12m	TREESURR	IMP open il 12m	M open il 12m		.INPUT	INTERVAL		264569
open il 24m	TREESURR	IMP open il 24m	M open il 24m		.INPUT	INTERVAL		264569
open rv 12m	TREESURR	IMP open rv 12m	M open rv 12m		.INPUT	INTERVAL		264569
open rv 24m	TREESURR	IMP open rv 24m	M open rv 24m		.INPUT	INTERVAL		264569
pct tl nvr dlq	TREESURR	IMP pct tl nvr dlq	M pct tl nvr dlq		.INPUT	INTERVAL		21796
percent bc gt 75	TREESURR	IMP percent bc gt 75	M percent bc gt 75		.INPUT	INTERVAL		20533
pub rec	TREESURR	IMP pub rec	M pub rec		.INPUT	INTERVAL		7
pub rec bankruptcies	TREESURR	IMP pub rec bankrupt...	M pub rec bankruptcies		.INPUT	INTERVAL		491
tax liens	TREESURR	IMP tax liens	M tax liens		.INPUT	INTERVAL		26
tot coll amt	TREESURR	IMP tot coll amt	M tot coll amt		.INPUT	INTERVAL		21744
tot cur bal	TREESURR	IMP tot cur bal	M tot cur bal		.INPUT	INTERVAL		21744
tot hi cred lim	TREESURR	IMP tot hi cred lim	M tot hi cred lim		.INPUT	INTERVAL		21744
total acc	TREESURR	IMP total acc	M total acc		.INPUT	INTERVAL		7
total bal ex mort	TREESURR	IMP total bal ex mort	M total bal ex mort		.INPUT	INTERVAL		15387
total bal il	TREESURR	IMP total bal il	M total bal il		.INPUT	INTERVAL		264569
total bc limit	TREESURR	IMP total bc limit	M total bc limit		.INPUT	INTERVAL		15387
total cu tl	TREESURR	IMP total cu tl	M total cu tl		.INPUT	INTERVAL		264569
total il high credit limit	TREESURR	IMP total il high credit...	M total il high credit li...		.INPUT	INTERVAL		21744
total rev hi lim	TREESURR	IMP total rev hi lim	M total rev hi lim		.INPUT	INTERVAL		21744

Display 1. Result of data impute node from SAS Enterprise Miner

DATA TRANSFORMATION:

For interval variables, ‘max. normal’ method was used in ‘Transform Variables’ node. Then each variable was transformed according to the suggestion provided by ‘max. normal’ method.

SAMPLING TECHNIQUE:

Level based options for rare level (‘bad’ in this case) has been applied by using Sample node. In the property, sample method = stratify, type = percentage, criterion = level based, level selection = rarest level was set. Following picture depicts the outcome:

Data=DATA

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
loan_status	.	Bad	295176	21.3532	
loan_status	.	Good	1087175	78.6468	

Data=SAMPLE

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
loan_status	.	Bad	295176	50	
loan_status	.	Good	295176	50	

Display 2. Result of sample node from SAS Enterprise Miner

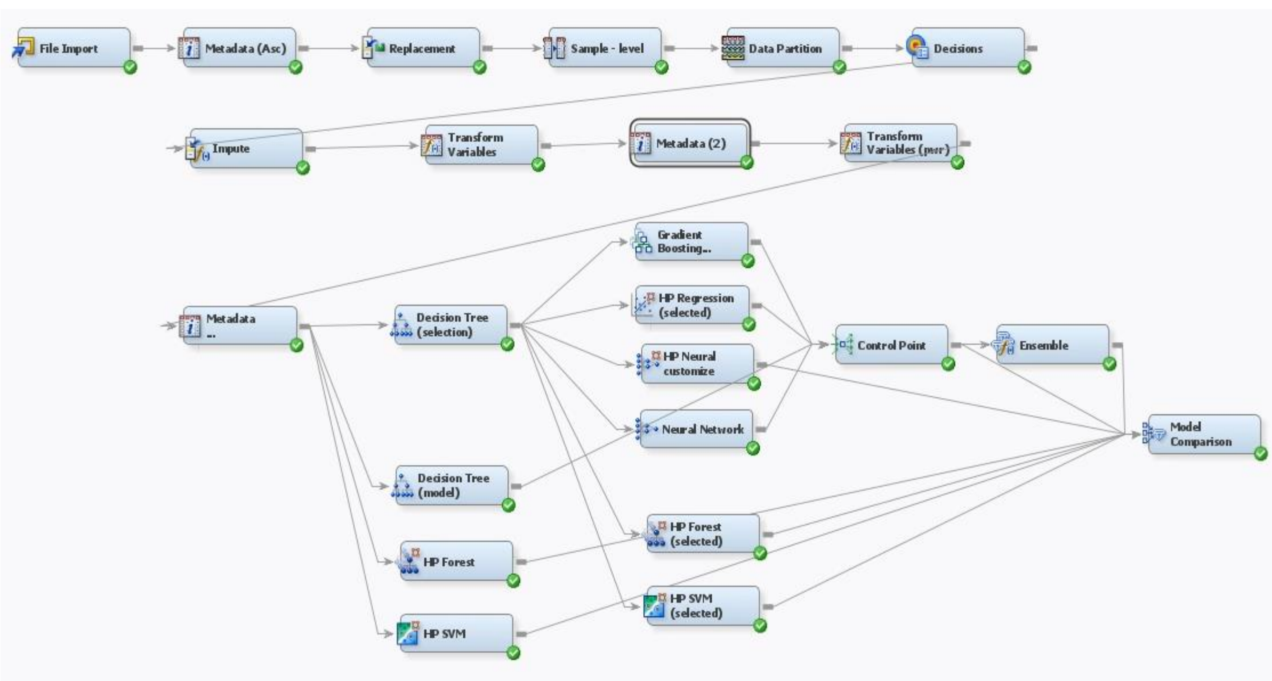
We can see proportion is now same for the two levels of loan status variable.

RESULT

Note: All the models were built with 80:20 training-validation split. Random seed was 12345 in all cases.

For model building, Target 1 = Bad was set.

Following diagram shows SAS Enterprise Miner nodes to build and choose our final model:



Display 3. Nodes and connections from SAS Enterprise Miner

We can see Random Forest comes out as the best model based on validation misclassification rate and F-1 score. In the following table, we also noted ROC values as supporting data.

Model	**Variable Selection Method**	**Validation Misclassification %**	**Validation F-1 score**	**Validation ROC**
(HP) Random Forest	Selected by model	34.19	0.67	0.735
Neural Network *Changed default settings in network: hiddn unit=5, direct = yes, in optimization: max iter 100	Preselected by Decision Tree	34.44	0.66	0.713
(HP) Neural *Changed default settings for number of hidden neurons from 3 to 5	Preselected by Decision Tree	34.47	0.66	0.717
(HP) Random Forest (2)	Preselected by Decision Tree	34.63	0.66	0.723
(HP) Regression	Preselected by Decision Tree	34.67	0.66	0.712
Gradient Boosting	Preselected by Decision Tree	34.84	0.66	0.711
(HP) SVM	Selected by model	36.44	0.65	0.711
(HP) SVM (2)	Preselected by Decision Tree	36.44	0.65	0.704
Decision Tree	Selected by model	36.51	0.67	0.646
Ensemble (Neural Network, HP Regression, Gradient	Preselected by Decision Tree	50.00	Undefined (everything classified	NA

Boosting, Decision Tree)			as bad)	

Table 2. Table showing results of different models

*Customized Neural Network and HP Neural yield better result.

We also tried to customize HP Random Forest by using HP Forest node. But while validation misclassification rate was not dropping that much, training misclassification rate starts dropping noticably (32% from 34%), that means model started to overfit. Hence, we did not show customized HP Forest in the final result.

VARIABLE IMPORTANCE:

According to our champion model HP Random Forest, the top 6 important variables are given below:

1. Interest rate
2. Subgrade
3. Grade
4. Term
5. Inq_fi (Number of personal finance inquiries)
6. DTI (debt-to-income ratio)

CONCLUSION

While the result does not provide a fancy 99%+ or 90%+ accuracy, it gives us a fairly good understanding of what are important predictors. From variable importance, we can get an idea about playing with values of which variable might change the outcome (I,e, bad vs good loan). Organizations like Lending Club can use those to modify a loan proposal to change/ increase the possibility of a 'bad' loan to 'good' one. By doing this, organizations can raise their profit level dramatically.

We also want to shed light on what could be done to improve model accuracy. Here is a list of those:

1. Location might have role in predicting good loanee. Combining states by using various factors (proximity, crime rate, literacy rate, and economic conditions) might become significant. We could not dig deeper into this.
2. Lending Club was established in 2006 and the dataset contains data from 2007. It might take a couple of years for financial organizations to learn know-how, and best practices to attract good clients and drop bad clients as well as default rate. Considering subset of data (such as when issue date is after 2012) might improve model accuracy, future researchers might explore this.
3. We tried to calculate profit/loss for each record, but to do that we needed to make many assumptions. Also, lack of a good data dictionary was an issue in this case. If researchers can get enough information to calculate profit/loss without making many assumptions, they will be able to confidently go for it. It will be more helpful to maximize profit of the organization than model based on misclassification rate.

Moreover, it opens door to future research on checking whether the same determinants can predict similarly for traditional banks' loanee.

REFERENCES

Dong, T. 2020. “A LightGBM model for predicting profitable Lending Club loans.” *DOI: 10.13140/RG.2.2.35406.08000*.

Teply, P. and Polena, M. 2020. “Best classification algorithms in peer-to-peer lending.” The North American Journal of Economics and Finance, 51. *https://doi.org/10.1016/j.najef.2019.01.001*

Nathan George. “All Lending Club loan data.” Accessed March 7, 2022. https://www.kaggle.com/datasets/wordsforthewise/lending-club.

Annie Millerbernd, Ronita Choudhuri-Wade. “LendingClub Personal Loans: 2022 Review.” Accessed June 15, 2022. https://www.nerdwallet.com/reviews/loans/personal-loans/lendingclub-personal-loans.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Rafae Abdullah
Oklahoma State University
405-780-2739
rafae.abdullah@okstate.edu, rafaeabdullah@gmail.com
Sucharitha Vallabhaneni
Oklahoma State University
314-528-1746
sucharitha.vallabhaneni@okstate.edu
Jaideep Muley
Walmart Global Tech
405-780-3832
jaideep.muley@okstate.edu