

Presented by

Final Report

for **the lending club**

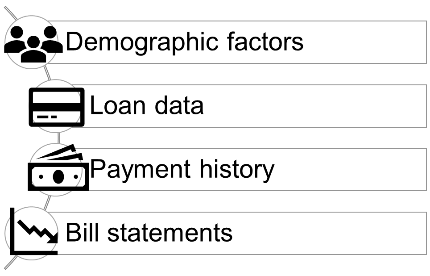
**EXECUTIVE SUMMARY**

Lending Club is a peer to peer U.S. based lending platform. The platform allows borrowers to create unsecured loans of up to $40,000. The platform makes money in regular circumstances by charging lenders a loan initiation fee and charging investors/lenders an account service fee. It also makes more money when a borrow gets troubled on their loan by deducting ~40% of collected money. The investors/lenders on the platform incur major losses when their loans are defaulted on, and thus we aimed at helping them make a better-informed decision when they accept to lend money or buy an existing loan.

**PROBLEM STATEMENT**

This analysis project aims at helping lenders at different stages of the lending process make an informed decision. We aim at providing answers in the below situations:

1. Initiating a loan 🡪 To assess whether the loan applicant is likely to pay the loan or default on it.
2. Purchasing an existing loan 🡪 To assess whether the loan vehicle is likely to default or not.

**DATA DESCRIPTION**

The dataset we used for analysis and prediction is provided by Lending Club and contains observations of accepted loans. The dataset has more than 2 million observations, 151 predictors and data ranging from 2007 till 2013.

**DATA EXPLORATORY ANALYSIS**

***Initial Analysis***

At the beginning of the analysis, we dropped all irrelevant data such as applicant’s member Id, profile URL, user entered job title, etc. We also dropped the predictors which their missing observations were more than 30% of their total. Then we dropped the missing observations that had any missing data. This left us with ~870,000 observations.

***Loan Initiation***

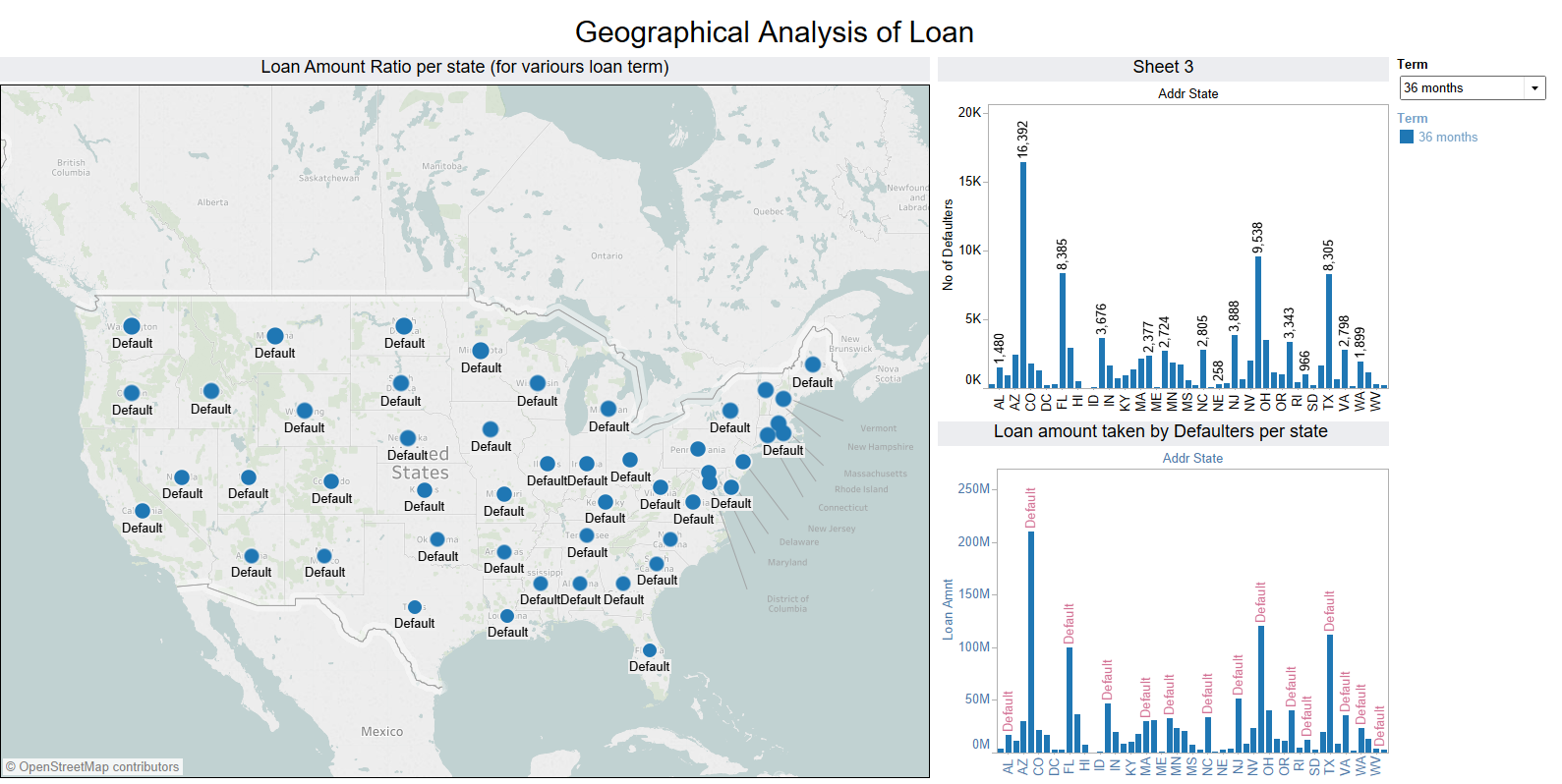
The dataset had observations about loans that were current, paid in full, delayed payments, and defaulted loans. It had data that was not available to lenders at the loan initiation stage, thus we focused only on data that is available to lenders and investors at initiation to be able to build a viable prediction model.

***Purchasing Loan***

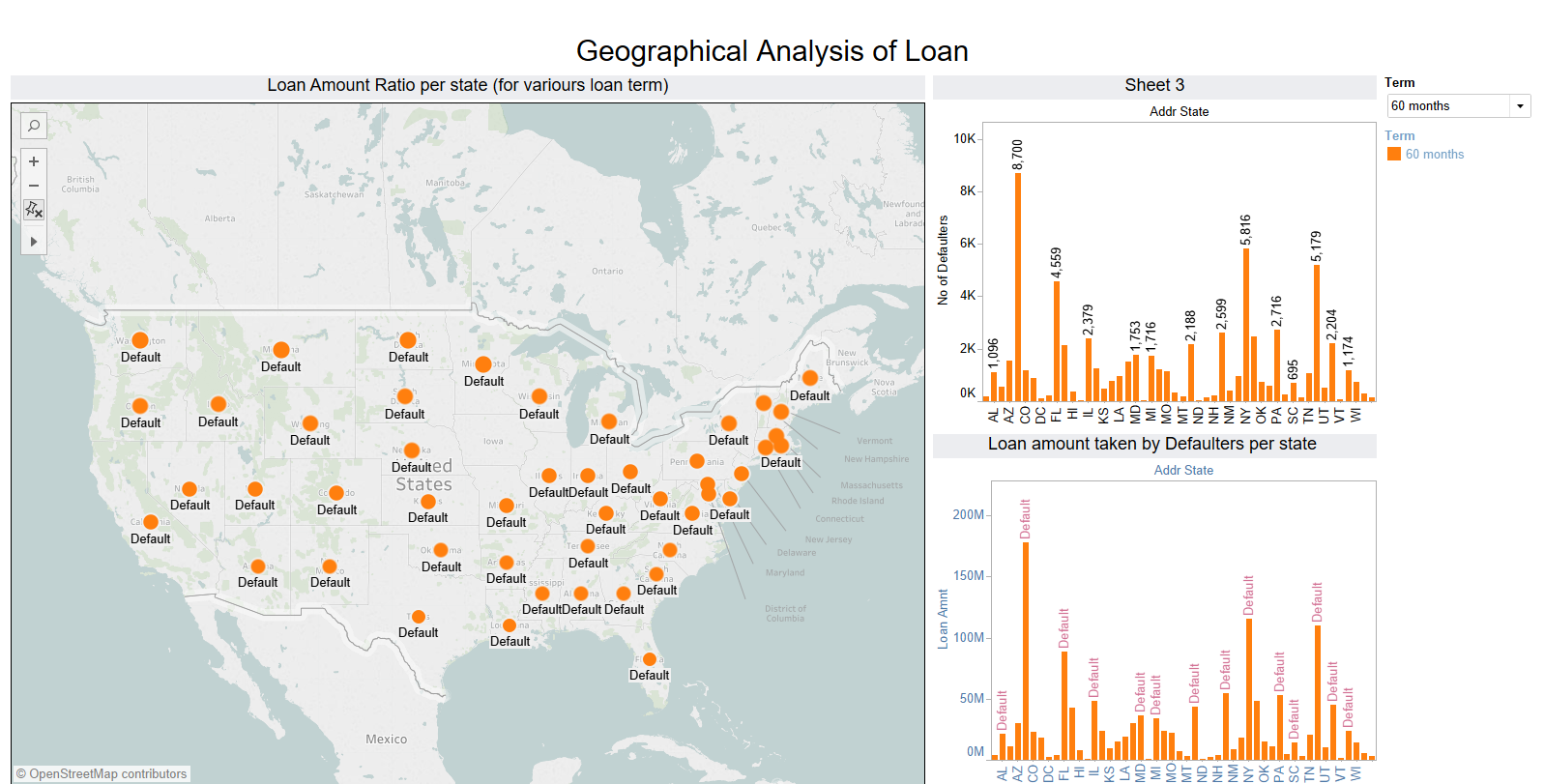
This analysis aims at helping lenders and investors who are about to buy an existing loan ongoing loan. This would help them assess the health of the loan and determine the likelihood of borrower to default or pay the remaining balance in full.

Performing exploratory analysis showed us the below statistical findings:

1. Average borrower’s FICO score is 699
2. Average debt to income ratio is 17.7%
3. Average loan amount is $16,500
4. Average return for investors 7.85%

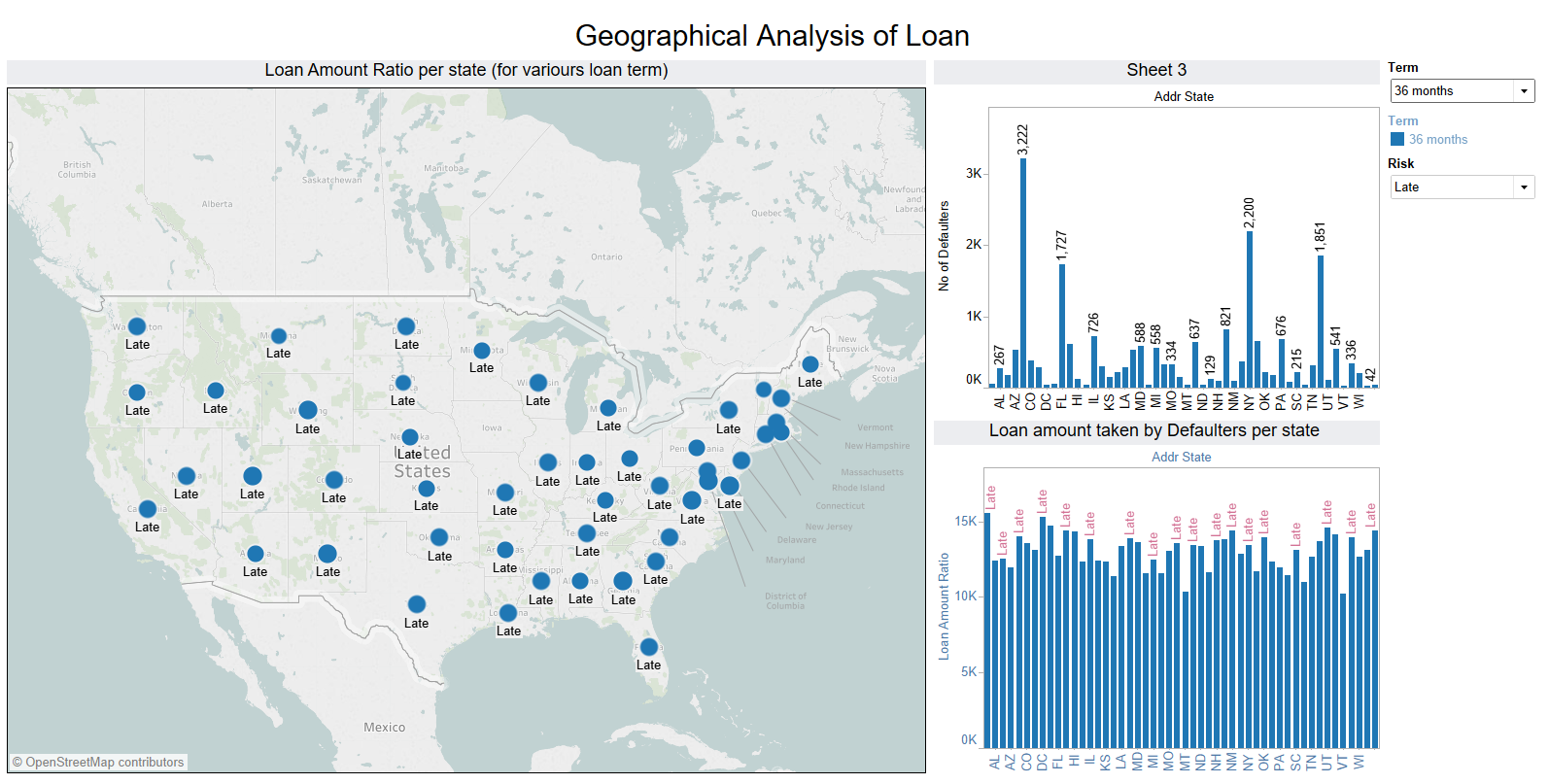


According to the picture above, the highest number of default customers for a 36-month loan term are in Arizona, Ohio, Texas and Florida.

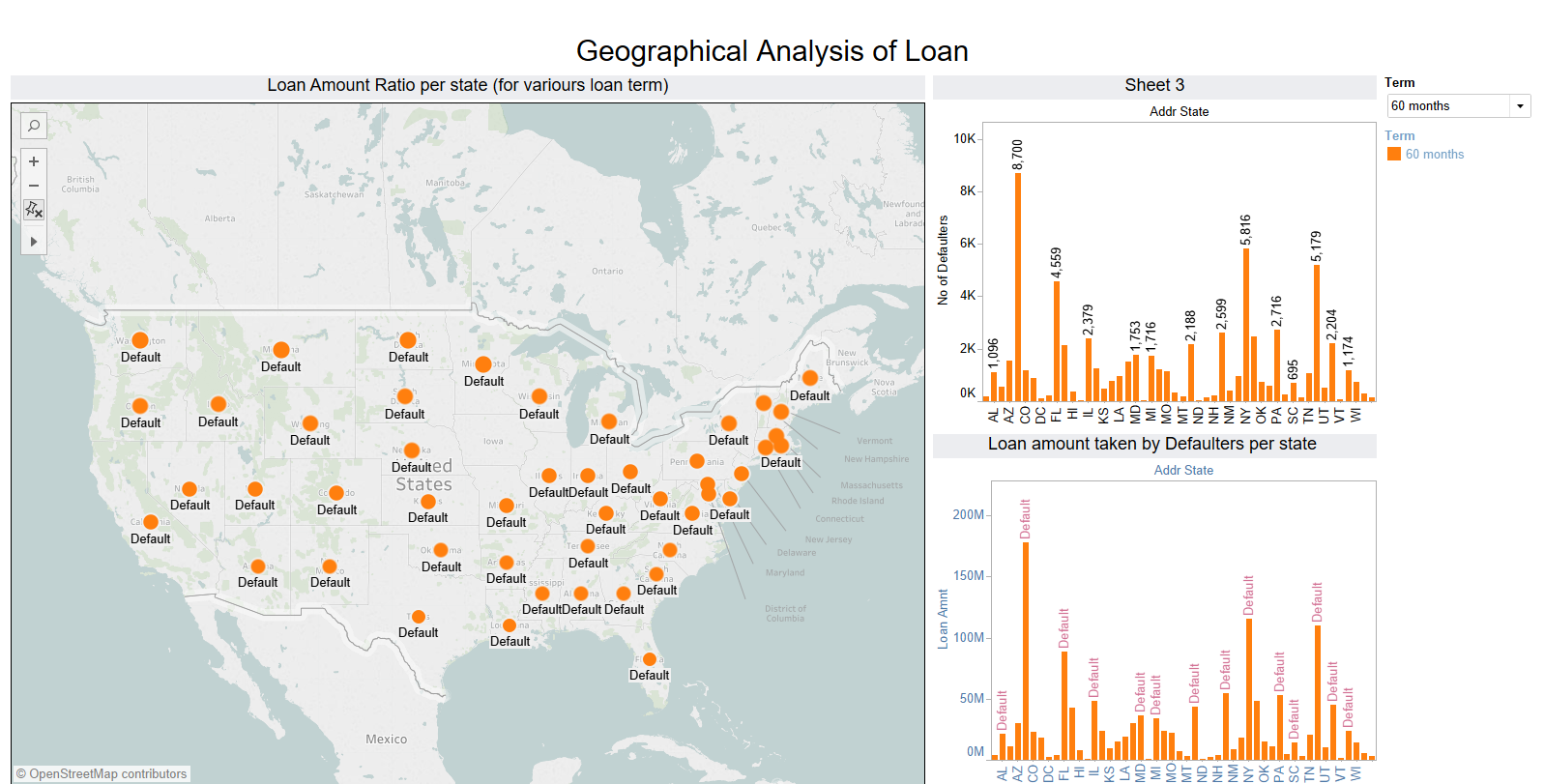


On the other hand, for a 60-month loan term, we see that the states with the highest number of defaulters are Arizona, Utah, New York and Florida.

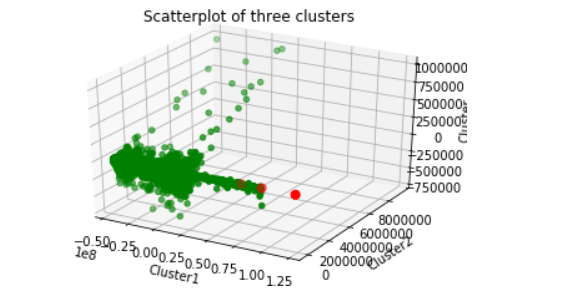
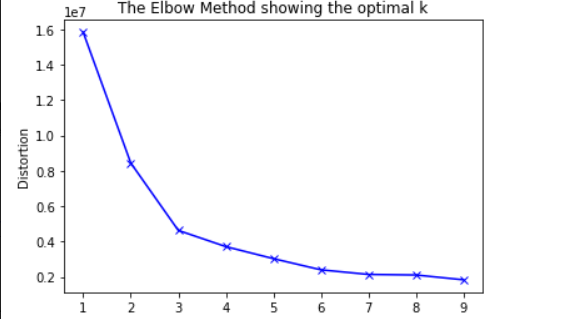
Moreover, in terms of late payments for a 36-month loan term, the picture below shows that the highest number of customers are in Arizona, New York, Utah, and Florida.



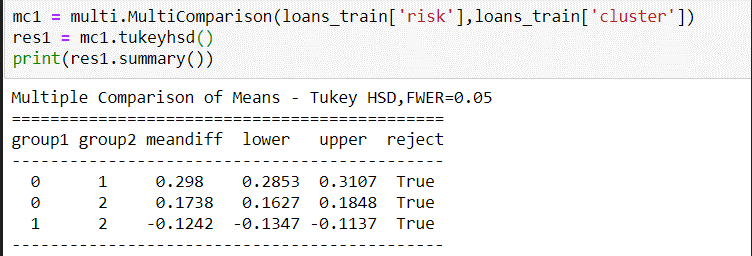
When the loan term increases from 36 to 60 months, the states with the highest number of late payment customers remain the same.



As part of earlier observation, we can see that there are lot of differences in the dataset. To better understand the dataset, we have run k-means clustering to see if there are any clusters within the data.

**\**

As per the elbow test, we have 3 clusters within. Second image shows the clusters and the dataset. To prove that there are clusters within the dataset, we ran Tukey test and the results are below:



Tukey Test confirms that there are 3 clusters as we can see al Trues in Reject column. Also, meandiff column shows the distance between them.

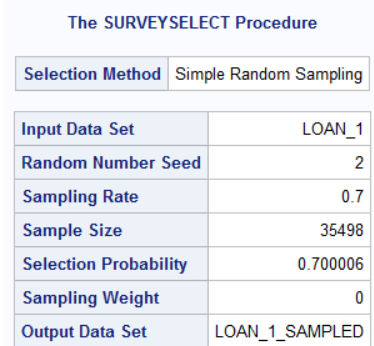
**TECHNIQUES AND METHODS**

To answer the aforementioned business questions, we followed this line of reasoning:

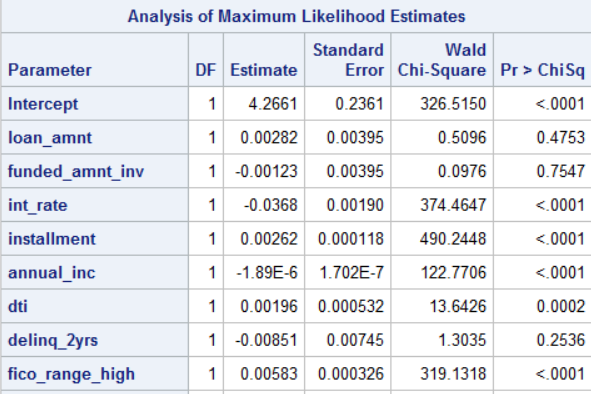
1. We can minimize the loan default risk by predicting customer’s behavior → Classification!
2. Using this insight, the business can take the necessary actions to avoid any unprecedented defaults.
3. A BI model that can predict default probability will mitigate the loss to business.
4. We are capturing only the variables that can be observed at the start of application and not observing the full history data that is provided to us

Since there are 3 clusters within the data, we will run probit, logistic, decision tree and random forest models for full data, as well as 3 clusters to see if we observe any change in error rate.

We have partitioned the data into training (with 70% of data) and test dataset (with 30% of data).



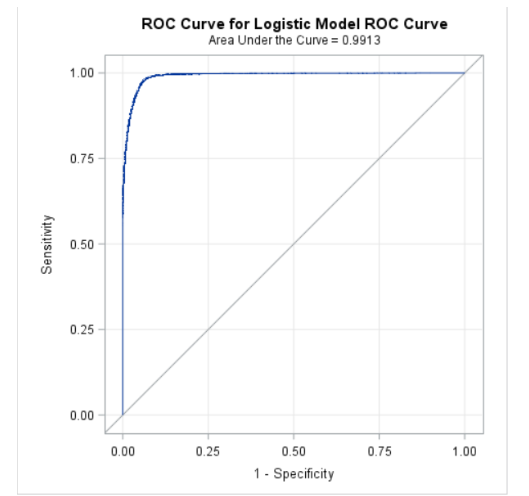
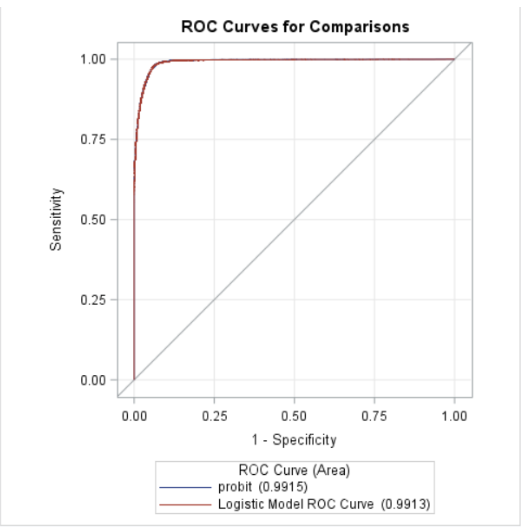
***Logistic regression***



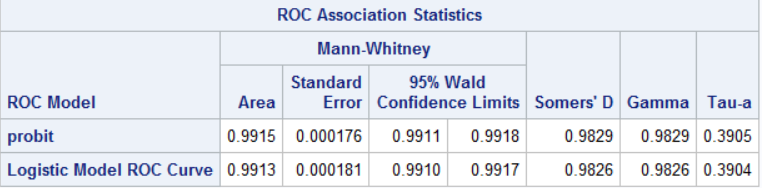
As per the Logistic regression,

* If we increase the loan amount taken, then risk of not getting defaulted increases by log of odds of ratio by 0.00282
* If we increase the interest rate taken, then risk of not getting defaulted decreases by log of odds of ratio by 0.0368
* If we increase the FICO score taken, then risk of not getting defaulted increases by log of odds of ratio by 0.00583

***Probit model***

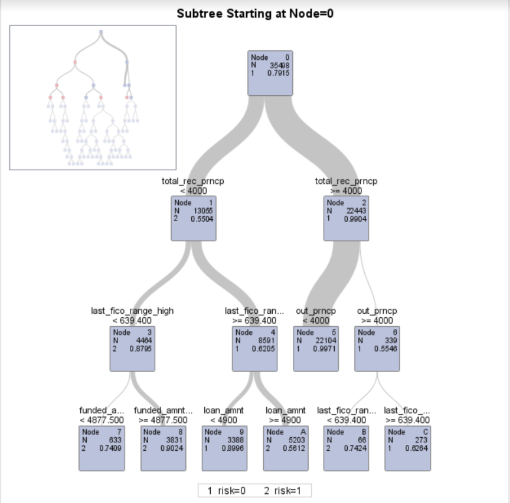
*Logistic Model Probit model*



On comparing, Logit and Probit model, we observe that logit has higher AUC curve which means that Probit is doing better job at classifying defaulters and non-defaulters than Logit Model.

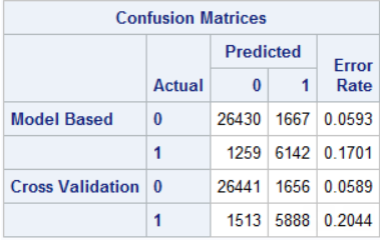
But the difference is very less between them

***Decision tree***



As pe the Decision Tree, we can see that, it is able to classify defaulters and non-defaulters easily.

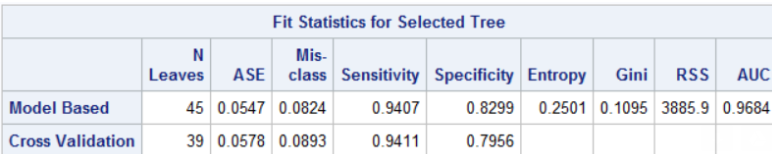
Below is the confusion matrix for the Decision Tree:



As per the confusion matrix, Decision tree model is classifying non-defaulters better than defaulters. As we are focusing on defaulters, we need to explore more and run other models.

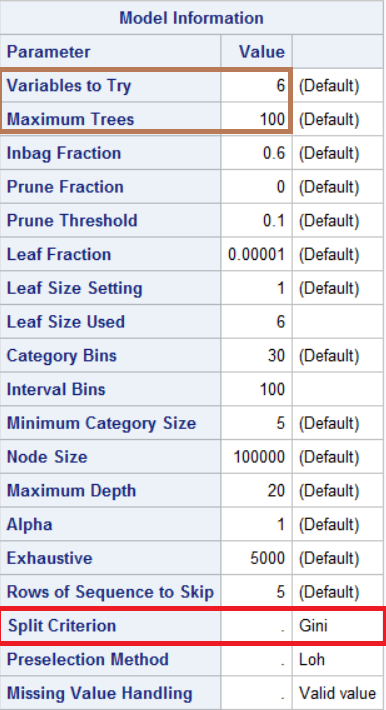


According to this model, last highest Credit Score (last\_fico\_range\_high) is the most important feature.



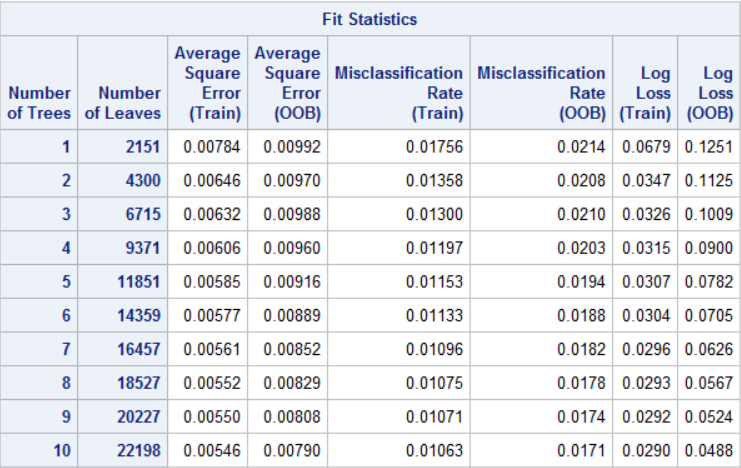
***Random forest***

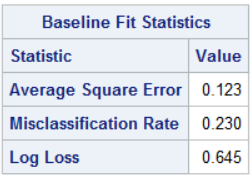
This is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification).



Inference on Random forest:

* we are using GINI index as our feature selection criteria
* After running grid search on python, we found 6 (that is defaulted value in SAS hpforest package) is better at classifying for this dataset.
* Maximum no of tree is capped at 100 for optimizing the result.





As we can see from above figure, error rate is decreasing as we are increasing the sample size. Also, it is working way better than the baseline model i.e. without sampling and just using base decision tree model.

We have run our models on full data and got impressive results. Now we run same models on or clustered data to see, if we get different and better results

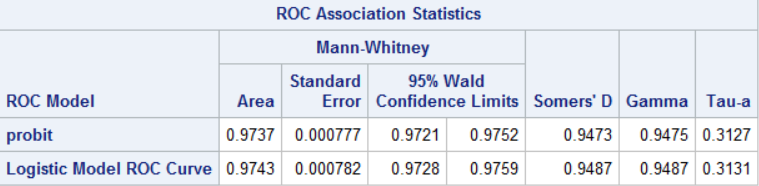
# Comparison between models among clusters

As per the elbow test and Tukey test, there are clusters in the dataset which means all the clusters might behave differently form each other and hence needs to run model separately.

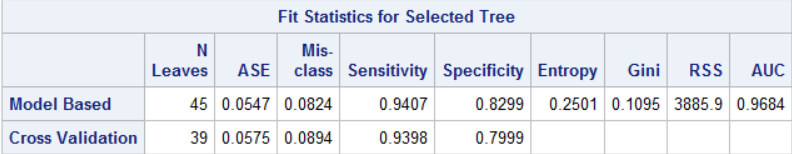
## Cluster 1

Cluster 1 inferences and Accuracy Scores

Logit probit



Decision tree

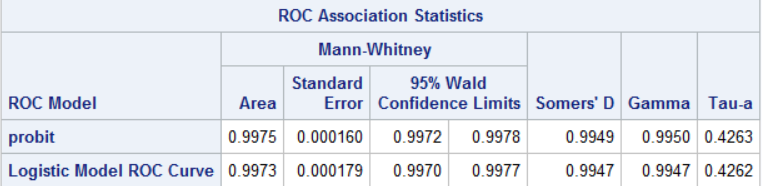


Random forest

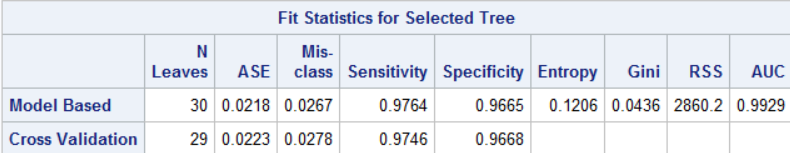


**Cluster 2**

Logit probit



Decision tree

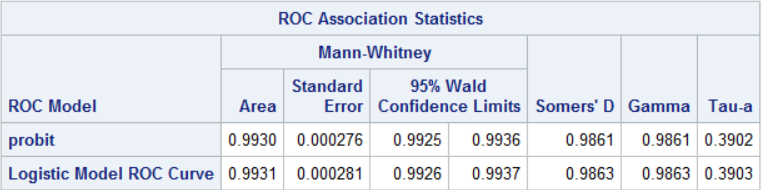


Random forest

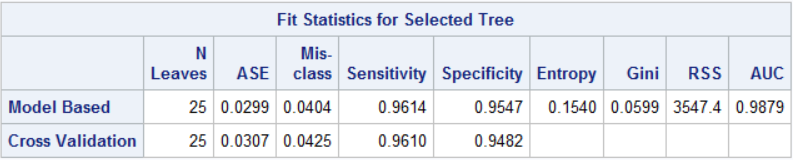


**Cluster 3**

Logit probit



Decision tree



Random forest

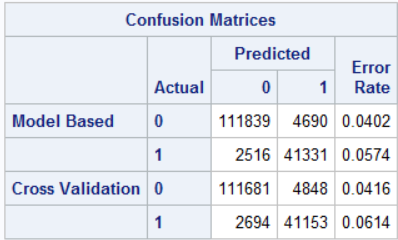


The below tables have the combined results of the all the models:

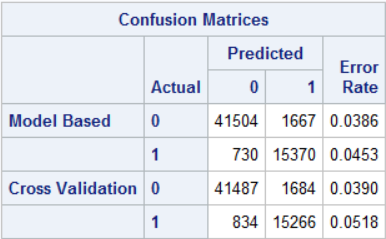
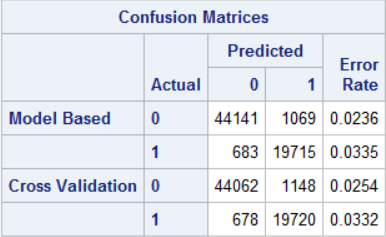
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model name | Model without Cluster | Cluster 1 | Cluster 2 | Cluster 3 |
|  |  |  |  |  |
| Logit | 99.15 | 97.37 | 99.75 | 99.3 |
| Probit | 99.13 | 97.43 | 99.73 | 99.31 |
| Decision Tree | 96.48 | 96.84 | 99.29 | 98.79 |
| Random Forest | 99.5 | 99.1 | 99.98 | 99.95 |

According to logistic regression, probit model, and decision tree, Cluster 2 provides the highest accuracy level. The random forest model is the only one where Cluster 2 is not the best option, instead Cluster 3 has the highest accuracy rate.

On average, we can observe minuscule improvements in the accuracies when model run the cluster.



Cluster 1

Cluster 2 Cluster 3

The above image is from confusion matrix of Decision tree for cluster1, cluster 3, and cluster 2.

The error rate for defaulters has been reduced significantly as compared to full dataset which means there are clusters in the dataset and behave differently from each other.

**RESULTS**

Based on the provided dataset we found out that the critical predictors for estimating likelihood for investors to incur loss on their investments:

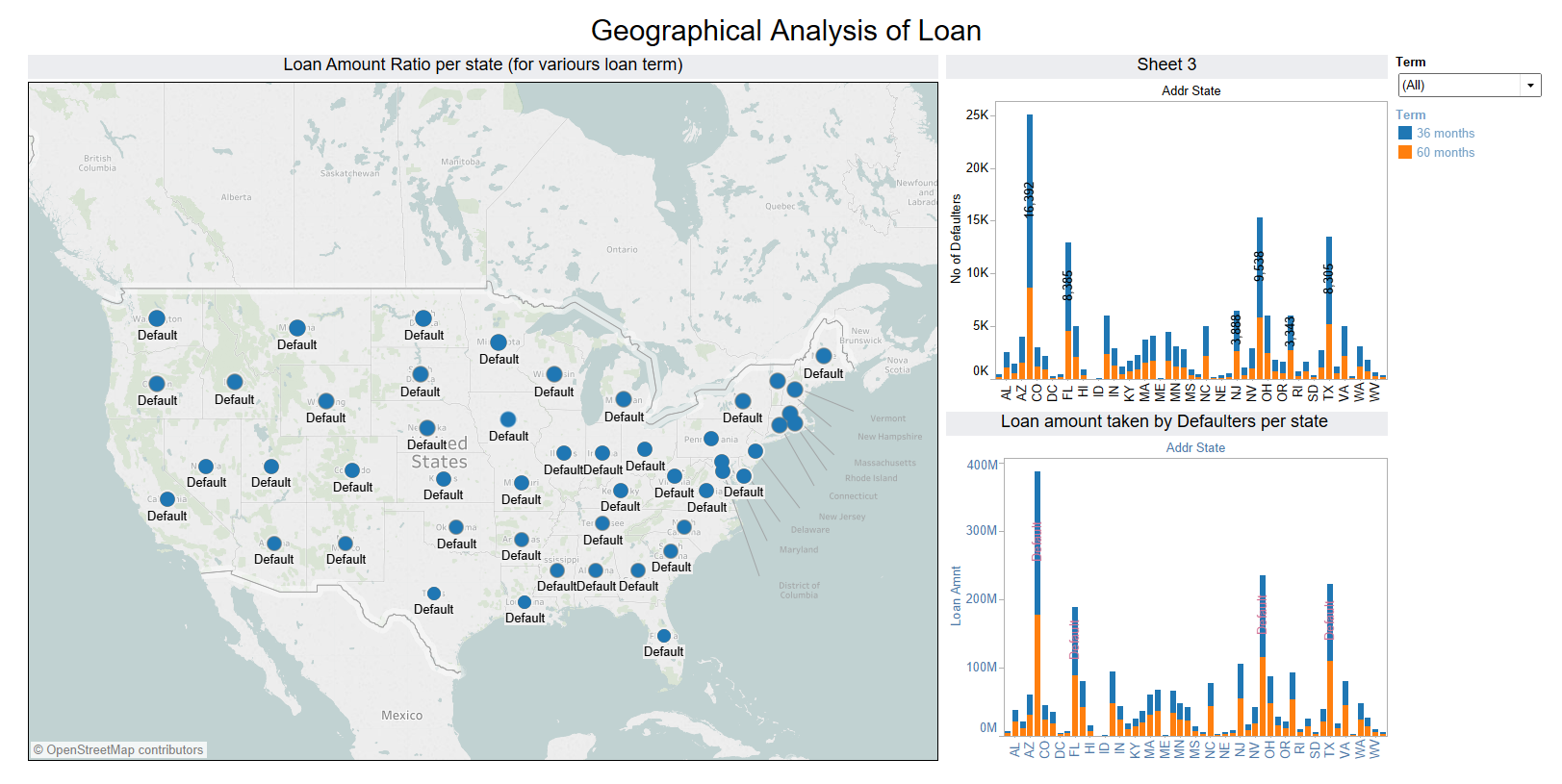
|  |
| --- |
| **Critical Predictors** |
| FICO Score |
| Ability to recover from mis-payments |
| Collection costs at the applicant’s location |
| Currently unpaid debt |
| Current monthly liabilities including mortgages and other loans |
| Any previously unpaid debt |
| Any existing debt settlement plans |
| Household income of loan applicant |
| Employment Status of loan applicant |

**CONCLUSION**

* Our classification model predicts investor´s likelihood of incurring loss with an error rate of 3.6% and accuracy of 96.4%.
* The cost of misclassification to investors would be the cost of collection costs plus the lost remaining balances.
* Collection costs on Lending Club is 40% of the amount collected.
* On average the loan amount is $15,390
* $6,156 is the expected loss for investors whose investments are defaulted on.

**APPENDIX**

Number of defaulters for 36- and 60- months loan terms.



Number of late payment customers for 36- and 60- months loan terms.

