RISING ANALYTICS
Data. Insight. Action.

2191 Lemay Ferry Rd
St. Louis, MO. 63125
(314-324-0695)

Propensity to Pay and Debt Management Programs

Use Case: Healthcare Services

Introduction

In today's world where there is economic uncertainty for every human being, it's quite necessary to understand the financial condition of an individual and have a proper Debt management program in order to avoid huge write-off amount and reduce number of defaulters. It becomes necessary to identify and divide customers in various segments where provider can roll out plans such as partially writing-off loan amount, reducing the interest rate, reducing or waiving off any penalties etc.

Table of contents

2191 Lemay Ferry Rd
St. Louis, MO. 63125
(314-324-0695)

Problem Statement

We worked for a client from healthcare domain in the US who provides credits to the customers that can be used for their treatments. They had 6.5 Million unique customer accounts. Through this study, we hoped to develop some insights that can help organization to hold a debt management plan in place that can work differently for different segment of customers.

Objective and Scope of the project

1. Objective
   The primary objectives of the study are:
   - Classify Good accounts and Bad accounts
   - Providing score/probability to good and bad accounts
   - Providing customer segmentation based on the behavior
2. Scope
   - The scope of the study covers 6.5 Million Credit Accounts.
   - The study covers 2 Years of data starting from Jan-2017 to Dec-2019
   - The study focuses only on the credit variables and demographic variables provided by client.

Data Source

The data were collected from the customer's RDBMS system.

Tools and Techniques

We have used following analytical techniques / methodology for analyzing data:

1. Summary of Statistics for each variable
2. EDA. Using Graphs and plots to visually represent all the variables.
3. Identification of significant variables through correlation matrix and WoE/IV.
4. Apply statistical as well as machine algorithms for classification problem.
5. Tools Used: R, Python and MS Excel
6. Techniques: Box Plot, Histogram, Bar Chart, Line Chart, Correlation Matrix, Logistic Regression, Random Forest, GBM

Analytics Approach

The Analytical approach will involve the following activities:

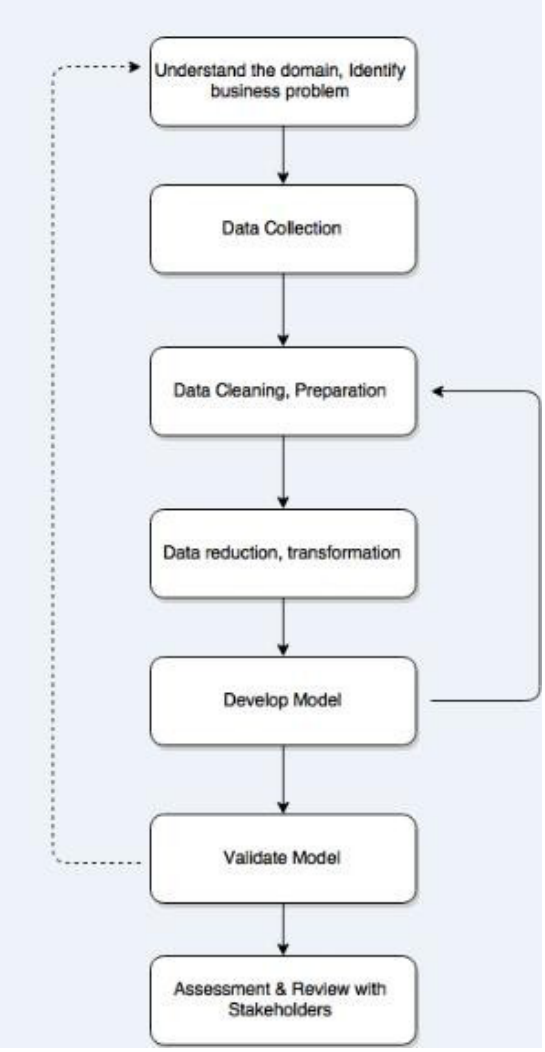
1. Data extraction from primary data source
2. Data quality check
3. Data cleaning and data preparation
4. Study each of the variables using EDA
5. Identifying / Generating Y variable

2191 Lemay Ferry Rd
St. Louis, MO. 63125
(314-324-0695)

6. Selecting the most significant variables by using combination of Correlation matrix and IV
7. Division of data into train and test
8. Model development

9. Stepwise regression and hyperparameter tuning
10. Finalizing model
11. Model validation on train and test data using Decile analysis, Gain and Lift Chart and KS Statistics
12. Verifying goodness of the model using ROC-AUC curve, confusion matrix, specificity and sensitivity checks and accuracy
13. Intervention strategies and recommendations

We plan to use the following Seven Step Analytical Approach for the Project



2191 Lemay Ferry Rd
St. Louis, MO. 63125
(314-324-0695)

2. Data Description and Preparation

Data Management

We were given 56 variables and more than 48 Million records. The dataset was around 24 GB in size. We used AWS server to process such a large amount of data.

Data Quality

However, the data structure was not very complex, quality of data was. Many numeric features were highly skewed. There were number of features having missing data and containing outliers.

Data Preparation

Variable transformation

1. We had dataset with patient's each visit and each transaction level. Hence, we had duplicate entry for each credit account. We had to go through lots of pre-processing and aggregation of some columns to make dataset where we can have per column one credit account.
2. Depending on the nature of data for numeric variables, we used either logarithmic or square root transformation of such variables.
3. For Categorical data, we converted them into dummy variables based on number of unique categories

Missing values and Outliers

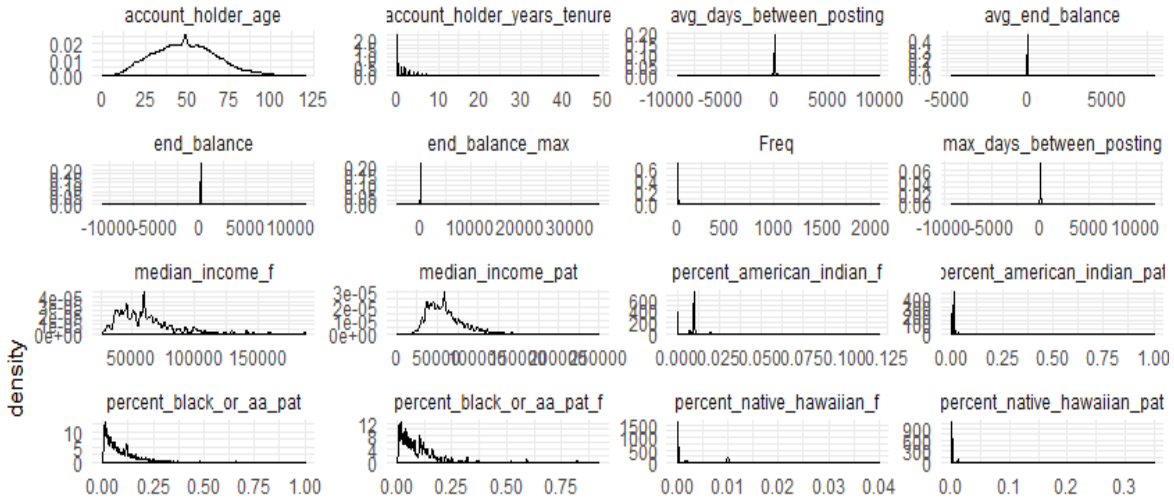
1. We discarded all the variables that had more than 50% of missing values. For remaining variables, we used mean value imputation technique for missing value imputation.
2. For Outliers, we capped the limit as mean +/- 3 * std formula.

2191 Lemay Ferry Rd
St. Louis, MO. 63125
(314-324-0695)

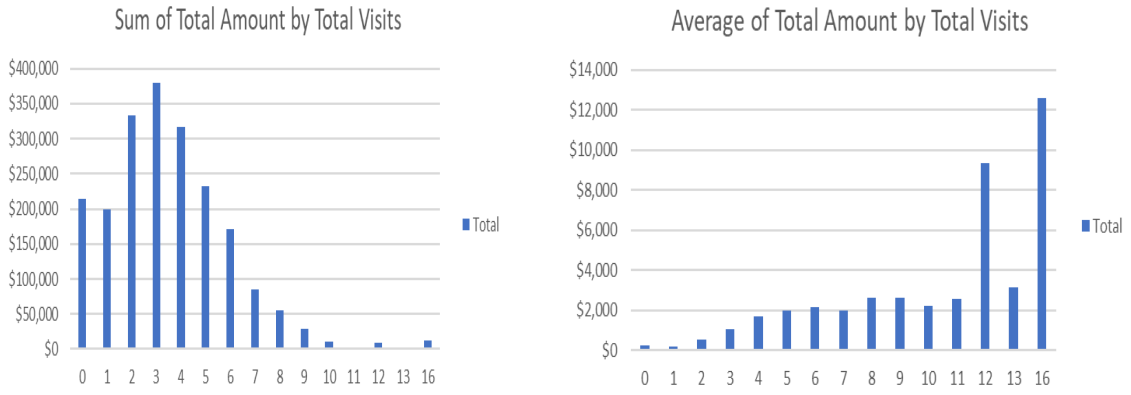
3. Exploratory Data Analysis

The exploratory data analysis is divided in major three parts. They are:

1. **Univariate analysis**: Here we use box plot / histogram / line graph etc. to check the distribution of numeric variables. In the below snapshot, we have shown density graph for all the numeric variables.



2. **Bivariate analysis**: Here we plot Bar graph to see the relationship between different continuous variables: Below graphs shows the relationship between number of visits and revenue generated:



3. **Bivariate Chi-Square test**: Here we perform chi-square test to check how dependent our target variable is on various categorical variables.

2191 Lemay Ferry Rd
St. Louis, MO. 63125
(314-324-0695)

Descriptive Stats:

Below is the snapshot of some of the numeric variables' descriptive statistics. We collect number of missing values, average, variance, standard deviation, minimum and maximum value and datapoints at different percentiles.

E	F	G	H	I	J	K	L	M	N	O	P	Q
mean	std	mean	mean	var	min		p1.1%	p5.5%	p10.1(	q1.25	q2.50	q3.75
48.9029	18.6069	-6.91791	104.724	346.218	0	FALSE	12	19	25	35	48.9129	61
1.63873	2.86741	-6.96351	10.241	8.22207	0	FALSE	0	0	0	0	0	2
15.464	8.68571	-10.5931	41.5211	75.4415	0	FALSE	2.7734	4.3255	5.7197	8.9519	13.9696	20.3315
205681	127127	-175700	587062	1.6E+10	0	FALSE	59300	78600	88500	117900	169400	258200
0.67385	0.19148	0.09942	1.24828	0.03666	0	TRUE	0.13	0.31	0.41	0.55	0.71	0.82
0.11607	0.14347	-0.31434	0.54648	0.02058	0	FALSE	0	0.01	0.01	0.03	0.07	0.14
0.05709	0.07117	-0.15642	0.27059	0.00506	0	FALSE	0	0	0	0.01	0.03	0.07
0.00151	0.00536	-0.01457	0.01759	2.87E-05	0	FALSE	0	0	0	0	0	0
0.10785	0.09593	-0.17995	0.39564	0.0092	0	FALSE	0	0.01	0.01	0.03	0.08	0.16
0.0346	0.01396	-0.00728	0.07649	0.00019	0	FALSE	0.01	0.02	0.02	0.03	0.03	0.04
0.00799	0.01157	-0.02673	0.04271	0.00013	0	FALSE	0	0	0	0	0.01	0.01
58362.5	21263.8	-5428.93	122154	4.5E+08	0	FALSE	24846	31212	35765	42853	55255	68707
14.8257	7.45037	-7.52541	37.1768	55.5081	1.6311	FALSE	2.9816	4.811	5.9695	9.0393	14.5622	19.3154
219314	129206	-168304	606932	1.7E+10	57900	FALSE	72800	86500	102600	132800	184000	272700
0.68302	0.16407	0.19081	1.17522	0.02692	0.04	TRUE	0.18	0.4	0.47	0.58	0.71	0.8
0.10424	0.1201	-0.25607	0.46455	0.01442	0	FALSE	0.01	0.01	0.02	0.03	0.07	0.13
0.06371	0.07113	-0.14969	0.27711	0.00506	0	FALSE	0	0	0.01	0.02	0.04	0.08
0.00169	0.00494	-0.01312	0.0165	2.44E-05	0	FALSE	0	0	0	0	0	0
0.10309	0.08835	-0.16197	0.36815	0.00781	0	FALSE	0	0.01	0.02	0.03	0.08	0.14
0.0356	0.01324	-0.00413	0.07533	0.00018	0.01	FALSE	0.01	0.02	0.02	0.03	0.03	0.04
0.00759	0.00774	-0.01564	0.03081	5.99E-05	0	FALSE	0	0	0	0	0.01	0.01
59335.7	22020	-6724.18	125396	4.8E+08	26702	FALSE	29601	35282	37041	43855	56555	68495
-1109.11	15045.2	-46244.7	44026.5	2.3E+08	-1.3E+07	TRUE	-8338.44	-4065.99	-2806	-1129	-299.25	-69
-1107.99	15045	-46243.1	44027.2	2.3E+08	-1.3E+07	TRUE	-8335	-4061.8	-2803	-1127.12	-299	-68.9
7.42501	13.5266	-33.1547	48.0048	182.968	1	FALSE	1	1	1	1	3	8
0.39742	18.57	-55.3127	56.1075	344.846	-4700	TRUE	-1.83811	0	0	0	0	0
21.4254	387.944	-1142.41	1185.26	150500	-9164	TRUE	-967.5	-312.6	-101.333	0	0	25.75

Data Insights and Derived variables

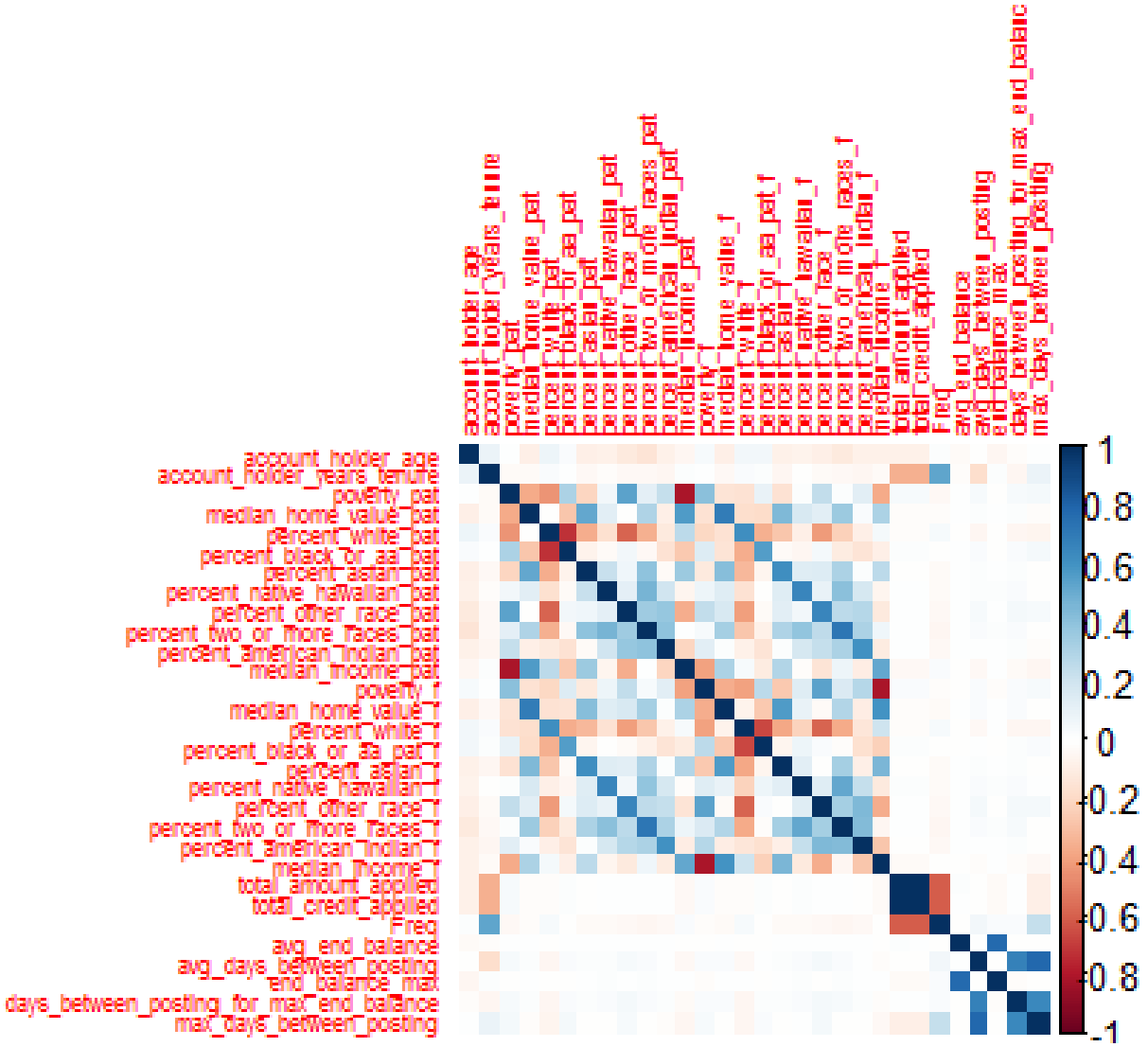
After completing EDA, we got lots of insights about the data. We could get some idea on variables that were very business specific and difficult to understand. We also felt the need to create some derived variables that might be better for final model development.

- For robust model development, it is a good practice to create derived variables and it is necessary to clean impurities in all the variables. b
- We created derived variables such as "avg days between posting", "avg end balance", "max days between posting", "max end balance", "frequency visits", "total amount applied", "total credit applied", etc..
- There were negative age and account_holder_tenure values. Corrected age variable by making it positive and made tenure to 0 for negative values.

2191 Lemay Ferry Rd
St. Louis, MO. 63125
(314-324-0695)

Correlation Matrix, WoE and IV

Correlation Matrix shows the relationship between all the continuous variable with one another. It helps to determine multicollinearity if at all exists in our dataset. Below is the correlation plot for some of the variables. Dark Blue suggests positive correlation and Dark Red suggests negative correlation.



WOE describes the relationship between a predictive variable and a binary target variable whereas IV measures the strength of that relationship.

Using IV and correlation matrix together, we can select limited number of independent variables that are statistically significant for our target variable.

2191 Lemay Ferry Rd
St. Louis, MO. 63125
(314-324-0695)

Model Development

K-Means cluster analysis

Since the data size was huge and it contained high variance, we first applied K-Means cluster analysis. K-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

It helped us dividing the data into 8 different clusters(groups) each group having different characteristics than each other. Now all these 8 groups of data had high variance between them and low variance within the group.

Logistic Regression, Decision Tree and Gradient Boosting

After data-preprocessing, we applied three different algorithms Logistic Regression, Decision tree and Gradient boosting for classification problem on different segments of customers which were generated by K-Means cluster analysis.

We always divide our dataset as 70% - training and 30% - validation. The model development was done at multiple levels to arrive at a most suitable model. The first one with actual variables, second one with some combination with derived variables and using different modelling techniques.

Since the objective is to predict bad loans(credit) accounts, we used binomial logistic regression, decision tree and gradient boosting techniques.

The data for the modeling was split into two parts train & Test data. The Split of the data is as follows:

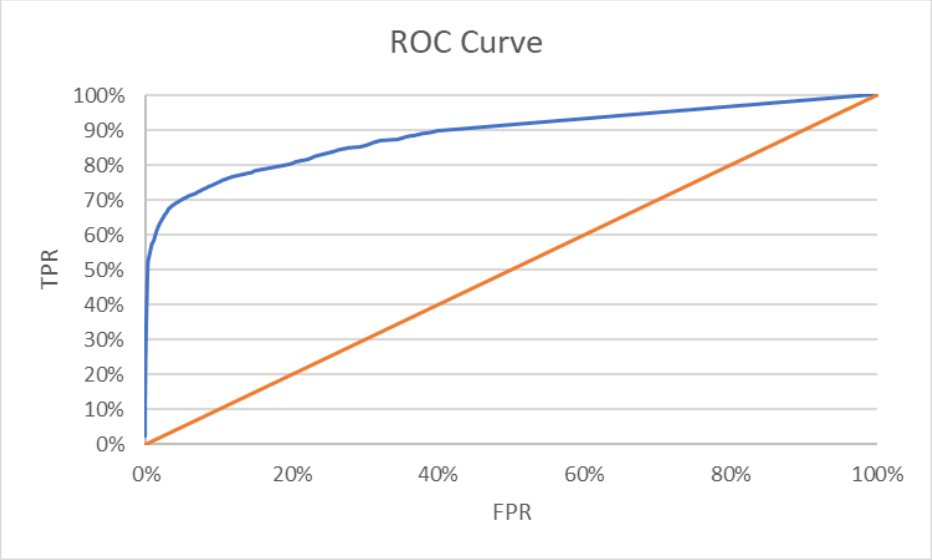
Modeling data using different filters			
Filter	Total Data Size	Training	Testing
All Records	6,548,112	4,583,678	1,964,434
Records having End balance > 0	66,821	46,774	20,047

2191 Lemay Ferry Rd
St. Louis, MO. 63125
(314-324-0695)

Inference:

Since we spent too much of time in variable selection, data preparation, data cleaning and initial clustering analysis exercise, we got very good initial result in terms of concordance / AUC.

We could achieve 0.86, 0.88 and 0.89 AUC from Logistic regression, Decision Tree and Gradient Boosting algorithms respectively. Below is the ROC curve for the same:



Validation

We used Decile analysis, Gain & Lift chart and KS statistics for initial validation of the model.

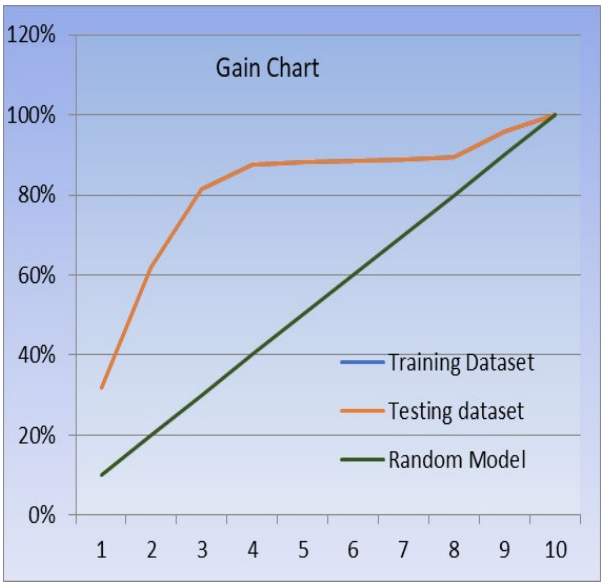
Decile Analysis:

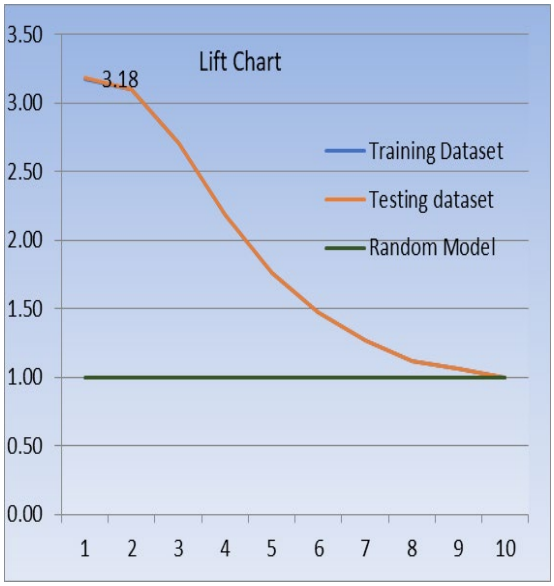
Training Dataset										
Decile	min_prob	max_prob	Good_cou	Bad_coun	Bad Rate	Bad%	CummBad	Good%	CummGod	KS
1	82.19%	100.00%	4729	450519	99%	32%	32%	0%	0%	0.315955
2	38.81%	82.19%	25638	429611	94%	30%	62%	1%	1%	0.610504
3	26.53%	38.81%	181125	274124	60%	19%	81%	6%	7%	0.745864
4	22.29%	26.53%	367735	87514	19%	6%	88%	12%	18%	0.690171
5	22.00%	22.29%	447397	7852	2%	1%	88%	14%	33%	0.55292
6	22.00%	22.00%	450029	5220	1%	0%	88%	14%	47%	0.412974
7	22.00%	22.00%	450164	5085	1%	0%	89%	14%	61%	0.272889
8	21.37%	22.00%	444431	10818	2%	1%	90%	14%	76%	0.138674
9	4.04%	21.37%	364820	90429	20%	6%	96%	12%	87%	0.085966
10	0.00%	4.04%	397304	57945	13%	4%	100%	13%	100%	1.11E-16
			3133372	1419117						

2191 Lemay Ferry Rd
St. Louis, MO. 63125
(314-324-0695)

Testing dataset										
Decile	min_prob	max_prob	Good_cou	Bad_coun	Bad Rate	Bad%	CummBad	Good%	CummGoo	KS
1	82.18%	100.00%	2041	193065	99%	32%	32%	0%	0%	0.316324
2	38.81%	82.18%	11294	183813	94%	30%	62%	1%	1%	0.61053
3	26.51%	38.81%	78064	117043	60%	19%	81%	6%	7%	0.74512
4	22.28%	26.51%	157585	37522	19%	6%	87%	12%	19%	0.689611
5	22.00%	22.28%	190681	4426	2%	1%	88%	14%	33%	0.554984
6	22.00%	22.00%	193295	1811	1%	0%	89%	14%	47%	0.414107
7	22.00%	22.00%	192851	2256	1%	0%	89%	14%	61%	0.274293
8	21.35%	22.00%	191024	4083	2%	1%	90%	14%	76%	0.138847
9	4.00%	21.35%	156536	38571	20%	6%	96%	12%	87%	0.085845
10	0.00%	4.00%	170275	24832	13%	4%	100%	13%	100%	0
			1343646	607422						

Gain & Lift Charts:





As we can see from Decile analysis outcome, we are getting almost similar results for our training and validation datasets. We are getting Maximum KS within first three deciles which is one of the indications of a good model. Our model is able to capture almost 81% of bad loan accounts in first 3 deciles.

If we refer Gain and Lift charts, we can see that we are getting similar gain and lift for both training and testing datasets.

2191 Lemay Ferry Rd
St. Louis, MO. 63125
(314-324-0695)

Confusion Matrix

After cross checking all three modeling techniques, we developed various confusion matrix based on different cut-offs. For all those cut-offs we checked various statistical parameters like precision, sensitivity, specificity and accuracy. Below is one of those examples:

TPR	0.813389
TNR	0.932501
FPR	0.067499
Precision	0.845145
Accuracy	0.895371

Confusion Matrix(CutOff - 0.2653319)			
		Predicted	
		0	1
Actual	0	2921872	211500
	1	264823	1154294

AUC	0.88
Gini	0.77

Model Deployment

Client wanted to calculate probability of all the accounts getting defaulters. We provided them probability incurred from the model. We also provided them the intercept and coefficient for all the parameters we used in the model so that they themselves can calculate the probability on their database.

Challenges and Recommendation

Since they used to keep track of each visit and each transaction in their database, their database size was very large, and it was increasing very fast. They wanted to scrap the old entries to maintain the size of the database. Now the challenge was to calculate some derived variables in case old data is scraped. We recommended the method to keep calculating average / median etc. even while old data is archived.

Phase 2 of the project was to calculate survival analysis on the data. Survival analysis gives us time to event estimation. In this case, we can predict time when an account may get default.