RISING ANALYTICS
Data. Insight. Action.

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

Credit Risk Modeling

Case Study: Dental Group in the US

Introduction

Credit risk modelling refers to the use of financial models to estimate losses a firm might suffer in the event of a borrower's default. This model helps to evaluate credit score of each borrower based on their demographics, past accounting variables and third-party data such as Vantage score and Fico score. The credit score helps to determine many factors such as what should be the limit, the interest rate, down payment for different segments of customers.

Table of contents

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

Problem Statement

We worked for a client from Dental healthcare domain in the US who provides credits to the customers that can be used for their treatments. They had around 100,000 unique customer accounts. Through this study, we hoped to develop credit score for each account 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 100,000 Credit Accounts.
   - The study covers 6 Years of data starting from Jan-2013 to Dec-2018
   - The study focuses on the credit variables, demographic variables and third-party variables fetched from Equifax.

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:

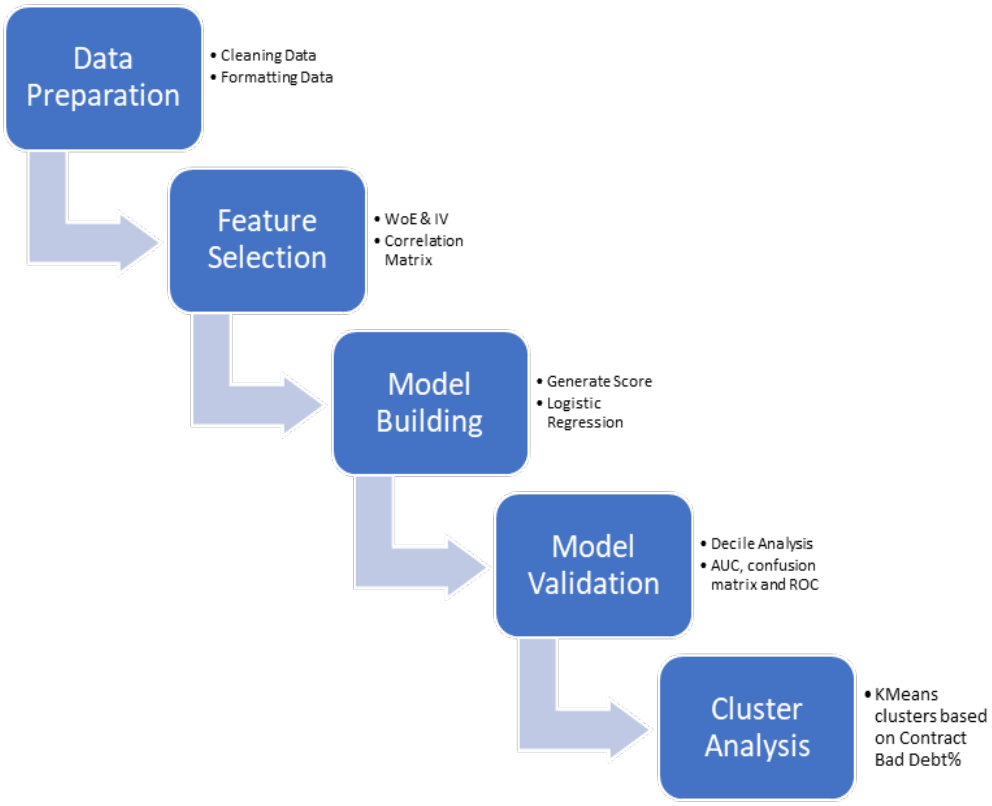
1. Data extraction from primary data source
2. Data quality check
3. Data cleaning and data preparation
4. Study each of the variables using EDA

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

5. Identifying / Generating Y variable
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 Five Step Analytical Approach for the Project



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

2. Data Description and Preparation

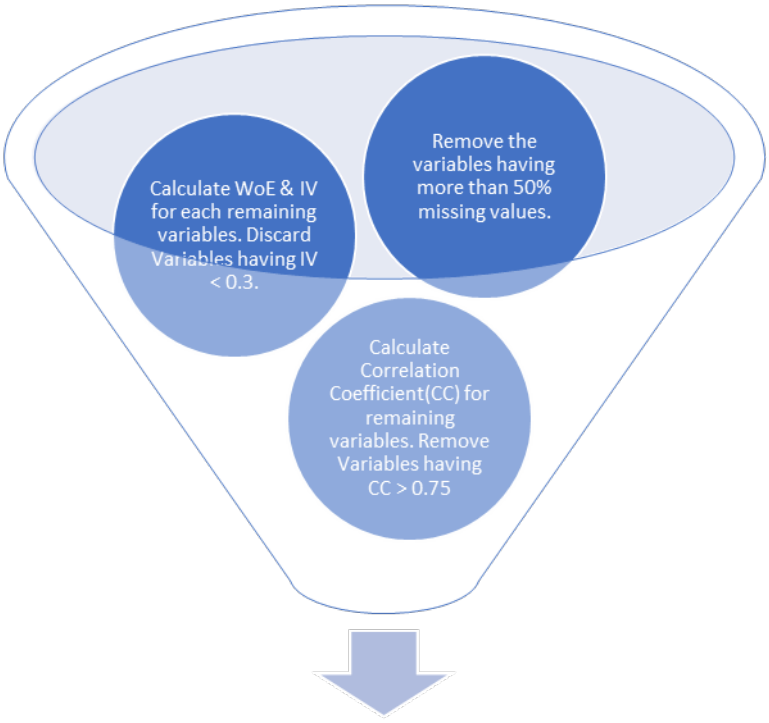
Data Management

We were given dataset with more than 1300 variables and more than 100,000 records.

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



Using above filtering process, we could eliminate more than 1300 variables and we came down to close to 10 variables for final model.

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

Variable transformation

1. Depending on the nature of data for numeric variables, we used either logarithmic or square root transformation of such variables.
2. For Categorical data, we converted them into dummy variables based on number of unique categories

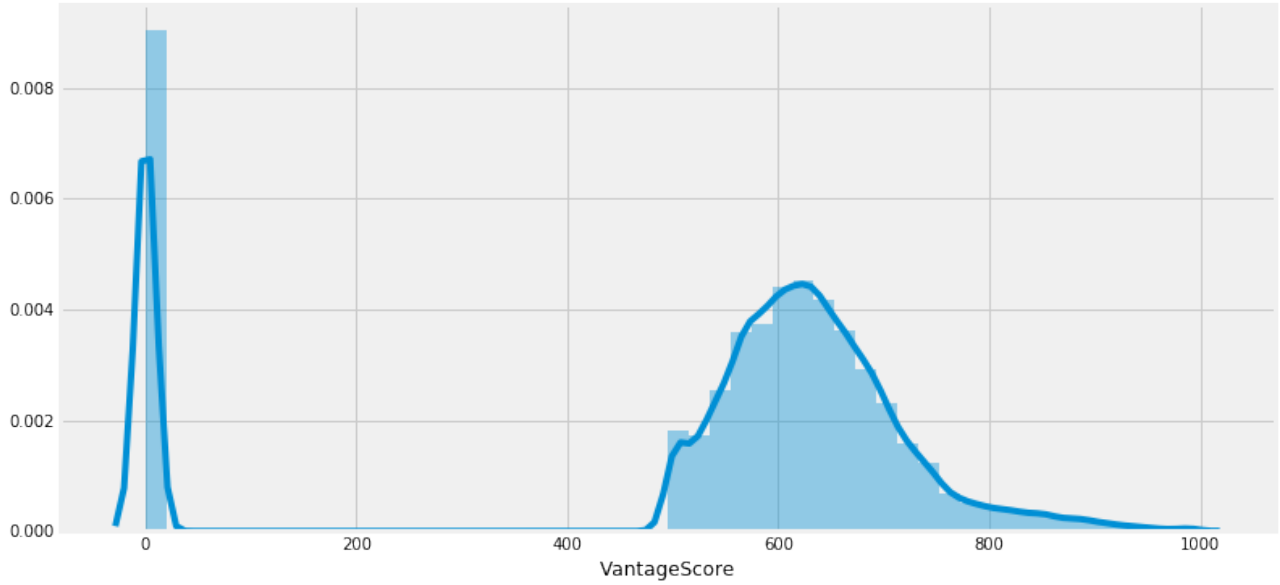
Missing values and Outliers

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.

3. Exploratory Data Analysis

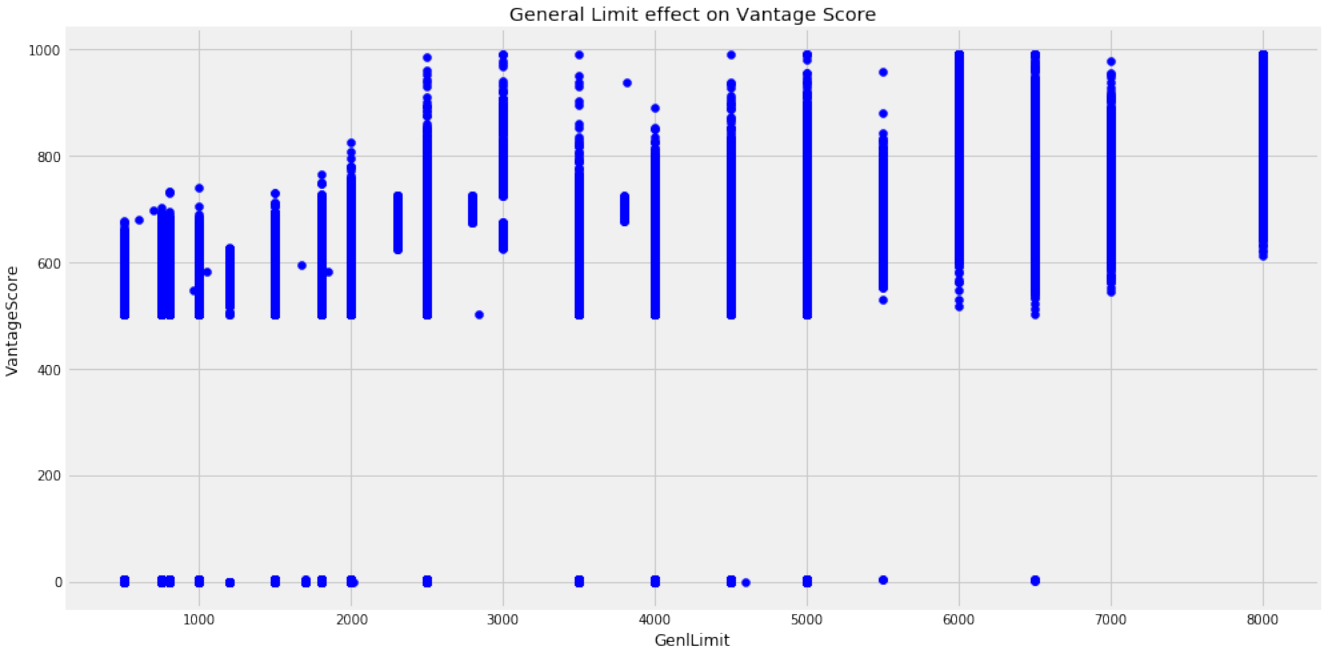
The exploratory data analysis is divided in major three parts. They are:

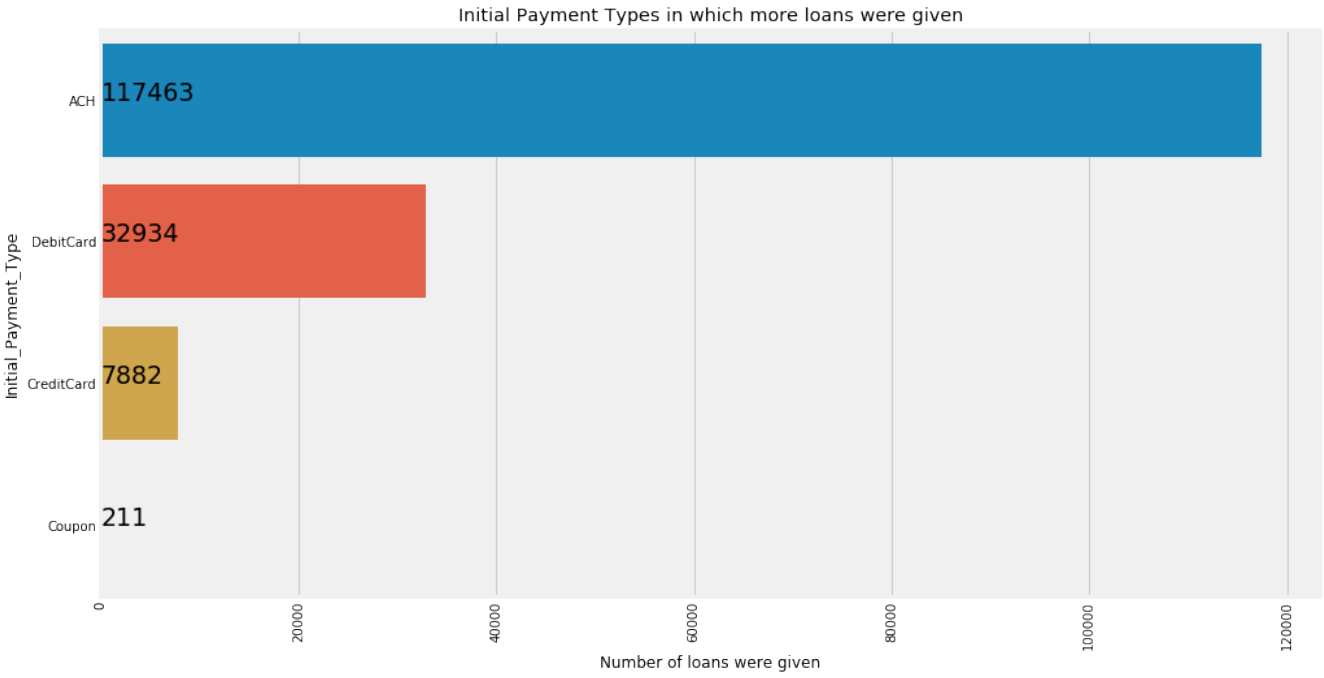
1. **Univariate analysis**: Here we use box plot / histogram / line graph etc. to check the distribution of numeric variables. In the below snapshot, we have shown density graph for all the numeric variables.



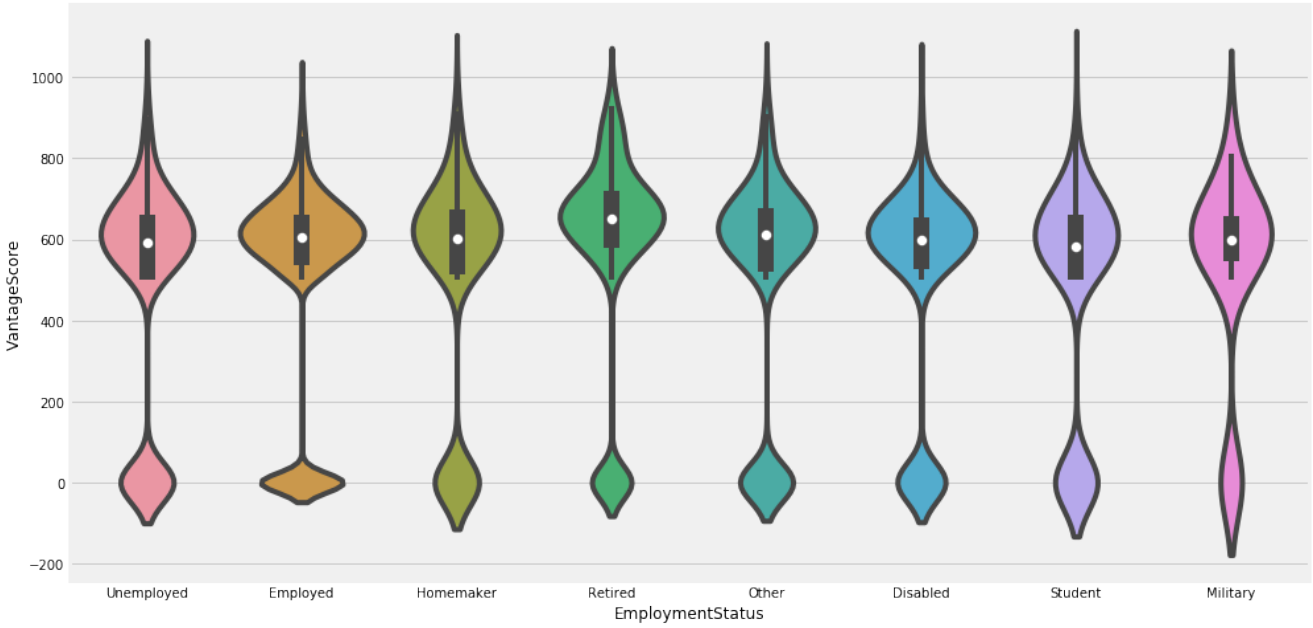
2191 Lemay Ferry Rd
St. Louis, MO. 63125
(314-324-0695)

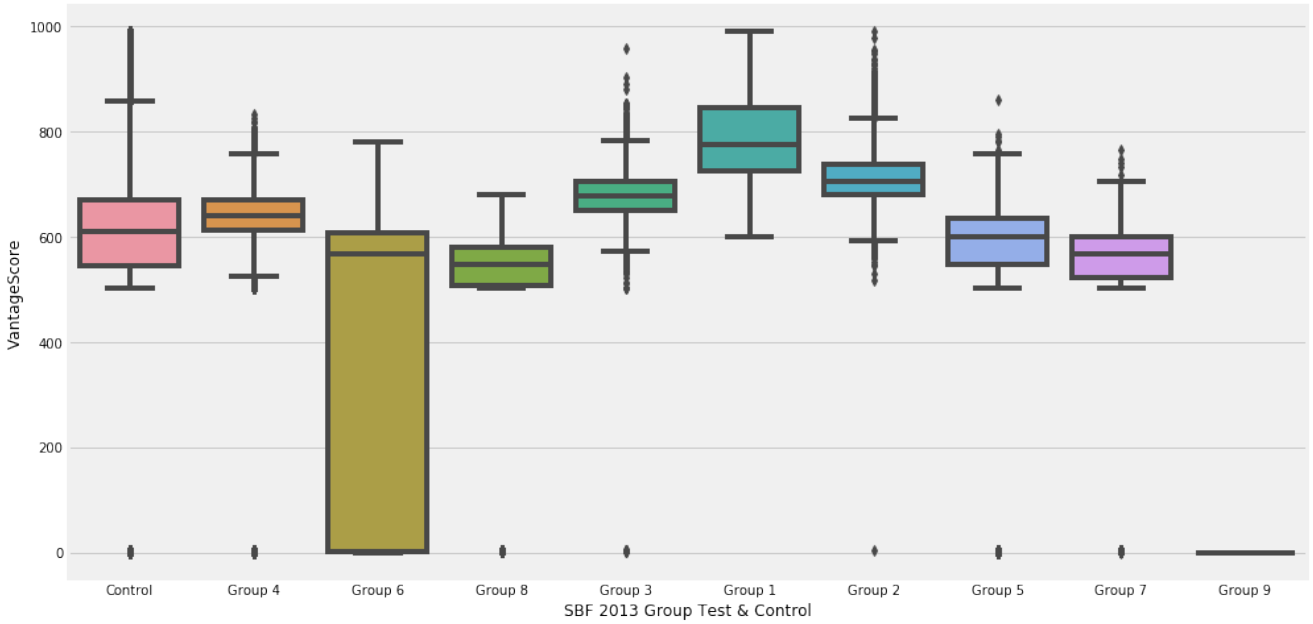
2. **Bivariate analysis**: Here we plot Bar graph to see the relationship between different continuous variables: Below graphs shows the relationship between number of visits and revenue generated:





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





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.

-

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.

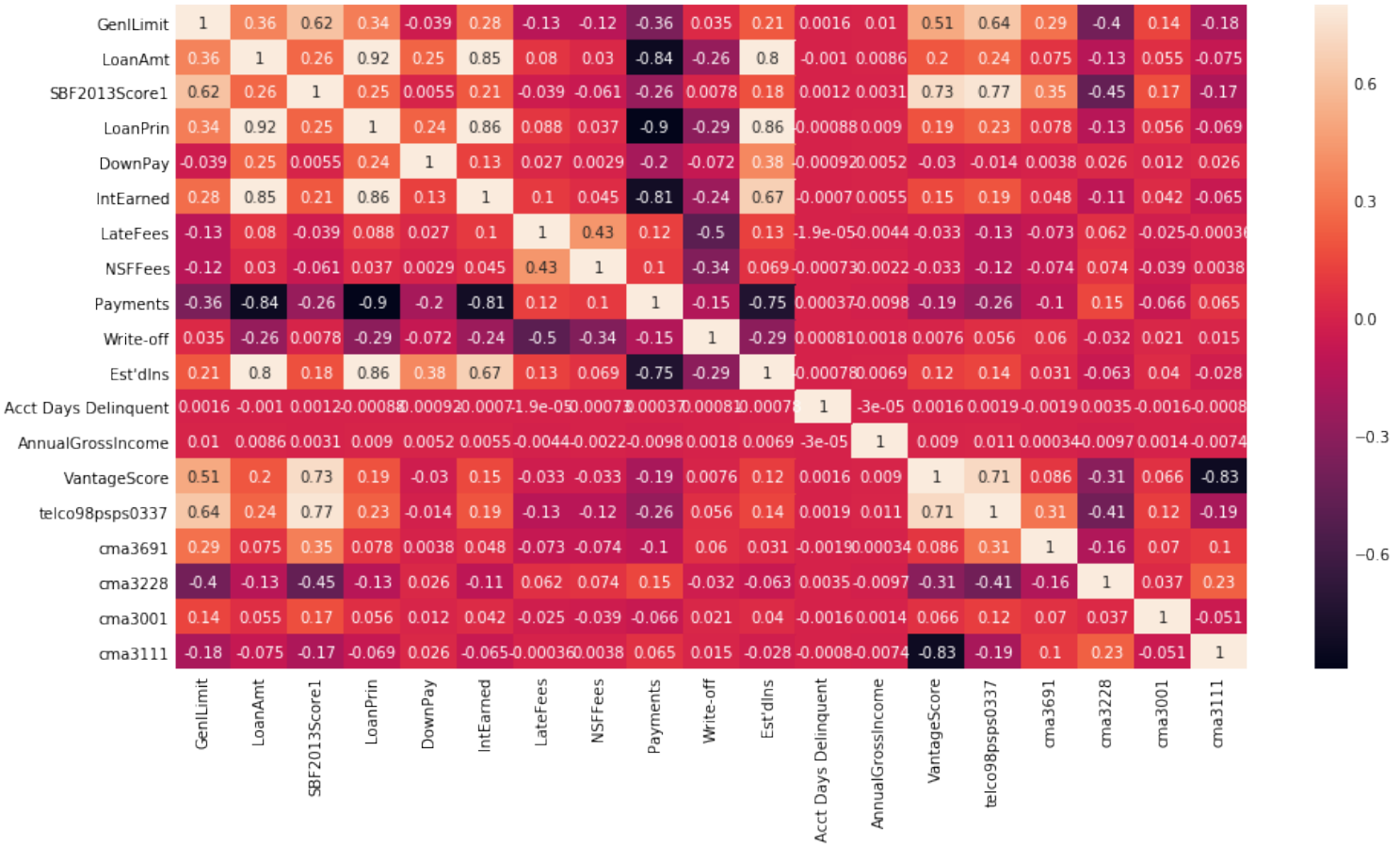
- For robust model development, it is a good practice to create derived variables and it is necessary to clean impurities in all the variables.
- We created derived variables such as "avg days between posting", "avg end balance", "max days between posting", "max end balance", "frequency visits", "total 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.

mean	std	var	min	p1.1%	p5.5%	p10.10%	q1.25%	q2.50%
3306.267	2210.068	4884401	500	500	500	800	1500	2500
5450690	4975859	2.48E+13	0	200	300	600	2845	9999996
57133.49	46958.35	2.21E+09	0	0	211	2061.4	6891	99996
3951.234	4875.842	23773836	0	0	0	0	0	57
5448167	4978017	2.48E+13	0	0	52	267	1337	9999996
27.19173	42.21659	1782.241	0	0	0	0	0	2
3700023	4825745	2.33E+13	0	56	168	335	1113.5	4799
27.48811	42.05177	1768.351	0	0	0	0	1	2
5.591193	16.38165	268.3585	0	0	0	0	0	1
30723.52	41580.86	1.73E+09	0	0	0	0	4444	9167
37.06313	45.38177	2059.505	0	0	1	1	2	5
67493.7	44472.8	1.98E+09	0	0	1333	4671.4	9549	99998
30.54705	42.83313	1834.677	0	0	0	1	2	4
31575.29	42986.95	1.85E+09	0	0	0	0	2222	7500
16.51648	30.63445	938.4696	0	0	0	0	1	4
52126.62	47713.76	2.28E+09	0	0	0	0	5000	10000
5.884041	16.4567	270.823	0	0	0	0	0	2
3356472	4719674	2.23E+13	0	56	161	319	1041	4050
3837.941	4843.77	23462105	0	0	2	4	17	54
614.4146	203.2708	41319.01	0	0	411	474	532	593
510.5996	261.3316	68294.21	0	0	1	2	526	603
514.812	202.1141	40850.1	0	0	1	4	510	563

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. Black color suggests negative correlation and light orange color suggests positive 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

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 training dataset.

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:

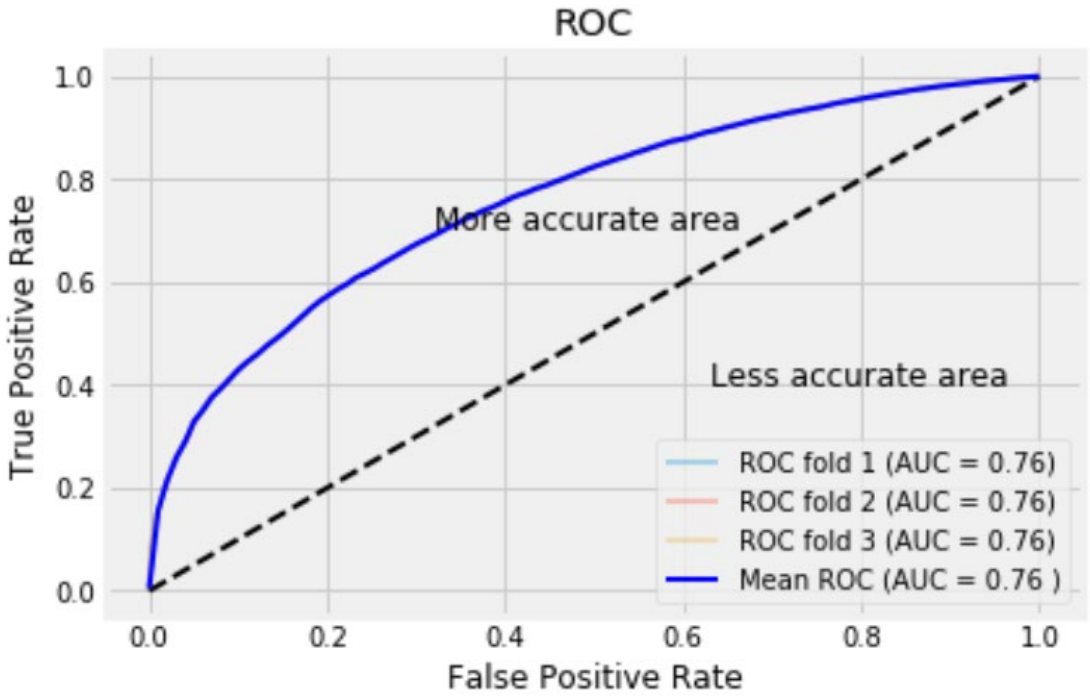
Modeling data using different filters			
Filter	Total Data Size	Training	Testing
General treatment	88,000	61,600	26,400
Ortho treatment	12,000	8,400	3,600

Inference:

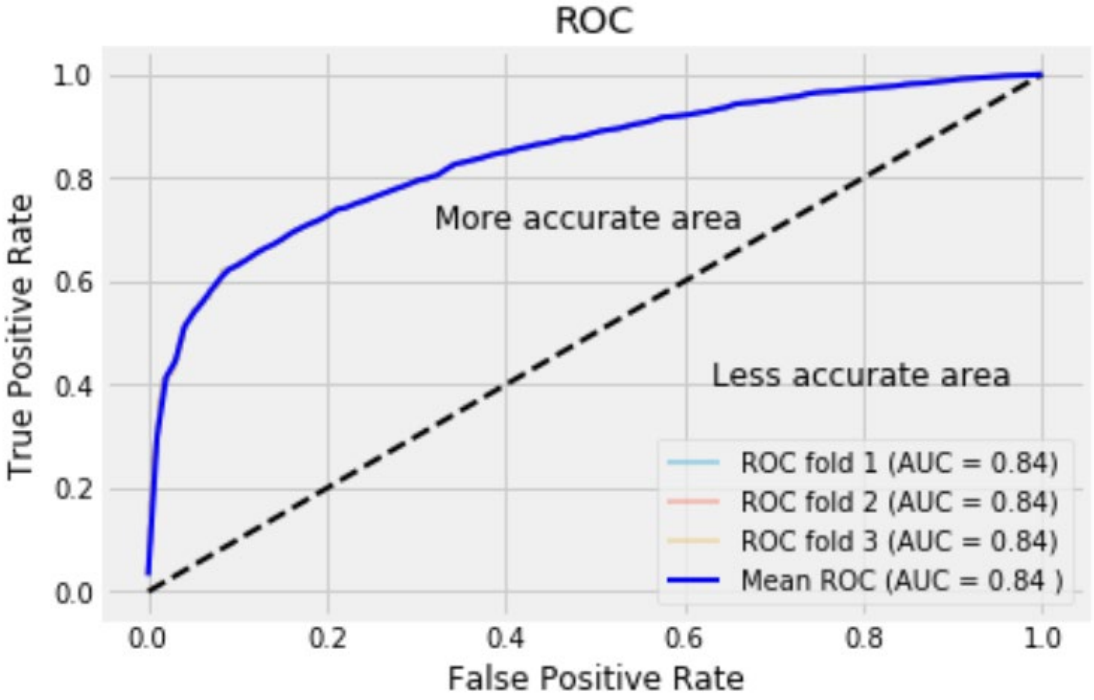
Since we spent too much of time in variable selection, data preparation, data cleaning and initial clustering analysis exercise, we got very good initial result in terms of concordance / AUC. We could achieve 0.76, 0.74 and 0.75 AUC from Logistic regression, Decision Tree and Gradient Boosting algorithms respectively for General records and 0.84 for Ortho records. Below is the ROC curves for the same:

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

General model ROC:



Ortho Model ROC:



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

Validation

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

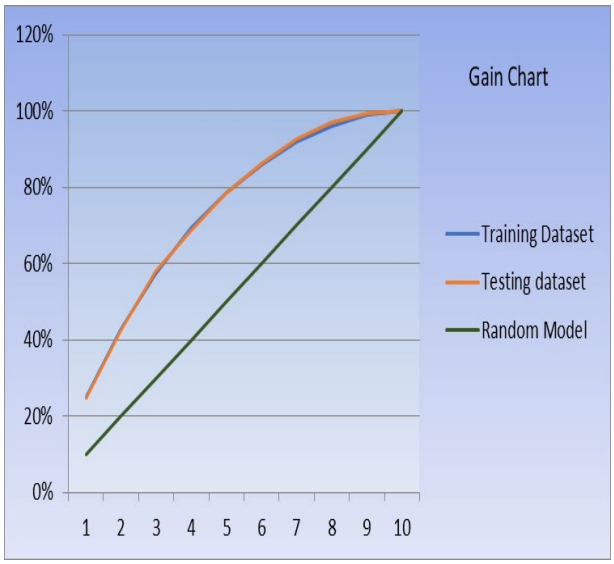
Decile Analysis(General model):

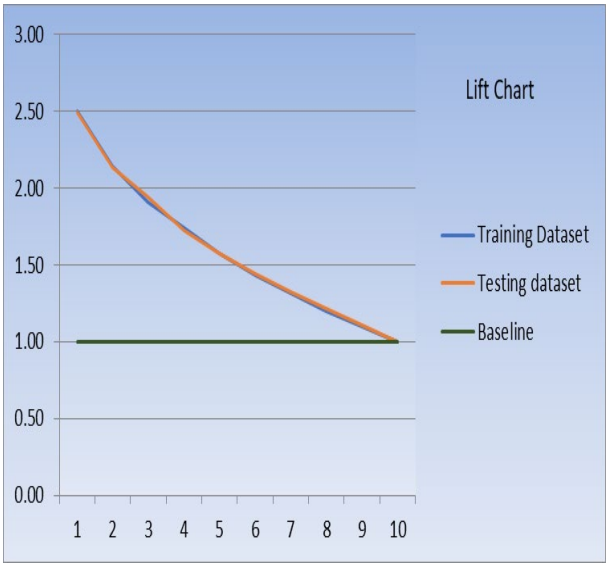
Training Dataset										
Decile	min_prob	max_prob	Good_cou	Bad_count	Bad Rate	Bad%	CummBad	Good%	CummGoo	KS
1	0.371378	0.662395	2549	2100	45%	25%	25%	7%	7%	0.183027
2	0.289624	0.371378	3155	1495	32%	18%	43%	8%	15%	0.278139
3	0.231037	0.2896	3435	1214	26%	14%	57%	9%	24%	0.332457
4	0.186676	0.231034	3618	1032	22%	12%	70%	9%	33%	0.360309
5	0.151847	0.186676	3890	760	16%	9%	79%	10%	44%	0.348648
6	0.123205	0.151847	4032	617	13%	7%	86%	11%	54%	0.31624
7	0.095855	0.123205	4150	500	11%	6%	92%	11%	65%	0.266808
8	0.063718	0.095783	4289	360	8%	4%	96%	11%	76%	0.197065
9	0.033547	0.063718	4397	253	5%	3%	99%	12%	88%	0.111752
10	0.005614	0.033547	4579	71	2%	1%	100%	12%	100%	1.11E-16
			38094	8402						

Testing dataset										
Decile	min_prob	max_prob	Good_cou	Bad_count	Bad Rate	Bad%	CummBad	Good%	CummGoo	KS
1	0.36945	0.649247	1097	895	45%	25%	25%	7%	7%	0.18238
2	0.286674	0.369404	1361	632	32%	18%	43%	8%	15%	0.275285
3	0.229631	0.286572	1435	558	28%	16%	58%	9%	24%	0.343031
4	0.185134	0.229607	1604	389	20%	11%	69%	10%	34%	0.35332
5	0.151498	0.185134	1650	343	17%	10%	79%	10%	44%	0.34797
6	0.123615	0.151498	1716	276	14%	8%	86%	11%	54%	0.319902
7	0.096333	0.123615	1759	234	12%	7%	93%	11%	65%	0.277495
8	0.063663	0.096333	1841	152	8%	4%	97%	11%	76%	0.207208
9	0.033547	0.063663	1912	81	4%	2%	99%	12%	88%	0.112784
10	0.006304	0.033547	1966	27	1%	1%	100%	12%	100%	1.11E-16
			16341	3587						

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

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 4 deciles which is one of the indications of a good model. Our model is able to capture almost 70% of bad loan accounts in first 4 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.

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	0.695073
TNR	0.665249
FPR	0.334751
Precision	0.314114
Accuracy	0.670638

Confusion Matrix(cutoff-0.18667576)			
		Predicted	
		0	1
Actual	0	25342	12752
	1	2562	5840

AUC	0.742
Ginni	0.484

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

Scorecard preparation

The final stage of this process is scorecard preparation using a standard logistic regression algorithm to estimate model parameters. We use coefficients generated from logistic regression model to calibrate the credit score. Here is the how it works:

Scaling – Factor

Factor is one of two scaling parameters used during scorecard calculation process. Factor can be expressed as:

- Factor = pdo / ln(2).

Where: pdo is points to double the odds – parameter given by the user

Scaling – Offset

Offset is one of two scaling parameters used during scorecard calculation process. Offset can be expressed as:

- Offset = Score – (Factor * ln(Odds))

Where: Score is "Scoring value for which you want to receive specific odds of loan repayment – parameter given by user
Odds is "odds of the loan repayment for specific scoring value – parameter given by user
Factor is "scaling parameter as calculated above"

Scorecard preparation – Scaling –Calculating score (WoE coding)

When WoE coding is selected for given characteristic, score for each bin (attribute) of such characteristic is calculated as:

$$Score = \left(\beta \cdot WoE + \frac{\alpha}{m} \right) \cdot factor + \frac{offset}{m}.$$

Where:
β logistic regression coefficient for characteristics that owns the given attribute
α logistic regression intercept term
WoE Weight of Evidence value for given attribute
m number of characteristics included in the model
factor scaling parameter based on formula presented previously
offset scaling parameter based on formula presented previously

Computation Details

Note: After computation is complete the resulting value is rounded to the nearest integer value.