Let’s check the character of the Data set.

1. we are going to draw statistical inferences.
2. Type of the independent data set
3. Missing Values in the data set
4. Multivariate and Bivariate analysis

* This dataset has 22 independent variable
* ID column is not any impact on the prediction it is a unique key
* This Dataset have no missing values
* This Dataset has some categoric variable which marked as int type need to convert into string like SEX ,MERRAGE
* Numeric continuous variables have suspicious outcome with respect to information value. All these variables need to delt with

77% of customer falls under non default category and 22% customer falls under default category

Overall 60% customer are Female and 40% are male category and out of 60% female 24% defaulted and out of 40% male 20% defaulted. Similar distribution found under this category

Majority of customer are graduated or above. There are other educational qualification but distribution is very minimal. All category more than 3 will be merged as these has very minimal contribution on the dataset.

52% unmarried customer and 45% married customer are using the credit card. Bad distribution is equal across this category. As 20% of bad customer lies under this two category so data cannot explain which segment most vulnerable for default

Distribution plotted with respect to target class. Both Good and Bad class has right skewed and med point found at 34. This describes maximum credit card issued for lower aged people. Bad distribution for age slightly flatterer with respect to good class

BILL\_AMT and PAY\_AMT are the bill amount and pay amount for the individual customer. Both independent variables are given for 6 consecutive months. The corelation checked with each other for both Bad and Good class. BILL\_AMT has high correlated with each other as the nature of the variable explains that the BILL\_AMT has an cumulative effect on each other for all across 6 months. This can lead to multicorrelation effect and can be a cause of mis judgement. Few variables needed to be dropped base on the predictive nature

For variable selection and feature selection I am considering Weight Of Evidence and Information value for primary criteria for choosing the independent variable

AGE: This independent variable has very less predictive power. Age group between 29 to 46 have more mass as most of the customer falls under this bucket(We saw before this is right skewed distribution). We can split this group further to get a uniformity and gather more WOE and IV from the bins

BILL\_AMT1 is not well distributed and need further split of the bins

Smaple1: I have applied the woe created from the initial bins

Sample2: This sample is created with the adjusted woe

Sample3: This sample is created from the original dataset I,e data without imputing woe values

There are multiple metrics by which the model performance can be evaluated. I have chosen accuracy score, precision score, recall score , f-score and auc score , Lets discuss with respect to sample :

For Sample1 or S1:

* Logistic Regression is giving accuracy of 0.819667 in train and 0.817222 in test, so it is clear that model is not over/under fit. Precision score is 0.672491 and 0.675222 for train and test respectively. It means ~67% class are actually positive out of all correctly predicted class( TP,TN). Recall score are 0.356989 and 0.3420 respectively. It means out of all ~35% positive class are truly predicted out of all positive class
* Decision Tree and Random Forest is giving better accuracy than logistic regression but test accuracy is not matching with test accuracy. So there has a overfitting issue, Precision and recall score also speak the same.
* KNN is giving me better accuracy then logistic regression 0.857429, but test set is 0.777111. though it also has overfitting issue but minimal than Decision Tree and Random Forest. Precision score is 0.777111 but this not incorporate with the test data, score achieved 0.498031. Recall score for test 0.3795 achieved with respective to 0.563632 achieved in train.
* Ada Boost accuracy is 0.819143 achieved in train set and 0.814333 in test. Clearly model is not over fit. Precision score is 0.684095 and 0.674814 for train and test respectively. Recall score is 0.335850, 0.3175 for train and test.

Noe out of the all algorithm Logistic Regression and Ada Boost has more robust result then then others .

For Sample1 or S1:

Similar result achieved from S2. So, this is clear binning adjustment is not adding up any special value

For Sample1 or S3:

S3 is the original dataset or data value not altered by woe bins. Logistic regression is giving very poor accuracy with 0.779143 and 0.777778 for train and test dataset. Precision and recall value is 0.000 so its means all class are predicted as negative . This module is not considered as an effective algorithm.

Ada Boost is giving is constant result as S! or S2 with respect to accuracy and Precision and recall score.

So I am choosing Logistic regression and Ada Boost for the next phase of modelling

Above confusion metrics tells that out of 2000 bad only 616 bad and 6707 good are predicted correctly. There are 1384 predicted incorrect as good and 293 predicted incorrect as Bad

From the logistic regression summary there are few parameters by which we can determine the best optimum model. Initially all the independent are considered as input variable. From the summary, few independent variables like ‘ PAY\_AMT5\_woe’,are not adding up values as P values of those variables are greater than 0.05(significance label), So all those variables will be drop in next phase of iteration. Notably this has been seen that accuracy, Precision and recall values are respectively better than other models. All these metrices are calculated considering 0.5 as threshold value. So to get the optimum value I will check accuracy, precision value with respect to different probability threshold value

After dropping the variables, the achieved result has been improved. Accuracy , precision, recall auc and gini index has been checked with respect to different threshold. I have applied probability 0.10 to 0.50 and 0.26 came out the optimal for the modelling with accuracy 0.787889, precision 0.522403, recall 0.5305, ruc 0.695964 and gini 0.391929. This is clearly seen that recall value has improved from 0.35 to 0.53 and AUC from 0.64 to 0.69

Same has been repeated with Ada boost algorithm. But result is not satisfactory as logistic regression. Number of Bad class prediction is very less so Precision and Recall score also came out not satisfactory and only limited probability band works here.