CredX Risk Analytics Approach Document

Batch : Jun-2018

Aditya Kumar
Raghuram Krishnamurthy
Rohit Saini
Veenu Bhanot

Business Objective

CredX is a leading credit card provider that gets thousands of credit card applicants every year. But in the past few years, it has experienced an increase in credit loss.

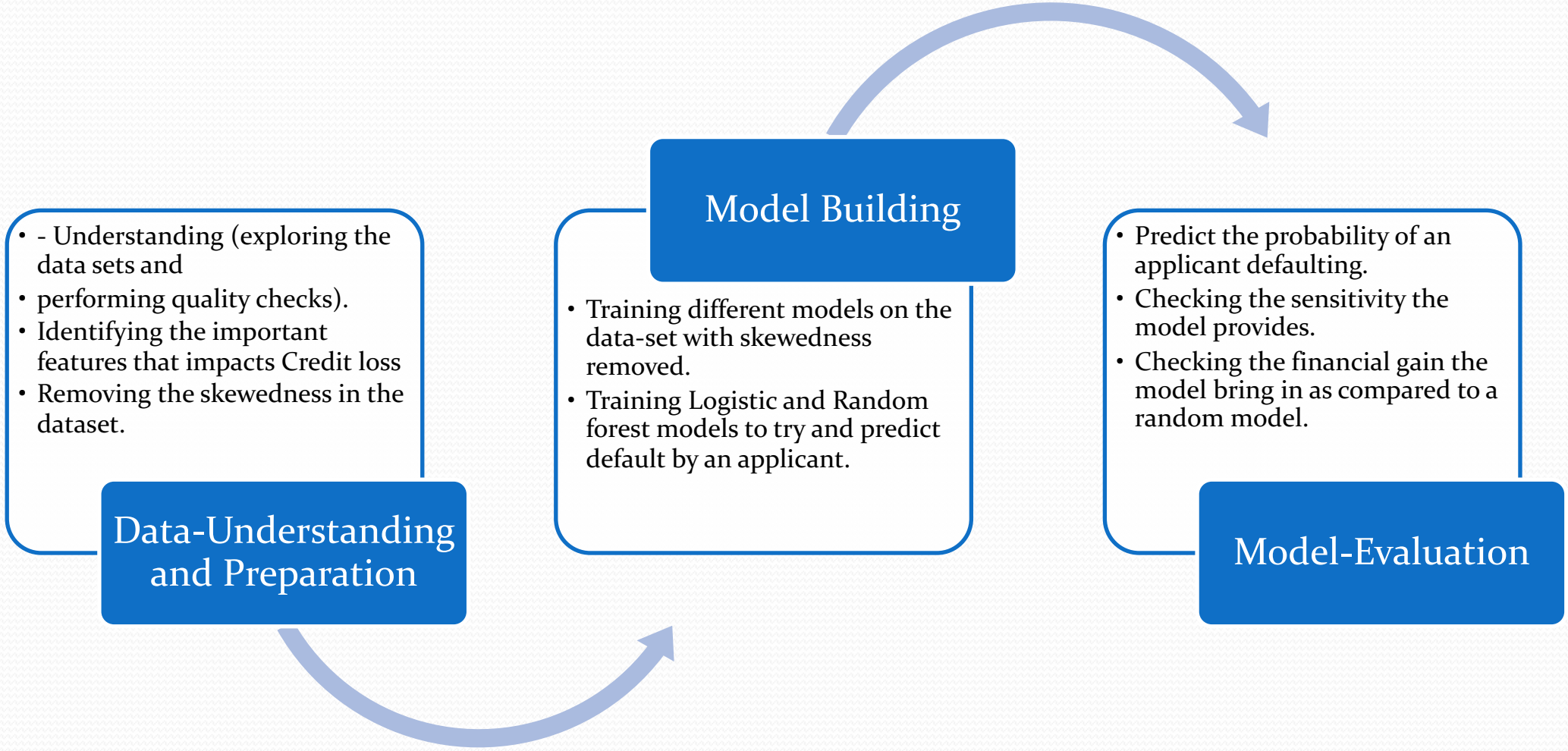
In this project, we will help CredX:

- Identify the right customers using predictive models.
- Using past data of the bank's applicants, we need to determine the factors affecting credit risk, create strategies to mitigate the acquisition risk.
- Assess the financial benefit of the project.

The datasets available to us are:

- *Demographic/application data*: This is obtained from the information provided by the applicants at the time of credit card application.
- *Credit bureau*: This is taken from the credit bureau.

Analysis Steps Followed



Data Understanding, Assumptions and EDA Summary

- Data is available in a structured format(csv), in two files for demographic and credit-bureau data.
- The Grain of data is at an individual customer level.
- There are duplicate Application IDs with different features in both demographic data and credit data.
- There are NA values in 'No. of dependents' and 'Performance Tag' in demographic data set.
- There are NA values in 'CC Utilization in 12 months', 'Trades opened in last 12 months', 'Open home loan', 'Outstanding balance' and 'Performance tag' in Credit bureau data set which has been imputed with WOE values.
- There are Age values in demographic data set that are less than 18, and has been dropped as erroneous data.
- There are negative values in Gender, MaritalStatus, No_of_dependents, Income, Education, Profession, Residence_type, No_of_mons_curr_residence, No_of_mons_curr_company and Performance_Tag which are dropped as erroneous data.
- The distribution of data in Performance tag is very skewed, the case of default is only around 4% of the data-set. The data is imbalanced.
- The inner join set of demographic and credit data is used for further data-preparation.

Data Preparation

Missing value treatment in Demographic Data	- The data points with missing values in the demographic data were dropped from the data set as they were only around 1% of the data set.
Erroneous Value Treatment	- The erroneous data points like, the negative value on income, age less than 18 , duplicate application IDs were removed from the data set as they were around 1% of the data.
Missing value treatment in Credit-bureau Data	- The data points with missing values in the credit-bureau data were replaced by WOE values.

Imbalance in the Data-set

- The data set has an inherent imbalance in the Performance Tag variable (3:66), the two categories available(Defaulters and non-defaulters) are to be predicted
- Due to defaulting being a rare case event, there is so much skewness in the dataset.
- This imbalance would impact the training of a model as the entropy would not change much when splitting on a variable that is a predictor of the defaulters case.
- Accuracy can not be used to evaluate the model as majority of the data is biased towards non-defaulters case.
- The combination of Sensitivity and Specificity or F1 score can be used to evaluate a model.

Overcoming Imbalance

The imbalance in the data-set is overcome using the following strategy:

- The data chunk of the non-defaulters is separated from the defaulters case.
- The non-defaulters data set is clustered into ‘n’ clusters.
- A random sample of datapoints is grabbed from each cluster formed, grabbing as many data-points as the defaulters case contained, from each cluster, hence down scaling the non-defaulters data-set.
- The distribution of defaulters and non-defaulters are now less skewed.

Important Predictors

The Information Value associated with the predictors are as follows, Hence the most strong predictors of credit loss are marked in green:

"No_of_PL_trades_opnd_L12M"	0.2685437	"Income"	0.03625863
"No_of_Inq_ex_HLAL_L12M"	0.2618248	"No.of.months.in.current.company"	0.02648455
"No_of_trades_opnd_L12M"	0.2532209	"woe.Presence_of_opn_HL.binned"	0.01716079
"woe.Avg_CC_Util_L12M.binned"	0.2478046	"woe.Outstanding_Bal.binned"	0.0146238
"No_of_30dpd_L6M"	0.234797	"Age"	0.003083984
"No_of_30dpd_L12M"	0.2142864	"No.of.dependents"	0.002494763
"No_of_PL_trades_opnd_L6M"	0.2122808	"ProfessionSE"	0.002280984
"No_of_90dpd_L12M"	0.2102182	"Presence_of_open_AL"	0.001561284
"No_of_60dpd_L6M"	0.2063916	"Type.of.residenceOthers"	0.0006247534
"Total_No_of_Trades"	0.1865789	"gender_dummy"	0.0003349801
"No_of_60dpd_L12M"	0.1855611	"Type.of.residenceLiving.with.Parents"	0.0001625963
"No_of_Inq_ex_HLAL_L6M"	0.1829944	"marital_status_dummy"	0.0001066424
"woe.No_of_trades_opnd_L6M.binned"	0.1686347	"ProfessionSE_PROF"	7.825915e-05
"No_of_90dpd_L6M"	0.1604187	"Type.of.residenceRented"	4.557127e-05
"No.of.months.in.current.residence"	0.09278225	"EducationMasters"	3.465606e-05
"Type.of.residenceOwned"	1.275249e-06		

Knowing the predictors

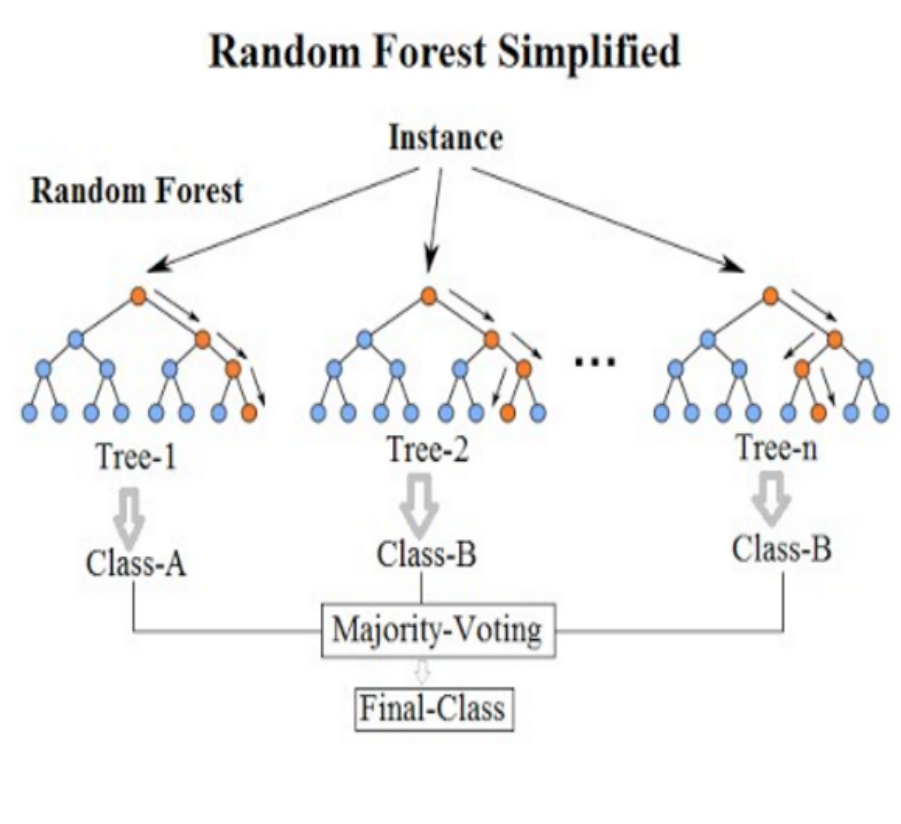
These are the important predictors used in predicting credit-loss in decreasing order of importance:

Feature name	Meaning
No_of_PL_trades_opnd_L12M	No of PL trades in last 12 month of customer
No_of_Inq_ex_HLAL_L12M	Number of times the customers has inquired in last 12 months
No_of_trades_opnd_L12M	Number of times the customer has done the trades in last 12 months
Avg_CC_Util_L12M	Average utilization of credit card by customer
No_of_30dpd_L6M	Number of times customer has not payed dues since 30 days days last 6 months
No_of_30dpd_L12M	Number of times customer has not payed dues since 30 days days last 12 months
No_of_PL_trades_opnd_L6M	No of PL trades in last 6 month of customer
No_of_90dpd_L12M	Number of times customer has not payed dues since 90 days days last 12 months
No_of_60dpd_L6M	Number of times customer has not payed dues since 60 days days last 6 months
Total_No_of_Trades	Number of times the customer has done total trades
No_of_60dpd_L12M	Number of times customer has not payed dues since 60 days days last 12 months
No_of_Inq_ex_HLAL_L6M	Number of times the customers has inquired in last 6 months
No_of_trades_opnd_L6M	No of PL trades in last 6 month of customer
No_of_90dpd_L6M	Number of times customer has not payed dues since 90days in last 6 months

The key Ingredients

The model used here is a Random-Forest model.

- This model is an ensemble of multiple decision trees, all trying to predict the case of a user defaulting, the resultant output is the probability generated by taking into account the output generated from all the incorporating trees.
- Using this model will enable CredX to instantaneously judge the financial-health of an applicant. And will enable CredX to accept or reject applications towards a financial gain.



Model Evaluation Metrics

The model chosen is a Random Forest.

The Model is evaluated on two broad terms:

- The sensitivity of the model, which is the power of the model to grab applicants that might default. Increasing the sensitivity will reduce the Credit loss incurred by CredX.
- The amount of financial gain this model is capable of bringing to CredX.

The sensitivity of the Random Forest model suggested for this problem is:

- *74.5% at 0.15 probability value*

Meaning that *74.5 % of the applicants that might cause credit-loss will be rejected* by the suggested model, where a random model would have rejected 50%.

The model *rejects 10.6% of the applicants that would not have defaulted*, where a random model would have rejected 50%.

Financial Gain from model

- Considering 100 applicants each applying for a credit of 1000 INR
- Out of which 4 applicants would actually Default (Considering only 4% of data were Default cases)
- 96 applicants would Not-default

Random Model	Suggested Random-Forest Model	Gain by suggested Random-Forest model
Since 50% of the actual non-defaulters are rejected, causes missed opportunity revenue of 48000 INR	Since 10.6% of the actual non-defaulters are rejected, causes missed opportunity revenue of 10176 INR	The suggested model reduces the missed opportunity revenue by 37824 INR
Since 50% of the defaulters are rejected, the amount of credit loss incurred is 2000 INR	Since 74.5% of defaulters are rejected, the amount of credit loss is reduced to 1020 INR	The suggested model reduces credit loss by 980 INR

The total financial gain for this use case by using this suggested model is 38804 INR.

The Application Scorecard

- The Random Forest model generates a probability value for any application, indicating the financial health of the applicant and the capability of the applicant to Default and hence causing credit loss.
- An Application-Scorecard is generated based on the results given the Random-Forest model. This score card can be used to easily decide the financial health of an applicant.
- The cut-off Application Score as per the model suggested is 383.6114.
- At the score of 400 the good to bad odds is 10 to 1.
- Every 20 points increase in the score doubles the good to bad odds of the applicant.

Summary

- The Random-Forest Model suggested is evidently a better model than a random model and can be used to reduce credit loss as well as gain opportunity revenue.
- The Application Score can be used to judge the financial health of an applicant.
- Using this suggested model will not only make it easy for CredX to judge the health of a customer but also result in financial gain with no additional man-power involved.