Acquisition Analytics Case Study

Problem Statement:

CredX is a leading credit card provider that gets thousands of credit card applications every year. But in the past few years, it has experienced an increase in credit loss. The CEO believes that the best strategy to mitigate credit risk is to 'acquire the right customers'.

In this project, your task is to help CredX identify the right customers using predictive models. Using past data of the bank's applicants, you need to determine the factors affecting credit risk, create strategies to mitigate the acquisition risk and assess the financial benefit of your project.

Methodology Employed

There were six major steps followed to complete this study as follows:

1) Data understanding and Exploratory data analysis of Demographic and Credit Bureau datasets
2) WOE and Information value analysis
3) Model building using only the Demographic dataset
4) Model building using both demographic and credit bureau dataset
5) Generating application scorecard using the best model
6) Financial benefit analysis.

Now lets see each one of the above one by one.

Data Understanding and Exploratory data analysis of Demographic and Credit Bureau datasets

During this step, we analysed each column of both the datasets one by one and imputed the null values with mean, median or mode wherever applicable.

Also we analysed each column and how it affected the default rate. We performed both univariate and bivariate analysis and following were the summary of the same:

1. Demographic variables are not very good predictors of defaulting. Only below 3 variables seems significant.

-Income

-No.of.months.in.current.residence

-No.of.months.in.current.company

2. credit bureau dataset has many variables which seems like good predictors of defaulters.

-No.of.times.90.DPD.or.worse.in.last.6.months

-No.of.times.60.DPD.or.worse.in.last.6.months

-No.of.times.30.DPD.or.worse.in.last.6.months

-No.of.times.90.DPD.or.worse.in.last.12.months

-No.of.times.60.DPD.or.worse.in.last.12.months

-No.of.times.30.DPD.or.worse.in.last.12.months

-No.of.trades.opened.in.last.6.months

-No.of.PL.trades.opened.in.last.6.months

-No.of.PL.trades.opened.in.last.12.months

After the above EDA was performed, we created a master dataframe by combining both the demographic and credit bureau datasets. While doing this, we further imputed null values of certain columns as applicable and also removed the duplicate applications from the data. We also created a separate dataframe for rejected candidates which had target column (Performance Tag) value as null

After this we moved on to next step of WOE and IV analysis

WOE and IV analysis

During this step we created functions for calculating woe and IV values for each column and eventually replaced original values of master and rejected dataframe with respective woe values.

We also analysed weak predictors in demographic and master datasets by analysing the IV values < 0.02.

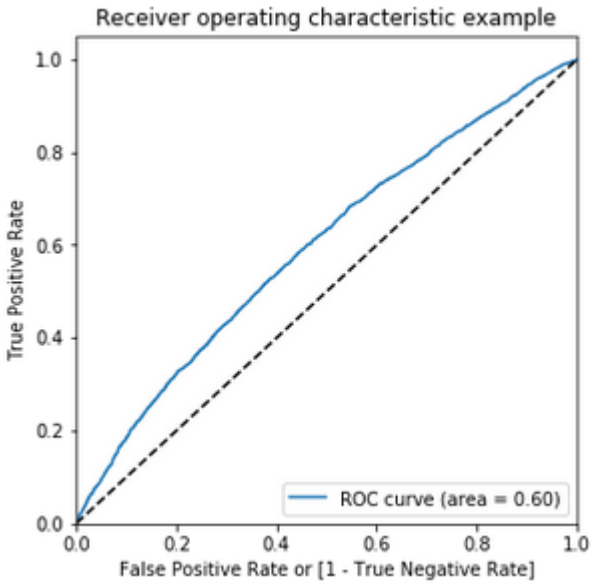
After this we went on to the next step of model building on demographic dataset

Model building using only demographic dataset

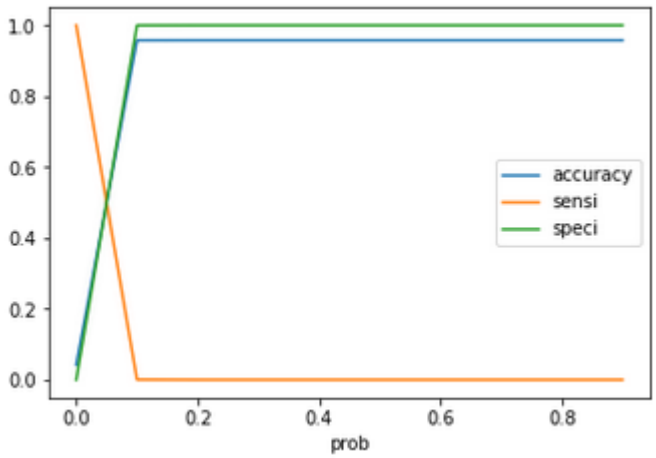
We first extracted the demographic data from woe converted master datset and built four models on this data:

Logistic regression with all variables – this model gave us too many insignificant variables and hence we chose to discard this

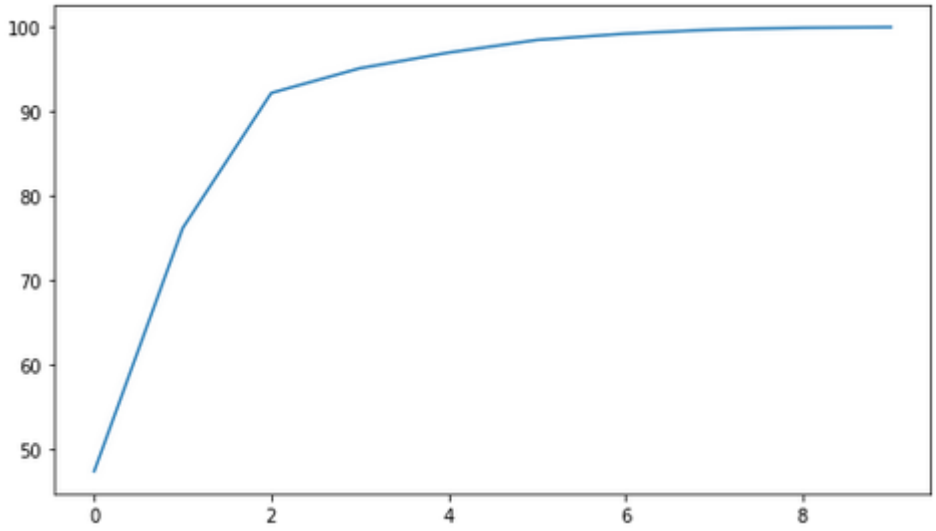
Logistic regression using RFE – This model gave us a ROC curve as below:



By using optimal cutoff as 0.05 as show below, we got accuracy of approx. 73% and sensitivity of approx. 38% which was quite low. Hence we moved on to Logistic regression with PCA.



Logistic regression with PCA



From above we can see that almost 95% variance in dataset can be captured by 4 variables. Hence we built a model with 4 PCs and got a accuracy of 59% and sensitivity of 56%

After this we also observed that this was an imbalanced dataset and hence we tried StartifiedKFOLD but couldn't really improve accuracy and sensitivity from above.

Finally we tried random forest

Random Forest

From random forest, although we got accuracy of approx. 95%, but our AUC score was just 0.5 and sensitivity was almost zero.

Hence looking at all the above models, we concluded that demographic dataset has a very low predictive power and is not a good data to make predictions. this was expected from what we already saw from IV values and EDA.

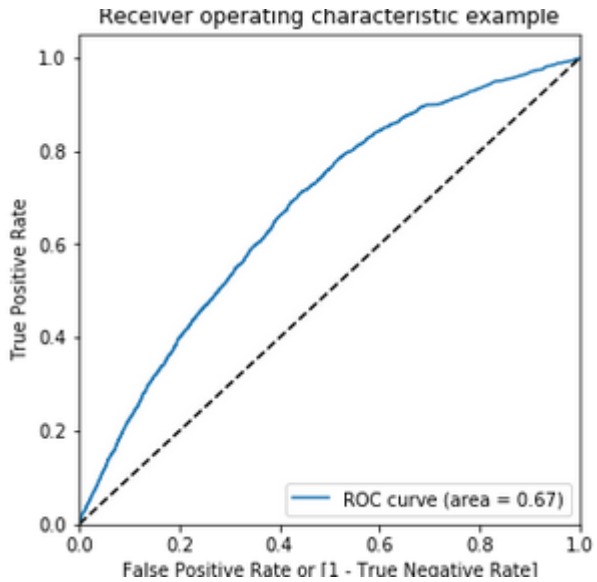
Model building using both demographic and credit bureau dataset

For this we used the master dataset and again built four models to finally settle for the best one:

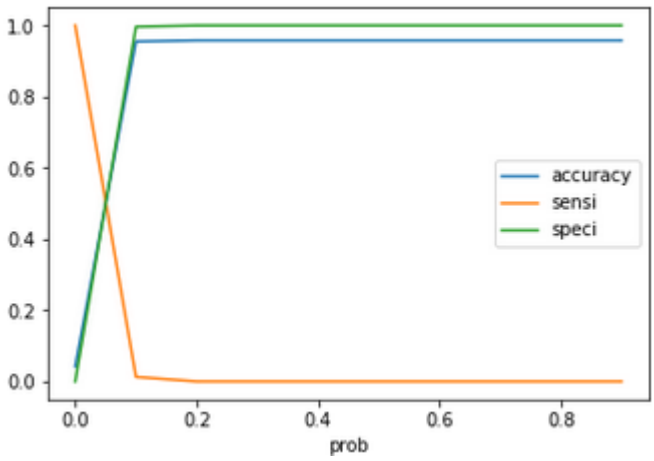
Logistic Regression with all the variables – this again gave us lot of insignificant variables and hence we chose to use RFE and PCA

Logistic regression with RFE

It gave us an ROC curve much better than we got with demographic data

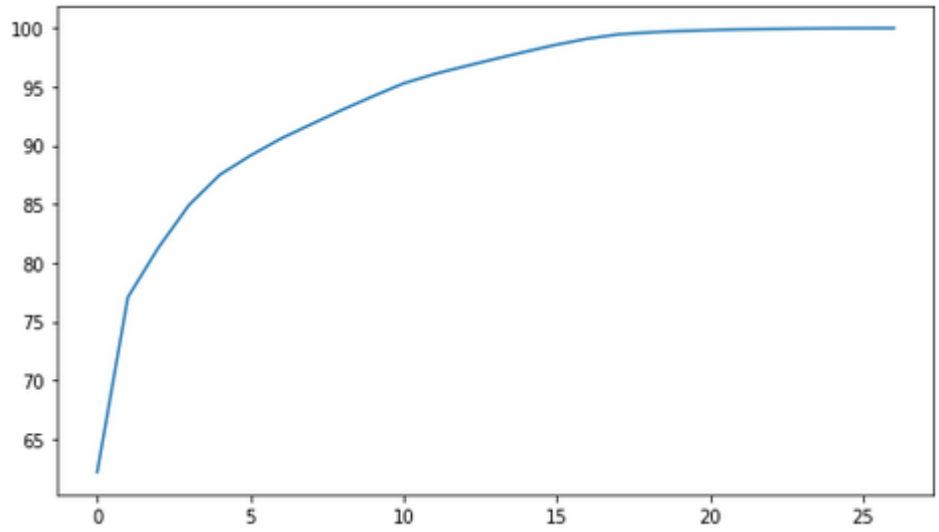


By finding optimal cutoff of 0.05 as shown below, we got an accuracy of 66% and sensitivity of 60% on the test set.



We then tried to improve this further by using PCA

Logistic regression with PCA



Almost 95% of the variance can be captured by 10 variables and hence we went ahead with 10 PCs.

This gave us an auc score of 0.68 and sensitivity of 0.69 on test set which was much better than logistic regression with RFE.

We tried improving this further with StratifiedKfold and handling imbalanced dataset.

After tuning the hyperparameters, we got an auc of 0.67 and sensitivity of 0.7 on test set which was okay.

We then tried to look at Random forest model

Random Forest

We built random forest with both default parameters and tuned hyperparameters but could not improve upon the scores that we got with Logistic Regression with regularized PCA.

Hence we finalized Logistic regression with regularized PCA as our final model and built our application scorecard based on the same

Generating Application Scorecard using best model

As seen above we used Logistic Regression with regularized PCA as our model to build our application scorecard.

We noticed the below score ranges from these scorecards:

1) Score range for approved candidates – 302 to 364
2) Score range for rejected candidates – 301 to 338
3) Score range of predicted defaulters from accepted candidates – 302 to 333
4) Score range of predicted non-defaulters from rejected candidates – 333 to 338

From above we concluded that our score cut off should be 335.

Financial Benefit assessment

We made two assumptions to calculate the benefits:

1) Bank will be able to retrieve 25% of outstanding balance per defaulter
2) Bank gains 8% of the total outstanding balance as interest income

Based on above, we calculated the following:

1) Average credit loss per default = 942903
2) Total Interest income = 2,577,404,854
3) Total credit loss = 2778735141

Now the financial benefits calculated were as follows:

1) Total savings by rejecting defaulters = 1891469436
2) Total loss of interest by rejecting non defaulters = 1068703452
3) If we use the model for auto approval/rejection, total saving = 822,765,984