Problem Statement

Business understanding

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, our task is to 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 and assess the financial benefit of our project.

Understanding the data

There are two data sets in this project: demographic and credit bureau data.

- Demographic/application data: This is obtained from the information provided by the applicants at the time of credit card application. It contains customer-level information on age, gender, income, marital status, etc.
- Credit bureau: This is taken from the credit bureau and contains variables such as 'number of times 30 DPD or worse in last 3/6/12 months', 'outstanding balance', 'number of trades', etc.

Both files contain a performance tag, which indicates whether the applicant has gone 90 days past due (DPD) or worse in the past 12 months (i.e. defaulted) after getting a credit card.

In some cases, you will find that all the variables in the credit bureau data are zero and credit card utilisation is missing. These represent cases in which there is a no-hit in the credit bureau. You will also find cases with missing credit card utilisation. These are the cases in which the applicant does not have any other credit card.

Data cleaning and preparation

Create a master file with all the relevant variables and perform the necessary data quality checks and cleaning. In credit risk analytics, the weight of evidence (WOE) (and, equivalently, information value analysis) is often used to identify the important variables. Apart from assessing variable importance, WOE is also used to impute missing values from the data. You'll note that some variables contain a significant number of missing values. Replace the actual values of all the variables by the corresponding WOE value and store the data in a separate file (e.g. woe_data) for further analysis.

Model-building

The two types of models you need to build are as follows:

- **Demographic data model:** Build a model to predict the likelihood of default using only the demographic data. This will give you a good idea about the predictive power of the application data. Obviously, the final model will use the credit bureau data as well, though this model is an important part of understanding the predictive power of application data.

- **Model using both demographic and credit bureau data:** Build a model to predict default using both the data sets. You may choose any type of model, though it is recommended to start with a logistic regression model first. Further, you can choose any type of model.

Model evaluation

Evaluate the models using relevant metrics and report the results. As part of model validation, predict the likelihood of default for the rejected candidates and assess whether the results correspond to your expectations.

Application scorecard

Build an application scorecard with the good to bad odds of 10 to 1 at a score of 400 doubling every 20 points.

- For the rejected population, calculate the application scores and assess the results. Compare the scores of the rejected population with the approved candidates and comment on the observations.
- On the basis of the scorecard, identify the cut-off score below which you would not grant credit cards to applicants.

Assessing the financial benefit of your project

You need to assess and explain the potential financial benefit of your project to the bank's management. From a P&L perspective, identify the metrics you are trying to optimise, explain (in simple terms) how the analysis and the model work, and share the results of the model. Finally, assess the financial benefit of the model and report the following:

- The implications of using the model for auto-approval or rejection, i.e., how many applicants on an average would the model automatically approve or reject
- The potential credit loss avoided with the help of the model
- Assumptions based on which the model was built

Make appropriate assumptions about the numbers wherever needed (e.g., the potential average credit loss per default, etc.). Present your analysis and recommendations in a PowerPoint presentation.