**Problem Statement**

**Please note that there are 2 submissions for the capstone project with different deadlines. Please check the details for the 2 submissions in the subsequent pages.**

**Business understanding**

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. The CEO believes that the best strategy to mitigate credit risk is to ‘acquire the right customers’.

In this project, you will 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.

**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.

You can find the **data dictionary**of both data sets [here](https://docs.google.com/spreadsheets/d/168hlx2UtNLLLw6EAqa8AsO0FveFaDG74uTt-lkT-kHE/edit?usp=sharing) and download the data sets from below.

**[Demographic Data](https://cdn.upgrad.com/UpGrad/temp/a790b665-7819-4490-87d9-b046340095fd/Demographic%20data.csv" \o "Demographic data.csv" \t "_blank)**

[file\_download](https://cdn.upgrad.com/UpGrad/temp/a790b665-7819-4490-87d9-b046340095fd/Demographic%20data.csv" \o "Demographic data.csv" \t "_blank)**[Download](https://cdn.upgrad.com/UpGrad/temp/a790b665-7819-4490-87d9-b046340095fd/Demographic%20data.csv" \o "Demographic data.csv" \t "_blank)**

**[Credit Bureau Data](https://cdn.upgrad.com/UpGrad/temp/3fa11a2f-f702-45e4-a7fd-17110148588d/Credit%20Bureau%20data.csv" \o "Credit Bureau data.csv" \t "_blank)**

[file\_download](https://cdn.upgrad.com/UpGrad/temp/3fa11a2f-f702-45e4-a7fd-17110148588d/Credit%20Bureau%20data.csv" \o "Credit Bureau data.csv" \t "_blank)**[Download](https://cdn.upgrad.com/UpGrad/temp/3fa11a2f-f702-45e4-a7fd-17110148588d/Credit%20Bureau%20data.csv" \o "Credit Bureau data.csv" \t "_blank)**

Both files contain a **performance tag**which represents whether the applicant has gone 90 days past due 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 credit card utilisation missing. 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 conduct 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.

You can read about**WOE and information value** **analysis** from the links provided in the **additional resources** section.

**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 of 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 a 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 management of the bank. From a P&L perspective, identify the metrics you are trying to optimise, explain (in simple terms) how the analysis and the model works, 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 has been 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. 

**Resources for the project**

* [Read more about Weight of evidence (Woe) and Information value (IV) Analysis](http://www.listendata.com/2015/03/weight-of-evidence-woe-and-information.html)
* [IV and Woe in R](https://cran.r-project.org/web/packages/Information/vignettes/Information-vignette.html)

**Mentorship**

Given below is the schedule for the mentorship sessions and the milestone you are expected to achieve before each session.

| Mentorship Session | Agenda |
| --- | --- |
| 3rd Feb | Understanding of the business problem, data understanding and structure of the solution. The approach to EDA and the type of models to be built may also be discussed. |
| 10th Feb | Results of EDA and model evaluation and techniques to improve the same. |
| 24th Feb | Choice of final model, iterations to improve the model and application scorecard. |
| 10th March | Financial benefit analysis and presentation to the business. |

**Do's and Don'ts for Mentorship Sessions**

* Be prepared for the session as per the agenda prescribed above
* You would need to drive these sessions and extract the most out of these sessions.
* Ask relevant and worthwhile queries that help you in clarifying the doubts
* Do not push the mentors to provide the solution
* Do not ask the mentors to review your R-codes
* Respect the time of mentors - if any mentorship session goes unattended by all the members of a group, no further mentor support will be available to the group

**Evaluation Rubric**

|  |  |  |
| --- | --- | --- |
| **Stage** | **Meets expectations** | **Does not meet expectations** |
| Data understanding, preparation and EDA | All data quality checks are performed and the reasons for missing/inconsistent data are mentioned, all data preparations steps are performed correctly.    Relevant EDA is done using plots and summaries. The insights from EDA should be clearly derived and explained.    The information value of variables is correctly compared and used to identify the important and unimportant variables. | All data quality checks are not performed, all data preparation steps are not/incorrectly performed. EDA is not conducted in sufficient detail/WOE analysis is incorrect or not performed. The information value interpretation is not done/is incorrect. Visualisations are not used. |
| Model building | Model hyperparameters are tuned using correct principles and the approach is explained clearly. Correct variable selection techniques are used.    A reasonable number of different models are attempted and the best one is chosen based on key performance metrics.    Model evaluation is conducted correctly and tests the accuracy, stability and generalisability of the model.    Model evaluation metrics are at par with the best possible models on this data set. The rejected population is used to assess model performance.    The application scorecard is correctly built using the final model and a suitable score cut-off is identified. The scores of rejected population are compared with those of the approved ones. | Model hyperparameters are tuned using correct principles of model selection. Interpretation of models is incorrect or unclear.    Correct evaluation metrics are not used and reported.    The final evaluation metrics are not at par with the best possible models on this data set.    The application scorecard is incorrectly built or a suitable cut-off is not identified.    The scores of rejected and approved population are not compared and interpreted. |
| Financial benefit assessment | The analysis is performed in simple terms and includes appropriate realistic assumptions.    Should include the implications of using the model, the incremental benefit in terms of credit loss avoided, and the potential loss of revenue due to rejection of good customers. | Incorrect or unrealistic assumptions are made, the benefits and potential losses are not analysed correctly. |
| Presentation of results | The presentation has a clear structure, is not too long, and explains the most important results concisely. The final recommendations include a brief explanation of the important variables, the model, the implications of using the model, and the financial assessment.    Overall, a clear positive impact on the business should be apparent. | The presentation lacks structure, is too long or does not put emphasis on the important observations. Does not include important variables, their interpretation or financial benefit analysis. |

**Mid-Submission**

**Note:**You are supposed to make two submissions as mentioned below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Submission** | **Deliverable** | **Weightage** | **Date** |
| 1 | Explain the approach to solve the problem in detail - describe how you have subsetted and cleaned the data, identified important predictor variables using EDA, etc. Also, explain how you will build and evaluate the models, choose the best one, and create an application scorecard using the model.  No R Code required. **Approach paper which also includes the future roadmap to be submitted in PDF format** | 25% | 17th  February 2019 |
| 2 | Final Submission - R code, final PPT **(In PDF format)** to the management | 65% | 17th March  2019 |