Problem Statement

Introduction

Solving this assignment will give you an idea about how real business problems are solved using EDA. In this case study, apart from applying the techniques you have learnt in EDA, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimise the risk of losing money while lending to customers.

Business Understanding

You work for a **consumer finance company** which specialises in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. Two **types of risks** are associated with the bank's decision:

- If the applicant is **likely to repay the loan**, then not approving the loan results in a **loss of business** to the company
- If the applicant is **not likely to repay the loan,** i.e. he/she is likely to default, then approving the loan may lead to a **financial loss** for the company

The data given below contains the information about past loan applicants and whether they 'defaulted' or not. The aim is to identify patterns which indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

In this case study, you will use EDA to understand how **consumer attributes** and **loan attributes** influence the tendency of default.

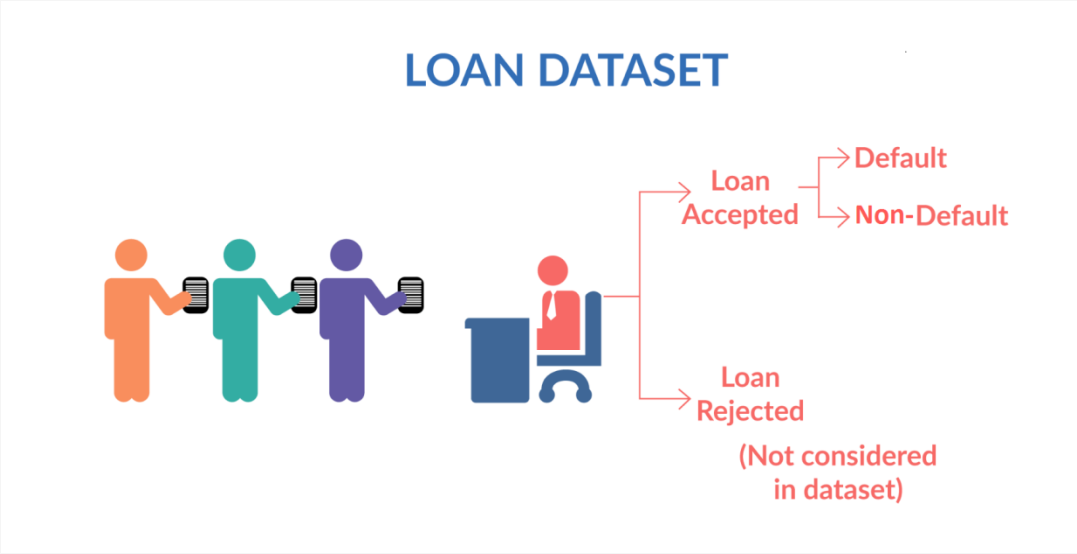

Figure 1. Loan Data Set

When a person applies for a loan, there are **two types of decisions** that could be taken by the company:

1. **Loan accepted:** If the company approves the loan, there are 3 possible scenarios described below:
    - **Fully paid**: Applicant has fully paid the loan (the principal and the interest rate)
    - **Current**: Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.
    - **Charged-off**: Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has **defaulted** on the loan
2. **Loan rejected**: The company had rejected the loan (because the candidate does not meet their requirements etc.). Since the loan was rejected, there is no transactional history of those applicants with the company and so this data is not available with the company (and thus in this dataset)

Business Objectives

This company is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface.

Like most other lending companies, lending loans to 'risky' applicants is the largest source of financial loss (called credit loss). The credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed. In other words, borrowers who **default** cause the largest amount of loss to the lenders. In this case, the customers labelled as 'charged-off' are the 'defaulters'.

If one is able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants using EDA is the aim of this case study.

In other words, the company wants to understand the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

To develop your understanding of the domain, you are advised to independently research a little about risk analytics (understanding the types of variables and their significance should be enough).

Data Understanding

Loan Dataset contains the complete loan data for all loans issued through the time period 2007 t0 2011.

You can access the data dictionary which describes the meaning of these variables from the Data Dictionary file.

Results Expected

1. Write all your code in one well-commented Python file; briefly mention the insights and observations from the analysis
2. Present the overall approach of the analysis in a presentation
    - Mention the problem statement and the analysis approach briefly
    - Explain the results of univariate, bivariate analysis etc. in business terms
    - Include visualisations and summarise the most important results in the presentation

You need to submit one Ipython notebook which clearly explains the thought process behind your analysis (either in comments of markdown text), code and relevant plots.

Evaluation Rubric

Criteria	Meets expectations	Does not meet expectations
Data understanding (5)	All data quality issues are correctly identified and reported. Wherever required, the meanings of the variables are correctly interpreted and written either in the comments or text.	Data quality issues are overlooked or are not identified correctly such as outliers, missing values and other data quality issues. The variables are interpreted incorrectly or the meaning of variables is not mentioned.
Data Cleaning and Manipulation (10)	Data quality issues are addressed in the right way (missing value imputation, outlier treatment and other kinds of data redundancies, etc.).	Data quality issues are not addressed correctly. The variables are not converted to an appropriate format for analysis. String and date manipulation is not

Criteria	Meets expectations	Does not meet expectations
	If applicable, data is converted to a suitable and convenient format to work with using the right methods. Manipulation of strings and dates is done correctly wherever required.	done correctly or is done using complex methods.
Data analysis (20)	The right problem is solved which is coherent with the needs of the business. The analysis has a clear structure and the flow is easy to understand. Univariate and segmented univariate analysis is done correctly and appropriate realistic assumptions are made wherever required. The analyses successfully identify at least the 5 important driver variables (i.e. variables which are strong indicators of default). Business-driven, type-driven and data-driven metrics are created for the important	The analyses do not address the right problem or deviate from the business objectives. The analysis lacks a clear structure and is not easy to follow. The univariate and bivariate analysis is not performed in sufficient detail and thus some crucial insights are missed out. The analyses are not able to identify enough important driver variables. New metrics are not derived wherever appropriate. The explanation for creating the derived metrics is either not mentioned or the metrics are not reasonable.

Criteria	Meets expectations	Does not meet expectations
	variables and utilised for analysis. The explanation for creating the derived metrics is mentioned and is reasonable. Bivariate analysis is performed correctly and is able to identify the important combinations of driver variables. The combinations of variables are chosen such that they make business or analytical sense. The most useful insights are explained correctly in the comments. Appropriate plots are created to present the results of the analysis. The choice of plots for respective cases is correct. The plots should clearly present the relevant insights and should be easy to read. The axes and important data points are labelled correctly.	Derived metrics are not analysed correctly/are insufficiently utilised. Important insights are not mentioned in the report or the Python file. Relevant plots are not created. The choice of plots is not ideal and the plots are either difficult to interpret or lack clarity or neatness. Relevant insights are not clearly presented by the plots. The axes and important data points are not labelled correctly/neatly.
Presentation and	The presentation has a clear	The presentation lacks structure, is

Criteria	Meets expectations	Does not meet expectations
Recommendations (10)	structure, is not too long, and explains the most important results concisely in simple language. The recommendations to solve the problems are realistic, actionable and coherent with the analysis. If any assumptions are made, they are stated clearly.	too long or does not put emphasis on the important observations. The language used is complicated for business people to understand. The recommendations to solve the problems are either unrealistic, non-actionable or incoherent with the analysis. Contains unnecessary details or lacks the important ones. Assumptions made, if any, are not stated clearly.
Conciseness and readability of the code (5)	The code is concise and syntactically correct. Wherever appropriate, built-in functions and standard libraries are used instead of writing long code (if-else statements, for loops, etc.). Custom functions are used to perform repetitive tasks. The code is readable with appropriately named variables and detailed comments are	Long and complex code used instead of shorter built-in functions. Custom functions are not used to perform repetitive tasks resulting in the same piece of code being repeated multiple times. Code readability is poor because of vaguely named variables or lack of comments wherever necessary.

Criteria	Meets expectations	Does not meet expectations
	written wherever necessary.	
VIVA (10)		

INSTRUCTIONS

1. Evaluation rubrics is given at the end of the case study kindly go through it properly.
2. Select a Group Leader among yourself who will be solely responsible for all submissions.
3. Each group has to do one submission, to be done by group leader.
4. Create a ZIP folder containing your Jupyter Notebook and PPT and upload on https://forms.gle/BM192TqEAZ2Aym7y5
5. Submission of case study data:

Class and Group	Date of Submission
18AITAIML2 Group A	04.Mar.2020 (before 2:40 P.M.)
18AITAIML2 Group B	06.Mar.2020 (before 9:30 A.M.)
18AITAIML3	05.Feb.2020 (before 9:30 A.M.)

If you fail to submit the documents on time, 50% of your marks obtained will be deducted.