Introduction

This assignment aims to give you an idea of applying EDA in a real business scenario. In this assignment, apart from applying the techniques that you have learnt in the EDA module, 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

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it to their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specialises in lending various types of loans to urban customers. You have to use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

When the company receives a loan application, the company has to decide 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 the loan application at the time of applying for the loan. It contains two types of scenarios:

- The client with payment difficulties: he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample,

- All other cases: All other cases when the payment is paid on time.

When a client applies for a loan, there are four types of decisions that could be taken by the client/company):

- **Approved:** The Company has approved loan Application

- **Cancelled:** The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client he received worse pricing which he did not want.

- **Refused:** The company had rejected the loan (because the client does not meet their requirements etc.).

- **Unused offer:** Loan has been cancelled by the client but on different stages of the process.

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

Business Objectives

This case study aims to identify patterns which indicate if a client has difficulty paying their installments 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. This will ensure that the consumers capable of repaying the loan are not rejected. 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 utilize 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).

This dataset has 3 files as explained below:

1. 'application_data.csv' contains all the information of the client at the time of application. The data is about whether a client has payment difficulties.

2. 'previous_application.csv' contains information about the client's previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.

3. 'columns_description.csv' is a data dictionary which describes the meaning of the variables.

Results Expected by Learners

- Present the overall approach of the analysis in a presentation. Mention the problem statement and the analysis approach briefly.

- Identify the missing data and use appropriate methods to deal with it. (Remove columns/or replace it with an appropriate value)

- Hint: Note that in EDA, since it is not necessary to replace the missing value, but if you have to replace the missing value, what should be the approach. Clearly mention the approach.

- Identify if there are outliers in the dataset. Also, mention why you think it is an outlier. Again, remember that for this exercise, it is not necessary to remove any data points.

- Identify if there is data imbalance in the data. Find the ratio of data imbalance.

- Hint: How will you analyse the data in case of data imbalance? You can plot more than one type of plot to analyse the different aspects due to data imbalance. For example, you can choose your own scale for the graphs, i.e. one can plot in terms of percentage or absolute value. Do this analysis for the 'Target variable' in the dataset (clients with payment difficulties and all other cases). Use a mix of univariate and bivariate analysis etc.

- Hint: Since there are a lot of columns, you can run your analysis in loops for the appropriate columns and find the insights.

- Explain the results of univariate, segmented univariate, bivariate analysis, etc. in business terms.

- Find the top 10 correlation for the Client with payment difficulties and all other cases (Target variable). Note that you have to find the top correlation by segmenting the data frame w.r.t to the target variable and then find the top correlation for each of the segmented data and find if any insight is there. Say, there are 5+1(target) variables in a dataset: Var1, Var2, Var3, Var4, Var5, Target. And if you have to find the top 3 correlation, it can be: Var1 & Var2, Var2 & Var3, Var1 & Var3. Target variable will not feature in this correlation as it is a categorical variable and not a continuous variable which is increasing or decreasing.

Include visualisations and summarise the most important results in the presentation. You are free to choose the graphs which explain the numerical/categorical variables. Insights should explain why the variable is important for differentiating the clients with payment difficulties with all other cases.

Evaluation Rubrics

Criteria	Meets expectations	Does not meet expectations
Data understanding (20%)	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 missing values, outliers 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 analysis and other kinds of data redundancies, etc.). 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	Data quality issues are not addressed correctly. The variables are not converted to an appropriate format for analysis. String and date manipulation is not done correctly or is done using complex methods
Data analysis (50%)	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	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.

	created for the important 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.
Conclusion and Recommendations (10%)	The presentation has a clear 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.	The presentation lacks structure, is 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 (10%)	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 written wherever necessary.	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.