Business Objectives

Project Brief

You are working for a consumer finance company. 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 to the person results in a loss of business to the company.
- If the applicant is not likely to repay the loan i.e. default, then approving the loan to the person results in a financial loss to the company.

In this case study, we consider only consumers whose loan application is approved. Here, our aim is to understand how consumer attributes and loan attributes influencing the tendency of defaulting.

LO

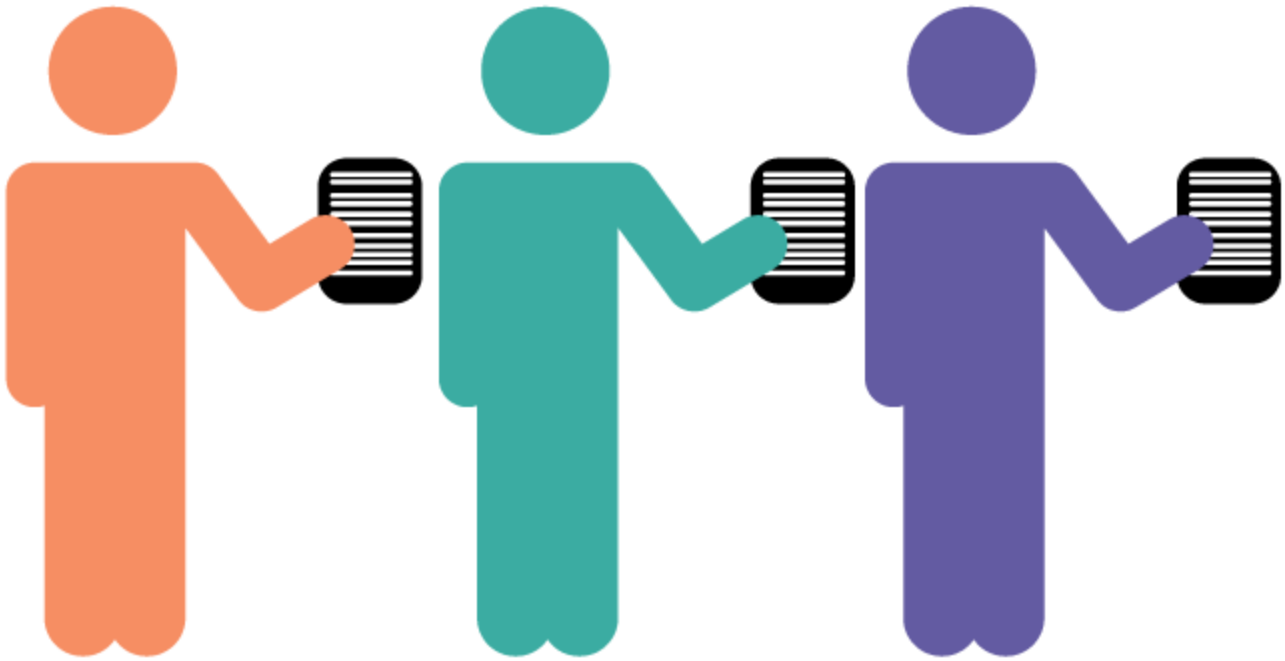

Fig1. Loan Dataset

Business and Data Understanding

1. Business Understanding?

This company is the largest online credit marketplace, facilitating personal loans, business loans, and financing for elective medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface. Investors provide the capital to enable many of the loans in exchange for earning interest.

2. Where did we get the data from?

Let's assume that the company has provided you the data set for analysis. It can be downloaded from the link below. It contains complete loan data for all loans issued through the time period 2007 t0 2011.

Loan Dataset
*file_download*Download

The company has come across some important attributes in order to understand behaviour of their approved loan customers w.r.t. loan default. Thus, the lending company has decided to work only on these variables to mitigate the future risk. The **driver variables** you need to consider for this case study are:

Attributes	**Definition**
annual_inc	Annual Income of applicant
loan_amnt	The listed amount of the loan applied for by the borrower
funded_amnt	The total amount committed to that loan at that point in time
int_rate	Interest Rate on the loan
grade	LC assigned loan grade
dti	*Debt to income ratio*
emp_length	Employment length in years
Purpose	A category provided by the borrower for the loan request.
home_ownership	The home ownership status provided by the borrower during registration

loan_status	Current status of the loan

You can access the data dictionary which describes the meaning of these variables from the provided link below:

Data Dictionary
*file_download*Download

3. What is company's business objective?

The business objectives and goals of data analysis are pretty simple. The company wants to understand the driving factors behind loan default (*loan_status_1)*. The company can utilise this knowledge for its portfolio and risk assessment. Specifically, the company wants to determine which driver variables are having the most influence on the tendency of loan default.

Your goal is divided into 3 main parts:

1. Data Preparation
2. Exploratory Data Analysis
3. Hypothesis testing

Problem Statement

How do we approach the case study? What should we deliver?

The entire case study is divided into three checkpoints to narrow down the problem statement. The checkpoints are interlinked with each other.

Checkpoint-1: Data Cleaning and preparation

- Load the file **loan.csv** into data frame *loan*.
- Impute the NA values for all driver variables.
- Remove rows with *loan_status* = "Fully Paid"
- Create a new column named *loan_status_1* with three levels current_new, default_new and late such that
    - rows with *loan_status* as "current" or "in grace period" are tagged as "current_new"
    - rows with *loan_status* as " default" or "charged off" are tagged as "default_new"
    - rows with *loan_status* as " late 16- 30 days" "late 31-120 days" are tagged as "late"
- Create new bin variables for int_rate and emp_length respectively as follows:
    - Create *int_rate_grp* such that *int_rate* < 10 is tagged "Low"; *int_rate* (10-18) is tagged "Medium"; *int_rate* (>18) is tagged "High"
    - Create *emp_len_grp* such that *emp_length* (0-4) is tagged as "Junior"; *emp_length* (5-8) is tagged as "Mid-level"; *emp_length* (>8) is tagged as "Senior".

Note: *int_rate* and *emp_length* should not be considered for further analysis. Use *int_rate_grp and emp_len_grp instead.*

Checkpoint 2: Exploratory Data Analysis

This is the second part of the analysis.

- **Univariate Analysis** : This analysis will show the distribution of driver variables**.** This analysis shall include the following for all driver variables:
    - Summary Statistics
    - Distribution plot
    - Outlier treatment

- **Multivariate Analysis:** This analysis will show how different variables interact with each other. This includes the following:
    - Finding correlations for all different pairs of **continuous variables** for e.g. *dti* v/s *annual_inc*.
    - Distribution of **all the driver variables** across different levels of two categorical variables:
    3. *loan_status_1:* Make plots to show how the continuous variables vary across the three levels of *loan_status_1; f*or e.g. how *annual_inc* is distributed across the levels **default_new, late and current_new.**
    4. *int_rate_grp:* Make plots to show how the continuous variables vary across the three levels of *int_rate_grp; f*or e.g. how *annual_inc* is distributed across the levels **Low, Medium and High**

Results Expected:

1. R commented file: Should include code for numerical and graphical analysis
2. A Tableau workbook consisting of dashboards and the story: Submit all the plots of univariate and multivariate analysis. The plots to be included will depend on which ones you think are most insightful.

Checkpoint 3: Hypothesis Testing

This is the third part of our analysis. Here you have to analyse if the continuous driver variables have **different mean values** for the two categorical variables *loan_status_1* and *int_rate_grp.*

1. Test hypotheses (95 % conf. level) for two levels of *loan_status_1 - default_new and current_new*
2. Test hypotheses (95 % conf. level) for two levels of *int_rate_grp - high* and *low*

For e.g. consider dti (debt to income ratio): You will calculate group mean of dti for “Current” and “Default” levels. Test the hypothesis (95% significance level) whether two group means are same or not.

Note: You have to run tests for all continuous driver variables.

This would help us understand which variables, in fact, are affecting the tendency to default.

Results Expected:

1. R code

2. Write your insights in attached doc (template provided on the previous page)

Important Note: You are supposed to code entirely in R. You are also required to showcase your visualization results. For every checkpoint, keep writing code in one well-commented R file which you can submit at the end. Please make sure to rename your R script with **"Group_Facilitator_Applicant_Id_main.R".**

Also, you have to write a brief explanation of your hypothesis results in the word document. The template of word doc. is provided here for your reference. Please make sure to rename the file as **Group_Facilitator_Applicant_Id_document.docx**

Hypotheses Insights
*file_download*Download

For Tableau, prepare visuals for the variables which are significant as per your analysis. Combine all visuals into the dashboard (each should contain a maximum of three charts) and prepare a story which would convey your business insights. The assignment would be graded on aesthetic properties of charts (choice of appropriate charts, color, the scale of axes, layout, readability etc.) and how effectively it conveys your insights. Please make sure to name the file as **Group_Facilitator_Applicant_Id_tableau.twbx**

At last, you have to prepare a short PPT to present the results of your analysis to credit manager of the company. This should briefly describe the important results and recommendations. Please name the PPT **"Group_Facilitator_Applicant_Id_main.pptx".** You can download the template of ppt from the provided link below.

Statistics Case Study
*file_download*Download

In total, there are four things you have to submit (along with the percentage of marks distribution):

1. R file with code: - 40%(Correctness of codes -30% & Comments -10%)
2. Powerpoint presentation: 25%

3. Tableau Dashboard: 15%
4. Word documents of your insights for goals 2 and 3: 20 %