You work for a **Fintech** which specializes 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 contains information about past loan applicants and whether they ‘defaulted’ or not. The aim is to identify patterns that 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.

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

* **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 installments, i.e. the tenure of the loan is not yet completed. These candidates are not labeled as 'defaulted'.
  + **Charged-off**: Applicant has not paid the installments in due time for a long period of time, i.e. he/she has defaulted on the loan
* **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.

**Dataset:**

* You have been provided with two datasets:
  + **Training data**: <https://filtereddatasets.s3.amazonaws.com/company-x/train.csv>
    - This should be used for EDA and Training the model
  + **Test data:** <https://filtereddatasets.s3.amazonaws.com/company-x/test.csv>
    - This should be used for generating predictions

**Data Dictionary:**

|  |  |  |
| --- | --- | --- |
| Col. No. | Column Name | Description |
| 0 | loan\_amnt | The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value. |
| 1 | term | The number of payments on the loan. Values are in months and can be either 36 or 60. |
| 2 | int\_rate | Interest Rate on the loan |
| 3 | installment | The monthly payment owed by the borrower if the loan originates. |
| 4 | grade | Assigned loan grade |
| 5 | sub\_grade | Assigned loan subgrade |
| 6 | emp\_title | The job title supplied by the Borrower when applying for the loan\* |
| 7 | emp\_length | Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years. |
| 8 | home\_ownership | The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER |
| 9 | annual\_inc | The self-reported annual income provided by the borrower during registration. |
| 10 | verification\_status | Indicates if income was verified, not verified, or if the income source was verified |
| 11 | issue\_d | The month which the loan was funded |
| 12 | loan\_status | Current status of the loan |
| 13 | purpose | A category provided by the borrower for the loan request. |
| 14 | title | The loan title provided by the borrower |
| 15 | zip\_code | The first 3 numbers of the zip code provided by the borrower in the loan application. |
| 16 | addr\_state | The state provided by the borrower in the loan application |
| 17 | dti | A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested loan, divided by the borrower's self-reported monthly income. |
| 18 | earliest\_cr\_line | The month the borrower's earliest reported credit line was opened |
| 19 | open\_acc | The number of open credit lines in the borrower's credit file. |
| 20 | pub\_rec | Number of derogatory public records |
| 21 | revol\_bal | Total credit revolving balance |
| 22 | revol\_util | Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit. |
| 23 | total\_acc | The total number of credit lines currently in the borrower's credit file |
| 24 | initial\_list\_status | The initial listing status of the loan. Possible values are — W, F |
| 25 | application\_type | Indicates whether the loan is an individual application or a joint application with two co-borrowers |
| 26 | mort\_acc | Number of mortgage accounts. |
| 27 | pub\_rec\_bankruptcies | Number of public record bankruptcies |