**Credit of Risk Modeling: Probability of loan defaults**

**Description**

The objective here is to identify customers who are likely to default over the next 6 months, for which we’ll develop a Risk Score to identify the risky customers and target them with remedial measures proactively.

**Datasets & Data Dictionary**

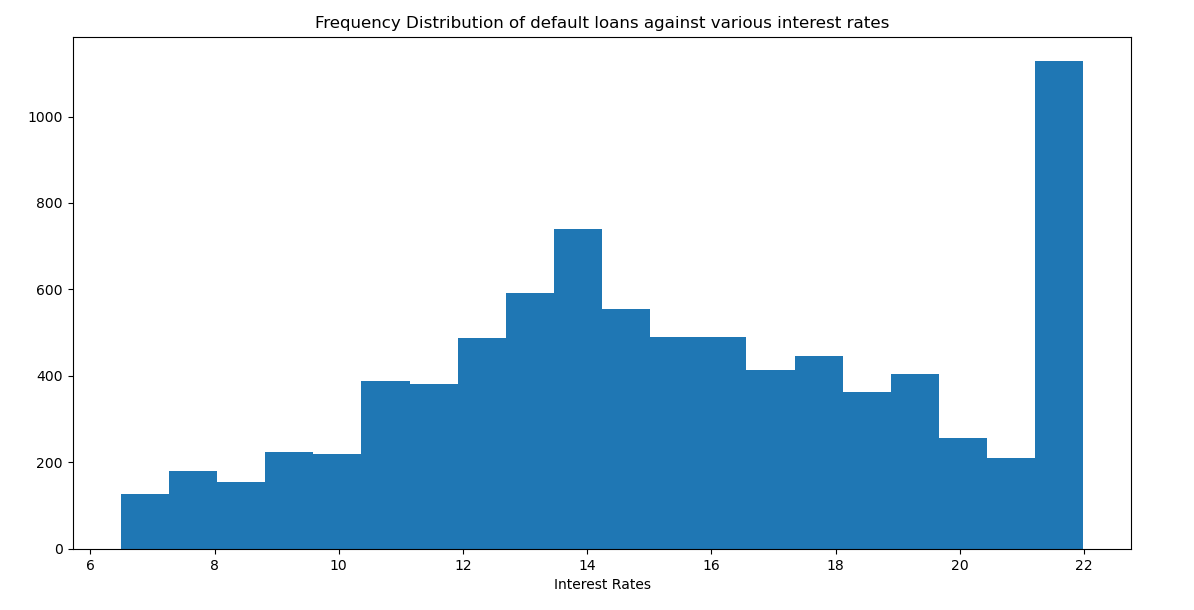
The data set has 57 columns with more than 50 K rows of data. **The flag Loan\_Status is the target variable where true indicates that the borrower has defaulted.**

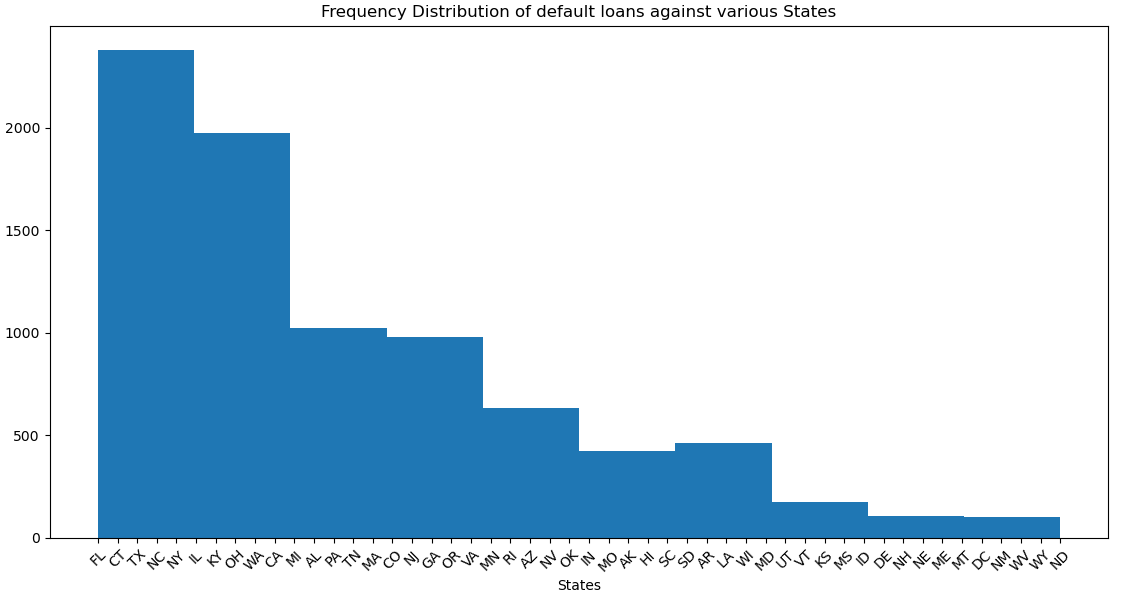
**Data Dictionary:**

|  |  |  |
| --- | --- | --- |
| **Feature** | **Description** | **Understanding for a layman** |
| Term | 36 or 60 term loan payment | Time period for repayment of the loan |
| Interest\_Pct | Interest Rate on the loan |  |
| Loan\_Grade | LC assigned loan grade | A classification of loan (kind of scoring a loan) |
| Loan\_Sub\_Grade | LC assigned loan subgrade |  |
| Emp\_Tenure | 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. |  |
| House\_Ownership | The home ownership status provided by the borrower.  Values are: RENT, OWN, MORTGAGE, OTHER |  |
| Is\_Verified | Indicates if income/income source was verified or not |  |
| **Loan\_Status** | Current status of the loan |  |
| State | The state provided by the borrower in the loan application |  |
| Fico | Average of FICO lower & upper range | It’s a credit score developed by Fair Issac Corporation |
| Rev\_Util\_Rate | Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit. |  |
| High\_Credit\_Ratio | Total high credit  limit |  |
| High\_BC\_Credit\_Ratio | Total bankcard high credit limit |  |
| High\_EMI\_Credit\_Ratio | Total installment high credit limit |  |
| DTI\_Ratio | A ratio of borrower’s total monthly debt payments to the  monthly income. |  |
| Total\_Revolving\_Ratio | Total revolving high credit/credit limit |  |
| Current\_Bal\_Avg | Average current balance of all accounts |  |
| Bank\_Card\_Utilization\_Ratio | Ratio of total current balance to high credit limit for all bankcard accounts. |  |
| Collection\_Fee | post charge off collection fee |  |
| Never\_Delinq\_Act\_Pct | Percent of trades never delinquent |  |
| EMI | The monthly payment owed by the borrower if the loan originates. |  |
| Payment\_Recieved | Payments received to date for total amount funded |  |
| Principal\_Recieved | Principal received to date |  |
| Late\_Fee\_Recieved | Late fees receive |  |
| Gross\_Recovery | post charge off gross recovery |  |
| Last\_Month\_Pmnt | Last total payment amount received |  |
| All\_Collection\_Amount | Total collection amounts ever owed |  |
| Cr\_Balance | Total credit balance excluding mortgage |  |
| IPA | Income per annum |  |
| Delinquencies\_2yr | The number of 30+ days  delinquency in the past 2 years | Criminal behavior |
| Open\_Cr\_Lines | The number of open credit lines in the borrower's credit file. |  |
| Public\_Derog\_Rec | Number of derogatory public records |  |
| Revolving\_Balance | Total credit revolving balance |  |
| Tot\_Cr\_Lines | The total number of credit lines currently in the borrower's credit file |  |
| All\_Current\_Balance | Total current balance of all accounts |  |
| Last\_24m\_Accounts | Number of trades opened in past 24 months. |  |
| Amount\_Owed\_Delinq | The past-due amount owed for the accounts on which the borrower is now delinquent. |  |
| Months\_Recent\_Rev\_Opened | Months since most recent revolving account opened |  |
| Accounts\_Having\_Mortgage | Number of mortgage accounts. |  |
| Satisfactory\_BC\_Accounts | Number of satisfactory bankcard accounts |  |
| Open\_Rev\_Accounts | Number of open revolving accounts |  |
| Satisfactory\_Accounts | Number of satisfactory accounts |  |
| Accounts\_120\_dpd\_Ever | Number of accounts ever 120 or more days past due |  |
| Accounts\_120\_dpd\_Cur | Number of accounts currently 120 days past due |  |
| Accounts\_90\_dpd\_Cur | Number of accounts 90 or more days past due in last 24 months |  |
| Last6m\_Inquiries | The number of inquiries in past 6 months |  |
| Delinquent\_Accounts | The number of accounts on which the borrower is now delinquent. |  |
| User\_ID | A unique LC assigned ID for the loan listing. |  |
| User\_ID2 | A unique LC assigned Id for the borrower member. |  |
| Loan\_Applied | The listed amount of the loan applied for by the borrower. |  |
| Loan\_Committed | The total amount committed to that loan at that point in time. |  |
| Disburse\_Mode | The method by which the borrower receives their loan. Possible values are: CASH, DIRECT\_PAY |  |
| Is\_Hardship | Flags whether or not the borrower is on a hardship plan |  |
| Loan\_Title | The loan title provided by the borrower |  |
| Is\_Purpose\_Credit\_Card | A category provided by the borrower for the loan request. |  |
| Is\_Ever\_Bankrupt | Number of public record bankruptcies |  |
| Tax\_Defaults | Number of tax liens |  |

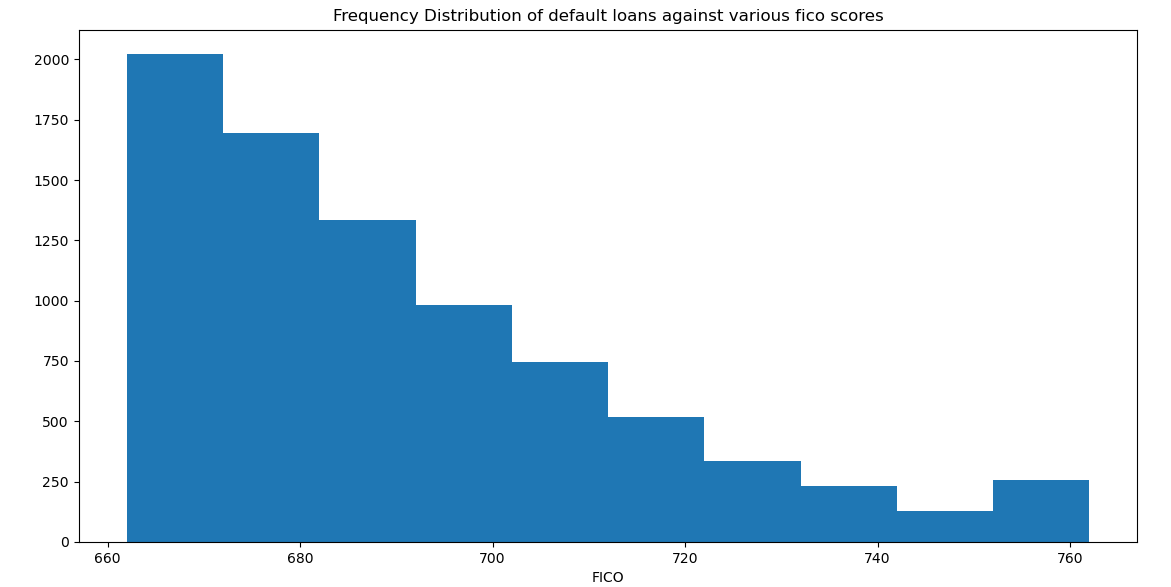
**Plotting of some predictors vs target to get a picture:-**

**Frequency Distribution of default loans against various interest rates**



**Frequency Distribution of default loans against various States**

**Frequency Distribution of default loans against various fico scores**



Since the target variable has 2 outcomes – 0 (Loan will not default) and 1 (Loan will default), we can conclude that this modeling comes under logistic regression.

Steps involved in making the model:-

1. **Check missing values**

After going through excel sheet, columns with missing values are - Emp\_Tenure

1. **Treat those missing values**

One way is to fill those missing values (imputation) – by either mean or median . But this is valid for continuous variables. When it’s a categorical variable:-

* Treat it as another category
* Replace it with most frequent observation
* Predict the observation

print(df.loc[df[**'Emp\_Tenure'**] == **'Missing'**])

# This gives out all those rows having **'Emp\_Tenure'** as ‘Missing’

There are total 3173 rows

Let’s predict the observation:-

Using decision classifier we can do that

1. **Check outliers**

An outlier is defined as an extremely high or extremely low value.

**Skewness:**

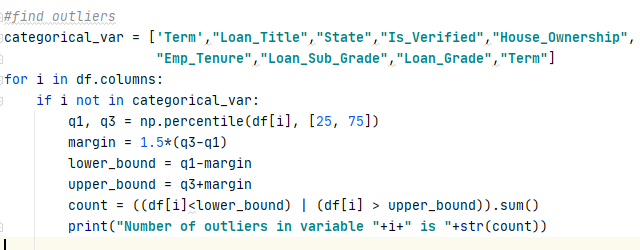
Skewness tells us the extent to which a data is normally distributed. Ideally, the Skewness value should be between -1 and +1, and any major deviation from this range indicates the presence of extreme values.

**IQR:**

Mathematically speaking:-

Any number falling outside below range is an outlier: [Min-IR\*1.5 , Max+ IR\*1.5]

IR = Q3 – Q1



1. **Treat those outliers**

Ways to treat outliers-

* Drop those values
* Capping and Flooring



1. **Convert categorical variables into numerical**

Most of the Machine learning algorithms cannot handle categorical variables unless we convert them to numerical values. Many algorithms’ performances vary based on how Categorical variables are encoded.

Category variables in our dataset:-

**['Term', 'Loan\_Grade', 'Loan\_Sub\_Grade', 'Emp\_Tenure', 'House\_Ownership', 'Is\_Verified', 'State', 'Loan\_Title']**

* **Term –** Has only 2 values (36 months and 60months). Let’s convert it into Boolean. If 36, then 1 else 0
* **Loan\_Grade –** Values (A, B, C, D, E, F, G). A being least risky and G most. We can map A as 1 (least) and G as 7 (most)
* **Loan\_Sub\_Grade –** Values (A1, A2, A3, A4, A5…….G4, G5). A2 is more risky than A1. Thus it is in order of increasing risks.We can map A1 as 1.1 and G5 as 7.5
* **Emp\_Tenure –**  This is basically integer values. <1 year = 0, 2 years = 2, >10years = 10
* **House\_Ownership -**

1. **Calculate information value of each variable:**

**We can use decision tree here but since there are too many predictors, we need to decide first which ones to keep**

IV of a variable helps us determine its significance in contributing towards the value of the target variable. IV is used to screen the potential predictors. It basically helps to rank variable based on their importance

[https://2.bp.blogspot.com/-hkTX-LJoANY/VPnv5Wd3UoI/AAAAAAAADk4/SZFPuuecbkg/s1600/IV.png](https://2.bp.blogspot.com/-hkTX-LJoANY/VPnv5Wd3UoI/AAAAAAAADk4/SZFPuuecbkg/s1600/IV.png)

|  |  |
| --- | --- |
| **Information Value** | **Variable Predictive Power** |
| Less than 0.02 | Not useful for prediction |
| 0.02 to 0.1 | Weak predictive Power |
| 0.1 to 0.3 | Medium predictive Power |
| 0.3 to 0.5 | Strong predictive Power |
| >0.5 | Suspicious Predictive Power |