**HOME LOAN PREDICTIONS**

**Executive summary**

1. Housing Finance company which provides home loans for the houses that are present across all urban, semi urban and rural areas for their valued customers.
2. The company validates the eligibility of the loan after customer applies for the loan. However, it consumes a lot of time for the manual validation of the eligibility process.
3. Hence, the company wants to automate the loan eligibility process based on the customer information and identify the factors/customer segments who are eligible for taking the loan.
4. As banks would give loans to only those customers who are eligible so that they can be assured of getting the money back.
5. Hence, the more accurate we are in predicting the eligible customers, the more beneficial it would be for the company.

**Detailed Overview of the Mortgage Approval & Funding Process:**

1. Pre-Assessment Discussion (15 minute conversation)
2. Pre-Approval Kick-Off (takes us no more than 1 day)
3. Opening a File (takes us no more than 1 day)
4. Lender Underwriting (takes 1 - 7 days from our formal submission)

* Credit history - Your lender will want to make sure when you've borrowed money, you've paid it back
* Capital - Ensuring you’ve accumulated assets
* Collateral - When it comes to a mortgage, you're putting your house up as collateral
* Capacity - In short, capacity is debt servicing. For instance, your housing cost shouldn't exceed 30 per cent to 32 percent of your gross income and all of your debts shouldn't exceed 40 per cent to 42 percent of your gross income
* Character - It’s an evaluation of all four previous C's as well as subjective and objective things such as how long have you been in your job, what type of job you have and how long you have lived in your current residence.

**Overall Time Consumed for single Loan Application:**

1. Conditional Commitment Processing (takes 2 - 4 days from lender approval)
2. Pre-Closing (takes 7 - 10 days from 'file complete')
3. Closing (typically by noon on the funding/possession date)

**Problem Statement**

1. Company wants to automate the loan eligibility process (real time) based on customer detail provided while filling an online application form.
2. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others.
3. To automate this process, the company has given us a problem to identify the customer segments that are eligible for loan amount so that they can specifically target these customers. Here, we have been provided a partial data set for further analysis.

**Structured Analysis Planning:** The SMART (Specific Measurable Assignable Relevant Time-based) objective was employed to analyze the data and understand the problem statement. The next step is to identify our independent variables and our dependent variable, the below map illustrates the process which was conducted to structure plan the project.

To this approach, we employed Exploratory data analysis which include univariate analysis and bivariate analysis.

**Assumptions for EDA:**

1. The customers whose salary is more can have a greater chance of loan approval.
2. The applicants who are graduates, have a better chance of loan approval than non-graduate applicants.
3. Married applicants would have upper hand than single or no-relationship applicants for loan approval.
4. The applicant who has less dependents has a high probability for loan approval.
5. The lesser the loan amount, the higher chances of loan getting approved.

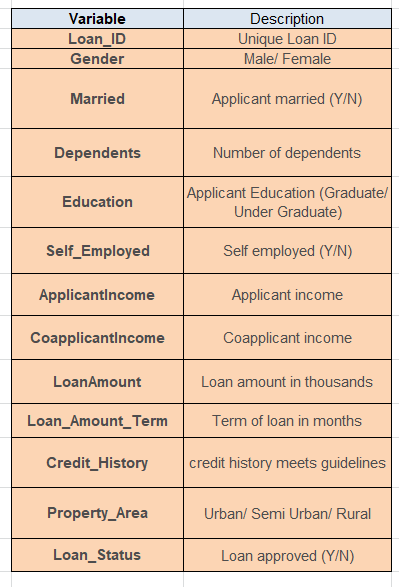
**Type of Problem:**

The above problem is a clear classification problem as we need to classify whether the Loan\_Status is yes or no. So this can be solved by any of the classification techniques like

1. Logistic Regression .
2. Decision Tree Algorithm.
3. Random Forest Technique.

**Description about the Data Columns:**

There are 2 data sets that are given. One is training data and one is testing data. It’s very useful to know about the data columns before getting into the actual problem for avoiding confusion at a later state. Now let us understand the data columns (that has been already given by the company itself ) first so that we will get a glance.



There are altogether 13 columns in our data set. Of them Loan\_Status is the response variable and rest all are the variables /factors that decide the approval of the loan or not.

Now let us look in to the each variable and can make some assumptions.(It’s just assumptions right, there is no harm in just assuming few statements)

* Loan ID -> As the name suggests each person should have a unique loan ID.
* Gender -> In general it is male or female. No offence for not including the third gender.
* Married -> Applicant who is married is represented by Y and not married is represented as N. The information regarding whether the applicant who is married is divorced or not has not been provided. So we don’t need to worry regarding all these.
* Dependents -> the number of people dependent on the applicant who has taken loan has been provided.
* Education -> It is either non -graduate or graduate. The assumption I can make is “ The probability of clearing the loan amount would be higher if the applicant is a graduate”.
* Self\_Employed -> As the name suggests Self Employed means , he/she is employed for himself/herself only. So freelancers or having their own business might come in this category. An applicant who is self employed is represented by Y and the one who is not is represented by N.
* Applicant Income -> Applicant Income suggests the income by Applicant.So the general assumption that i can make would be “The one who earns more have a high probability of clearing loan amount and would be highly eligible for loan **”**
* Co Applicant income -> this represents the income of co-applicant. I can also assume that “ If co applicant income is higher , the probability of being eligible would be higher “
* Loan Amount -> This amount represents the loan amount in thousands. One assumption I can make is that “ If Loan amount is higher , the probability of repaying would be lesser and vice versa”
* Loan\_Amount\_Term -> This represents the number of months required to repay the loan.
* Credit\_History -> When I googled it , I got this information. A **credit history** is a record of a borrower’s responsible repayment of debts. It suggests → 1 denotes that the credit history is good and 0 otherwise.
* Property\_Area -> The area where they belong to is my general assumption as nothing more is told. Here there can be three types. Urban or Semi Urban or Rural
* Loan\_Status -> If the applicant is eligible for loan it’s yes represented by Y else it’s no represented by N**.**