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.

Variable	Description
Loan_ID	Unique Loan ID
Gender	Male/ Female
Married	Applicant married (Y/N)
Dependents	Number of dependents
Education	Applicant Education (Graduate/ Under Graduate)
Self_Employed	Self employed (Y/N)
ApplicantIncome	Applicant income
CoapplicantIncome	Coapplicant income
LoanAmount	Loan amount in thousands
Loan_Amount_Term	Term of loan in months
Credit_History	credit history meets guidelines
Property_Area	Urban/ Semi Urban/ Rural
Loan_Status	Loan approved (Y/N)

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 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**.**

Tidying the data

Now that we've identified several errors in the data set, we need to fix them before we continue with our analysis. Let's review the issues:

There are missing values in some variables. Based on the importance of the variables, we will decide on the method to use.

Looking at the distributions of the data, we noticed that ApplicantIncome and LoanAmount have outliers.

Fixing outliers can be tricky. It's hard to tell if they were caused by measurement error, errors while recording, or if the outliers are real anomalies. If we decide to remove records, we have to document the reason behind this decision.

In this data set, we will assume that missing values are systematic because the missing data are coming in certain variables in a random manner. Also, we note that missing values are on both numerical and categorical data, therefore, we will be creating different functions to handle these scenarios. These functions help in imputing missing values with plausible data values. These values are inferred from a distribution that is designed for each missing data point.

Type of Variables:

1. Input variable (Predictor): Gender, Married, Education, Self_Employed, ApplicantIncome, CoapplicantIncome, LoanAmount, Loan_Amount_Term, Credit_History
2. Output variable (Target): Loan_Status

Variable category:

1. Categorical variables: Loan_ID, Gender, Married, Dependents, Education, Self_Employed, Property_Area, Loan_Status
2. Continuous variables: ApplicantIncome, CoapplicantIncome, LoanAmount, Loan_Amount_Term

Data Pre-processing Steps:

1. Handling missing values
2. Creation of required variables
3. Replacing the data-values

Handling Missing Values:

1. Defining function to fetch “most_common” value of the feature
2. Defining function to “replace_missing” values with most_common/mean values
3. Calling “replace_missing” function for all features with null values by passing the feature name to the function.

```
In [6]: def most_common(col):
            most_cmn=pd.get_dummies(col).sum().sort_values(ascending = False).index[0]
            return most_cmn
```

```
In [7]: def replace_missing(col):
            most_cmn = most_common(col)
            for i in range(len(col)):
                if pd.isnull(col[i]):
                    col[i] = most_cmn
                    inplace=True
```

```
data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 614 entries, 0 to 613
Data columns (total 13 columns):
Loan_ID              614 non-null object
Gender               601 non-null object
Married              611 non-null object
Dependents           599 non-null object
Education            614 non-null object
Self_Employed        582 non-null object
ApplicantIncome      614 non-null int64
CoapplicantIncome    614 non-null float64
LoanAmount           592 non-null float64
Loan_Amount_Term     600 non-null float64
Credit_History       564 non-null float64
Property_Area        614 non-null object
Loan_Status          614 non-null object
dtypes: float64(4), int64(1), object(8)
memory usage: 62.4+ KB
```

Before Handling Null Values

```
data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 614 entries, 0 to 613
Data columns (total 13 columns):
Loan_ID              614 non-null object
Gender               614 non-null object
Married              614 non-null object
Dependents           614 non-null object
Education            614 non-null object
Self_Employed        614 non-null object
ApplicantIncome      614 non-null int64
CoapplicantIncome    614 non-null float64
LoanAmount           614 non-null float64
Loan_Amount_Term     614 non-null float64
Credit_History       614 non-null float64
Property_Area        614 non-null object
Loan_Status          614 non-null object
dtypes: float64(4), int64(1), object(8)
memory usage: 62.4+ KB
```

After Handling Null Values

Creation of required variables:

1. Creating “Total_Income” variable by adding “ApplicantIncome” to “CoapplicantIncome”
2. Calculating the “EMI” variable

   By considering the above link, I have found that on an average it would be around 8.5% to 9.5%. Hence for the safe-side I am assuming that 9% is the interest rate.

   - A = P * R * (1+R)^N
   - B = (1+R)^(N-1)
   - EMI = A/B

```
data["Total_Income"] = data["ApplicantIncome"]+data["CoapplicantIncome"]
```

By considering the above link, I have found that on an average it would be around 8.5% to 9.5%. Hence for safe-side I am assuming that 9% is the interest rate.

A = P * R * (1+R)^N

B = (1+R)^(N-1)

EMI = A/B.

```
data["EMI"] = (data["LoanAmount"]*0.09*(1.09**data["Loan_Amount_Term"]))/(1.09**(data["Loan_Amount_Term"]-1))
```

Replacing the data-values:

1. “Dependents” variable has some data-values marked as “**3+**”, so changing them as value “**3**”.

```
data['Dependents'].replace('3+','3' ,inplace=True)
```

Data Storytelling

This step is to explore the dataset to discover certain trends.

The goal of the exploratory data analysis was to find out whether a certain demographic is more likely to get approved compared to others.

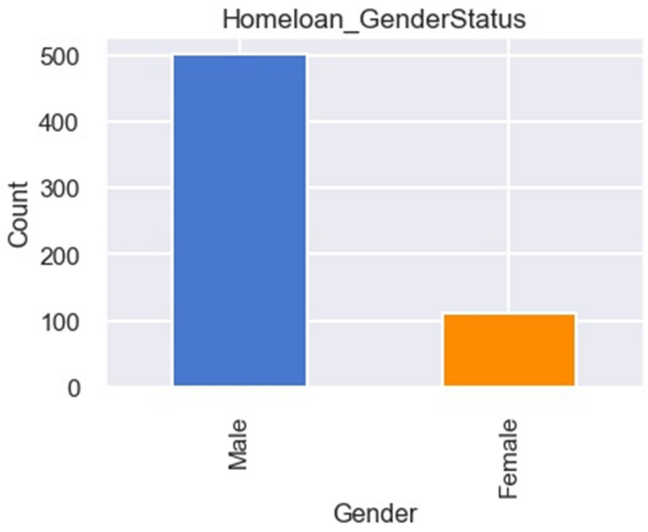
The dataset was truncated to create a dataset centered around gender and education. In order to compare the demographics perfectly, different histograms were used to explore these newly-formed dataset.

Plotting graphs for different features includes repetitive line of codes, so to overcome this scenario different plotting functions were defined. These functions were called as per the need of visualization.

```
def feature__countplot(col):
    plt.figure(figsize=(6, 4))
    data[col].value_counts().plot(kind='bar',color=('b','darkorange'))
    plt.xlabel(col, fontsize=16)
    plt.ylabel('Count', fontsize=16)
    plt.title("Homeloan_"+col+"Status")
    plt.savefig("Homeloan_"+col+"Status.jpeg",bbox_inches = 'tight')
    plt.show()
```

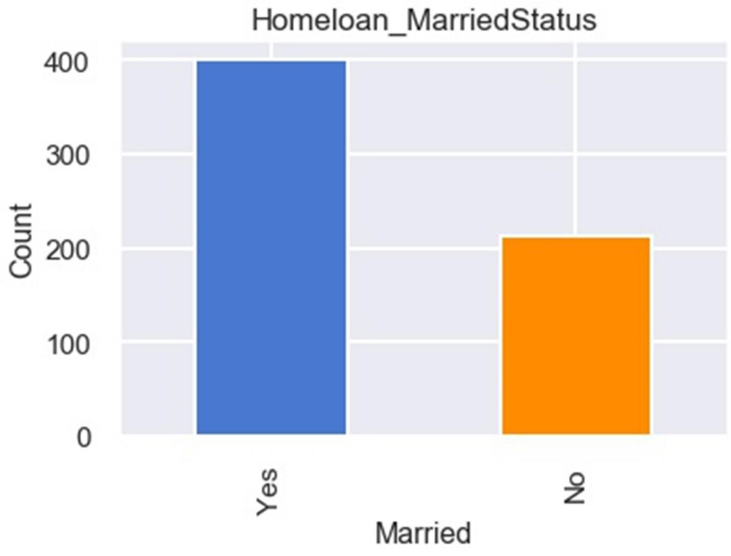
```
def feature_comp_plot(col1,col2):

    fig, ax = plt.subplots(figsize=(6,4))

    x2 = data.groupby([col1,col2])['Loan_ID'].count()
    x2 = x2.reset_index()

    stats = x2[col2].drop_duplicates()
    margin_bottom = np.zeros(len(x2[col1].drop_duplicates()))
    colors = ["#006D2C", "#31A354","#74C476"]

    for num, status in enumerate(stats):
        values = list(x2[x2[col2] == status].loc[:, 'Loan_ID'])
        x2[x2[col2] == status].plot.bar(x=col1,y='Loan_ID', ax=ax, stacked=True,
                                        bottom = margin_bottom, color=colors[num], label=status)
        margin_bottom += values
    ax.set_title(col1+" vs. "+col2)
    plt.savefig(col1+" vs. "+col2+".jpeg",bbox_inches = 'tight')
    plt.show()
```

Single Variable Analysis



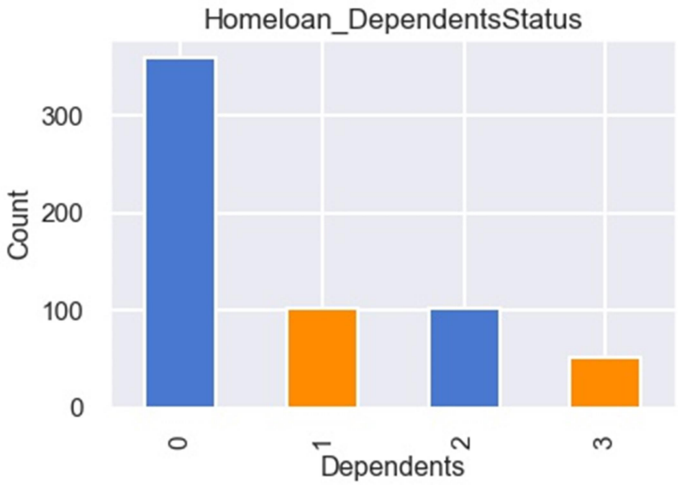
Gender Column:

According to our analysis, Gender may influence home loan approval. As we can conclude that, mortgage lenders were more inclined towards men than women expecting men to be the lead borrowers on single applications.



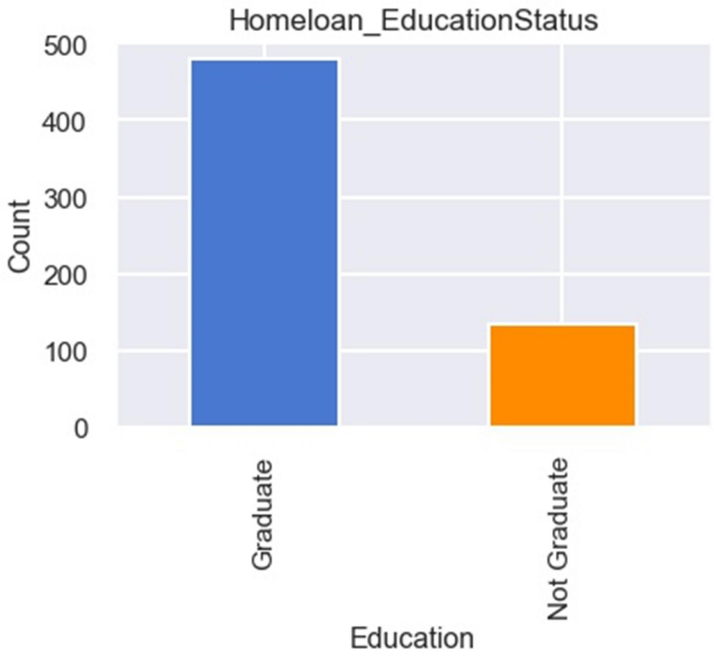
Marital Status:

From the above results, we can conclude that most of the home loans were approved to married couples compared to persons who are single or with no relationship.



Dependents:

From the analysis, we can conclude that the number of dependents may automatically affected the approvals of home loans. There is a higher chance of getting home loan approval for applicants who have less number of dependents or no dependents.



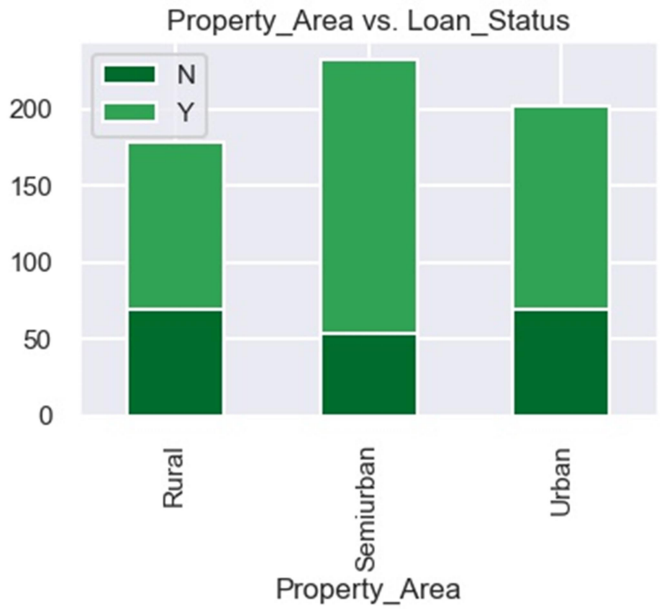
Education:

From the analysis, we can conclude that the educational status may automatically affect the approvals of home loans. There is a higher chance of getting home loan approval for applicants who are graduates.

Employment:

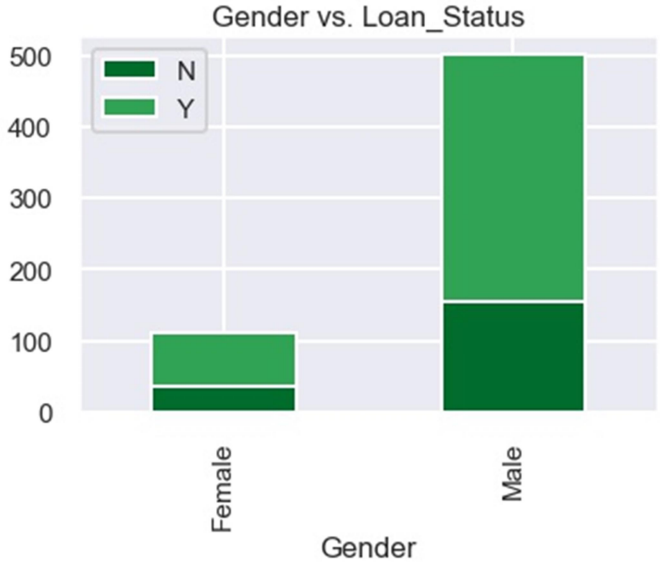
From the analysis, we can conclude that the employment status may automatically affected the approvals of home loans. There is a higher chance of getting home loan approval for applicants who are self_employed.

Multiple Variable Analysis



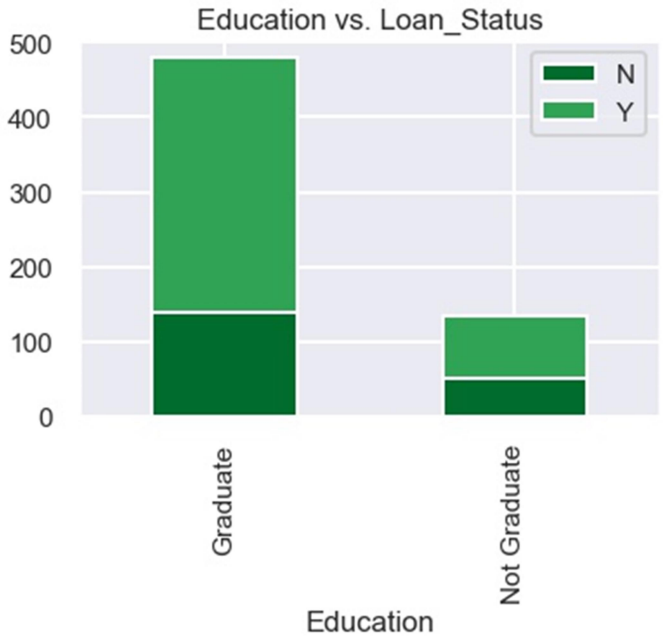
Relationship between Property area and loan status:

From the above results we can infer that the higher percentage of loan approval is for semi-urban houses followed by urban and rural houses.



Relationship between Gender and Loan status:

From the data analysis, we can conclude that male gender as primary applicants have a higher percentage of loan approval than female as primary applicants.



Relationship between education vs Loan status:

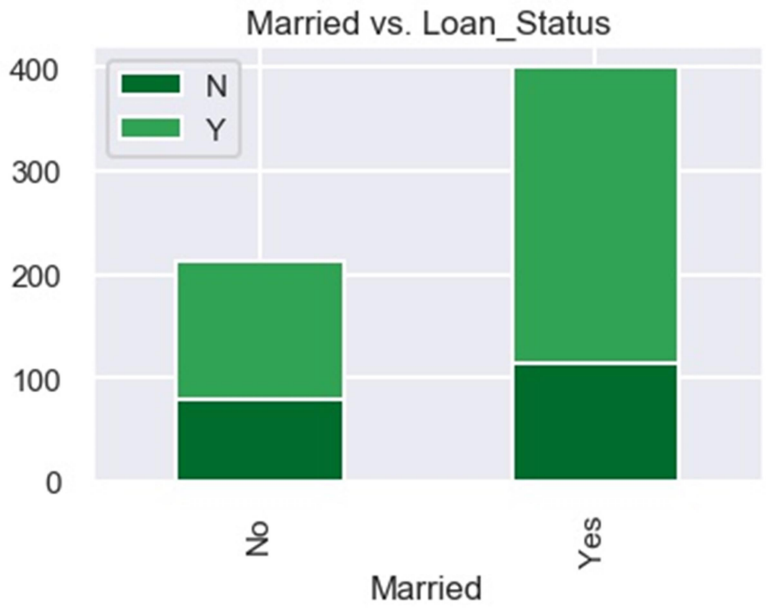
From the analysis, we can conclude that the applicants who are graduate were in a higher percentage of loan approval than non-graduate applicants.

Relationship between Self-Employed vs Loan_Status:

From the data analysis, we can conclude that home-ownership rates for self-employed households were more declined than for salaried households.

Relationship between Dependents vs Loan status:

From the analysis, we can conclude that the number of dependents may automatically affected the approvals of home loans. There is a higher chance of getting home loan approval for applicants who have less number of dependents or no dependents.



Relationship between Marital status vs Loan status:

From the analysis, we can conclude that the highest number of customers are married who were eligible for the home loan approval than single customers.

Correlation Heatmap