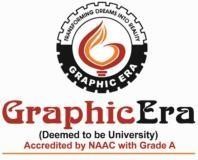
# Mini Project Report

**on**

**HOME LOAN PREDICTION**

**2021-2022**



**Submitted to: Submitted by:**

**Dr. Narayan Chaturvedi/Dr. Jyoti Agarwal Mr. Abhishek Jindal**

**( Evaluators)**

**Roll. No.: 2015143**

**Resource Person: Dr. Narayan Chaturvedi CSE-DS-AI-VI-Sem**  **Session: 2021-2022**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**GRAPHIC ERA (DEEMED TO BE) UNVERSITY,**

**DEHRADUN**

## CONTENTS

1. **Abstract**

1. **Introduction**

1. **Motivation**

1. **Libraries and Modules**

1. **METHODOLOGY**

1. **DEPLOYMENT**

1. **CONCLUSION**
2. **FUTURE IMPROVEMENTS**
3. **REFRENCES**

# Abstract

In our banking system, banks have many products to sell but main source of income of any banks is on its credit line. So they can earn from interest of those loans which they credits. A bank’s profit or a loss depends to a large extent on loans i.e. whether the customers are paying back the loan or defaulting. By predicting the loan approval we can actually solve a big problem. But as the right predictions are very important for the maximization of profits, it is essential to study the nature of the different methods and their comparison. A very important approach in predictive analytics is used to study the problem of predicting loan approvals

1. Collection of Data
2. Data Cleaning
3. Performance Evaluation.

**Introduction**

House Loan Prediction is very helpful for employee of banks as well as for the

applicant also. The aim of this Project is to provide quick, immediate and easy

way to choose the deserving applicants.

Customer first apply for loan after that company or bank validates the customer eligibility for loan. We can automate the loan eligibility process (real time) based on customer details provided while filling application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and other.

This project has taken the data of previous customers of various banks to whom on a set of parameters loan were approved. So the machine learning model is trained on that record to get accurate results. Our main objective of this project is to predict the safety of loan issue.

## Motivation

Being extremely interested in everything having a relation with the Machine Learning and Data Science, the independent project was a great occasion to give me the time to learn and confirm my interest for this field. The fact that we can make estimations, predictions and give the ability for machines to learn by themselves is both powerful and limitless in term of application possibilities.

Home Loan approval is a very important process for banking organizations. The system approved or reject the loan applications. Recovery of loans is a major contributing parameter in the financial statements of a bank. It is very difficult to predict the possibility of payment of loan by the customer. Using Machine learning we predict the loan approval.

# Libraries and Modules

# 

**Pandas**

## NumPy

NumPy is a Python library for creating and manipulating vectors and matrices.

**Seaborn**

Seaborn is an open-source Python library built on top of [matplotlib.](https://www.section.io/engineering-education/matplotlib-visualization-python/) It is used for data visualization and exploratory data analysis. Seaborn works easily with dataframes and the Pandas library.

.

## Matplotlib

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter.

**Softwares Required**

* Windows 7/8/10
* Google chrome
* Anaconda (Jupyter Notebook)
* Colab
* Jupyter notebook

**Hardware Required**

* GPU-8 to 16 GB
* RAM- 8GB(Min.)

# METHODOLOGY

## Problem Statement

Home Loan Prediction and predict the person will receive the loan or not according to the details provided.

## EXPLORATORY DATA ANALYSIS

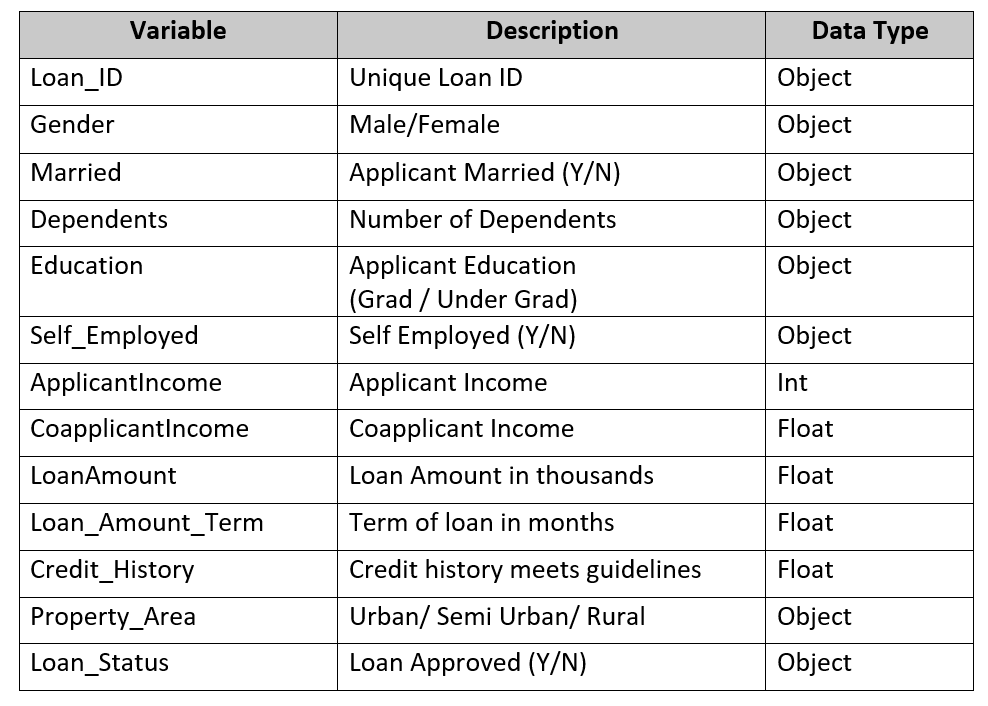
This system predicts whether the loan is approved or reject . This System refers the following things or ways.

**Data Collection:**

The crucial element in machine learning task for which a particular attention

should be clearly taken is the data. Indeed, the results will be highly influenced by the data based on where we found them, how are they formatted, are they consistent, is there any outlier and so on. At this step, many questions should be answered to guarantee that the learning algorithm will be efficient and accurate.

I have taken the data from Kaggle( https://www.kaggle.com/datasets/sazid28/home-loan)

****

1. We have imported all the necessary libraries:

*import pandas as pd*

*import numpy as np*

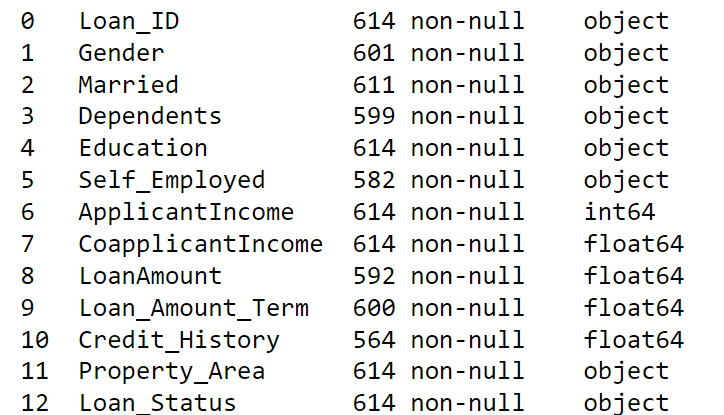
*import seaborn as sns*

*from matplotlib import pyplot as plt*

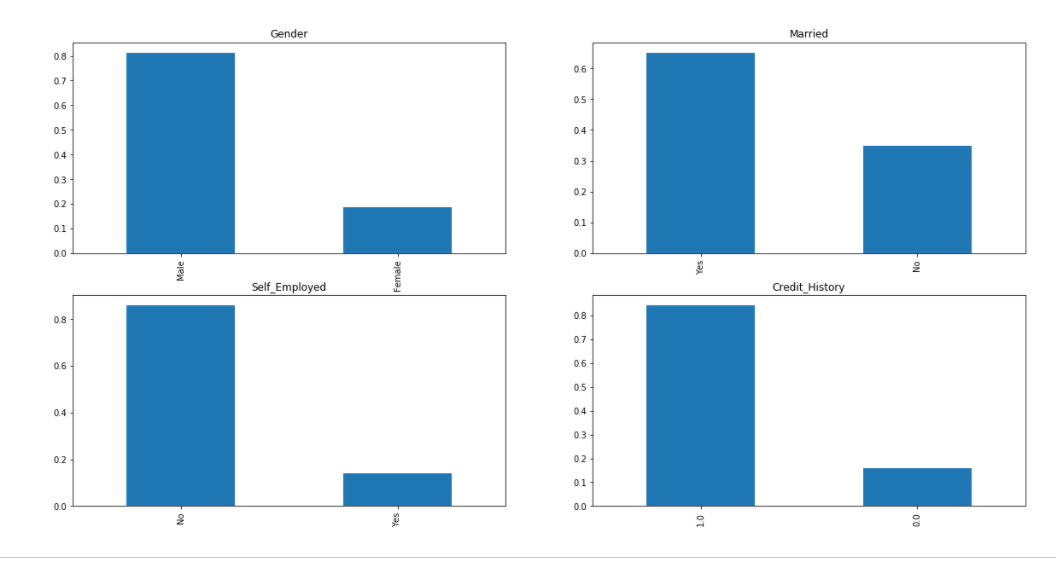
*import matplotlib*

*%matplotlib inline*

We have found info of our dataset



## DATA VISUALIZATION

Examining each variable individually. For categorical features, we use frequency table or bar plots which will calculate the number of each category in a particular variable.

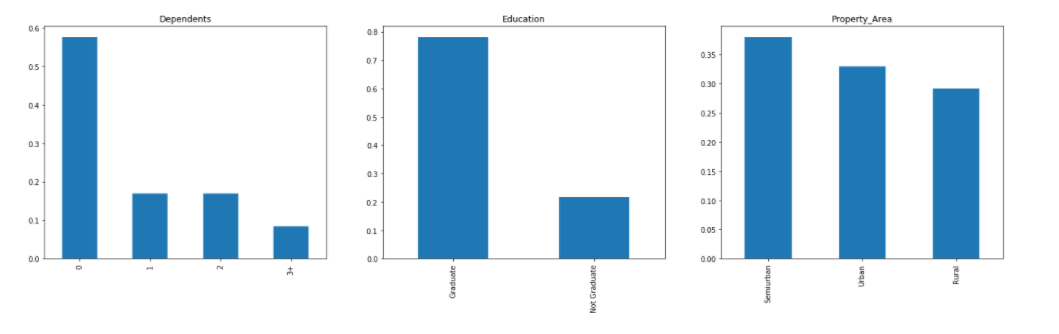
It can be inferred from the above bar plots that:

• 80% applicant in the dataset are male.

• Around 65% of the applicants in the dataset are married.

• Around 15% of the applicants in the dataset are self-employed.

• Around 85% applicants have repaid their debts.



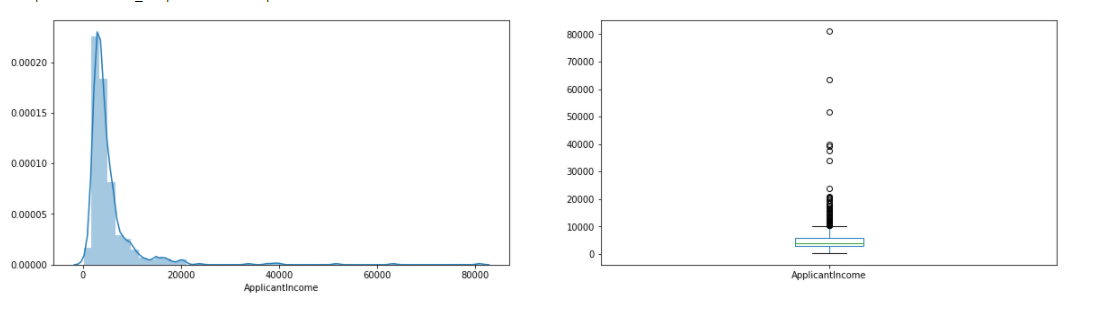
Following inferences can be made from the above bar plots:

• Most of the applicant do not have any dependents.

• Around 80% of the applicants are graduates.

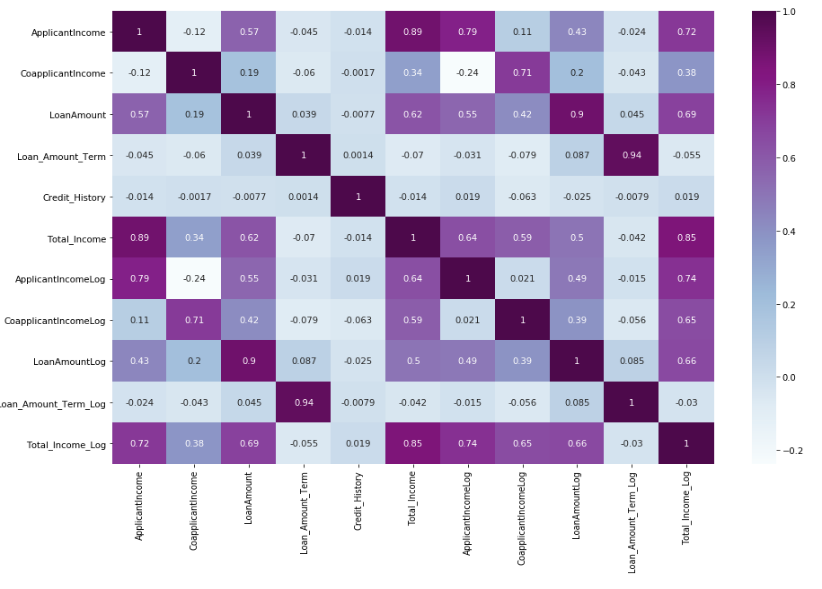
• Most of the applicant are from Semiurban area.

**For numerical variables, probability density plots can be used to look at the distribution of the variable.**



It can be inferred that most of the data in the distribution of applicant income is towards left which is not normally distributed, and the boxplot confirms the presence of a lot of extreme values/outliers. This can be attributed to the income disparity in the society.

**Coorelation matrix**

The following heatmap shows the correlation between all the numerical variables. The variable with darker color means their correlation is more.

**DATA PREPROCESSING**

The quality of the inputs in the model will decide the quality of your output. The following steps were taken to pre-process the data to feed into the prediction model.

**Missing Value Imputation**

After understanding all the variable in the data, we can now impute the missing values and treat the outliers because missing data and outliers can have adverse effect on the model performance.

We consider the following values in all the features one by one.

For categorical variables: impute using mode.

For numerical variable: imputation using mean or median.

Here, I have used median to impute the missing values as evident from Exploratory Data Analysis that loan amount has outliers

Outlier Treatment:

As LoanAmount contains outliers, it is rightly skewed. One way to remove this skewness is by doing the log transformation. As a result, we get a distribution like the normal distribution and does no affect the smaller values much but reduces the larger values

**MODEL TRAINING**

1.Logistic Regression



2.Decision Tree Classifier



3.Random Forest Classifier



The following actions were taken to achieve this result:

• Explored the dataset to understand the data.

• Perform different tests of statistical significance to uncover hidden data relationships that our predictive model could learn from and leverage when predicting unseen instances.

• Employed statistical analysis using various python libraries to identify the number and names of features that could more likely help in identifying the potential customers by predicting whether a loan would be approved or not.

**Conclusion**

So here, it can be concluded with confidence that the Random Forest Classifier is extremely efficient and gives a better result when compared to other models. It works correctly and fulfills all requirements of bankers. This system properly and accurately calculate the result. It predicts the loan is approve or reject to loan applicant or customer very accuratly.

**Future Improvements:**

• In upcoming years, as the new data keeps coming in, it will be important for data science team to deal with missing values, imbalanced data sets and additional features.

• It is important to notice this fact that the default loans are only about 10% of the total loans, thus during the training process, the model will favor predicting more negatives than positive results.

## References

1.Kaggle(for dataset)( https://www.kaggle.com/datasets/sazid28/home-loan)

2.Youtube-(Krish naik)

3.Github link- (<https://github.com/Abhijndl/HOME-LOAN-PREDICTION.git>)

4. https://pandas.pydata.org/

5. <https://numpy.org/>

6. https://matplotlib.org/

7.<https://seaborn.pydata.org/>