Project Proposal

Introduction:

With the enhancement in the banking sector, many people are applying for bank loans, for a variety of purposes. But all these applicants are not genuine, and everyone cannot be approved based on reliability. So, to reduce this risk factor behind selecting the safe person to save lots of bank efforts and assets. This can be achieved by mining the data of the previous records of the people to whom the loan was approved and based on these records/experiences the machine was trained using the machine learning model which gives the best accurate results.

Through this system we can predict whether that applicant is safe or not and the whole process of validation of features is automated by machine learning technique. Loan Prediction is very helpful for employees of banks as well as for the applicant also. The aim of this project is to provide a quick, immediate, and easy way to choose the deserving applicants. This Loan Prediction System can automatically calculate the weight of each feature field taking part in loan processing and compares the new data with its associated weight. It checks the applicant's data and provides whether his/her loan can be approved or not as a result.

Objectives and Research Questions:

The purpose of this project is to build a regression model that determines the maximum loan amount for which a person is eligible based on several factors such as income, credit history, loan amount, loan term, and so on. In this study, we would like to explore the following research questions:

- What are the most important factors that determine a customer's loan eligibility?
- For customers who are not eligible for the required loan amount and duration, what is the maximum amount they can borrow for the given duration?
- To determine whether a customer is eligible for a loan, check their eligibility?

Motivation:

The purpose behind exploring these questions is to try to understand the factors that play a vital role in determining how much loan an applicant is eligible for and what will be the most effective method of determining the loan amount. The information obtained from the loan provider can be useful for the loan seeker as well as for the loan provider. There are many advantages associated with this type of tool for both loan providers and loan seekers. On the one hand, a loan provider can use this tool to make informed decisions about loan approval, while on the other hand, a loan seeker can use it to understand what factors can influence their loan eligibility and take the necessary steps to improve their chances of being approved.

Hypothesis:

Here are some possible hypotheses for the research questions:

1. Most important factors for loan eligibility: The income of the applicant and the co-applicant, the credit history, and the loan amount requested are the most important factors that determine a customer's loan eligibility. Additionally, factors such as marital status, number of dependents, education, and property area may also play a role in determining loan eligibility.

2. Maximum loan amount for given duration: For customers who are not eligible for the required loan amount and duration, the maximum amount they can borrow for the given duration depends on their income, credit history, loan amount requested, loan amount term, and other factors. A regression model based on these factors can predict the maximum loan amount for a given duration.
3. Eligibility determination: A machine learning model based on the customer details such as gender, marital status, education, number of dependents, income of self and co-applicant, loan amount requested, loan amount term, credit history, and property area can accurately determine a customer's loan eligibility. The model can achieve an accuracy of at least 80% in predicting loan eligibility. Additionally, the model can identify the most important factors that contribute to loan eligibility and provide explanations for the model's predictions.

To test the hypothesis, a statistical model such as logistic regression or decision tree can be trained on the loan eligibility dataset to identify the most important predictors of loan eligibility and estimate their effect on the response variable.

The model can then be evaluated using metrics such as accuracy, precision, recall, and F1-score to determine its performance in predicting the loan eligibility status of new applicants.

If the model shows a significant relationship between the predictor variables and the loan eligibility status, it can be used by the company to automate the loan eligibility process and reduce the risk of bias or error in manual processing.

If the model shows no significant relationship, further analysis may be required to identify other factors that affect loan eligibility or to refine the predictor variables used in the model.

Analysis plan:

A dataset we have chosen for this project is that of Loan Prediction, which indicates the past record of an applicant based on which we can estimate the likelihood of the applicant receiving the loan.

- The Dataset includes the client loan data and whether their loan got approved or not.
- The main goal is to find out loan approval prediction over testing data using model.
- The training dataset contains 13 columns.

There are 13 different variables in the dataset and the differentiation between them is shown below:

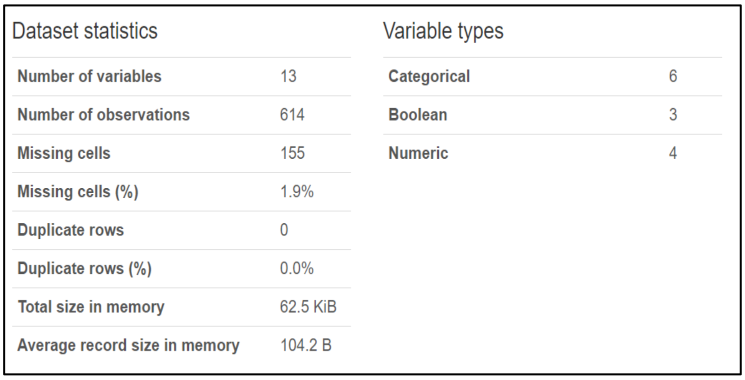
- **Observed variable:** The observed variable in this dataset is the loan eligibility status, which can take on two values: "Y" if the loan is approved or "N" if the loan is not approved.
- **Predictor variables:** Gender, Married, Dependents, Education, Self_Employed, Applicant_Income, Coapplicant_Income, Loan_Amount, Loan_Amount_Term, Credit_History, Property_Area are the predictor variables. These variables can help the model identify patterns and relationships that can be used to determine whether an applicant is eligible for a loan or not.
- **Quantitative variables**: ApplicantIncome, CoapplicantIncome, LoanAmount, and Loan_Amount_Term are quantitative variables, which can take on a range of continuous or discrete numerical values.
- **Qualitative variables:** Gender, Married, Education, Self_Employed, Dependents, Credit_History, and Property_Area are qualitative variables, which take on a limited number of discrete values.

The model which we are going to use are Linear regression model, Logistic regression model and Decision tree model. Linear regression model is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data points. Logistic Regression is a classification algorithm that predicts the probability of a binary outcome, in this case, whether a loan will default or not. A decision tree is a flowchart-like model that maps out different courses of action and their potential outcomes to aid decision-making.

Data:

In this dataset we have 614 observations, 13 variables and there are 155 missing cells.

Dataset



Dataset statistics

Number of variables	13
Number of observations	614
Missing cells	155
Missing cells (%)	1.9%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	62.5 KiB
Average record size in memory	104.2 B

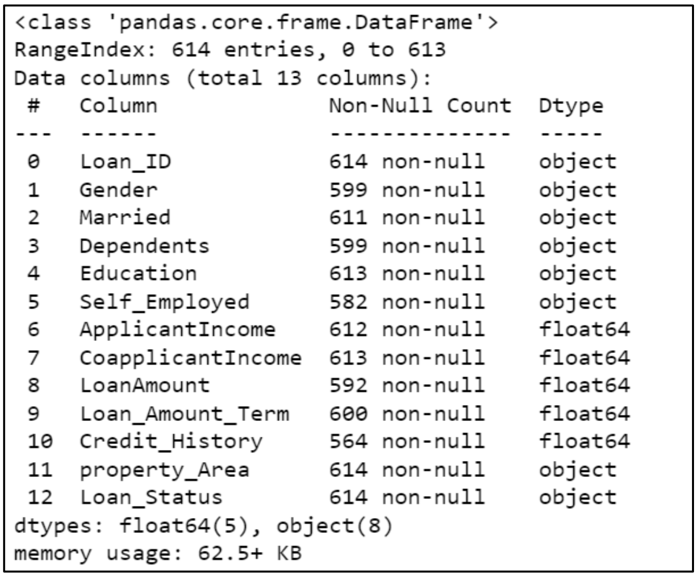
Variable types

Categorical	6
Boolean	3
Numeric	4

The dataset description is given below:

Column Name	Description
Loan_ID	Every time a loan request is made, it gives the information of the Unique ID for that loan request.
Gender	This section identifies the gender of the applicant.
Married	This information is used to identify the Marital status of the applicant.
Dependents	Amount of people who depend on the applicant and his/her immediate family as a main source of income.
Education	This column identifies the education of the applicant.
Self_Employed	Whether the applicant is an employee or an owner of a company
ApplicantIncome	It provides the details of the Applicant's income and expenses that have been recorded.
CoapplicantIncome	In it, the Applicant provides a detailed explanation of the Co-applicant income.
LoanAmount	In this, the amount of loan that the applicant requires is shown.
Loan_Amount_Term	The time frame in which the applicant wants to repay the amount of the loan.
Credit_History	Is the candidate's credit history good or bad over the past few years?
Property_Area	Type of property that the candidate has used as a mortgage. This type tells us a lot about the value of the property.

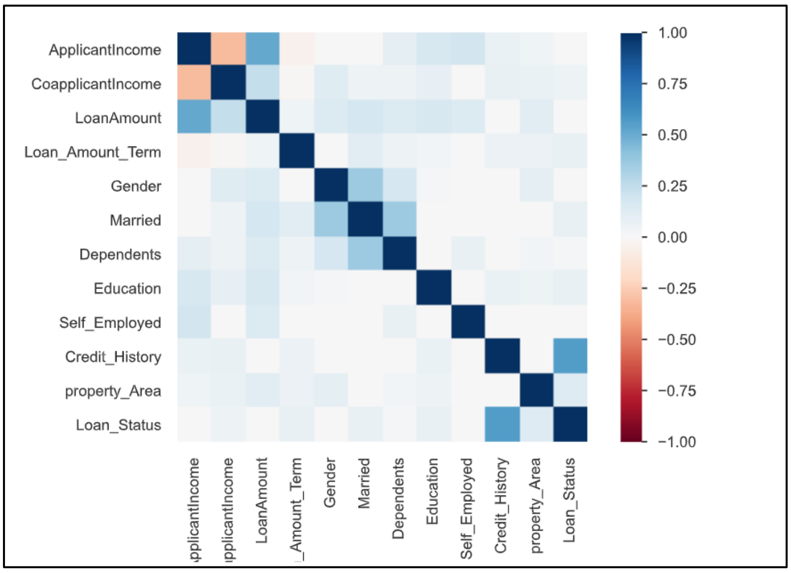
Dataset



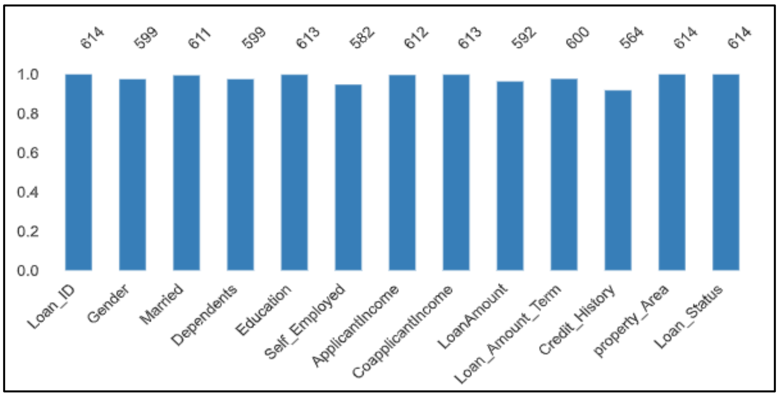
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 614 entries, 0 to 613
Data columns (total 13 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Loan_ID            614 non-null    object 
 1   Gender             599 non-null    object 
 2   Married            611 non-null    object 
 3   Dependents         599 non-null    object 
 4   Education          613 non-null    object 
 5   Self_Employed      582 non-null    object 
 6   ApplicantIncome    612 non-null    float64
 7   CoapplicantIncome  613 non-null    float64
 8   LoanAmount         592 non-null    float64
 9   Loan_Amount_Term   600 non-null    float64
 10  Credit_History     564 non-null    float64
 11  property_Area      614 non-null    object 
 12  Loan_Status        614 non-null    object 
dtypes: float64(5), object(8)
memory usage: 62.5+ KB
```

We have analysed the dataset through different plots and it is shown below:

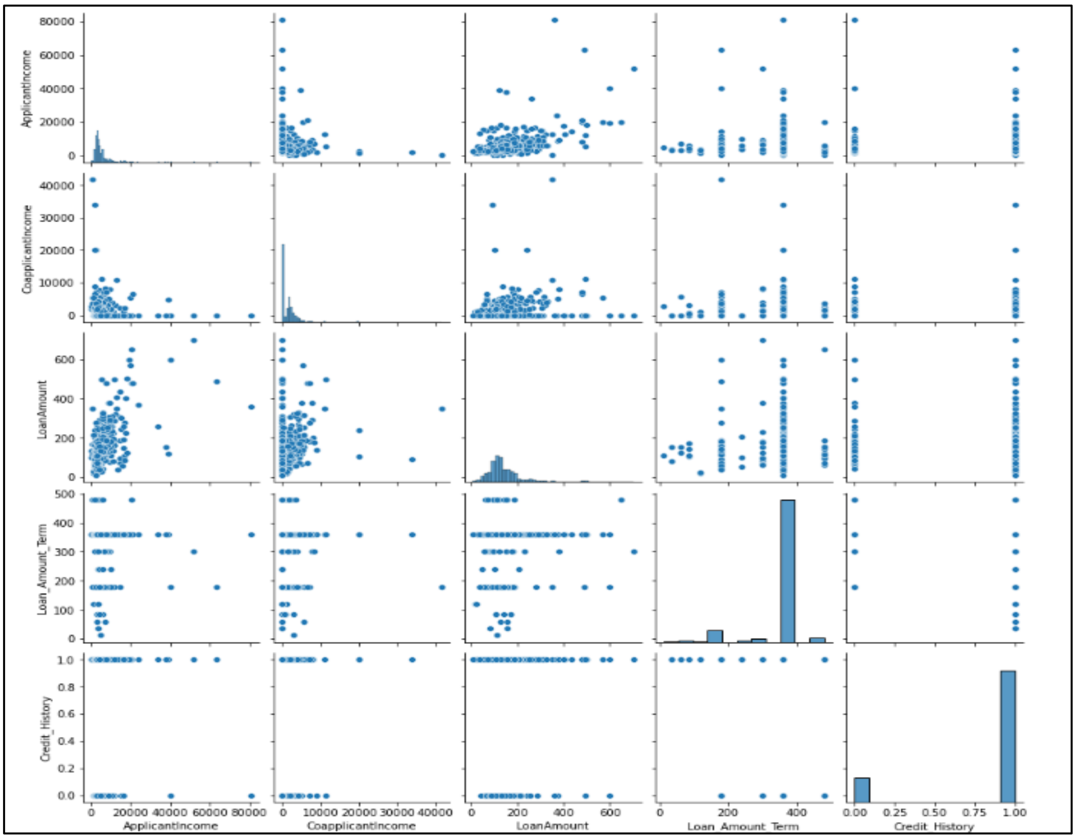
Correlations Heatmap



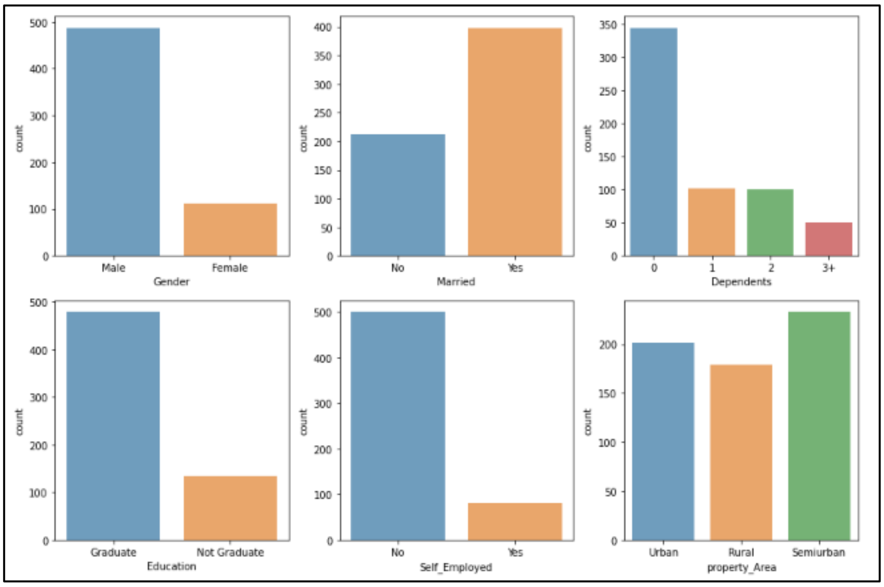
Column Count Bar Chart



Numerical Column Pair Plot



Categorical Column Bar Chart



References:

- EXPLORATORY DATA ANALYSIS FOR LOAN PREDICTION: https://www.irjmets.com/uploadedfiles/paper//issue_5_may_2022/22400/final/fin_irjmets1651993759.pdf, https://www.ijrte.org/wp-content/uploads/papers/v7i4s/E2026017519.pdf

Team Member Worked:

- Vatsal Rameshbhai Gohel
- Venkatesh Dhanuskodi
- Manan Chandrakant Davey
- Shraddha Kamath Barkur
- Meghana Dodda