Home Loan Prediction

project done by

Subhajyoti Maity

Department of Computer Science
Ramakrishna Mission Vivekananda Educational and Research Institute
Belur Math, Howrah
Pin - 711 202, West Bengal

Contents

1 introduction:

Home loan prediction is a smart technology that uses data to quickly decide if someone can get a loan to buy a home. The dataset used in this analysis includes various attributes related to home loan applicants, such as their income, credit score, Education, employment status, loan amount, Property Area, and other relevant financial indicators. Additionally, the dataset contains a binary target variable indicating whether a home loan application was approved (1) or denied (0).

The home loan approval status analysis is a data-driven exploration aimed at understanding the factors influencing the approval or denial of loan applications. The availability of vast amounts of data in the financial industry has made it possible to utilize advanced analytics and machine learning techniques to gain valuable insights into the loan approval process.

The primary objective of this analysis is to uncover patterns and relationships between different applicant characteristics and loan approval outcomes. By doing so, we can develop a predictive model that accurately estimates the likelihood of home loan approval for future applicants based on their financial profiles.

2 Data collection

I collected the data from the site
`https://www.kaggle.com/datasets/rishikeshkonapure/home-loan-approval`

3 Data Description:

Dataset has 614 rows and 13 columns.All values are not non-null.Data type of most of the columns is object.ApplicantIncome, CoapplicantIncome, LoanAmount,`Loan_Amount_Term` and `Credit_History` are in int64 and float64..Out of 13 columns there are 8 categorical columns and 5 numeric columns.some of the columns' description given below:

- `Loan_ID:`A unique identifier for each loan application, facilitating easy tracking and referencing
- Gender : The gender of the applicant, indicating whether they are male or female.
- Married : A binary variable denoting whether the applicant is married (Yes) or not (No).
- Dependents : This feature indicates the number of dependents the applicant has, such as children or other family members financially supported by the applicant.
- Education : Categorizes the applicant's education level as either Graduate or Not Graduate.
- `Self_Employed:`A binary variable indicating whether the applicant is self-employed (Yes) or not (No).
- ApplicantIncome : The income of the applicant, representing their earning capacity.
- CoapplicantIncome :The income of the co-applicant, if applicable.
- LoanAmount : The amount of the loan requested by the applicant.
- `Loan_Amount_Term:` The term or duration of the loan(in month), specifying the time within which the loan must be repaid.
- `Credit_History:` A binary variable indicating the credit history of the applicant, whether they have a good credit history yes (1) or not (0).
- `Property_Area:` Categorizes the property area of the applicant as Urban, Semiurban, or Rural.
- `Loan_Status:` The target variable representing the loan approval status, where 'Y' indicates the loan was approved, and 'N' indicates it was denied.

4 Data Preprocessing:

Data preprocessing is an essential step in the data analysis and machine learning pipeline. It involves cleaning, transforming, and organizing raw data into a format suitable for further analysis or model training.
The process of data preprocessing typically includes the following steps:

4.1 Data Cleaning:

This step involves handling missing values, noisy data, and inconsistencies in the dataset. Missing values can be filled or removed based on various techniques, such as mean imputation, median imputation.

- In the dataset 'Gender', 'Dependents', and `'Self_Employed'` variables have missing values. I can not interprate them, as that would be giving the model wrong information. I'll just drop the rows with missing values.After removing those rows from the dataset which contain null value in columns 'Gender','Dependents' and 'Self_Employed ,new Dataframe has 554 rows and 13 columns.

- In the dataset 'LoanAmount', `'Loan_Amount_Term'`, and `'Credit_History'` variables have missing values.replace the missing values of numerical features by median value.

4.2 Removing duplicates:

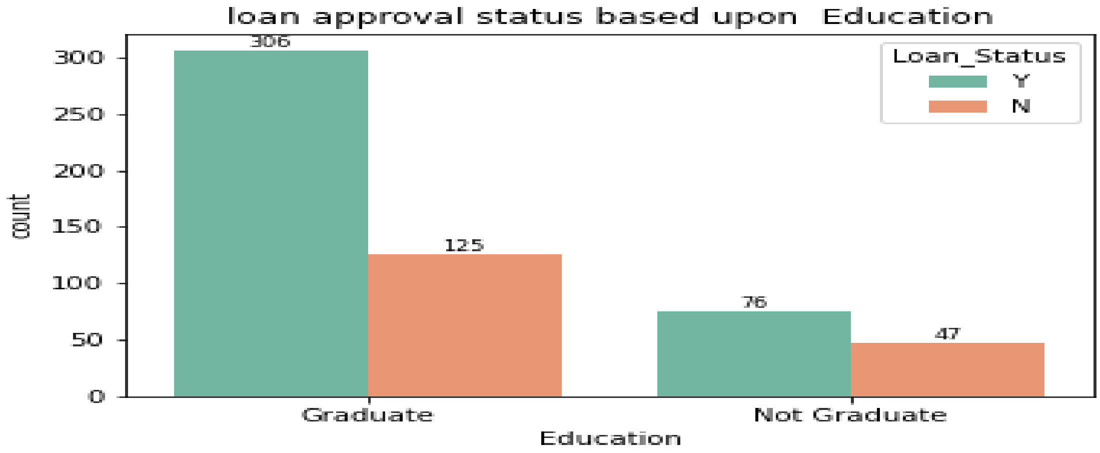
Identifying and removing identical rows or observations.In the given dataset there does not exist no duplicate rows.we do our Exploratory Data Analysis over new dataframe with 554 rows.

5 Exploratory Data Analysis(EDA):

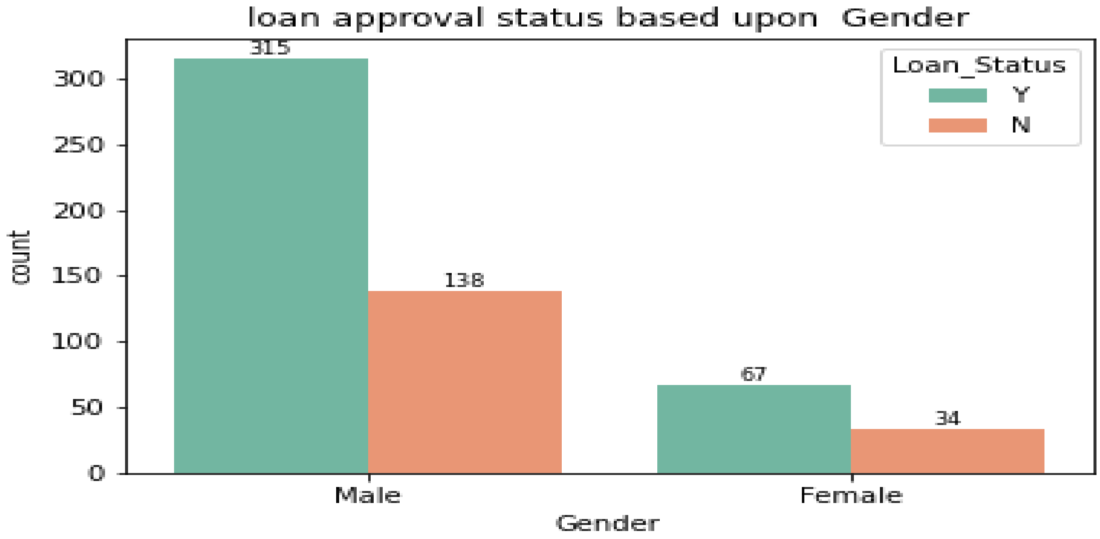
EDA is a critical phase that involves visualizing and analyzing the dataset to gain insights into the distribution of features and their impact on household electic amount paid .

EDA helps to uncover insights, identify data quality issues, and guide further analysis or modeling tasks. Create visualizations to explore and understand the data better. Commonly used plots include histograms and bar plots.

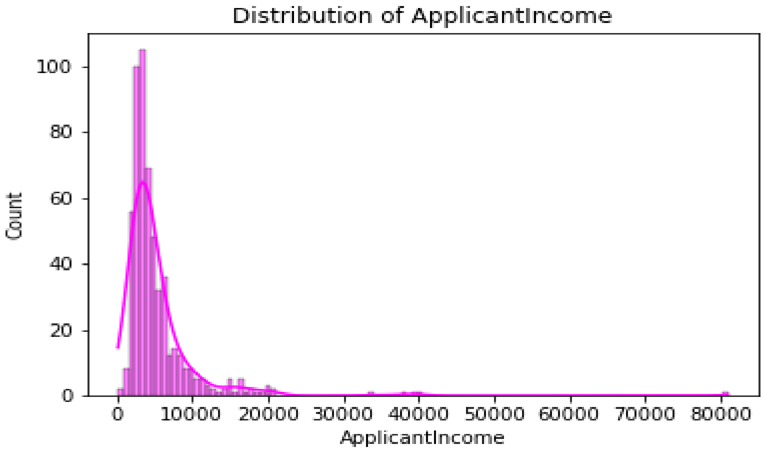
- some following insights are given below:



from the above picture we can say that percentage of loan approval among the graduate students is greater than percentage of loan approval among the non graduate students.



from the above picture we can say that percentage of loan approval among the Male candidate is greater than percentage of loan approval among the Female candidate.



from this picture we can say that most of the family income in the dataset lies between 2000 and 4000.

6 Feature Engineering:

Feature engineering is the process of transforming or creating new features from existing ones to improve the model's performance. This step aims to extract relevant information from the data that can enhance loan approval predictions.

6.1 Data Transformation:

Data transformation involves converting data into a suitable format for analysis or modeling. Common transformations include:.

- Not all features (variables) in the dataset may be relevant or important for the analysis or modeling.So we drop 'Loan ID' column.
- Encoding categorical variables:Converting categorical variables into numerical representations (e.g., one-hot encoding or label encoding) for the model to process. `get_dummies()` is a function in pandas that converts categorical variables into binary columns, also known as dummy variables. It creates a new DataFrame `df_encoded` where each categorical column in the original DataFrame is replaced with a set of binary columns representing each unique category in that column.

6.2 Splitting the Data:

- The scikit learn algorithms take two separate arguments. This means they need independent variables separately and the dependent variable (or target variable) separately. But since in the train dataset both independent and dependent variables are present together so I need to separate them out.
- Firstly, I'll create a set of independent variables from the train dataset. So I'm dropping the 'target' variable from it using axis=1. This axis=1 specifies that the drop shall happen from the column.
- The data set will be split into 80% train and 20% test.

7 model selection:

Various classification algorithms, such as logistic regression, Categorical Naive Bayes and support vector machines, will be employed to build the predictive model.

8 Model Evaluation :

The model's performance will be evaluated using appropriate metrics like accuracy, precision, recall, F1 score.

1. Logistic regression : We fit the whole dataset into logistic regression model. we get the value of R2 score is 81.981982.

2. Categorical Naive Bayes : We fit the whole dataset into Categorical Naive Bayes model. we get the value of R2 score is 79.279279.

3. Support Vector Machines: We fit the whole dataset into Support Vector Machines model. we get the value of R2 score is 71.171171 .

The best-performing model will be selected for deployment to ensure accurate predictions.We chose the model 'Logistic Regression' because of maximum accuracy value of this model.

9 Real-time Deployment:

The final model will be integrated into the home loan application system to provide real-time loan approval predictions based on applicant details.

In the new data where the target variable is unknown to us we predict loan approval status .The rows in the new data where 'Gender', 'Dependents', and `'Self_Employed'` variables have missing values, I can not predicted those rows, as that would be giving the wrong information. I'll just drop the rows with missing values.And we fill the target variable 'loan approval status' values in the dataframe with 'Pending'.

10 Conclusion:

However, it is essential to acknowledge the limitations of the predictive model. The accuracy of loan approval predictions may be influenced by factors beyond the dataset, such as changing economic conditions.
In summary, the home loan approval prediction analysis has contributed to a more efficient, transparent, and equitable lending process. By leveraging machine learning, this predictive model empowers lenders and borrowers alike, facilitating responsible lending practices and supporting the dream of homeownership for millions of individuals and families.