**Data Set:**

ABC company is a NBFC & deals in all kinds of home loans. They have presence across all urban, semi urban and rural areas. Customer first applies for home loan and after that company validates the customer eligibility for loan.

ABC wants to automate the loan eligibility process (real time) based on customer detail provided while filling online application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. To automate this process, they have provided a dataset to identify the customers segments that are eligible for loan amount so that they can specifically target these customers.

**Data Dictionary**

**Train file:**CSVcontaining the customers for whom loan eligibility is known as 'Loan\_Status'

|  |  |
| --- | --- |
| **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 | (Target) Loan approved (Y/N) |

**Validation file:** CSVcontaining the customer information for whom loan eligibility is to be predicted

|  |  |
| --- | --- |
| **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 |

**Submission file format:**

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Loan\_ID | Unique Loan ID |
| Loan\_Status | (Target) Loan approved (Y/N) |

**Task:**

Design a model that will be able to predict if customer should be eligible for loan or not.

**Deliverables:**

Provide the following:

• The source code you used to build the model and make predictions. (You are

free to use any language and any open-source package/library)

• A .csv file containing the predictions of the validation data. You can add the target

column (‘Loan\_Status) to the validation data or simply provide it alone with the Loan\_ID

column.

• Briefly answer the following questions:

o Describe your model and why did you choose this model over other

types of models?

* I tried to predict on test data using various model. But as compared to data size and imbalance in data chosen the Decision tree model. I divided the train test split using the stratify which helped to handle the imbalance to some extent.

o Describe any other models you have tried and why do you think this

model performs better?

* Same as above. Looking at metrics RF seems better performing model.

o How did you handle missing data?

* Filled missing values for categorical columns like Gender, Married etc filled using mode value of particular columns
* Filled missing values in Loan amount column with the LR model technique where we tried to predict missing values using applicant income and co-applicant income.
* Filled missing values in credit score using the KNN imputer techniques.

o How did you handle categorical (string) data?

* Gender, Married are handled using one hot encoding.
* Dependents, Education, Property\_Area, Self\_Employed and Loan\_Amount\_Term handled using dictionary mapping.

o How did you handle unbalanced data?

* Using stratify while train test split and applying class weight at modelling.

How did you test your model ?

* Using the metrics and confusion metrics.