Loan Credit-Worthiness Classification

-Prajwal Waykos

- +917249542810

[Pwaykos1@gmail.com](mailto:Pwaykos1@gmail.com)

Project Repo = <https://github.com/Praj-17/Loan-Creaditworthiness-classification>

Kaggle = <https://github.com/Praj-17/Loan-Creaditworthiness-classification>

Drive = <https://github.com/Praj-17/Loan-Creaditworthiness-classification>

Problem Statement

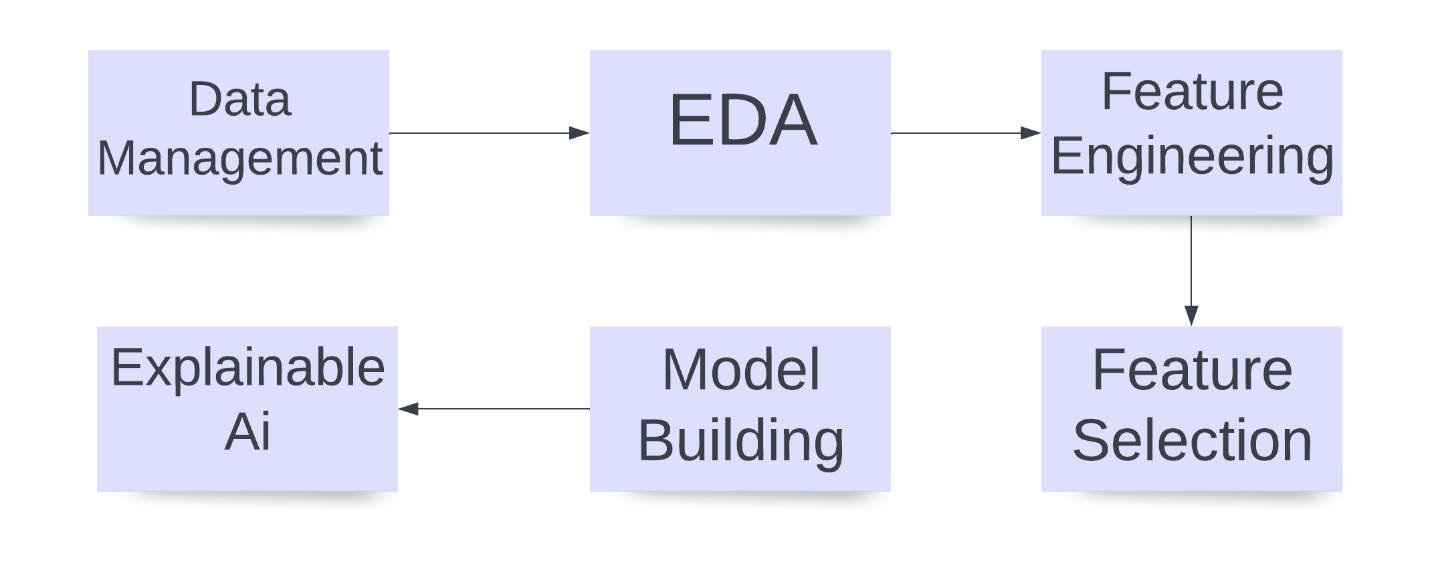
The Problem statement is to prepare a model based upon the given dataset which classifies a new applicant as risky or creditworthy.

Abstract

The dataset presents us with 2 tables namely applicant and loans which contain applicant and loan data respectively. The Dataset is available in the repository.

I have performed all the major Machine Learning procedures such as , Exploratory Data Analysis, Data Analytics, Data Preprocessing, Feature Engineering, Feature Selection, Modeling and Explainable AI on the dataset.

Procedure



Methodology

1. Data Management – [visit](https://github.com/Praj-17/Loan-Creaditworthiness-classification/blob/main/data_manager.ipynb) :

This file is used to merge the given two tables and also to generate, Pandas Profiling report everytime needed.

***Components***:

1. Merging of 2 tables
2. Generating reports of the dataset.
3. **Merging of 2 tables:**

* There are 2 tables given, Applicant and Loan which are separated logically by the data they contain.
* But for EDA and Better understanding of the model we will be joining the 2 tables based upon the relationships in them.
* In the Applicant Table we can see the first column as applicant\_id which is the **Primary key**of it.
* Moreover, in the Loan Table we can see the Second column as applicant\_id which is the **Foreign Key**of it.
* Based upon this common column we will be joining the 2 datasets and store in a csv file called [data\_merged.csv](https://github.com/Praj-17/Loan-Creaditworthiness-classification/blob/main/data/data_merged.csv).

1. **Generating Reports of the dataset.**

Pandas profiling report is library is an automated python tool for basic EDA.

* This report provides basic EDA and an overview of the dataset.
* The following important data such as
  + Missing Data
  + Types of Variables
  + Classes in Ordinal variables
  + Range of Numeric Variables
  + Correlations and Sample Data Visualizations

You can visit all the reports created [here](https://github.com/Praj-17/Loan-Creaditworthiness-classification/tree/main/Reports).

**Some Inferences from the generated report**

* There are a total of 27 columns and 1000 records of each column.
* Luckily there are no duplicate reocords or columns in the dataset.
* There are 4 numeric and 23 categorical features in the dataset as of now
* he Balance of the dataset is 7:3 which makes it critical to think weather it should be balanced or imbalanced. Since we lack enough data, we would be considering it as imbalanced and performing oversampling in the later phases.
* Around 12% of the data is missing.
* There are some columns which have more than 60% percent of missing cells. We might need to eliminate them or they will add unnecessary bias in case we try imputing them
* All sorts of Features are present within the dataset and hence, we need to take special care while handling each variable
* The Variable "Telephone" is constant (No Use)

2. Feature Engineering:

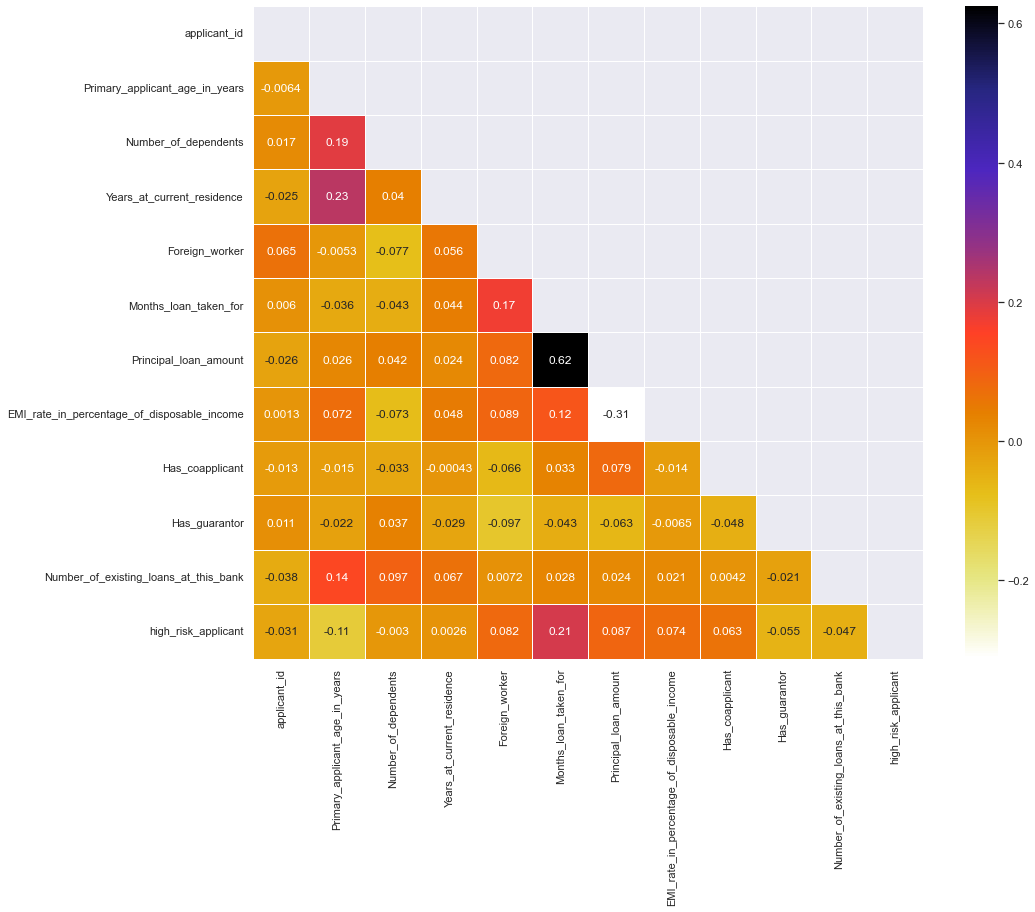
EDA or Exploratory data analysis is a procedure in Machine Learning where we analyze the given data both statistically and logically to generate insights that could solve difficult business problems.

I have performed EDA in 2 phases

1. EDA and FE [notebook](https://github.com/Praj-17/Loan-Creaditworthiness-classification/blob/main/eda_and_fe.ipynb) –
2. Advance EDA [notebook](https://github.com/Praj-17/Loan-Creaditworthiness-classification/blob/main/eda_advance.ipynb) –
3. **EDA and FE -**

* Since the basic , eda has been covered by the generated report we will see some visualizations. To better understand the spread and variance of data.

Correlations

****

The chart above states various correlations among the variables. But there is one thing to note here that it consists only 13 cols and not the remaining columns. This is because the other cols might be having some string value or datetime values within them. Hence, We should consider Data Cleaning/ and Feature Engineering before thoroughly understanding the data.

### **Insights of the correlations from raw data.**

 Duration for which the loan was taken is highly correlated with applicant being a defaulter.

 The more is the age of the applicant the more loans he might take

 The more is the principal amount the Longer it gets.

 The more is the age of the applicant the longer he is supposed to live at the same residence.

 Foreign Workers usually take bigger and longer loans.

 The More the dependents the more loans a person takes

 The more the applicants age the more people depend on him.

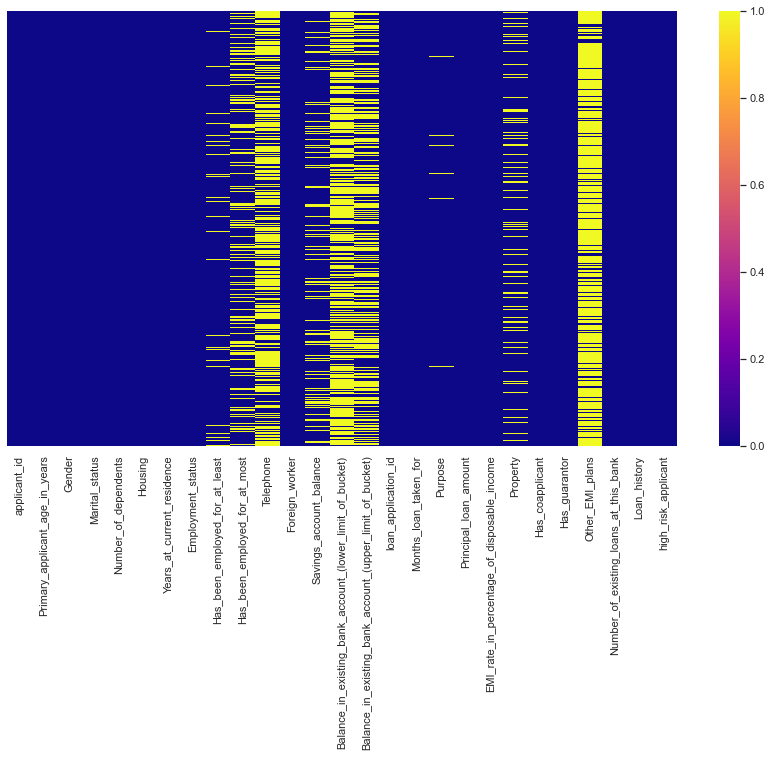
 We must also observe that there are 12 columns in the above plot which means there are 15 columns with no-numeric datatype.

**All the above made conclusions are logically correct and hence, we can also conclude that the data is natural and not artificially made.**

**Feature Engineering**

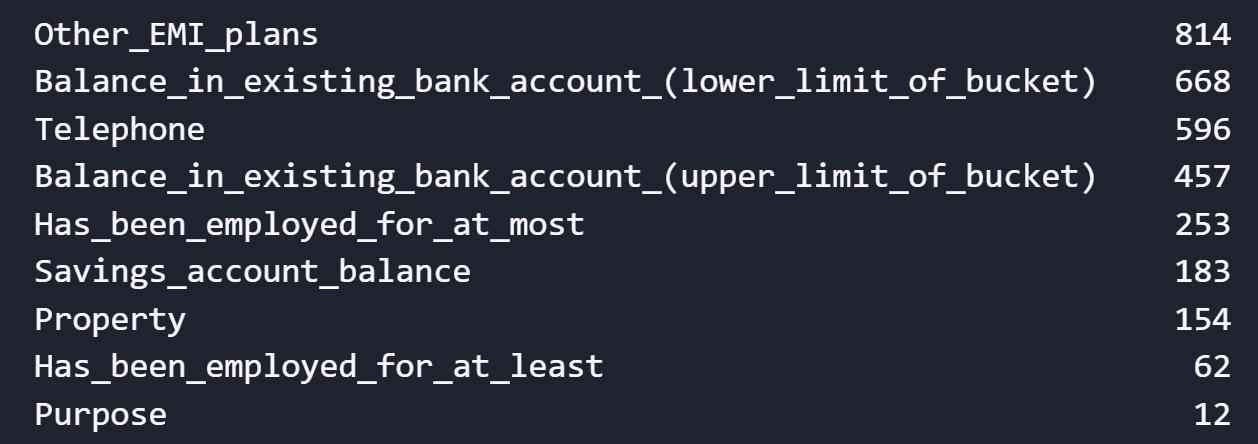
1. **Missing Data –**

The following plot gives us insights on which columns have missing cells and with what quantity they are missing.



From the above Data we can conclude that

* There is total 9 columns with null values
* Out of the 9 columns 2 Columns namely 'Telephone' and 'Other\_EMI\_plans' needs to be eliminated entirely. Since, Telephone is a constant and Other\_EMI\_Plans have more than 80% null Values
* The 2 columns 'Existing Bank Balance' also needs to be eliminated and they also do not enough variance among the data.



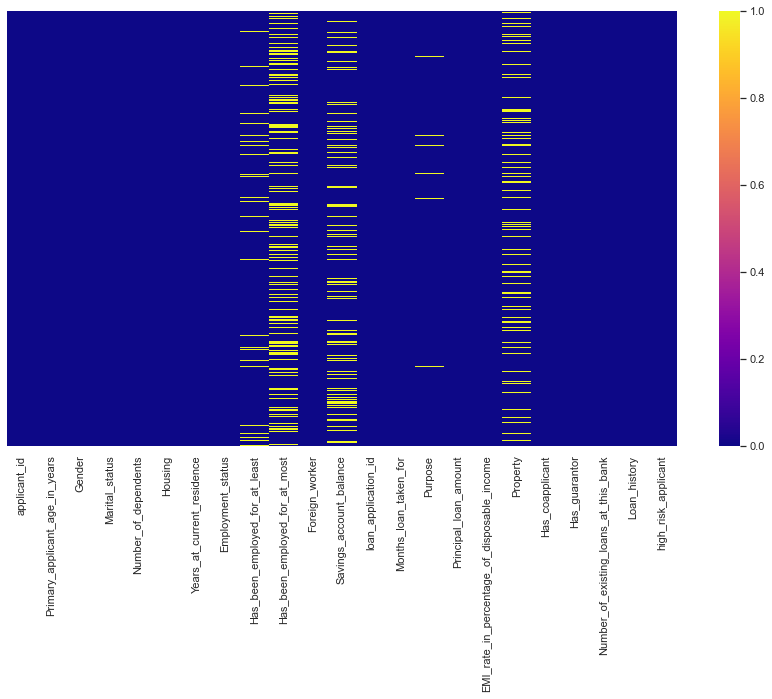
The Picture above enlists the columns having null values in descending order.

As you can see there are 9 features with missing values most of which are non-Numeric.

Hence we would need to apply Imputation where we have at least 60% of the data available and drop the columns with more than 40% of the missing Data.

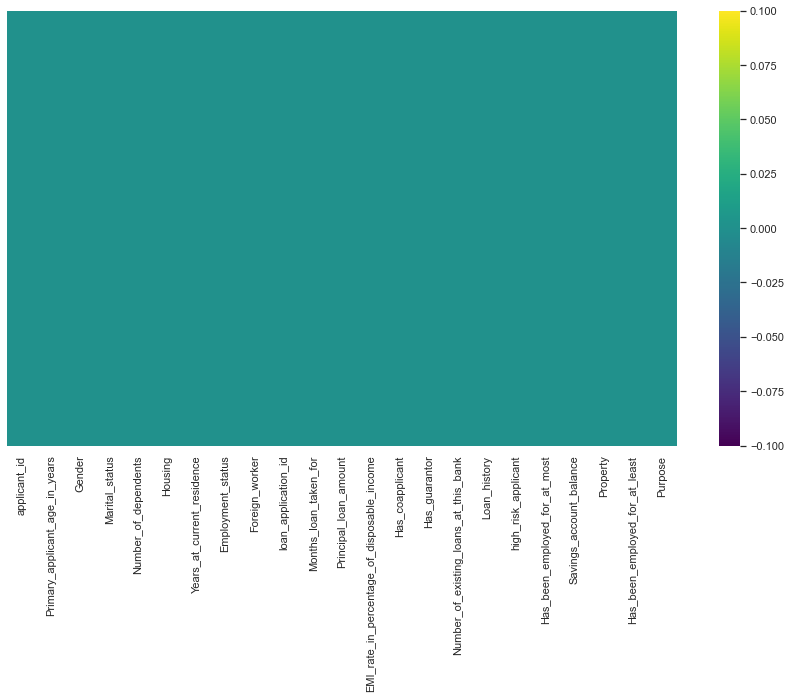
As you can see there are 9 features with missing values most of which are non-Numeric.

Hence we would need to apply Imputation where we have at least 60% of the data available and drop the columns with more than 40% of the missing Data.



After Eliminating top 4 columns with missing values

The Picture above depicts that after eliminating the top 4 columns the density of missing values has certainly decreased.



The chart above represents that there are no null values in the dataset.

**Data Imputation –**

I have used SK Learns KNN imputer class for imputing the missing cells.

The value of k is 5 which mean is will take the weighted average of 5 nearest neighbors as the missing value and also, fill it in the missing cells.