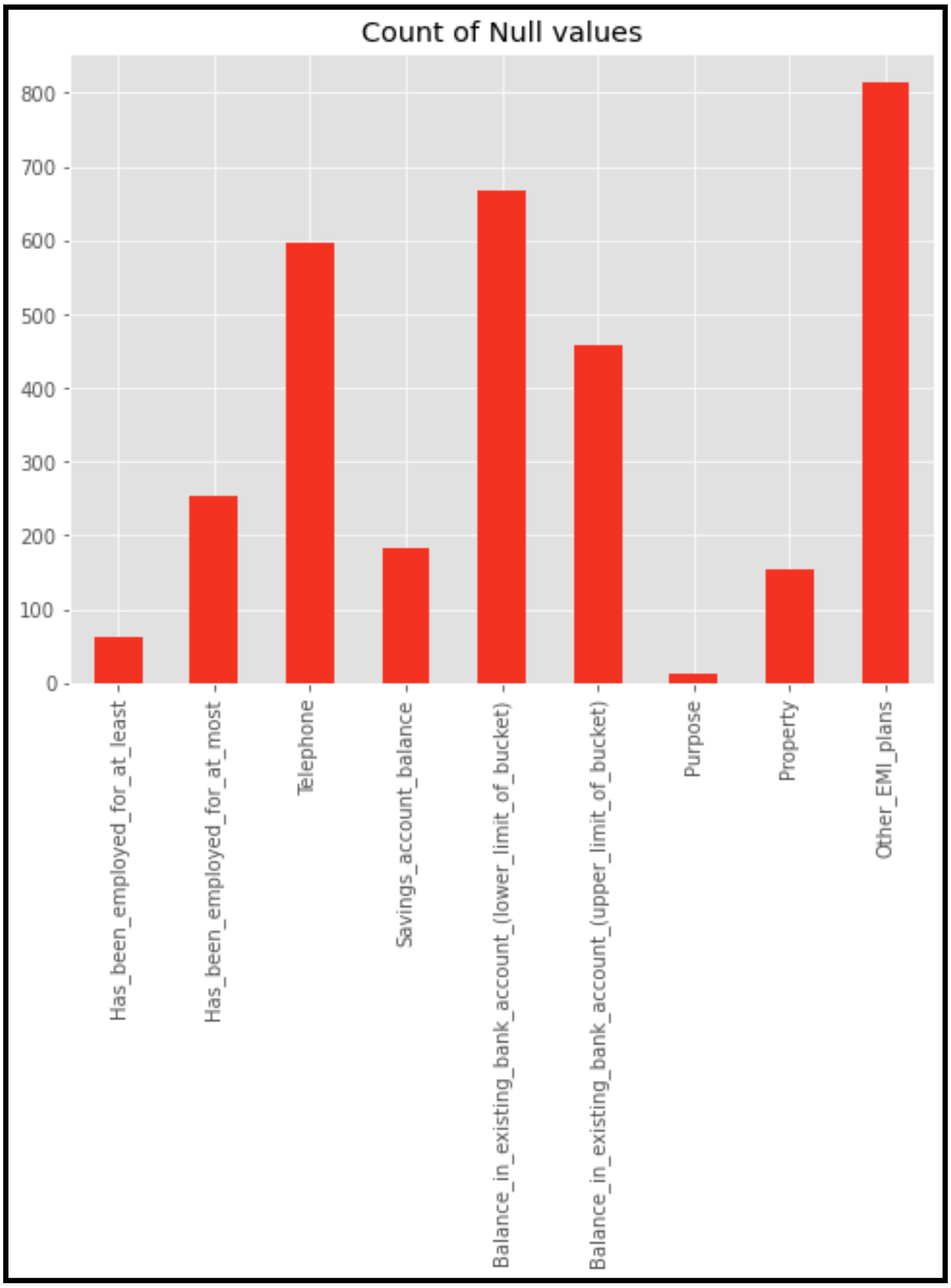
Documentation for the tasks

TASK 1- EDA

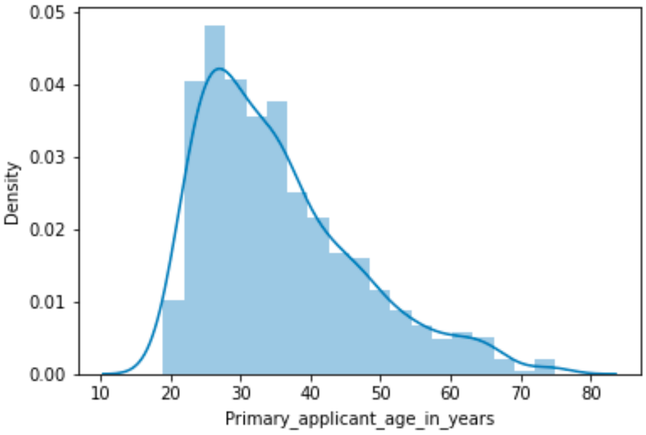
- Merged the two datasets into a single one by inner joining them with primary key value applicant id.

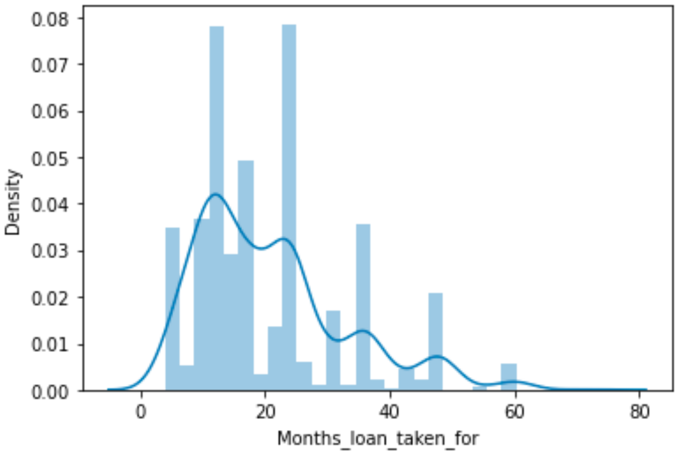
Interesting insights from the data those are found are the following:

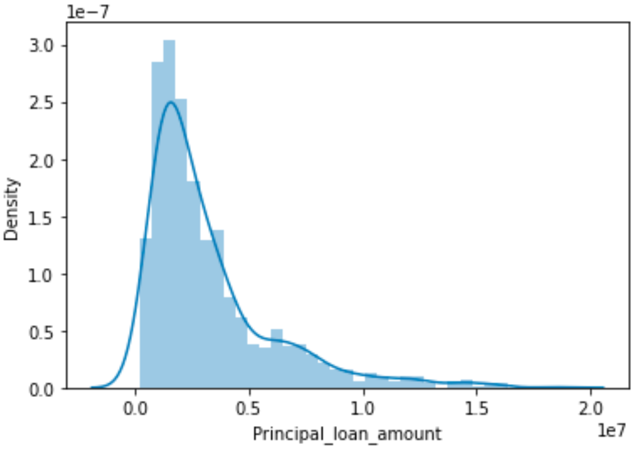
- There are null values present in the dataset. Roughly 11.84 % of the data is null.



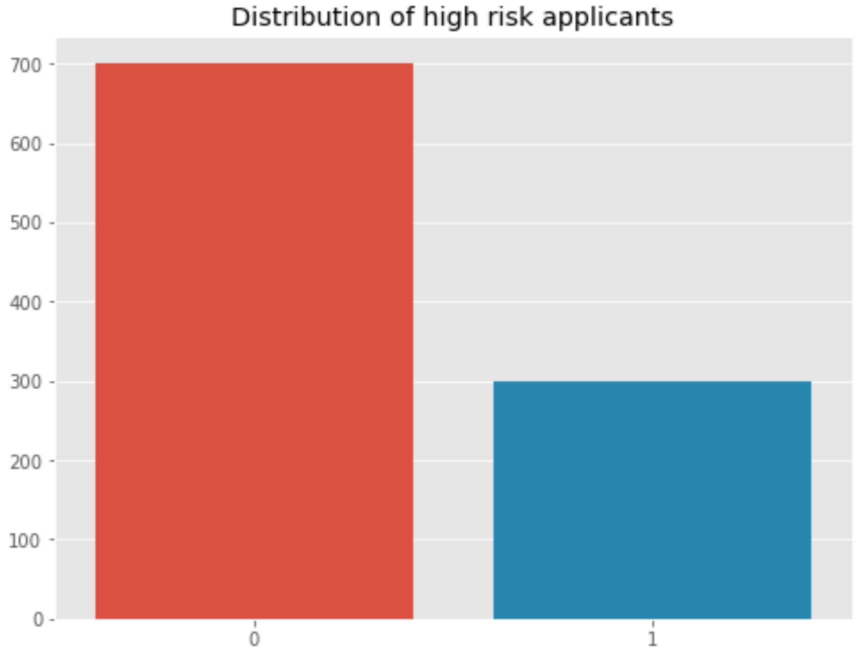
- The continuous variables of the data are rightly skewed.



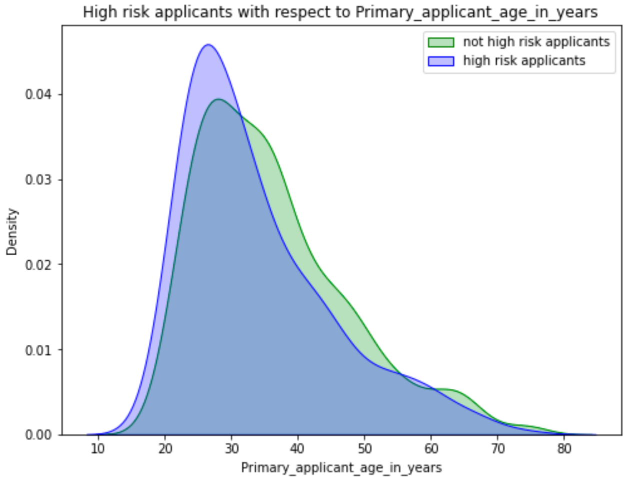


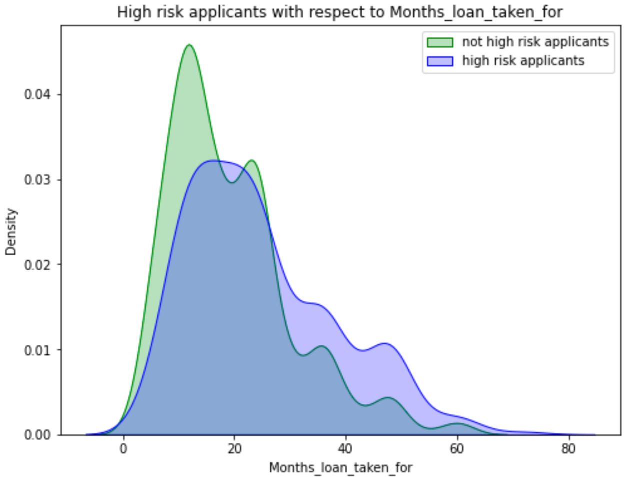


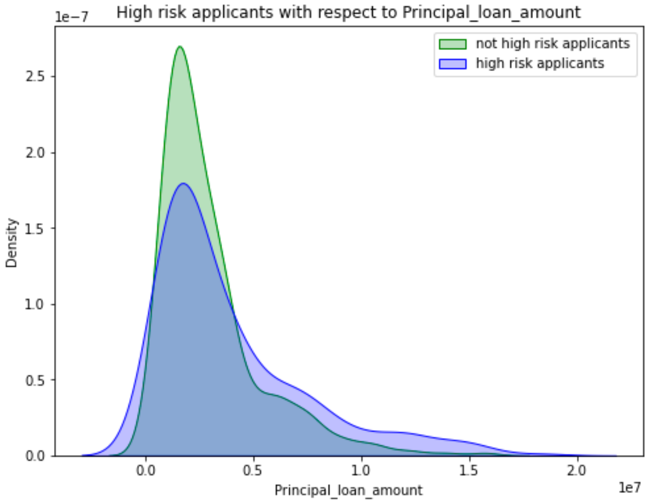
- The distribution of the target variable is moderately imbalanced with 70:30 ratio.



- The density distribution plot for the continuous variables segmenting the high risk and low risk candidates are given below.







- Applicants with marital status single has more low risk candidates.

- Applicants who own their housing has more low risk candidates.

- Most of the Skilled employee / Official employee belong to low risk.

- Those who have critical/pending loans at the other bank and existing loans paid back only till now have more applicants. These two sub groups also interestingly have more low risk applicants than high risk.

- Most of the applicants do not have co-applicant or guarantor.

- Telephone column only have one value.

- loan purpose for electronic equipment shows largest amount of applicants with lower credit risk.
- loan purpose for new vehicle shows largest amount of applicants with higher credit risk.

[I have not included all of the visualisations from Jupyter Notebook]

Task 2- Building ML Classifier

- Merged the two datasets into a single one by inner joining them with primary key value applicant id.
- Counted the null value and null value percentage for each column.
  - Dropped the column with more than 80% missing values
  - For columns with more than 30% missing values performed the chi square test to see how relevant those columns are for the target variable. The test fail to reject the null hypothesis that the columns are independent. Hence we dropped the columns.
  - Imputed missing values for rest of the columns.
- Calculated the outliers for continuous variables and capped them.
- Measured the Correlation coefficient of the numeric variables.
  - Measured VIF for the variables for further feature selection.
  - Dropped the columns until the VIF was within acceptable range (< 10)
- Removed redundant features (ids)
- Label encoded the categorical variables
- To handle the imbalanced dataset SMOTE was used in the Model training pipeline.
- Recall value was given more priority for reducing the False Negatives as per the business constraints.
- Fitted 3 ML models:
  - Decision Tree Classifier
  - Random Forest Classifier
  - XGBoost Classifier
- Used RandomizedSearchCV for hyper parameter tuning.
- Recall score of XGBoost model outperformed the other two.