Domain Case Study Assignment on BFSI Bank Institution Default Problem

Submitted By,

Gayatri S. Gera

Problem Statement

- To acquire the right customers for the home loan department by applying the end to end process by model development.
- To assist Home Credit in deciding which loan applications should be disbursed or rejected based on applicant's past behavior and application information.
- To determine the factors that are affecting the home credit risk.
- To generate the strategies so that the financial benefits and credit risk will be alleviated.
- Lastly to identify the individuals that has the willingness, ability and/or integrity to pay it back.

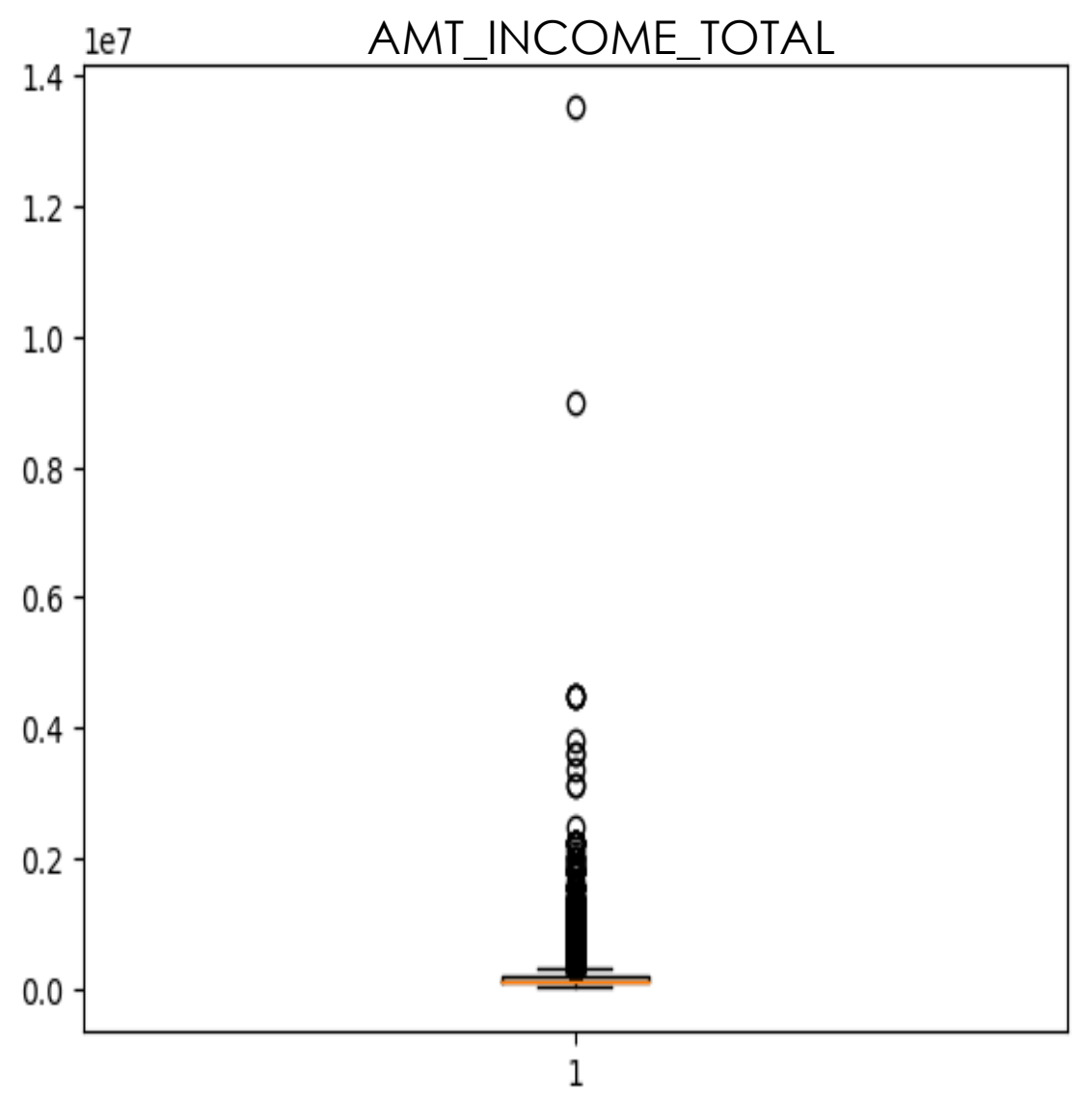
Data Understanding

- There are two datasets given viz. application.csv & bureau.csv
- Application data: This data is obtained from the information provided by the individuals at the time of applying for the home credit loans. It consist of customer level information like age, marital status, job , income type etc.
- Bureau data: The bureau information is at trade level, each individual trade level information is provided. Banks and non-banking financial companies are bound by a periodic mandate to report all of the trades of its borrowers to the Credit Bureaus.

Data Preparation

- Checking all the columns for the missing values
- Checking all the columns for Nan
- Checking the necessary columns for duplicate values
- Outliers detection using the boxplot and quartiles.
- Dummy variable creation for the necessary features.

Outliers Detection using Box plot

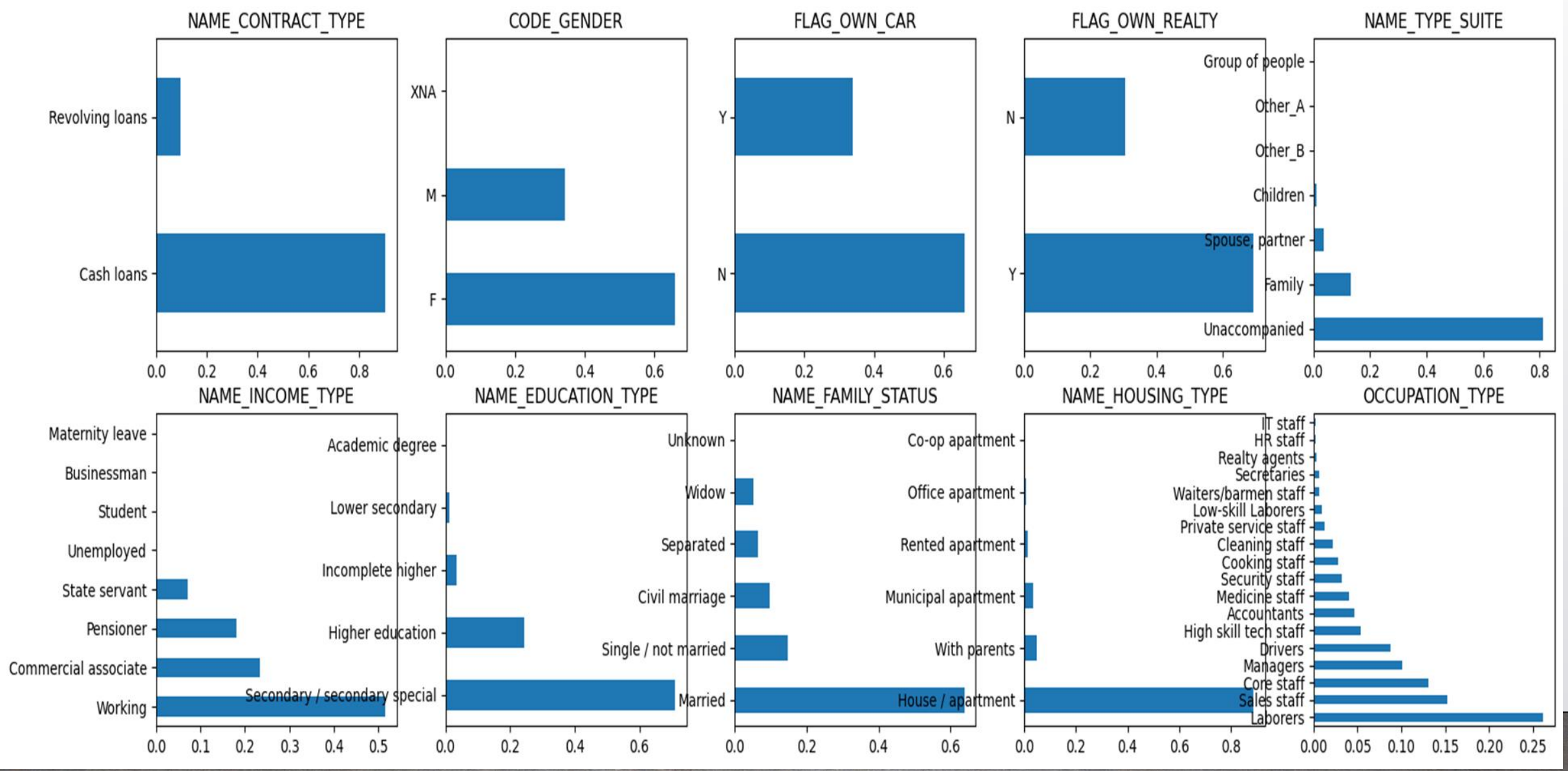


- There are outliers in the Total income amount.
- These difference in the quartiles are expected as it is based on income sources which vary according to the individual.
- Other columns does not contain outliers as only few values lies outside the quartiles.

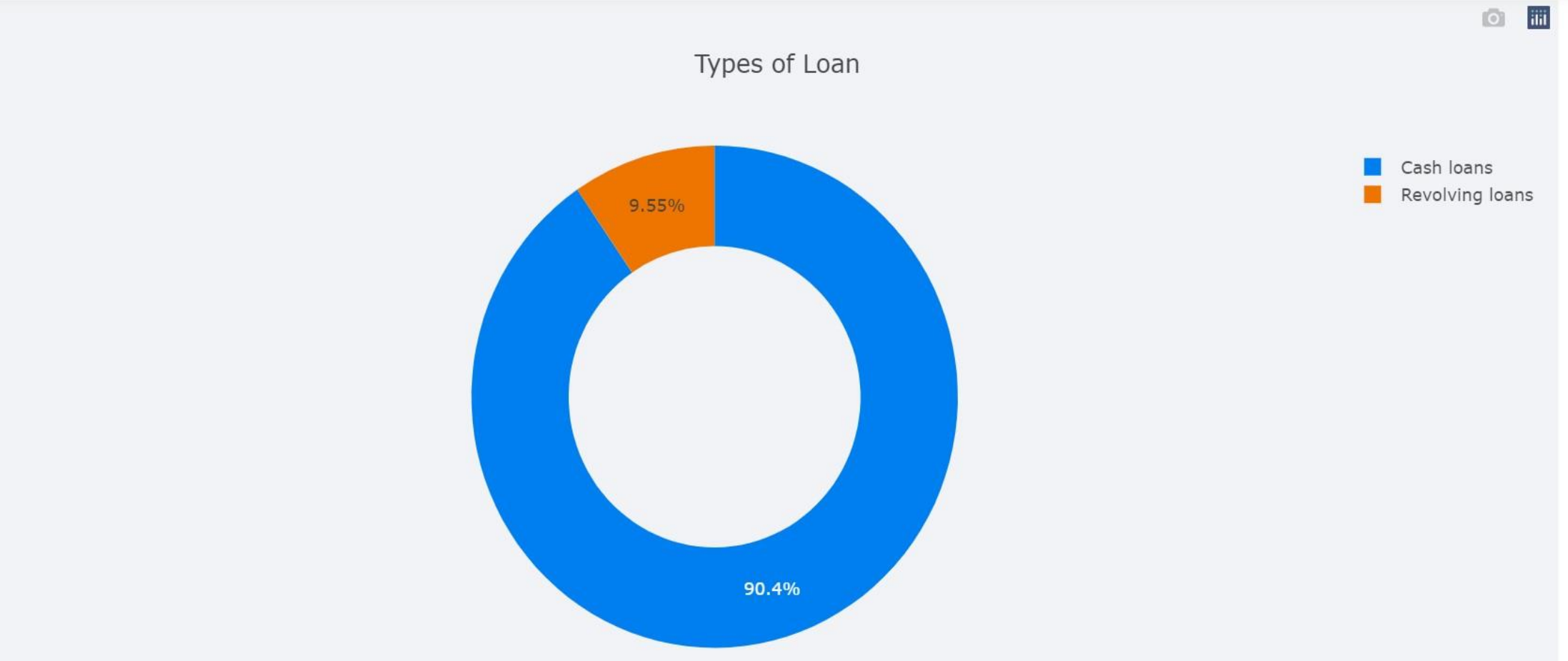
Data Cleaning

- There were no duplicate values found in the entire dataset.
- TARGET is the dependent variable and it was converted into the integer form as it was represented as a string value.
- Imputed the empty values in the columns of the dataset such as contract type, marital status etc.
- Removed all the NAN values from the columns such as 'NAME_HOUSING_TYPE', 'OCCUPATION_TYPE','WEEKDAY_APPR_PROCESS_START','ORGANIZATION_TYPE','FONDKAPREMONT_MODE', 'HOUSETYPE_MODE', 'WALLSMATERIAL_MODE', 'EMERGENCYSTATE_MODE’ etc.

Exploratory Data Analysis

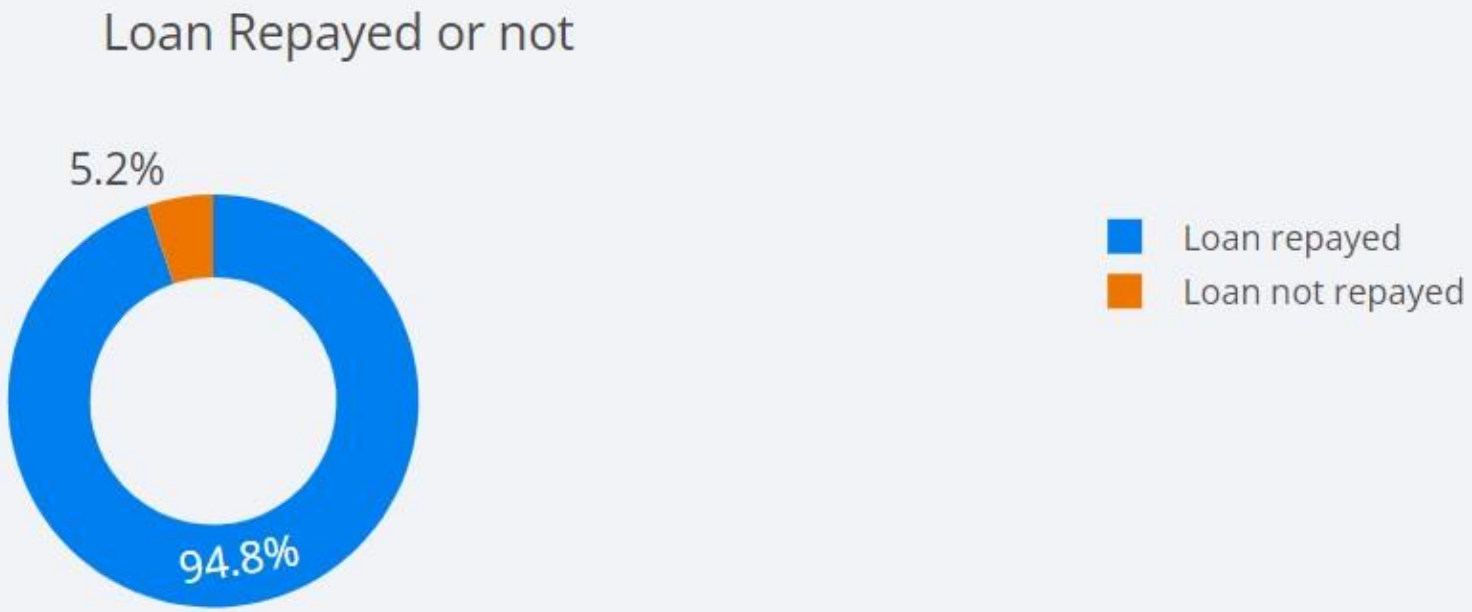


Exploratory Data Analysis



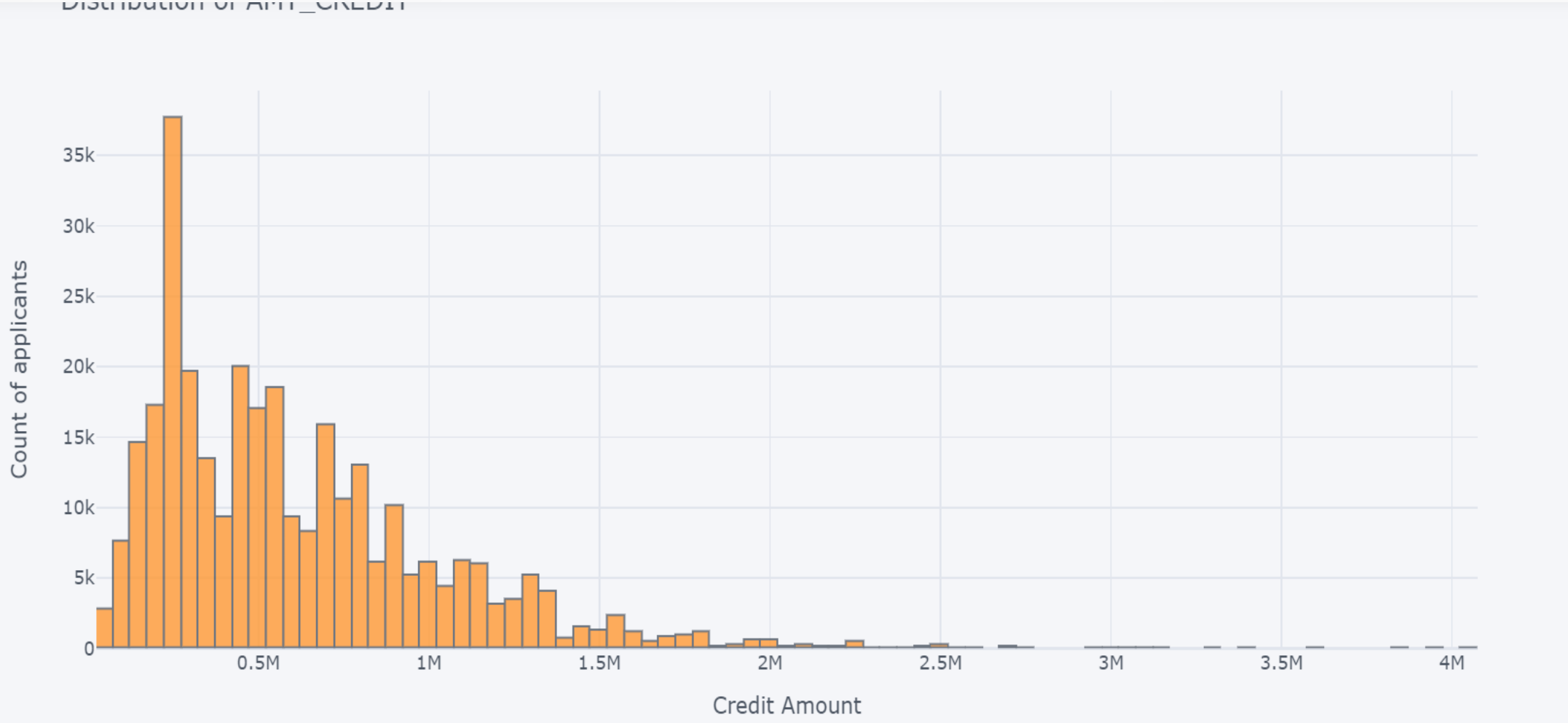
It can be seen from the above pie chart that consumers prefer cash loans instead of revolving loans

Exploratory Data Analysis



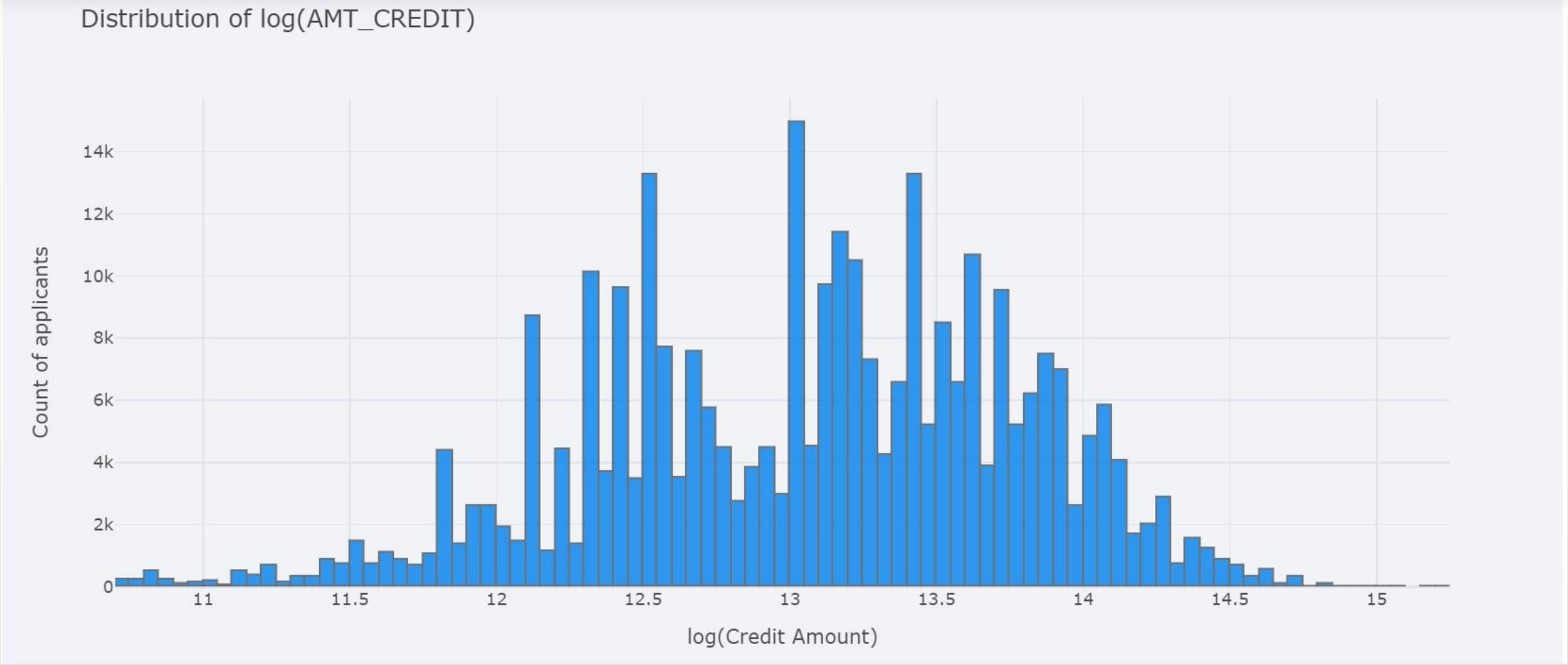
The data is imbalanced for dependent target variable (94.8%(Loan repayed-0) and 5.2%(Loan not repayed-1)) and this need to be tackled

Exploratory Data Analysis



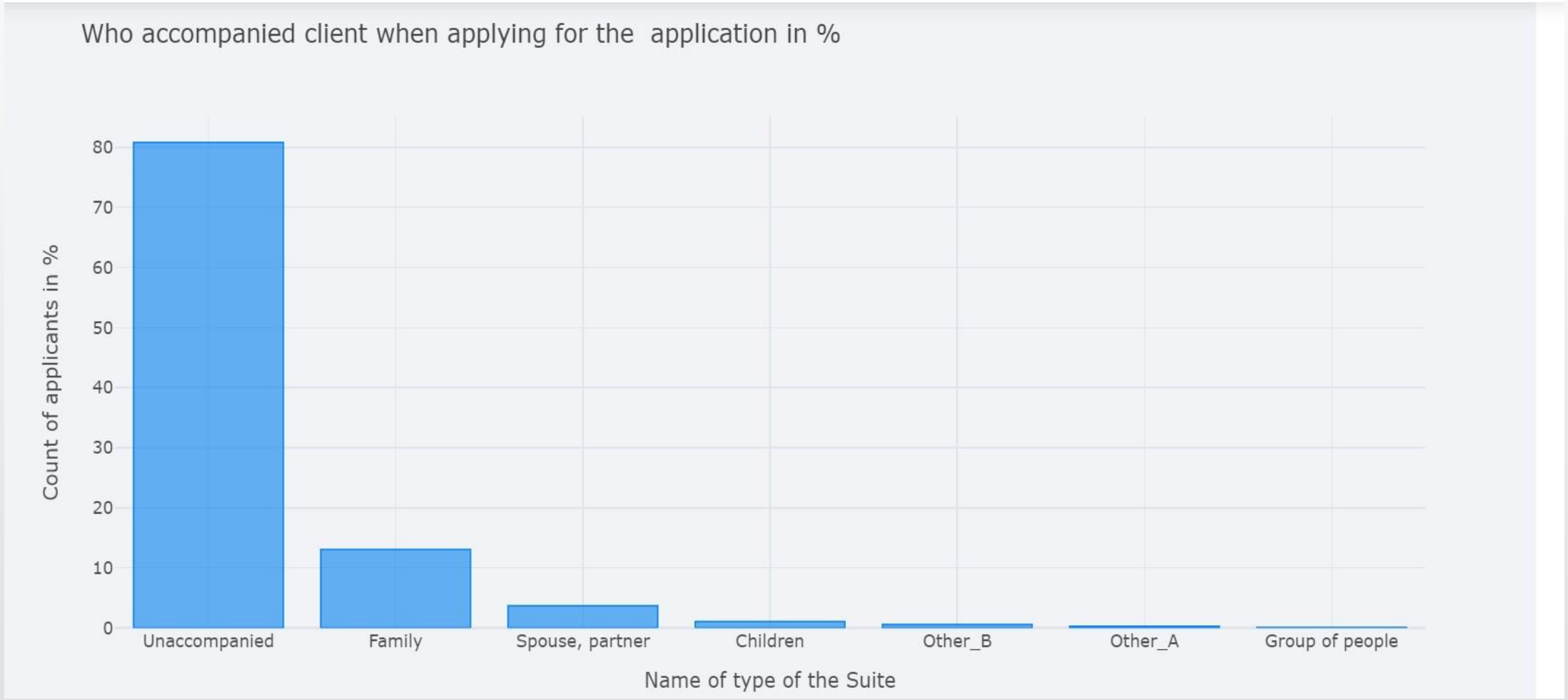
- **The proportion is right skewed and there are extreme values, we can apply log distribution.**
- **Consumer with high income(>1000000) are likely to repay the loan.**

Exploratory Data Analysis

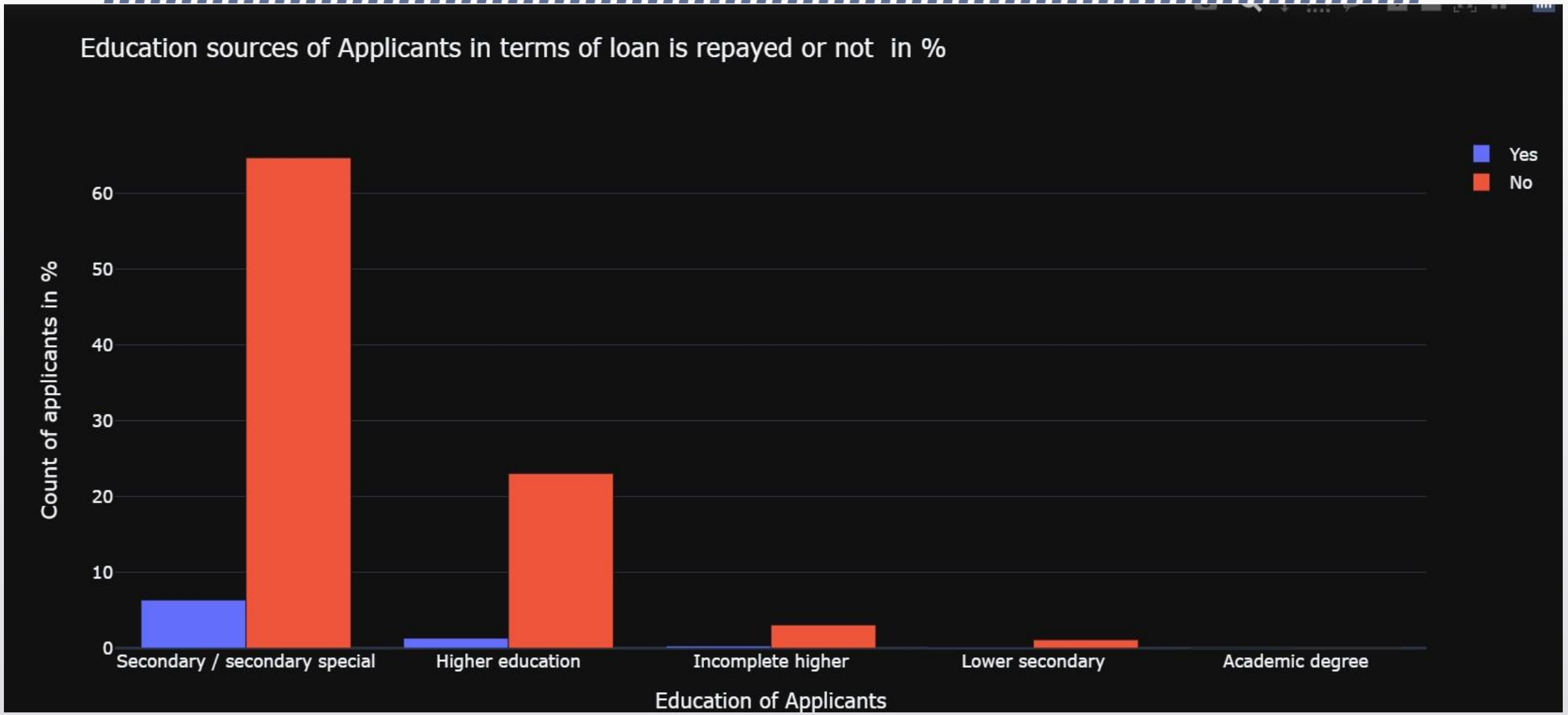


- **Now the distribution can be seen as normally distributed after log transformation.**
- **Consumer who are taking credit for large amount are very likely to repay the loan.**

Exploratory Data Analysis

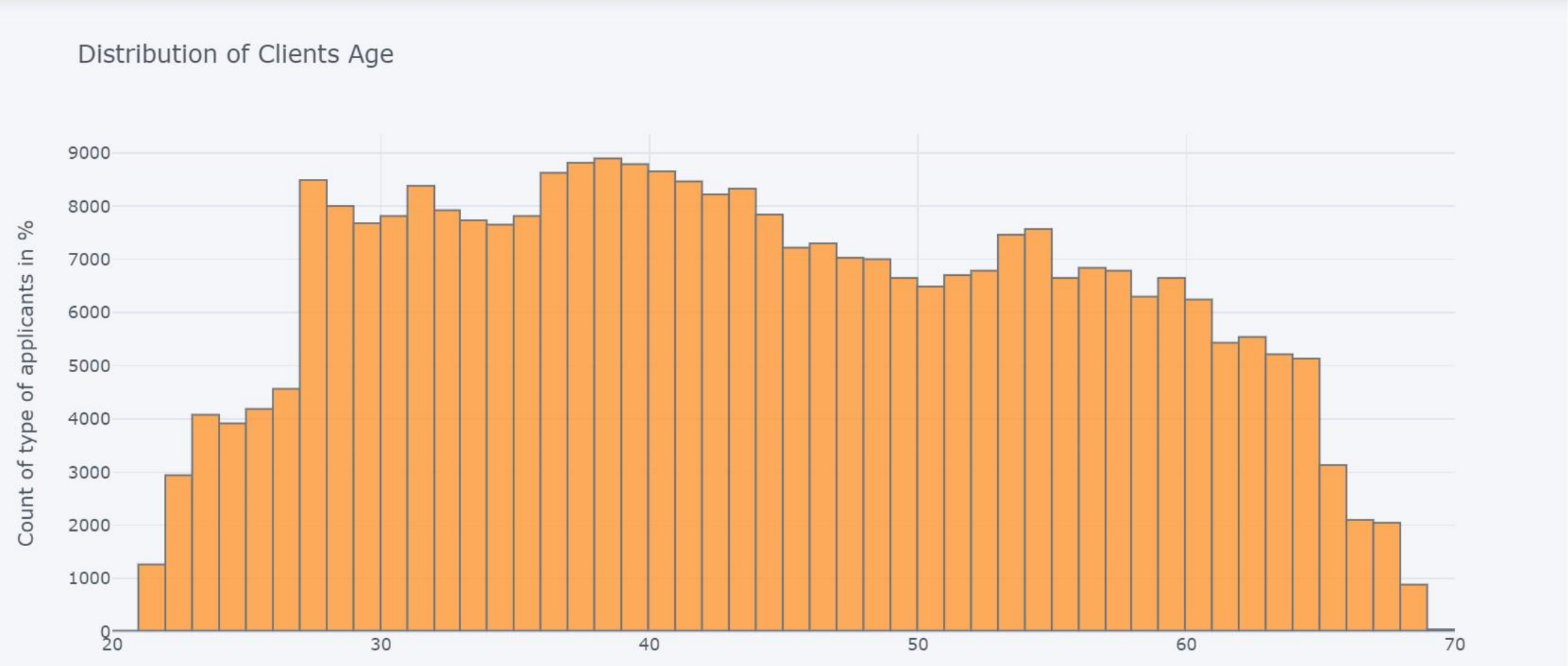


Exploratory Data Analysis



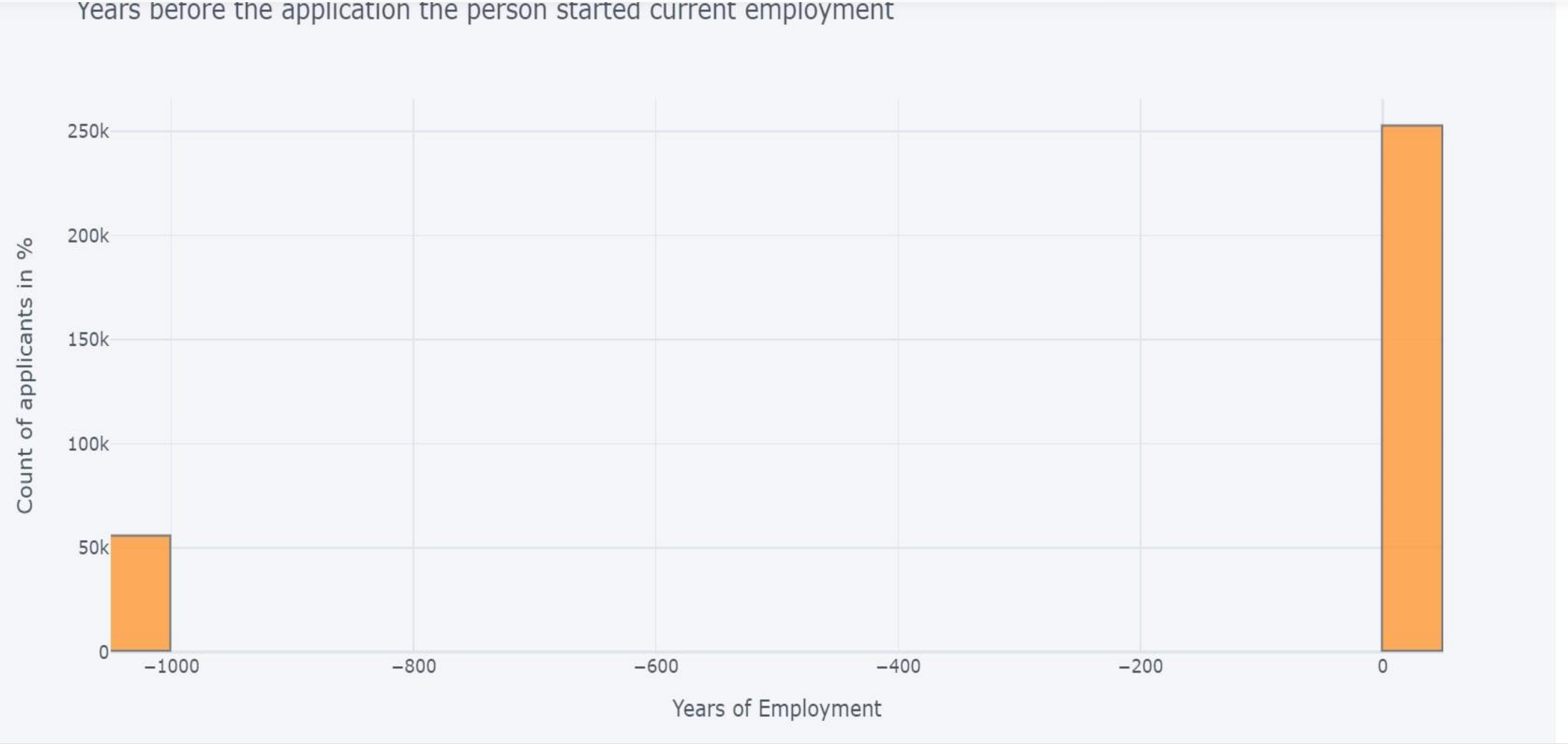
All the applicants having higher education are most likely to repay the loan

Distribution of Applicants age

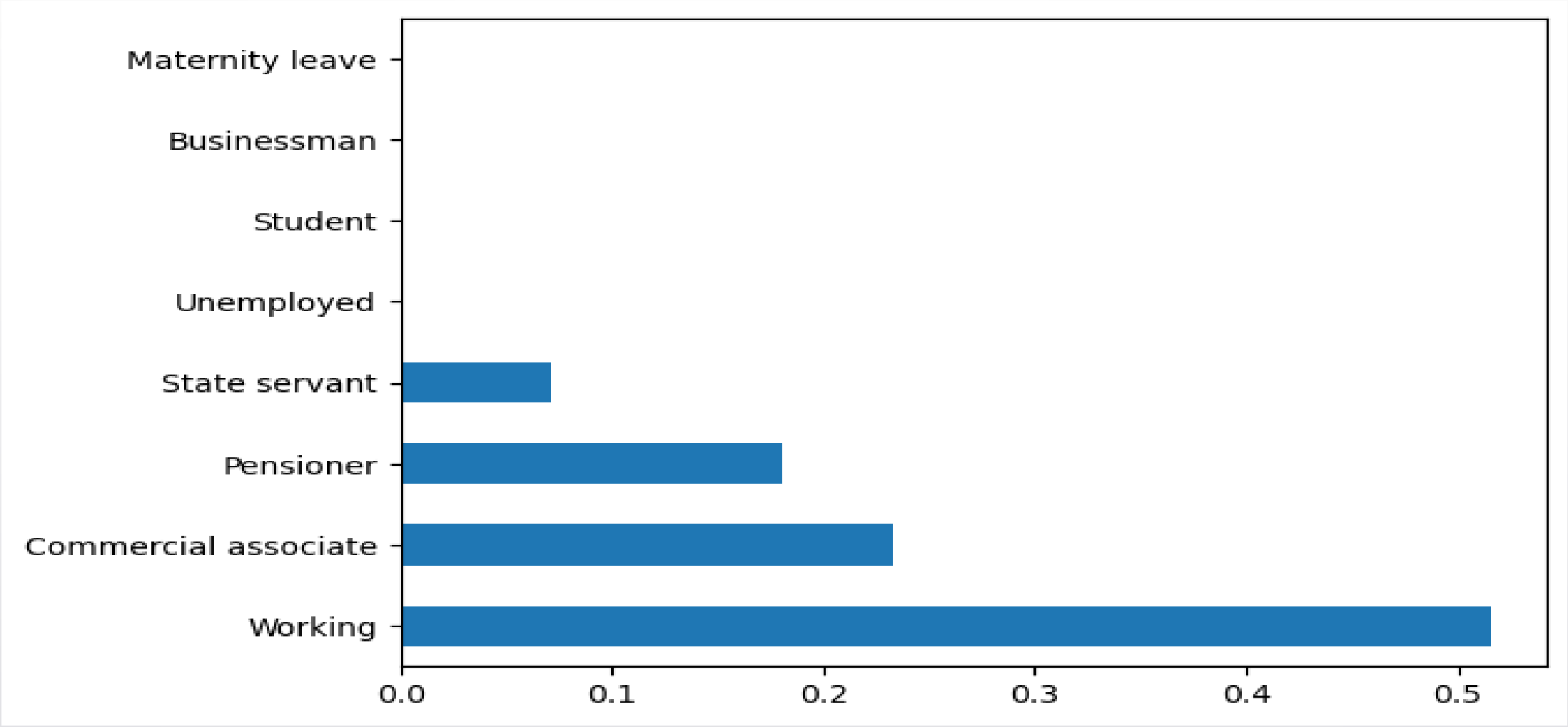


- **Its wide rage of age groups of applicants seeking or took the loan**
- **From the graph it is evident that age group of 25-60 years are the maximum number of applicants.**

Current Employment of the applicants in days



The data looks bizzare it seems there is some data entry error



From the graph we can see that the working class and commercial associate are most likely to repay the loan

Concluding remarks after performing EDA

Following factors seems to contribute in building the model for performance evaluation-

- Amount credited in the customers account
- Total annual income
- Marital status
- Source of Income (Type of employment)
- Educational qualification
- Type of loan
- No. of Dependent
- Age

Building predictive model using logistic Regression

- Considering the type of problem this can be classified into two categories – Defaulters and Non-defaulters.
- For this we can use the following two different models-
  - **Logistic Regression**
  - **Random Forest**
- Differentiating the data into train and test sets.
- In random forest we need to vary the number of trees, minimum number of leaves and buckets.

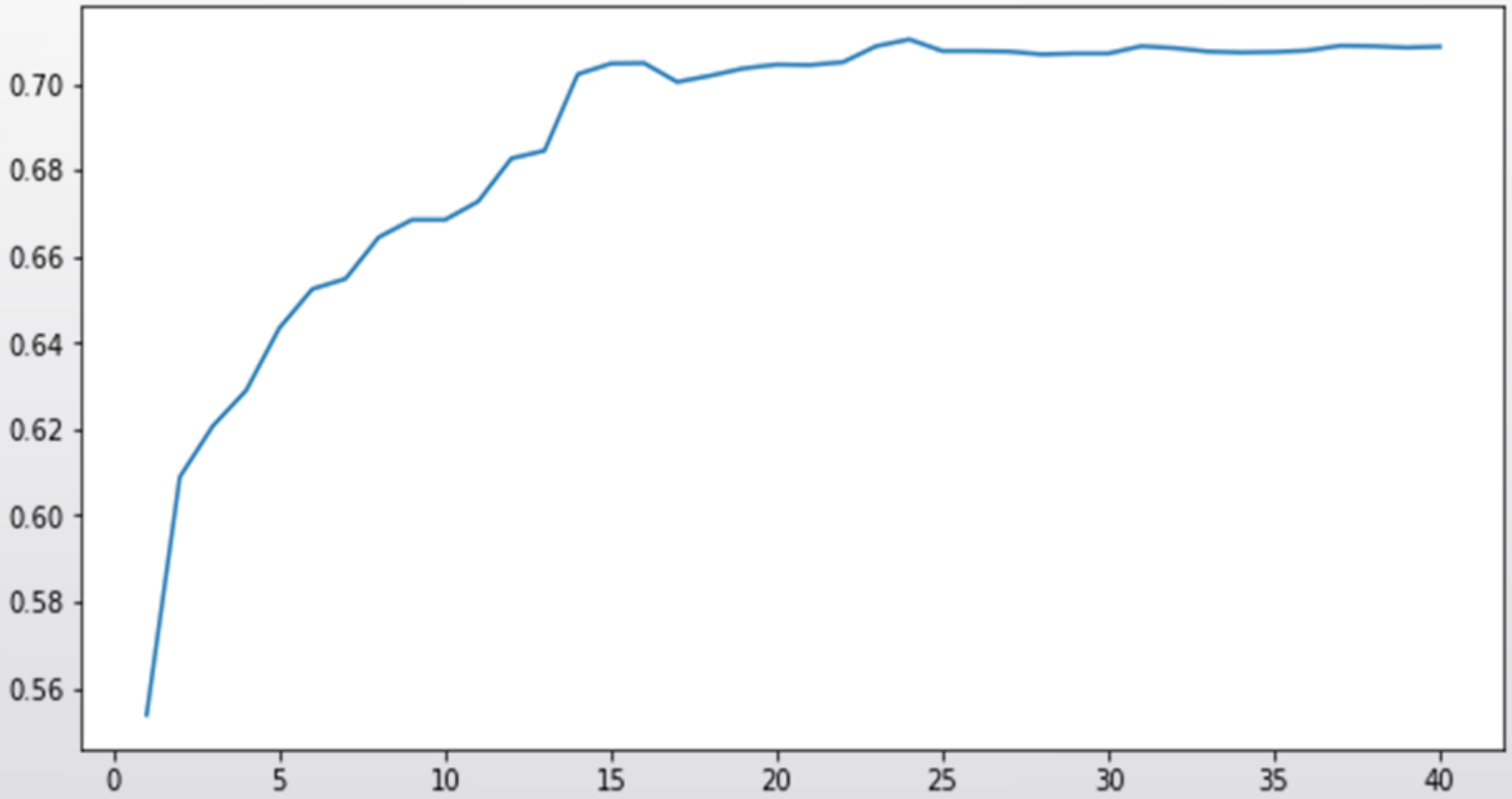
Logistic Regression model Classification chart

	precision	recall	f1-score	support
0	0.94	0.99	0.97	226148
1	0.73	0.28	0.41	19860
accuracy			0.93	246008
macro avg	0.84	0.64	0.69	246008
weighted avg	0.92	0.93	0.92	246008

Using Random Forest

- Accuracy Score (on train set)= 0.998
- Accuracy Score (on unseen data)=0.693
- Plotting the confusion matrix for the best cut off values.

Cross Validation with Recursive Feature Elimination



In this case the model is fine tuned using RFE

Conclusion

- Logistic Regression provides the supporting boundary which foremost separates the given positive and negative data points.
- Random forest gives a better conclusion as compare to other models to take the end node.
- In the conclusion we can say that the bank should not only target the moneyed clients but also consider the other applicants characteristics while sanctioning the loan amount.
- This study utilizes the predictive power of different modelling technique to asses the critical features which guides the bank in credit granting and predicting loan default.