Credit EDA

Assignment

By:- Anjali Jain

Problem Statement

- This dataset is of customers who applied for a bank loan.
- This dataset has a columns “TARGET” which shows which customers has payment difficulties.
- This dataset to by analysed and factors to be identified which helps finding the defaulters and good customers.
  - Bank don’t want to give loan to defaulters.
  - Bank want to give load to good customers.
- The previous application dataset has the loan approved/rejected in past, this data to be analysed to check which category customers were given the loan previously and should not be given loan now as they have defaulted or not paying timely.

Objective

- Perform EDA process on the datasets.
- Find some useful insights, patterns and variable which clear indicates
  - User is likely to default the loan or having payment difficulties.
  - User will pay their instalments on time.
- Insights from these datasets to help bank for identifying the defaulters to mitigate the business loss and good customers to have good business.

Objective

- Perform EDA process on the datasets.
- Find some useful insights, patterns and variable which clear indicates
  - User is likely to default the loan or having payment difficulties.
  - User will pay their instalments on time.
- Insights from these datasets to help bank for identifying the defaulters to mitigate the business loss and good customers to have good business.

Data Understanding

- There are 3 files
  - application_data.csv : The dataset having loan applications currently applied by customers.
  - previous_applidation.csv : The dataset having information on the applications in past, if those were approved, rejected, refused or unused.
  - columns_description.csv : This file is a data dictionary having information on each column of both files mentioned above.

Approach

- Data Cleaning
  - Columns having more than 40% columns removed.
  - Then columns of no use identified and dropped.
- Missing Values Treatment
  - Category type columns are imputed with mode in missing values.
  - Numerical columns are imputed with median in missing values.
- Outlier Treatment
  - Outliers are replaced with the upper or lower limit values in numerical column types.
- Data Standardization
  - Column types are converted to object/category which have less unique values i.e. flag type values having 1,0 which are similar to True,False.
- Binning of the fields
  - Few columns which has multiple values like org type, are grouped in common categories.

Data Analysis

- **Univariate analysis**
  - Used count plot in loops to check frequency distribution of categorical variables.
  - Used box plot to see distribution of numerical variables

- **Bivariate Analysis**
  - Used pair plot for numerical values relational distribution
  - Used bar graph for by grouping on categorical variable and aggregating numerical variable
  - For both categorical variables
    - Create barplot with hue of TARGET variable to see which category has payment difficulties more.

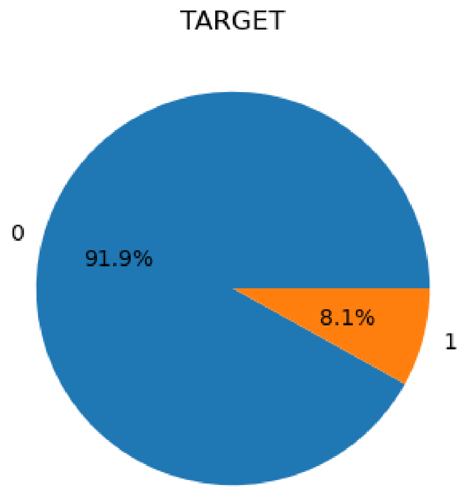
- **Multivariate Analysis**
  - Used heat map by creating pivot table for 2 columns and aggregating target on them

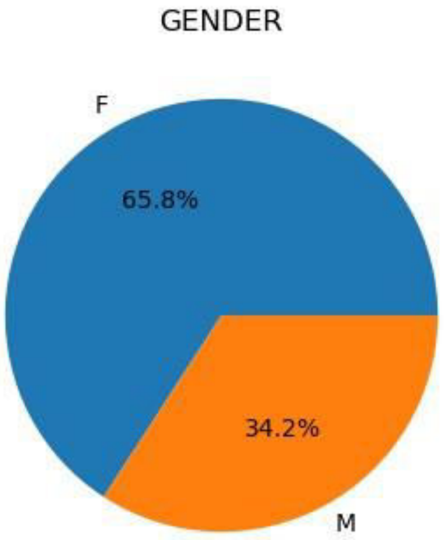
Insights

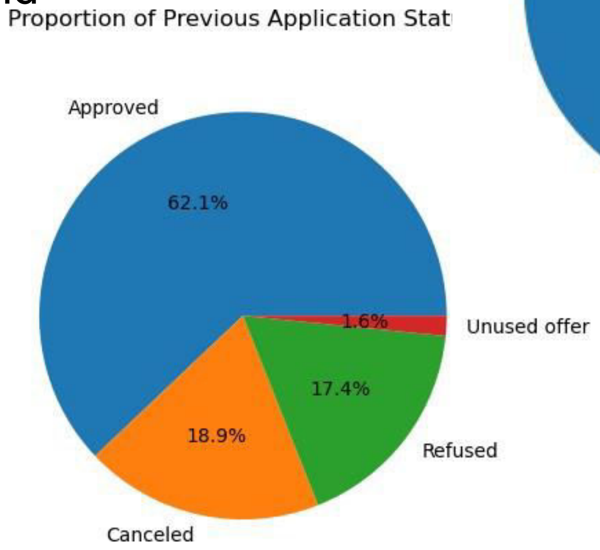
- Target data is not balanced, only 8% clients are shown who defaults.

- Ratio of Male Vs Female who applied for loan is 65% and 35% respectively.

- Almost 62% applications are approved in previous application data and 18% and 17% are Cancelled and refused respectively.

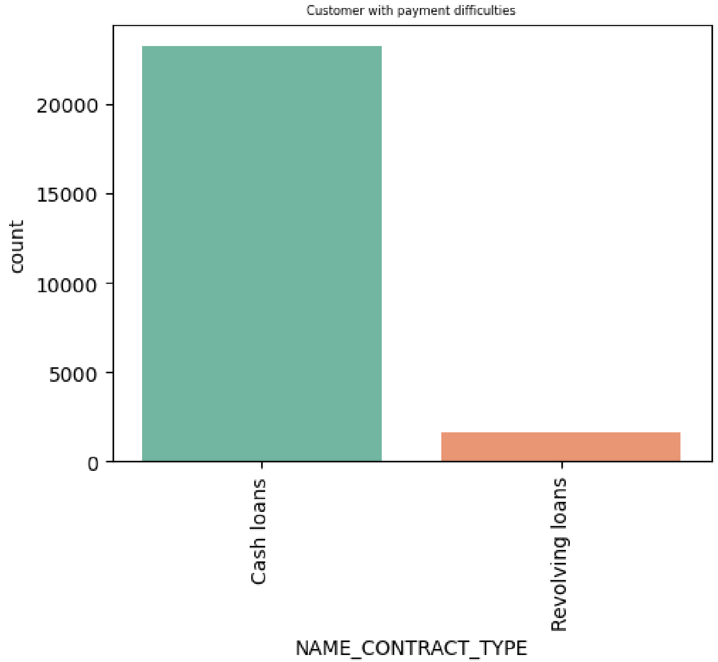




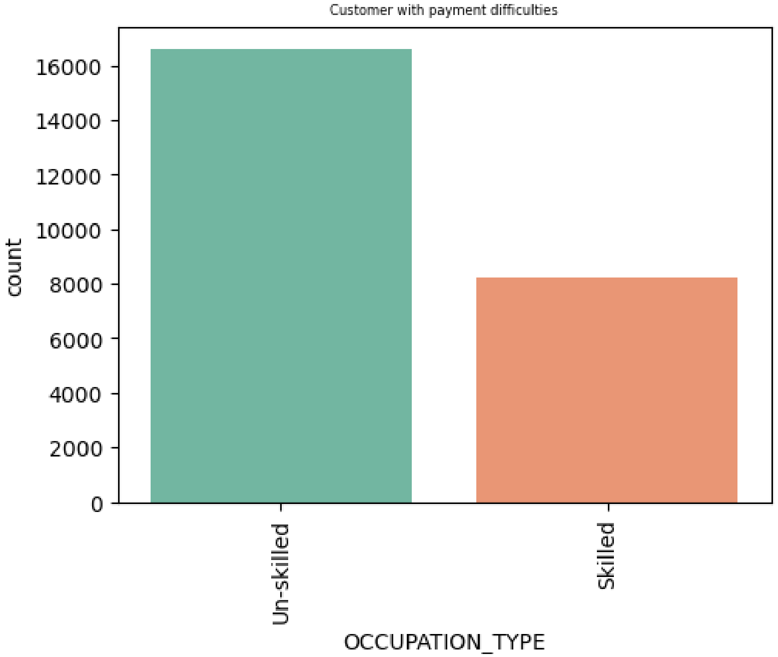


Insights

- Cash loan Contract type clients are majorly doing defaults compared to Revolving loan type.

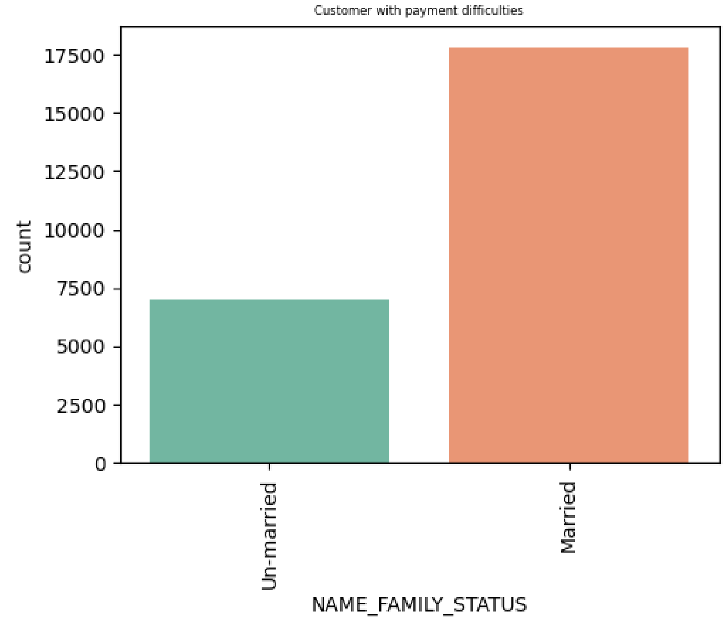


- Unskilled Occupation type clients do maximum default compared to skilled ones.

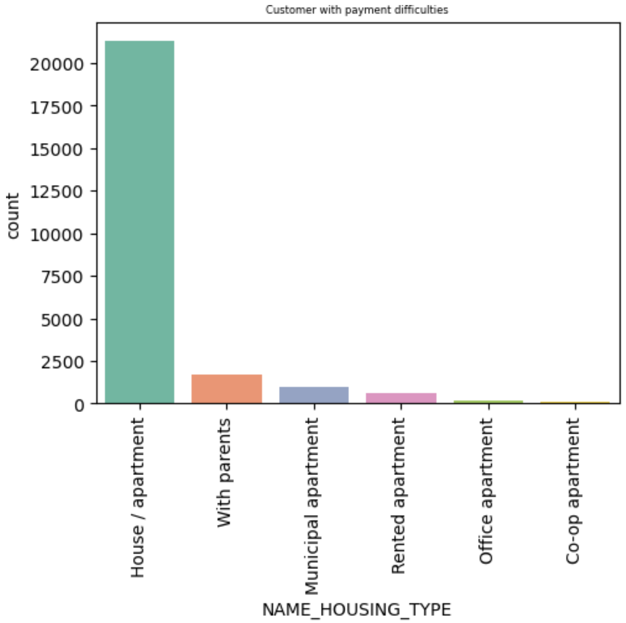


Insights

- Married clients are doing more default compared to Unmarried.

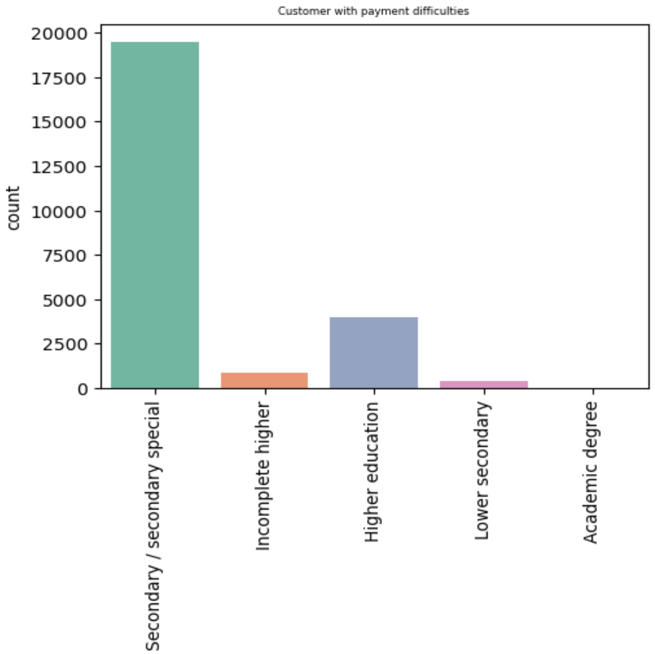


- Clients who owns house/apartment are doing more default compared to other categories.

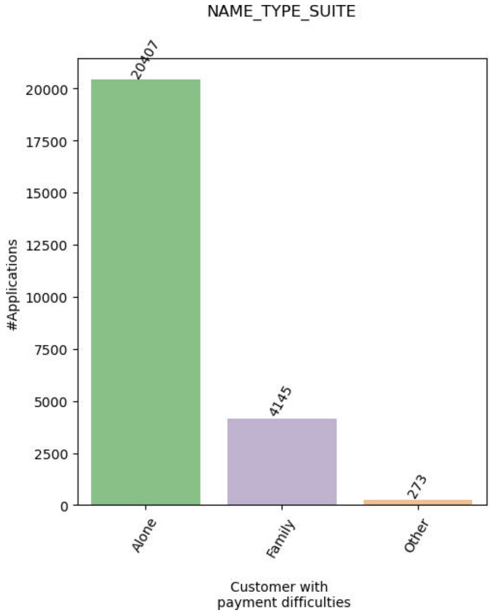


Insights

- Secondary Educated type clients are the maximum who have payment difficulties.

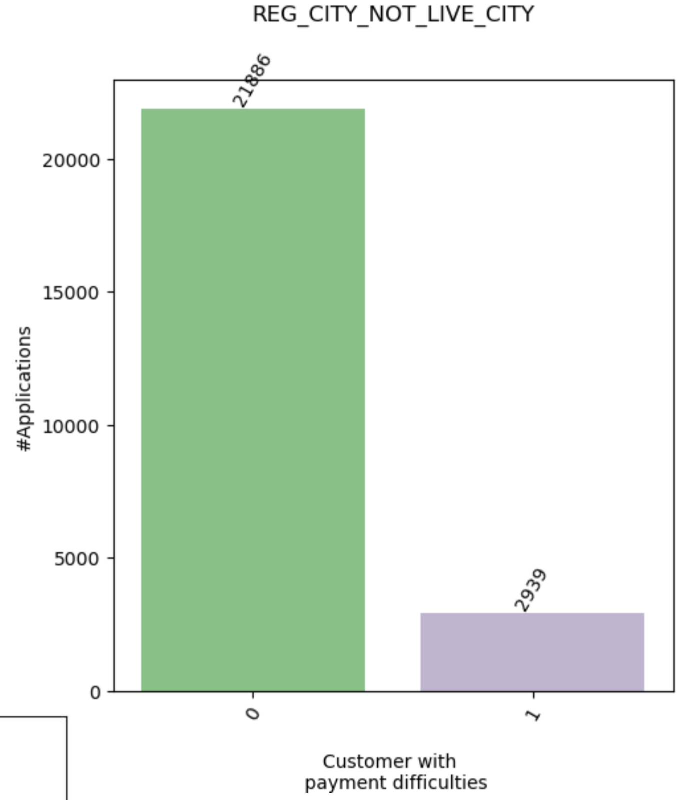


- Clients who are unaccompanied while applying for loan does maximum defaults compared to ones who visited along with family member or anyone else.

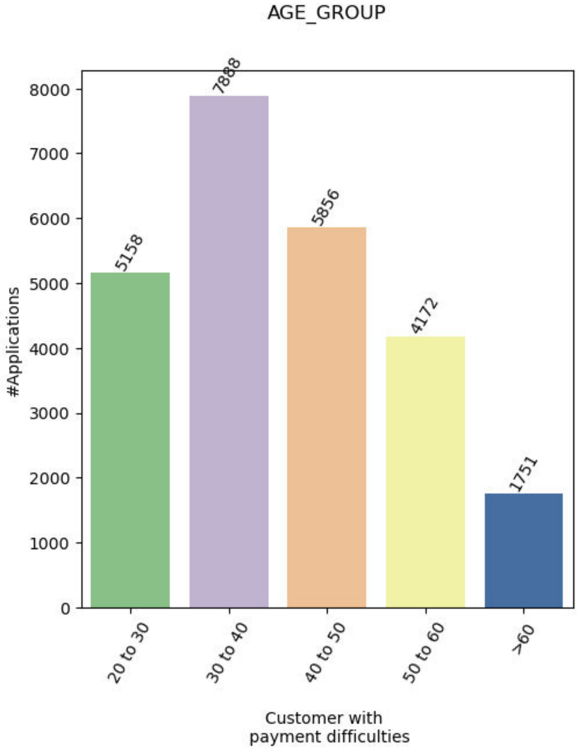


Insights

- Clients whose live city is different than the reg city are doing more defaults.



- Young clients of age group 30-40 are having more payment difficulties.



Conclusion

Factors which indicate possibility of loan default

- Education level : Less educated people are more likely to default their loan.
- Occupation : Unemployed, unskilled labors and people with less stable jobs are more likely to default
- Applicants have more defaulter in their social circle are more likely to default their loan.
- Loan history: Applicant whose previous loans are refused/cancelled are more likely to default their loan.
- Loan frequency: Frequent borrowers are also more likely to default.

Thanks