CREDIT EDA

ASSIGNMENT

PROBLEM STATEMENT

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it to their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specialises in lending various types of loans to urban customers. You have to use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

When the company receives a loan application, the company has to decide for loan approval based on the applicants profile. Two types of risks are associated with the banks decision:

- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

PROBLEM STATEMENT

The data given below contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:

- **The client with payment difficulties:** he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample,
- **All other cases:** All other cases when the payment is paid on time.

Business Objectives

This case study aims to identify patterns which indicate if a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough.

Assumptions

- Dropping columns have null values greater than or equal to 30%
- Dropping below mentioned columns

EXT_SOURCE_2
EXT_SOURCE_3FLAG_DOCUMENT_2
FLAG_DOCUMENT_3
FLAG_DOCUMENT_4
FLAG_DOCUMENT_5
FLAG_DOCUMENT_6
FLAG_DOCUMENT_7
FLAG_DOCUMENT_8
FLAG_DOCUMENT_9
FLAG_DOCUMENT_10
FLAG_DOCUMENT_11
FLAG_DOCUMENT_12
FLAG_DOCUMENT_13

FLAG_DOCUMENT_13
FLAG_DOCUMENT_14
FLAG_DOCUMENT_15
FLAG_DOCUMENT_16
FLAG_DOCUMENT_17
FLAG_DOCUMENT_18
FLAG_DOCUMENT_19
FLAG_DOCUMENT_20
FLAG_DOCUMENT_21
AMT_REQ_CREDIT_BUREAU_HOUR
AMT_REQ_CREDIT_BUREAU_DAY
AMT_REQ_CREDIT_BUREAU_WEEK
AMT_REQ_CREDIT_BUREAU_MON
AMT_REQ_CREDIT_BUREAU_QRT
AMT_REQ_CREDIT_BUREAU_YEAR

- Filled negative employment values in Employment (in Yrs) column with null
- Filled 'XNA' in CODE_GENDER column with null

Approach

Importing the csv file of application data and previous application data

Understanding the column description file

Checking the structure of the data

Missing value check

Handling missing values

- Dropping columns have null values greater than or equal to 30%.
- Dropping unnecessary columns.
- Dropping rows containing missing values.

Approach

Checking and Handling Outlier

- Removing row having salary greater than 100000.

Performing Univariate and Bivariate Analysis

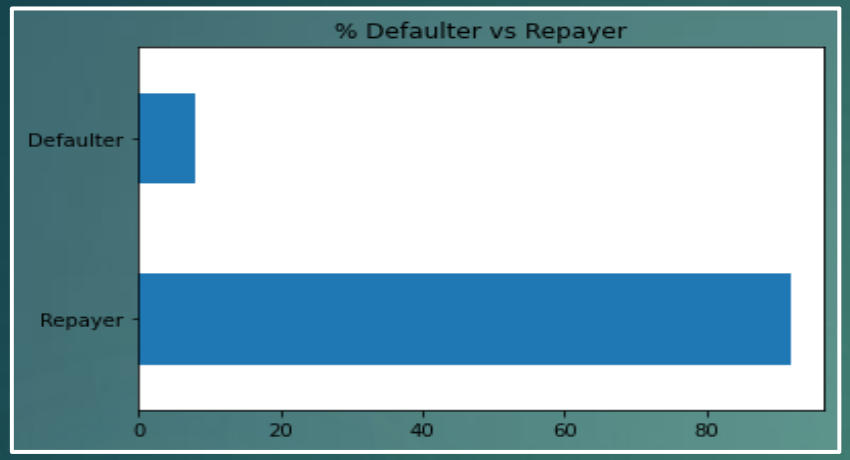
- Analyzing the data using bar chart, pie chart and line chart.

Performing Multivariate Analysis

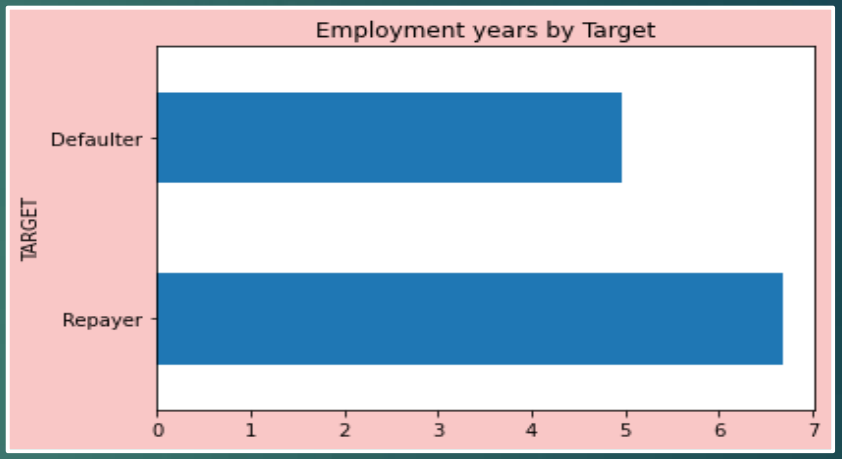
- Analyzing using bar chart, pie chart, scatter plot, pair plot and heatmap.

Analyzing Previous application data

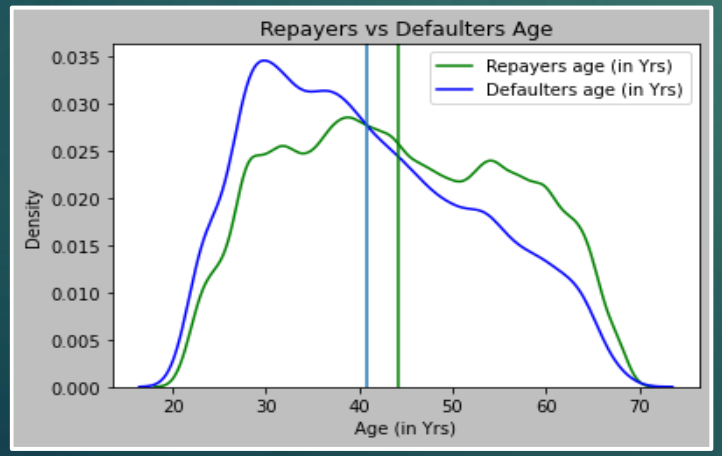
RESULTS



More than 90% of the Clients are repayers



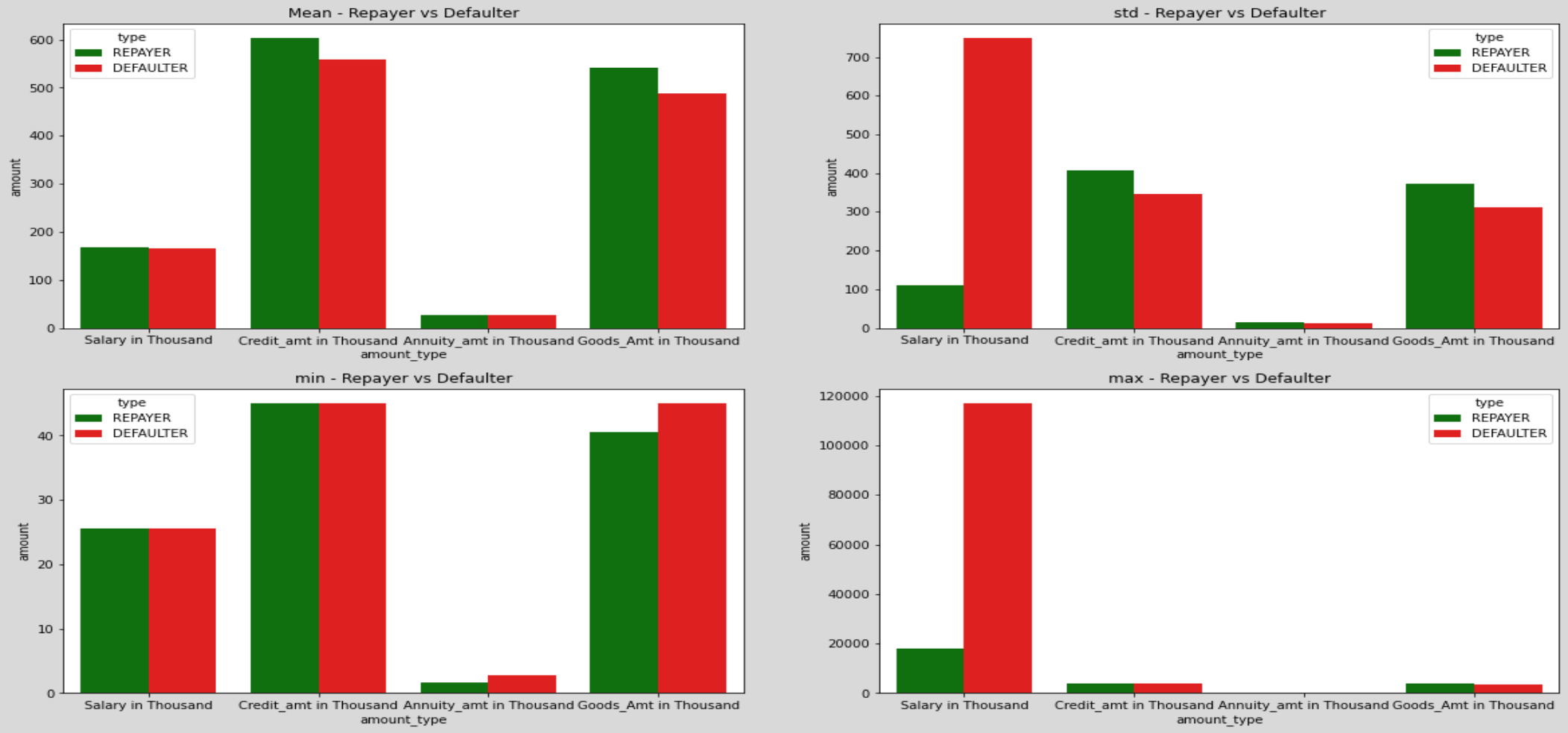
Average of employment years is less in case of defaulters as compared to repayers



- Maximum defaulters have the age between 28 to 33
- Average age of defaulters is less as compared to repayers
- Defaulters tend to decrease as the age increases

RESULTS

- Average income of defaulter and non-defaulter client are same
- Standard deviation of income of defaulter is very high as compared to non-defaulter
- Max income of defaulter is very high as compared to non-defaulter

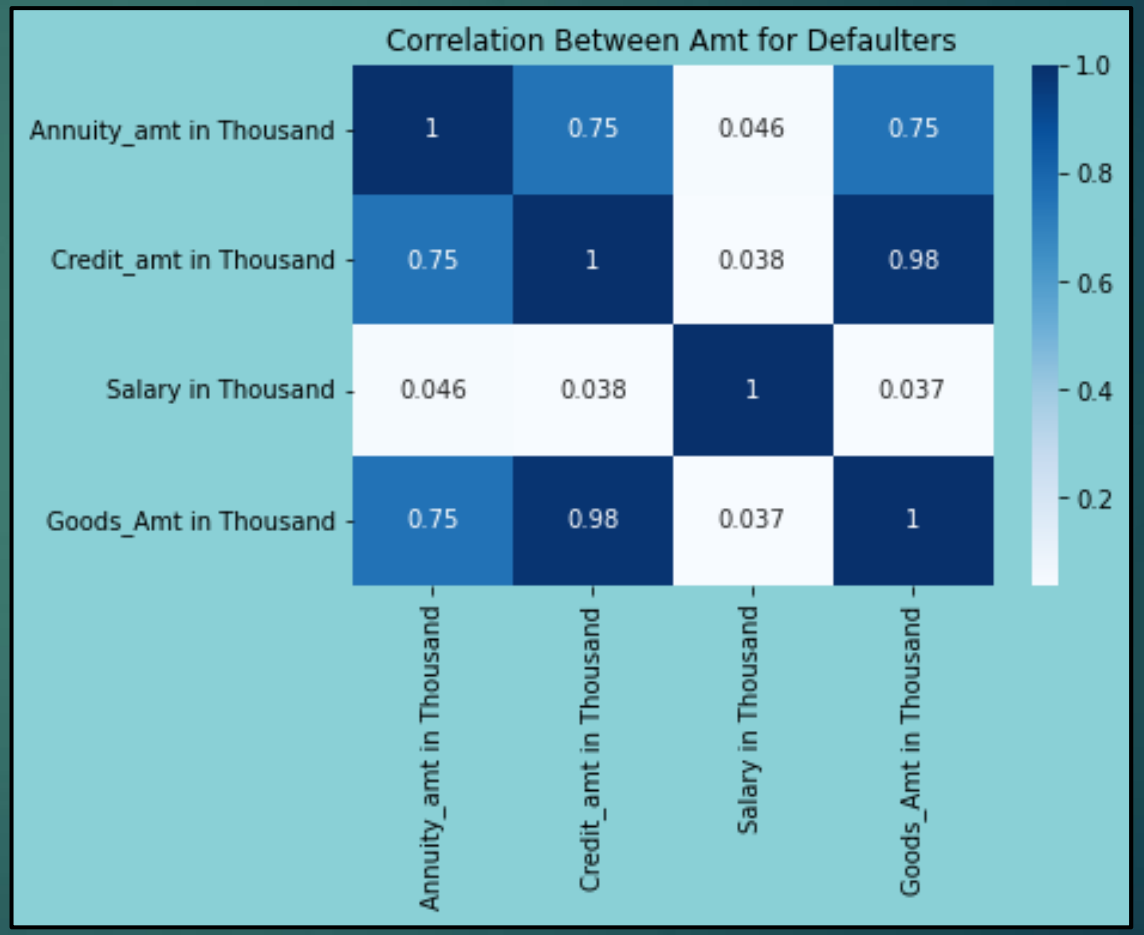


Results

There is a strong positive linear relationship between credit amount and annuity amount for defaulters

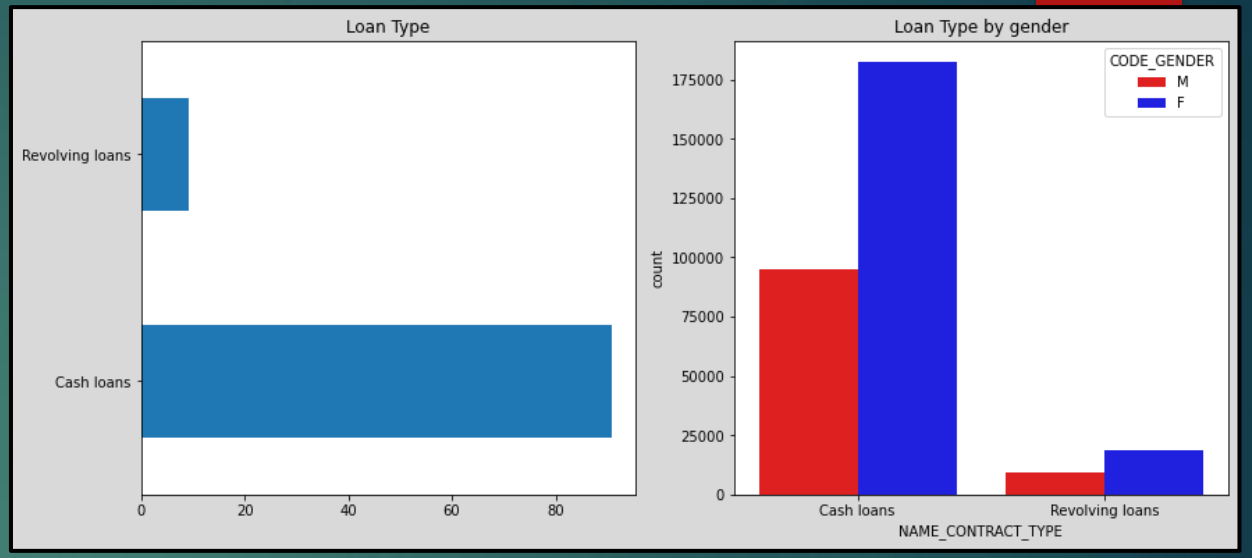
There is a strong positive linear relationship between credit amount and goods amount for defaulters

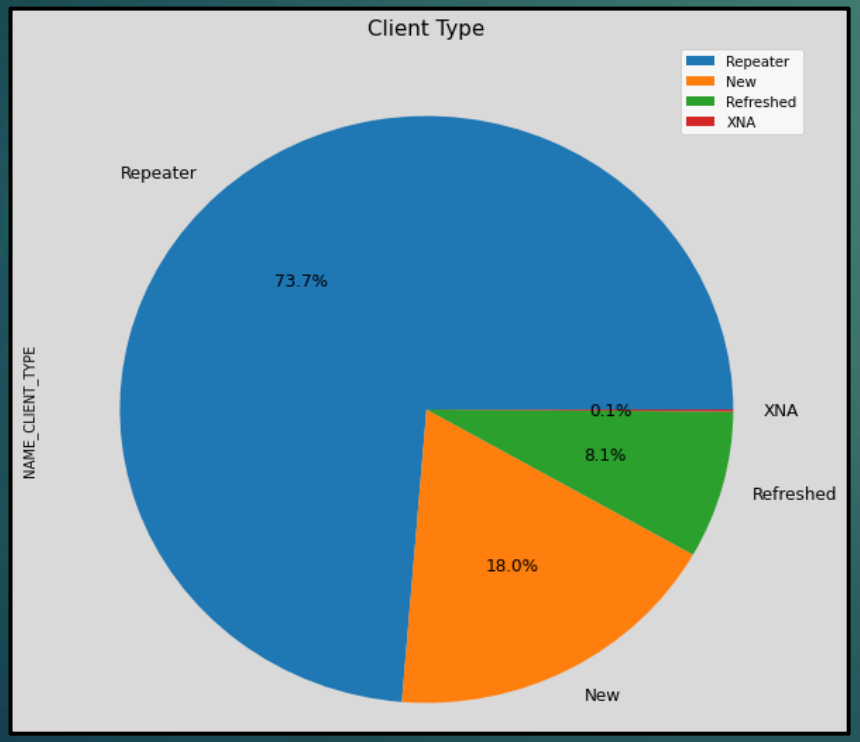
There is a strong positive linear relationship between goods amount and annuity amount for defaulters



Results

- Majority of the clients prefer cash loans
- Both the gender prefer cash loans over revolving loans
- Female prefers cash loan more as compared to male

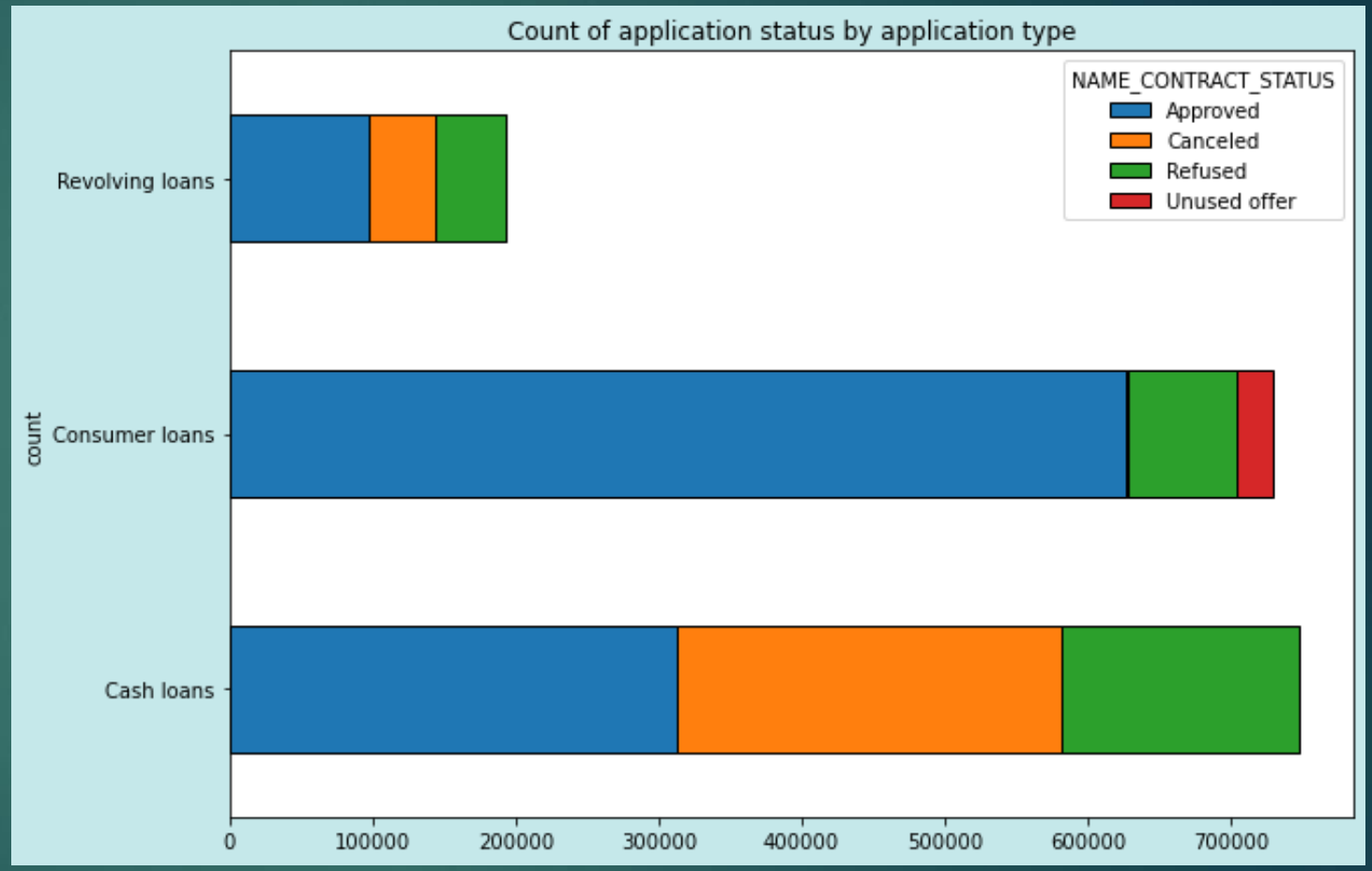




73.7% Clients are old clients whereas 18% Clients are new

RESULTS

- Most Approved loans are Consumer loans
- Most Canceled and Refused loans are Cash loans
- Unused offer is more in case of Consumer loans



CONCLUSION

- There is a high chance of refusal for the customers applying the loan for the purpose of paying the other loan.
- Average salary for defaulters is less as compared to non-defaulters.
- Self-employed customers have a high chance of defaulting due to non-stable income