Credit EDA Case Study

Submitted by:

Ningareddy Modase

Rudin Bose

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 as 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 applicant's profile. Two types of risks are associated with the bank's decision:

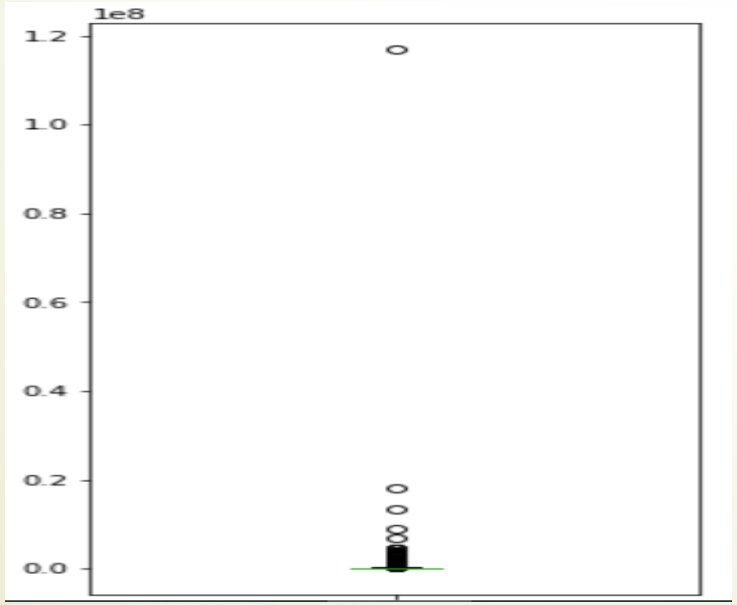
- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company (**Interest Loss**)
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.(**Credit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

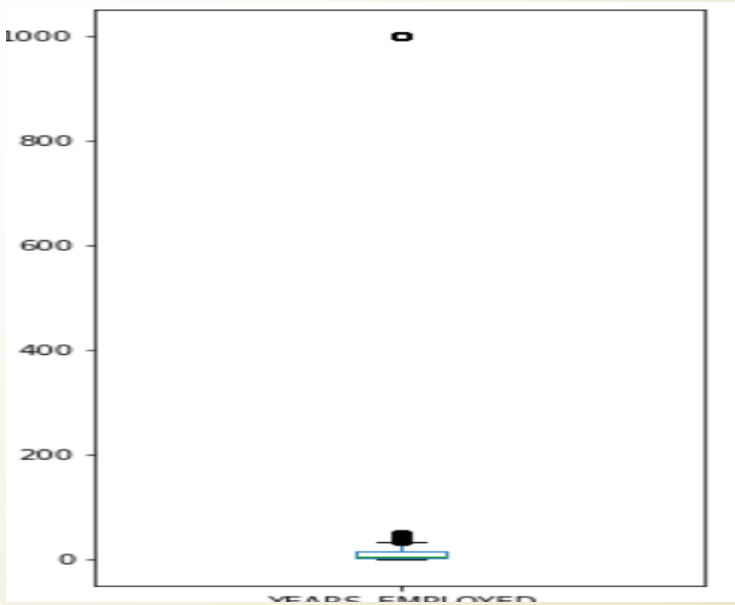
Approach

- Data Cleaning
  - Raw Data
    - Understand the data set as is (i.e. get to know number of variables, records etc) present in the given data set
  - Structure Data
    - Understand the different datatypes used
  - Data Processing
    - Missing value identification
    - Dropping columns whose missing value percentage is > 50 %
    - Possible way to handle missing values for columns whose missing value percentage is <15%
    - Missing value treatment for both continuous and categorical fields (“Median” value for continuous and “mode” value for categorical field is considered as treatment)
    - Variable classification based on their unique type (Assumption is made here to come to this conclusion based on number of unique values. If number of unique values are more than 50, then just variables are considered as “Continuous”)
    - Datatype check and conversion to the most appropriate ones
    - Outliners detection
- EDA
- Insights, Visual Graphics

Findings - Current Application

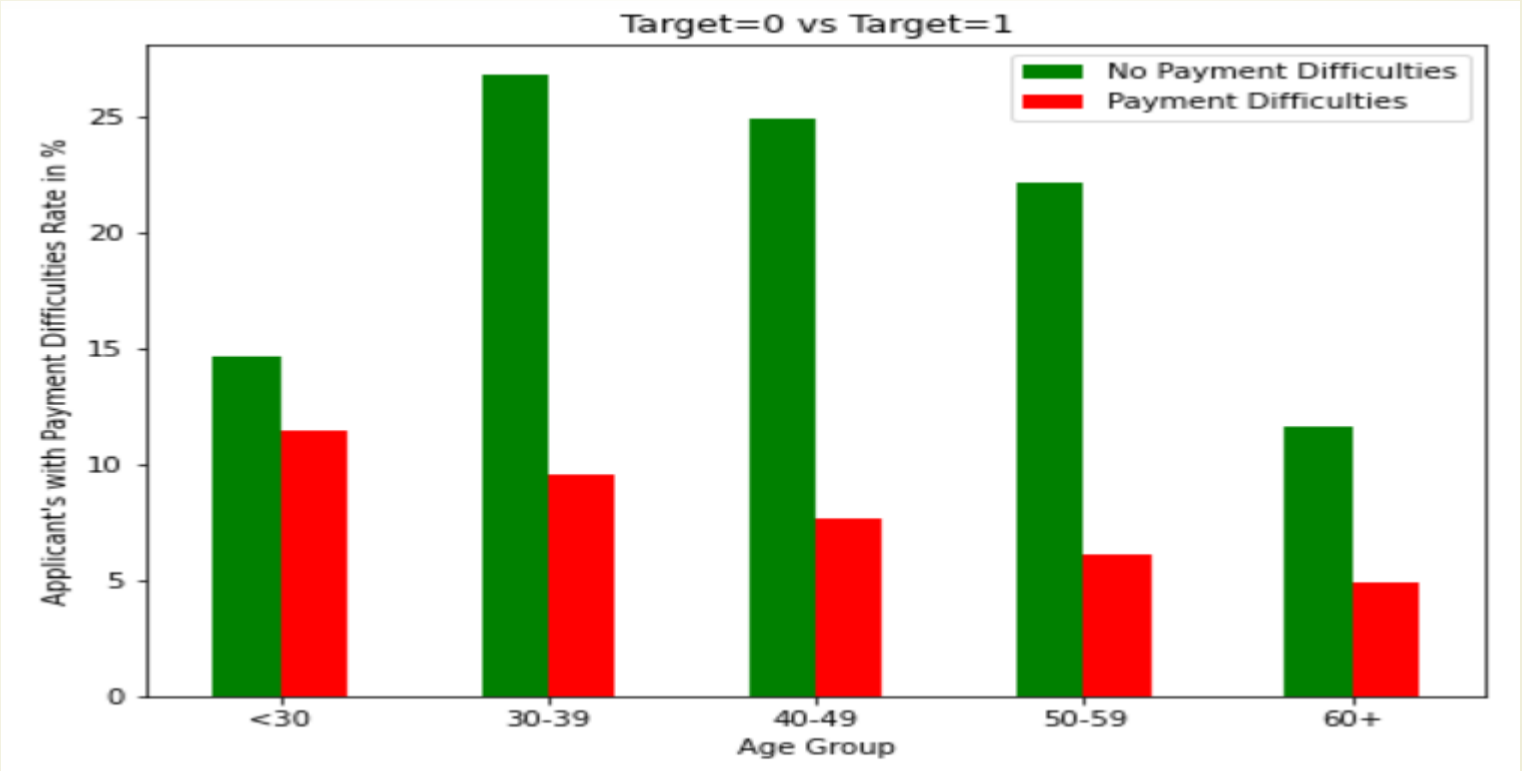
- Missing values Report
  - **55 columns have No missing values**
  - **16 columns have missing value percentage is between 1 and 15**
  - **10 columns have missing value percentage is between 16 and 50**
  - **41 columns have missing value percentage is greater than 50**
- INCOME_TOTAL and YEARS_EMPLOYED are having outliers because of some erroneous data





Findings - Current Application(Contd…)

- Binning : Converting continuous variable to categorical variable

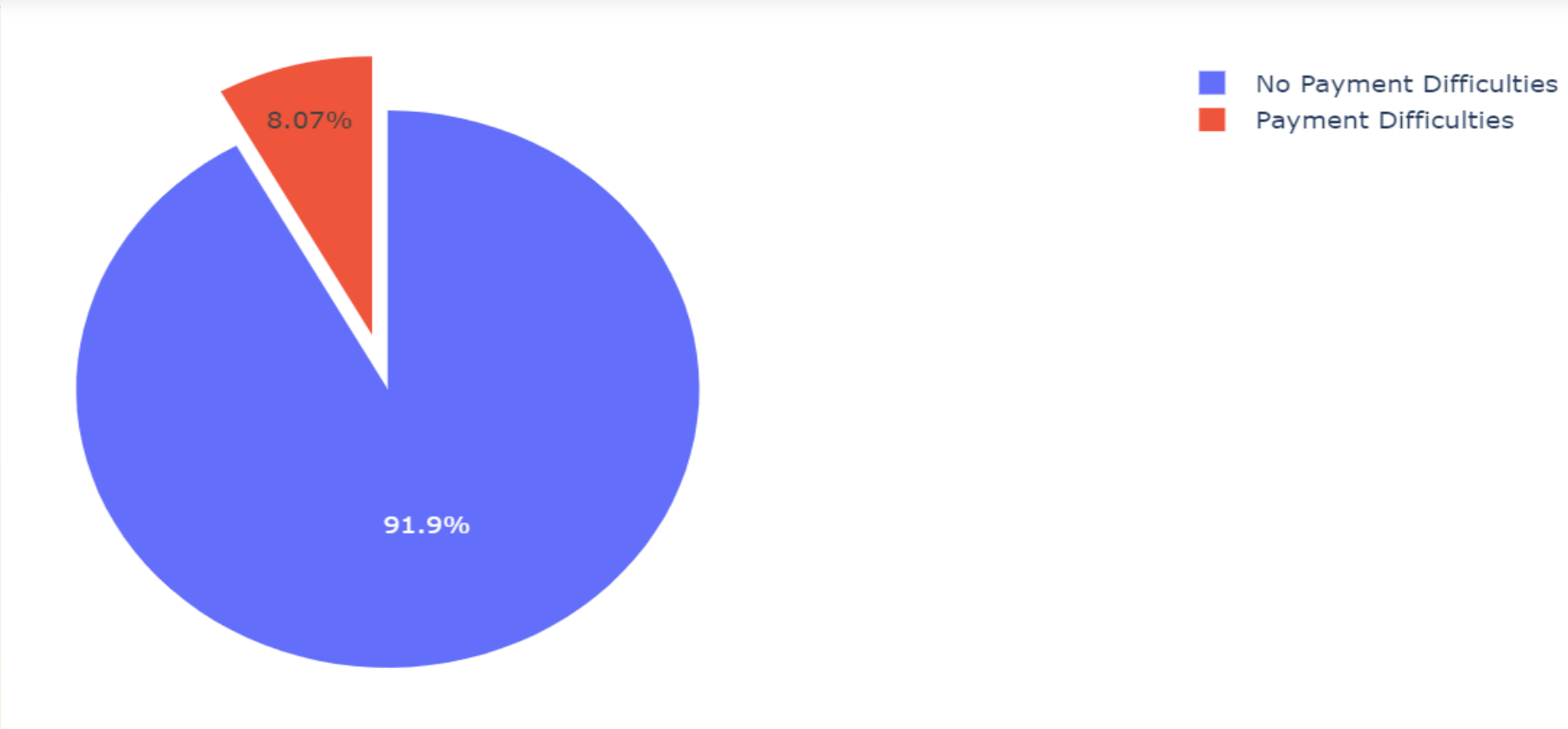


About 28% of the applicants aged between 30 - 40 years have no difficulties in payment.

Hence there is a good chance that we can make profit when we lend loan to them.

Findings - Current Application(Contd…)

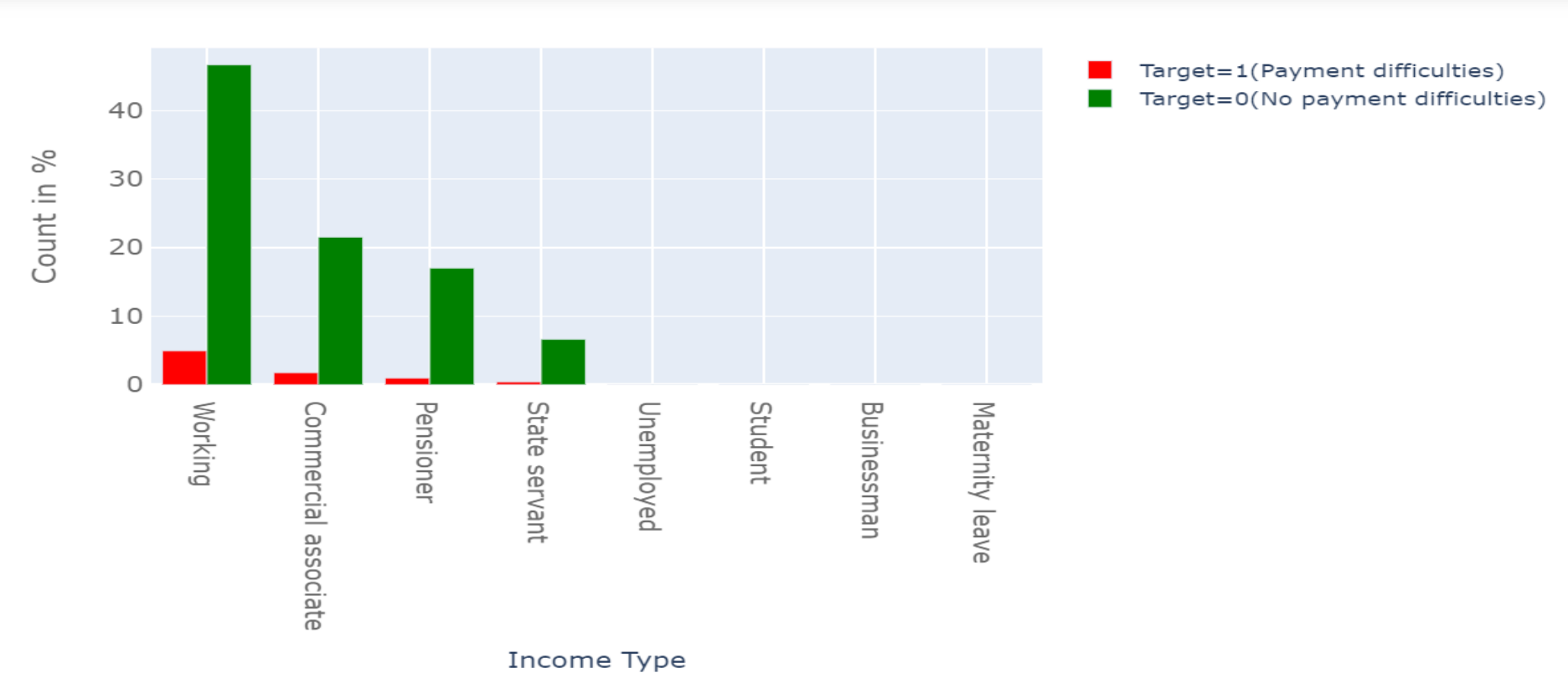
- Visualize imbalance data



Application data is highly imbalanced.

Findings - Current Application(Contd…)

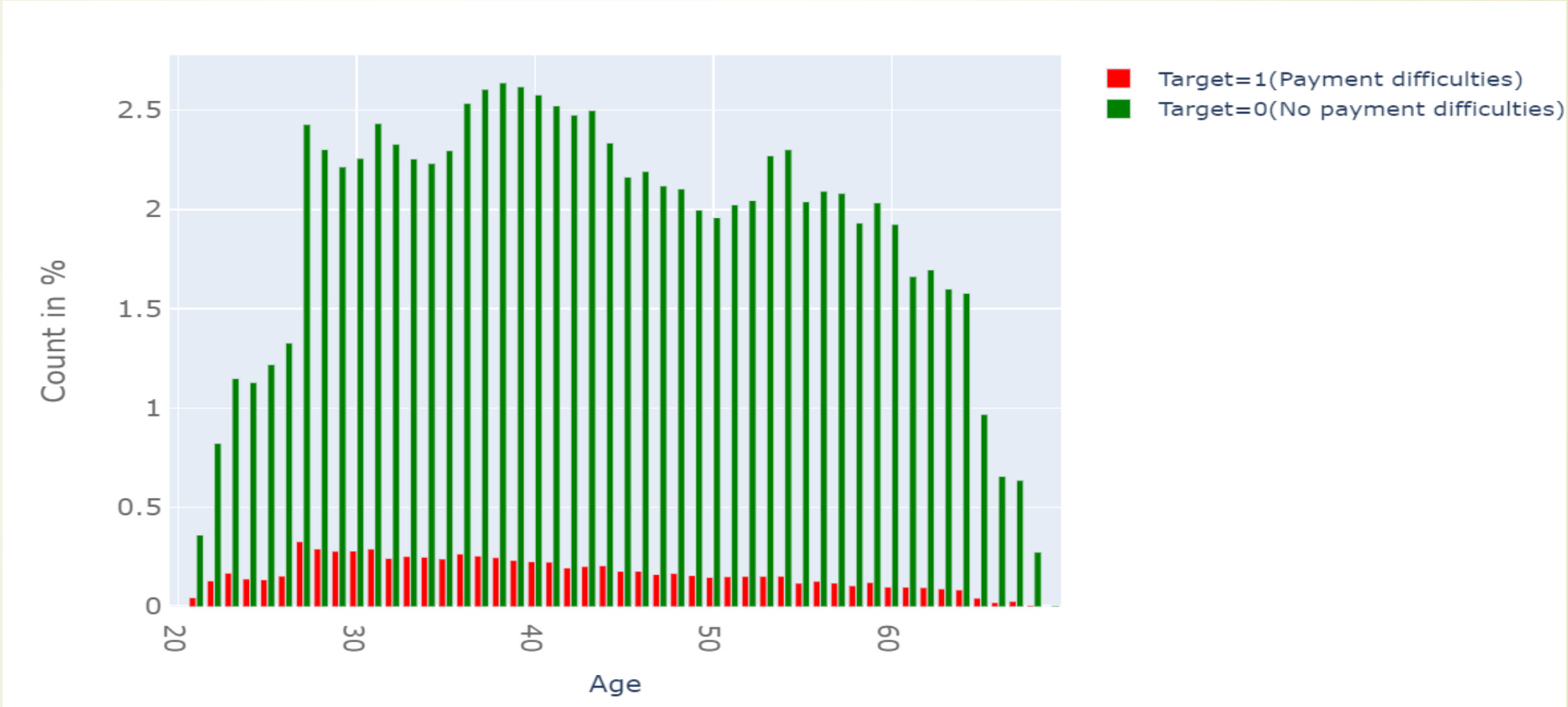
- Visualize NAME_INCOME_TYPE



Applicants with income type "Working" are the highest who are having issues in payment, they are also the highest when it comes to having no difficulties which are more than double of the income type "Commercial associate.

Findings - Current Application(Contd…)

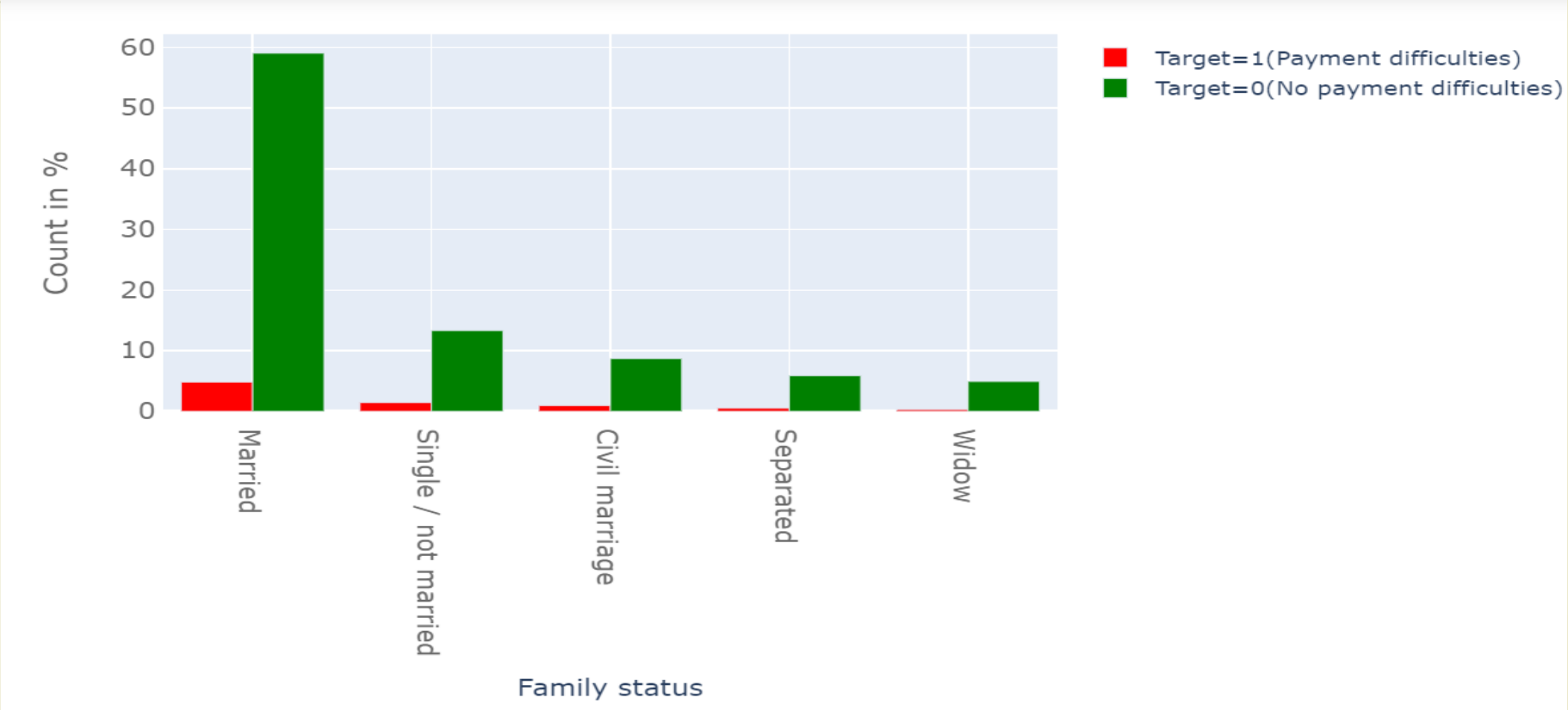
- Visualize AGE



It looks like the highest percentage of applicants not having issues with payment belong to the age group of 35-45, however, the tendency of the people to default looks to decrease as the age decreases from 28.

Findings - Current Application(Contd…)

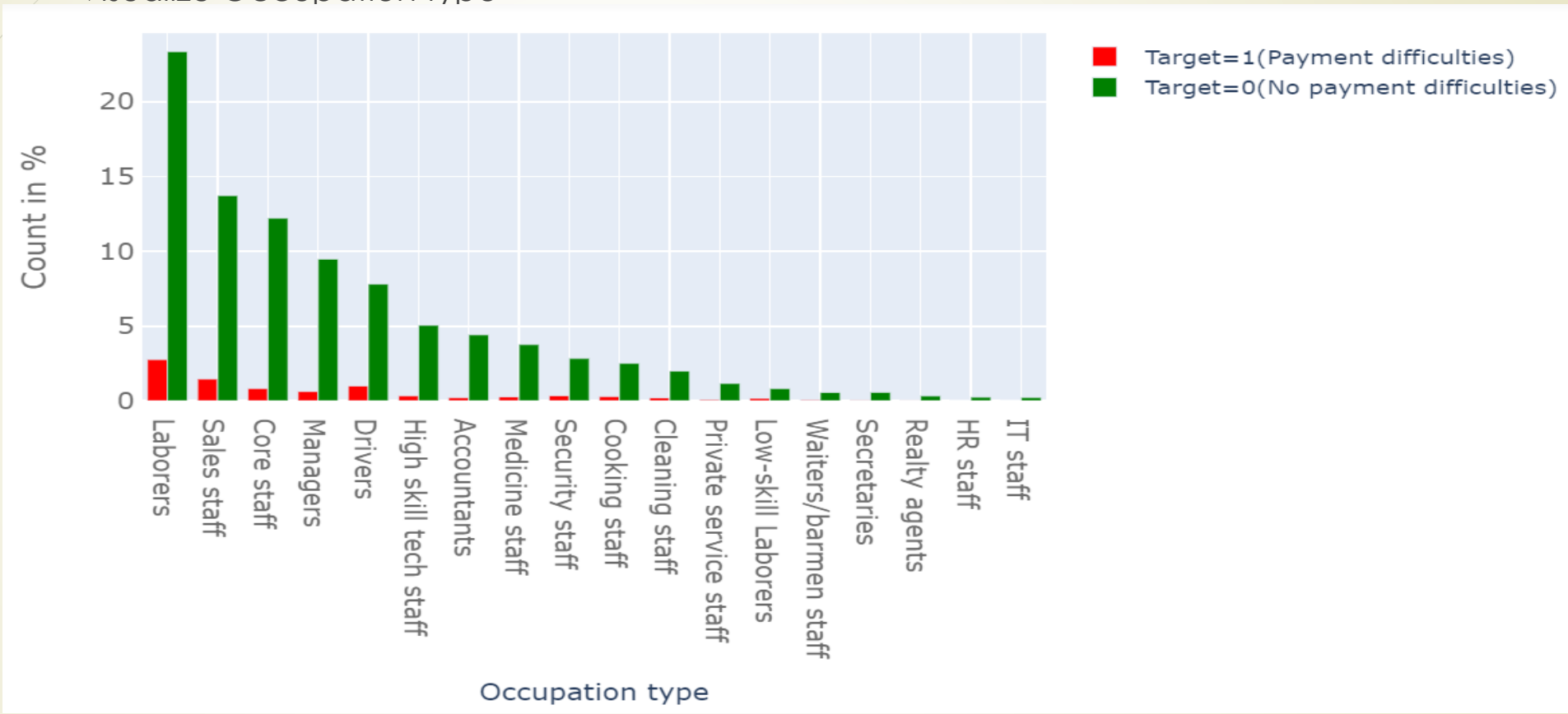
- Visualize Family Status



Married Applicants are having the highest payment difficulties, however, most applicants who are not having issues in payment belong to the same category.

Findings - Current Application(Contd…)

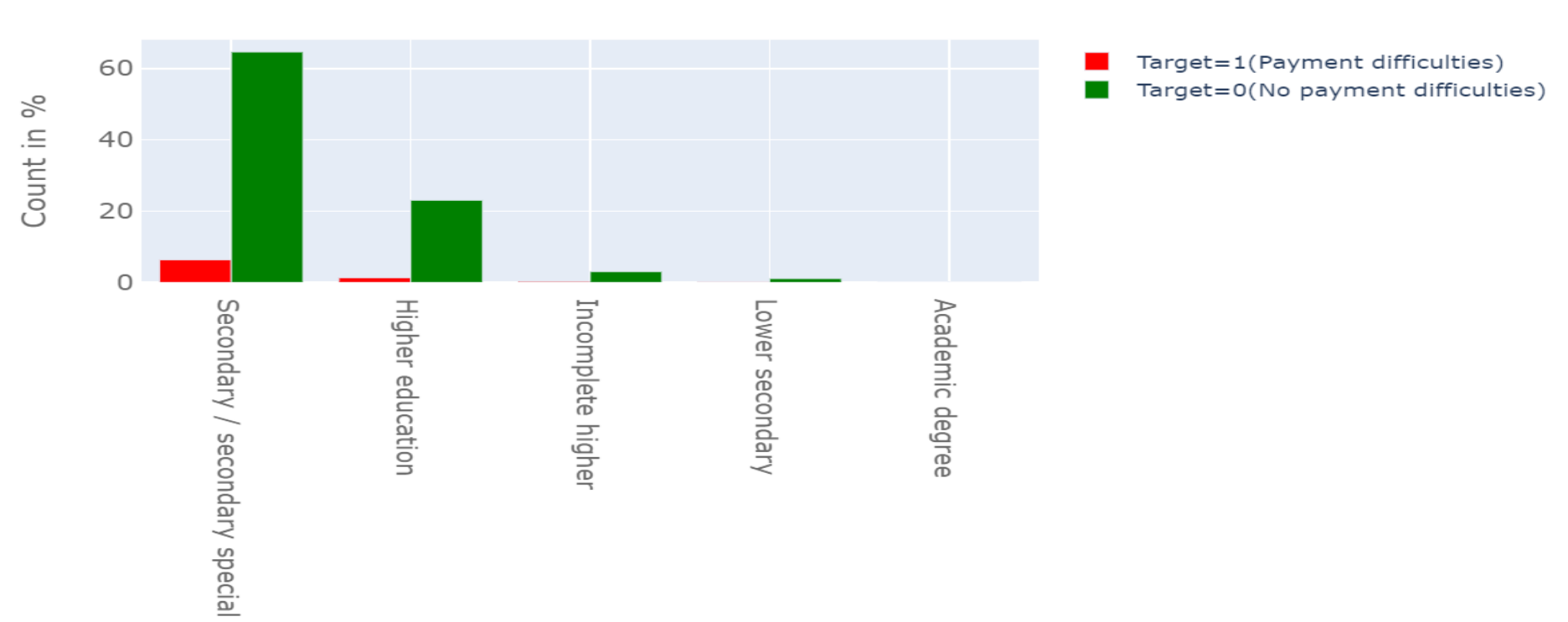
- Visualize Occupation type



Laborers are the ones who have taken most of the loans and also the highest in having difficulties, which is almost double of the occupation type "sales staff" which 2nd in position.

Findings - Current Application(Contd…)

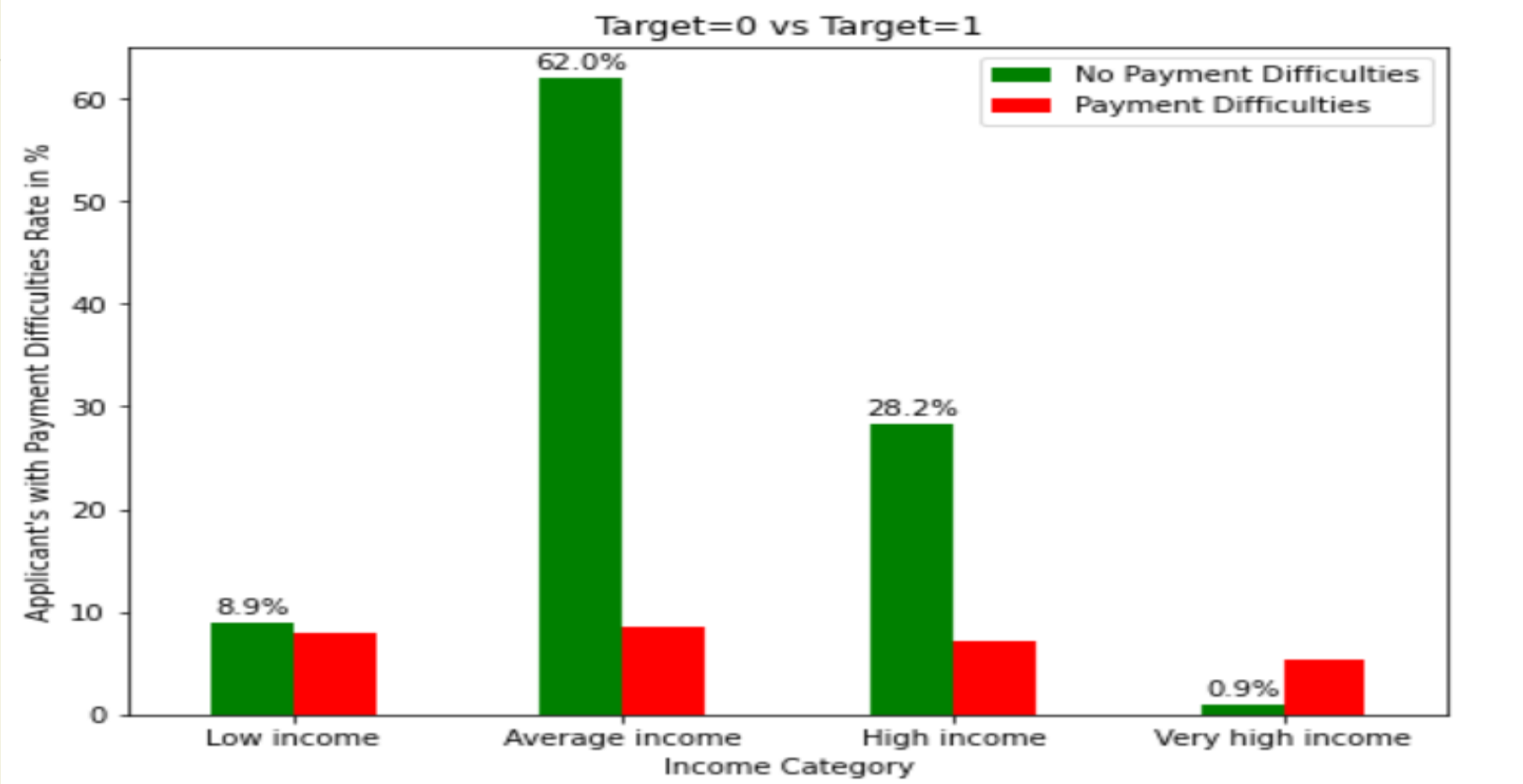
- Visualize education type



We have very few to none applicants who are having issues with payment difficulties with an academic degree as their education type.

The highest counts of the applicants lies with those having is with education "Secondary / secondary special" however, they are also the ones taking the majority of the loans.

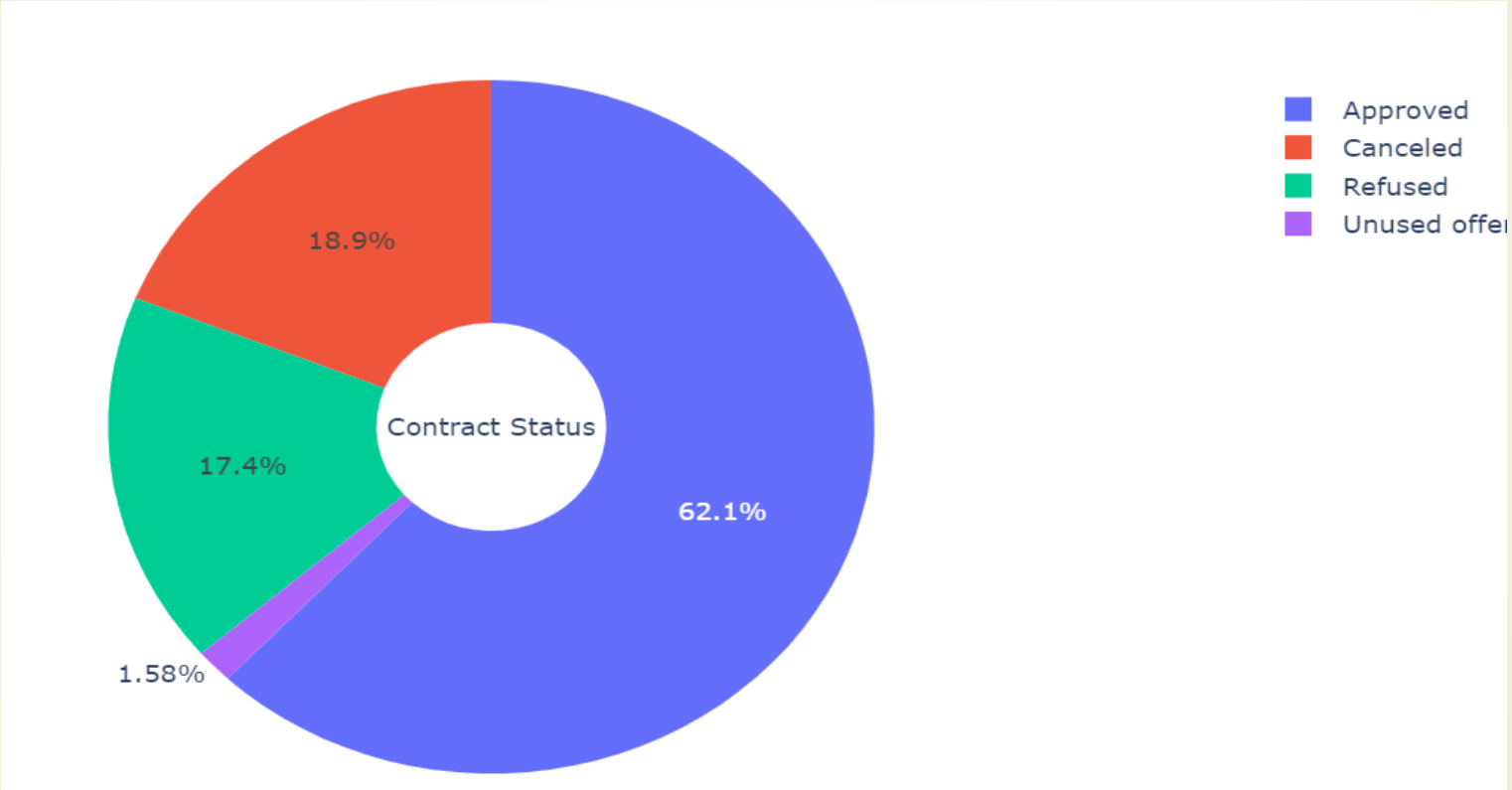
Findings - Current Application(Contd…)



Count of the applicants of all income category are having issues in payment of loans that is comparable to each other

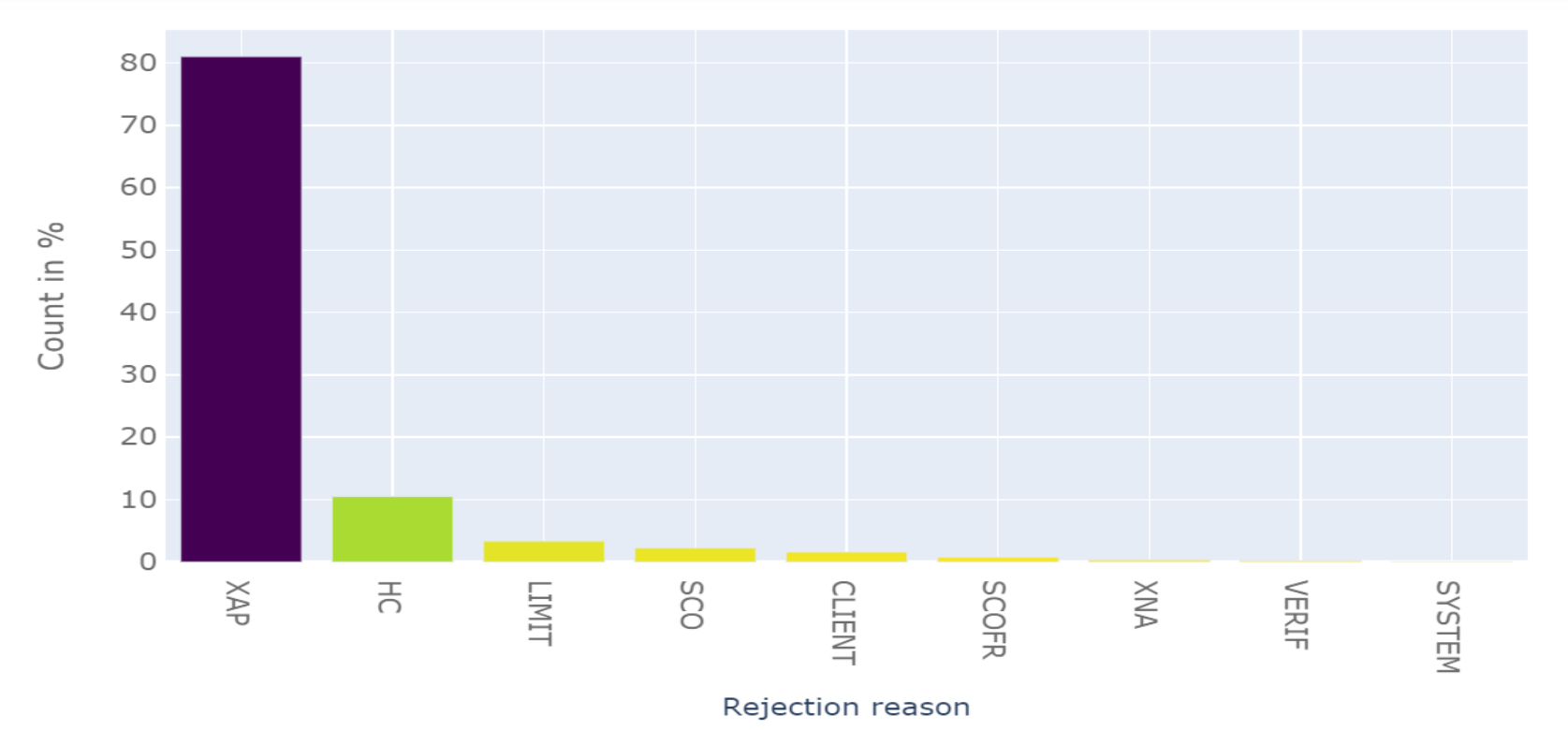
Average Income category seems to have the highest counts for having no issues in repayment.

Findings - Previous Application



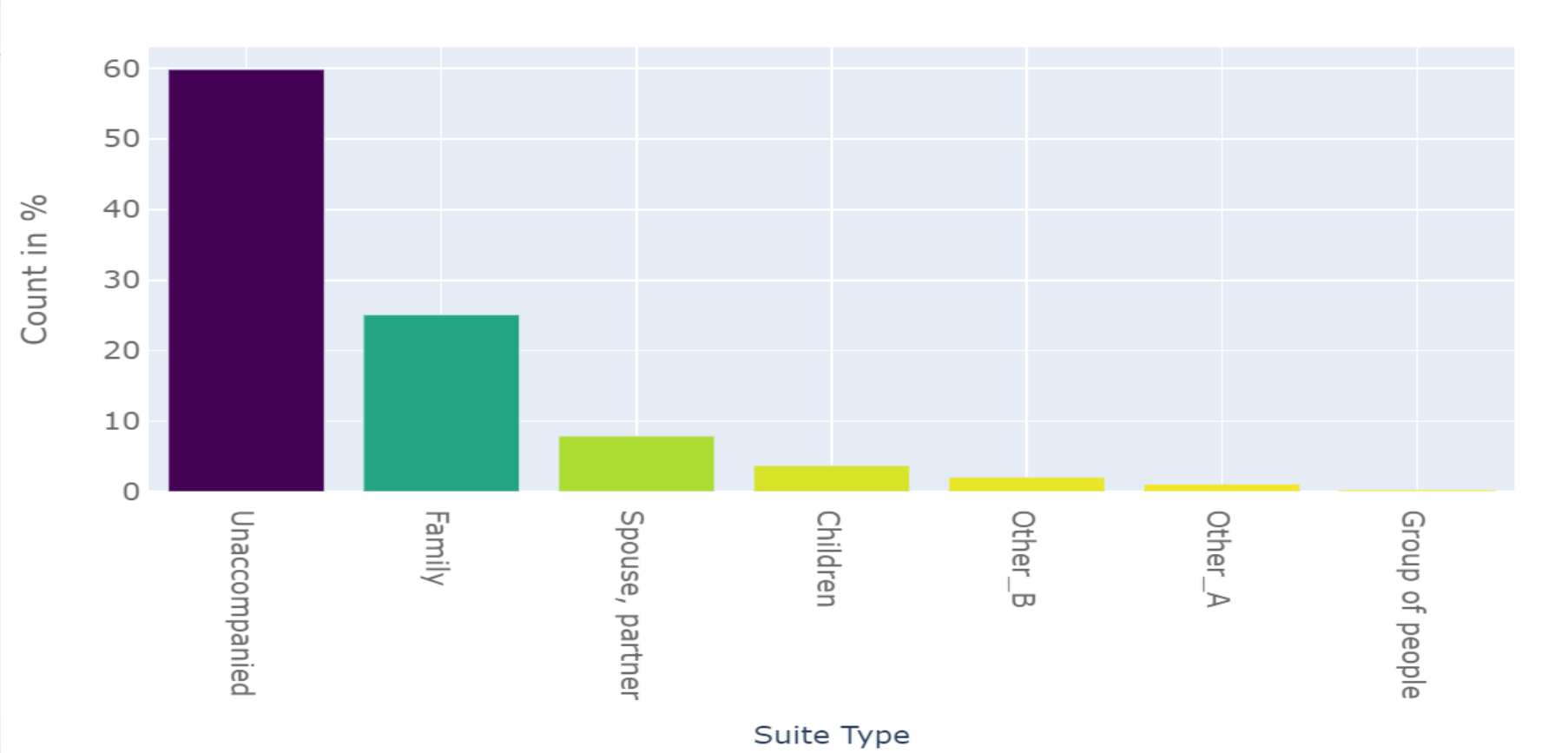
62% of the applications previous approved

Findings - Previous Application (Contd…)



Rejection rate is very high. Approx.81% of the previous applications were rejected. Reason not clearly specified (XAP)

Findings - Previous Application (Contd…)



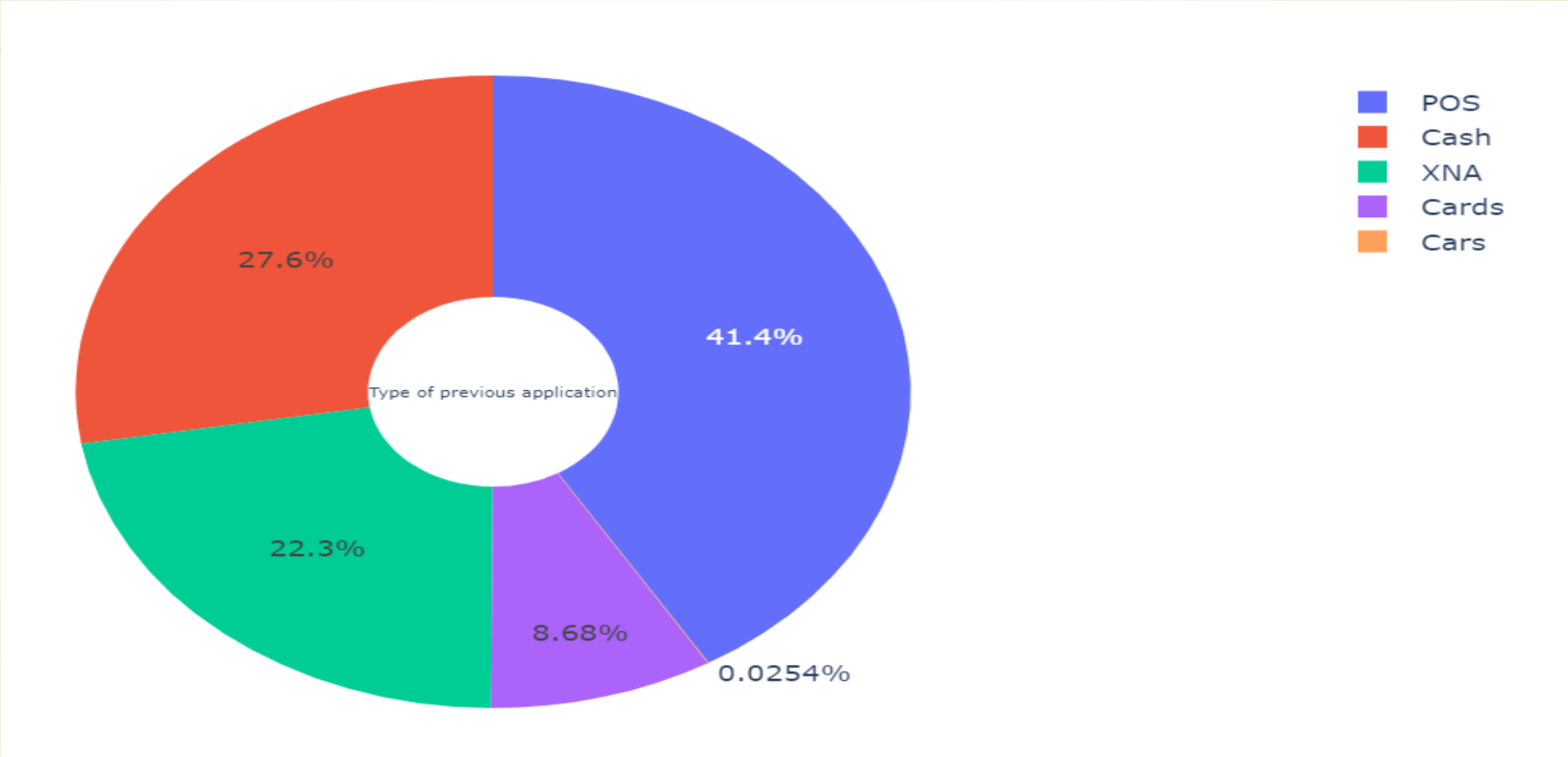
Previously 60% of applicants were applied for loan individually

Findings - Previous Application (Contd…)

74% of applicants who applied for loan were repeaters

18% were new applicants approached for loan previously

Findings - Previous Application (Contd…)

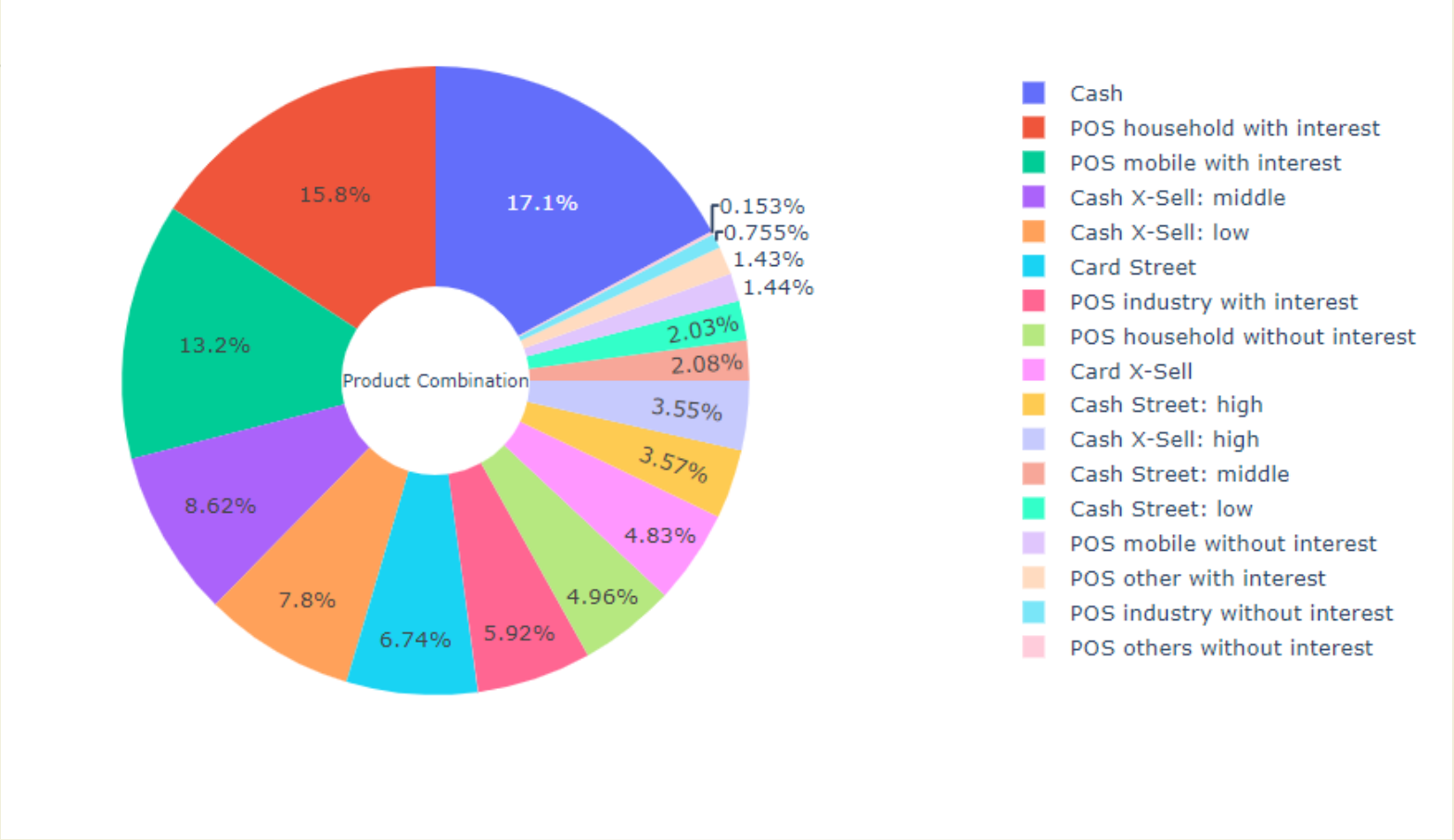


41% applicants were applied loan for POS

28% applicants were applied loan for CASH

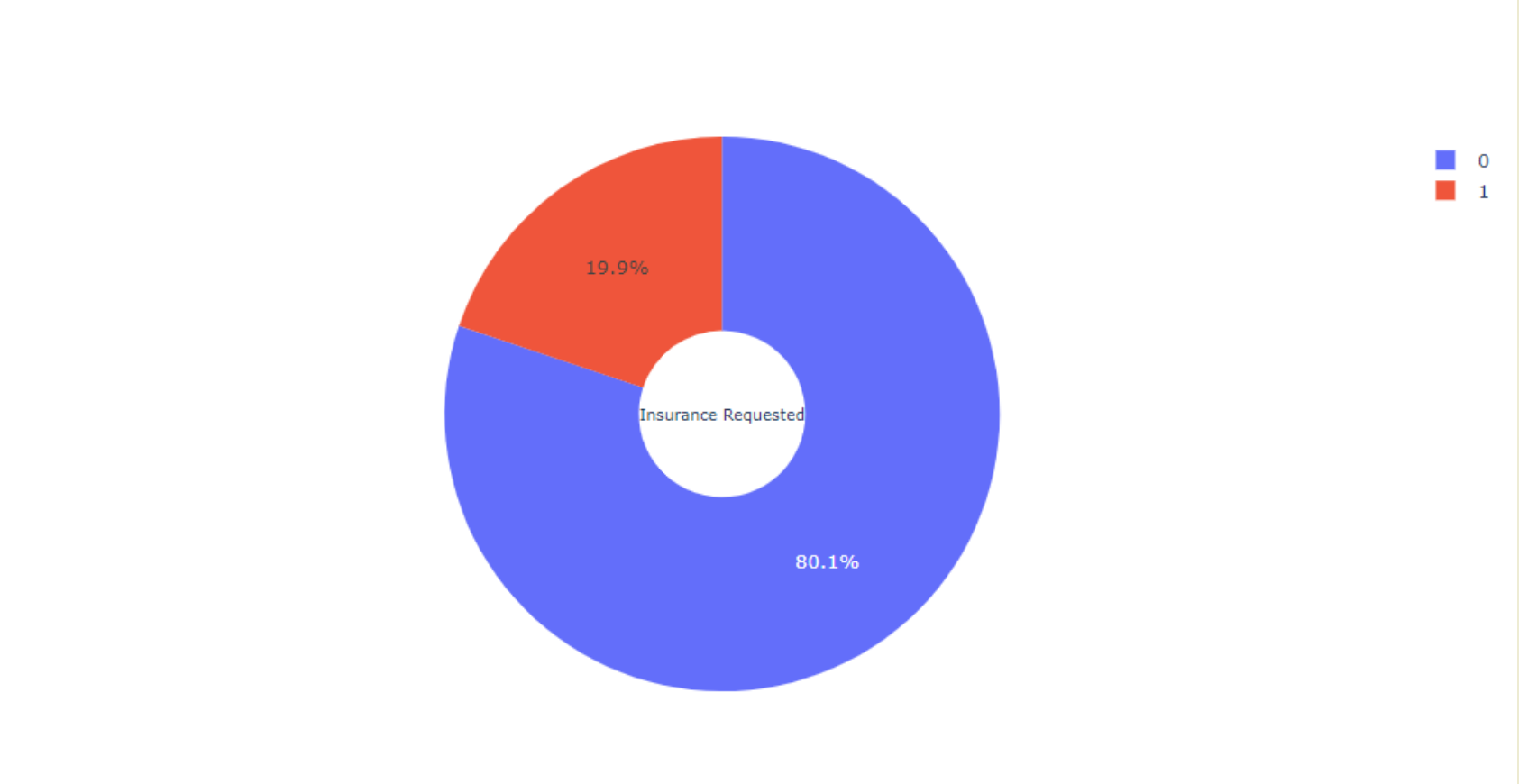
None of the applicants applied for CAR loan

Findings - Previous Application (Contd…)



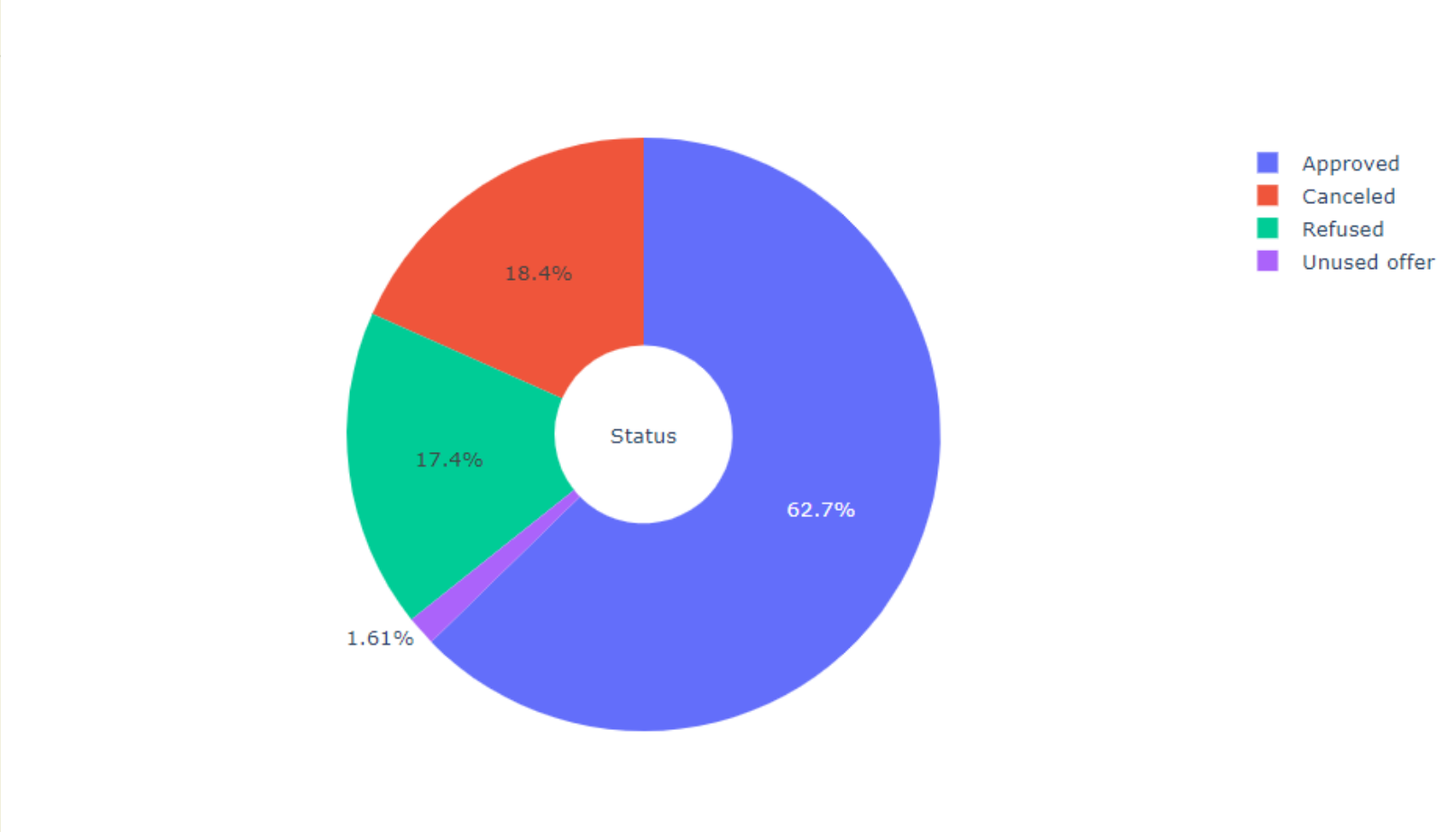
Most applicants(17%) were previously applied for CASH product

Findings - Previous Application (Contd…)



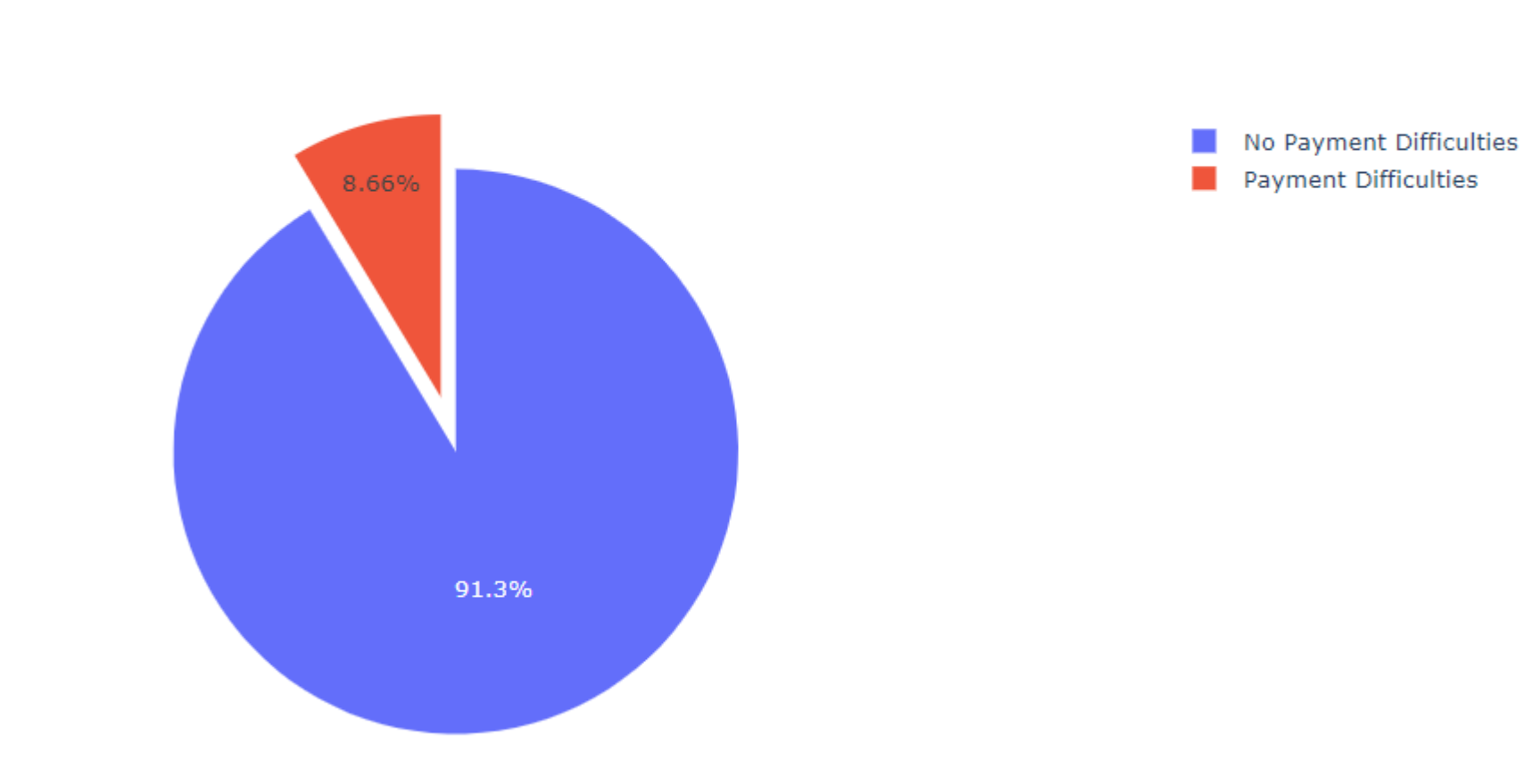
Only 20% applicants requested for insurance previously

Findings - Previous Application (Contd…)



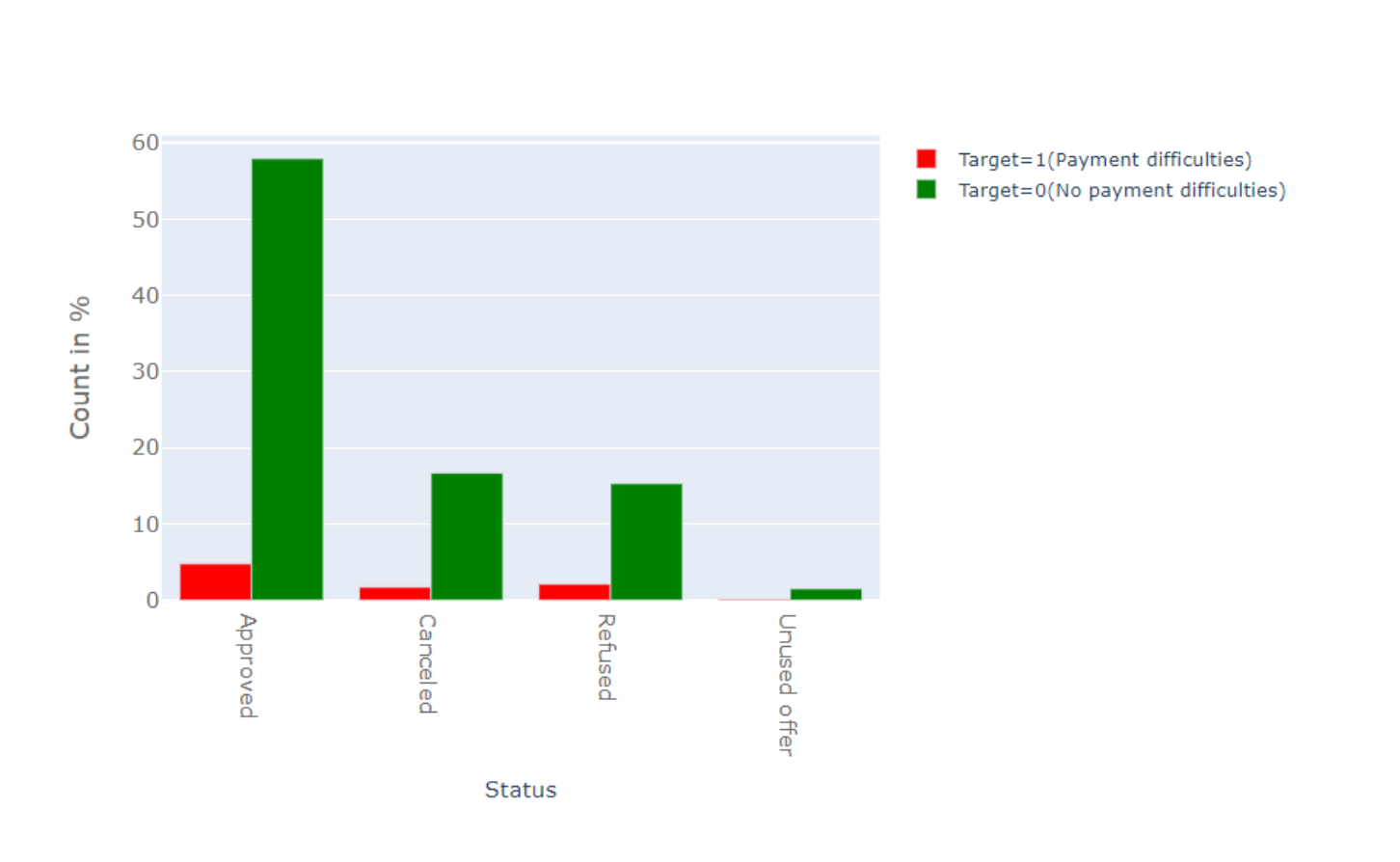
More than 60% of the Applications were approved.

Findings – Divided Dataframe(t=1,0)



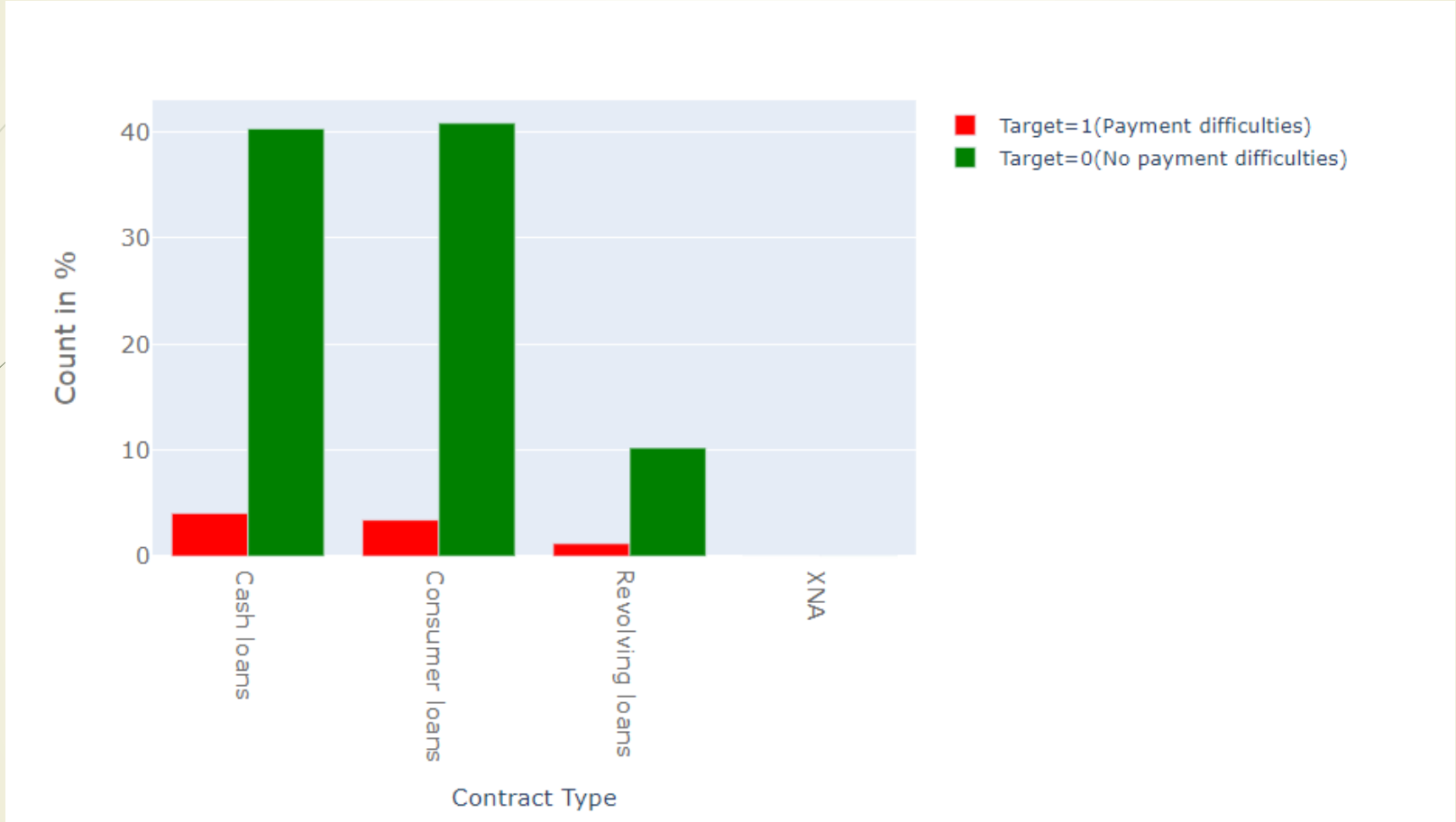
Data is highly imbalanced

Findings – Divided Dataframe..(Contd..)



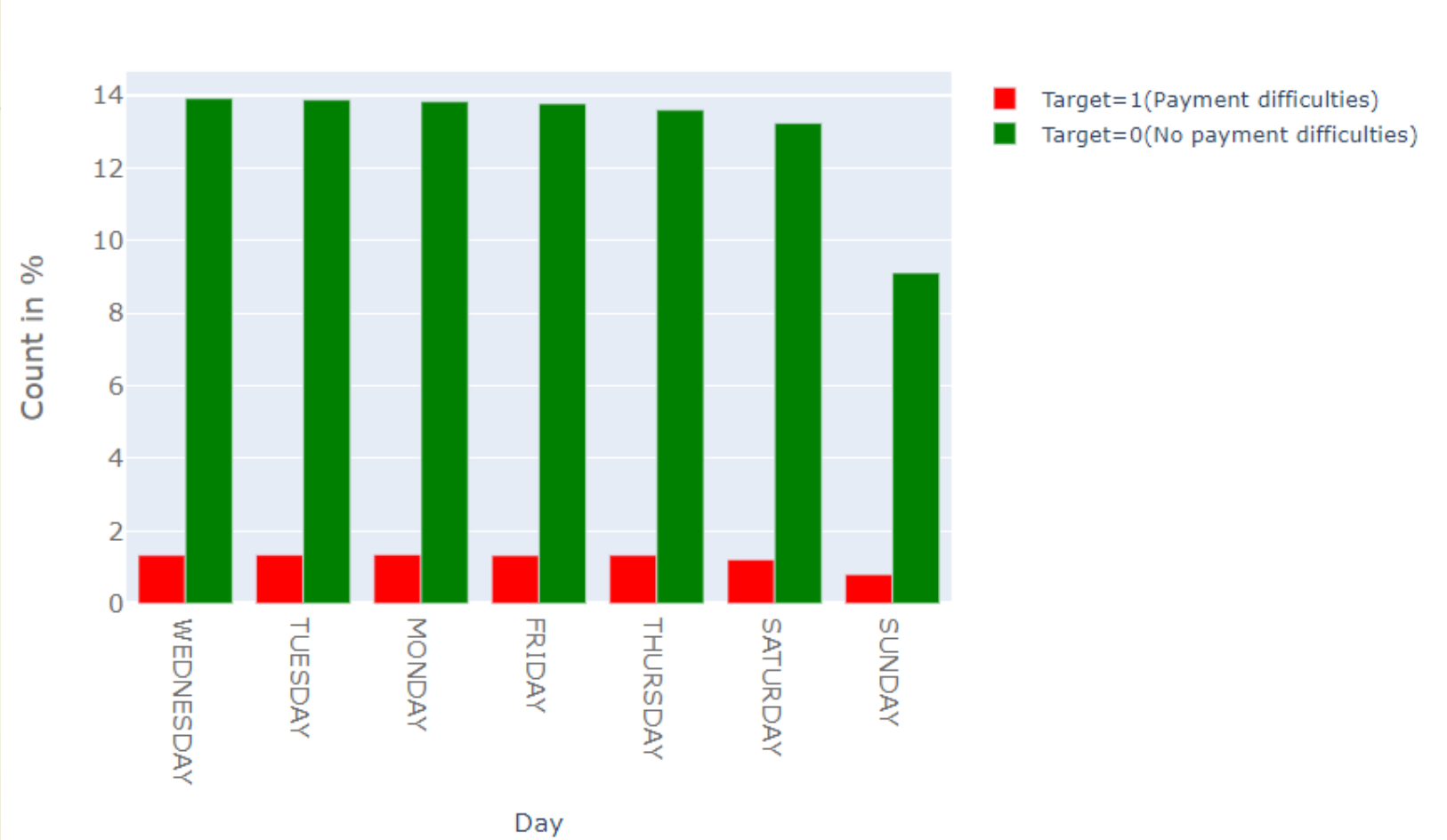
60% of the previous application were approved to applicants who do not have any payment difficulties.

Findings – Divided Dataframe..(Contd..)



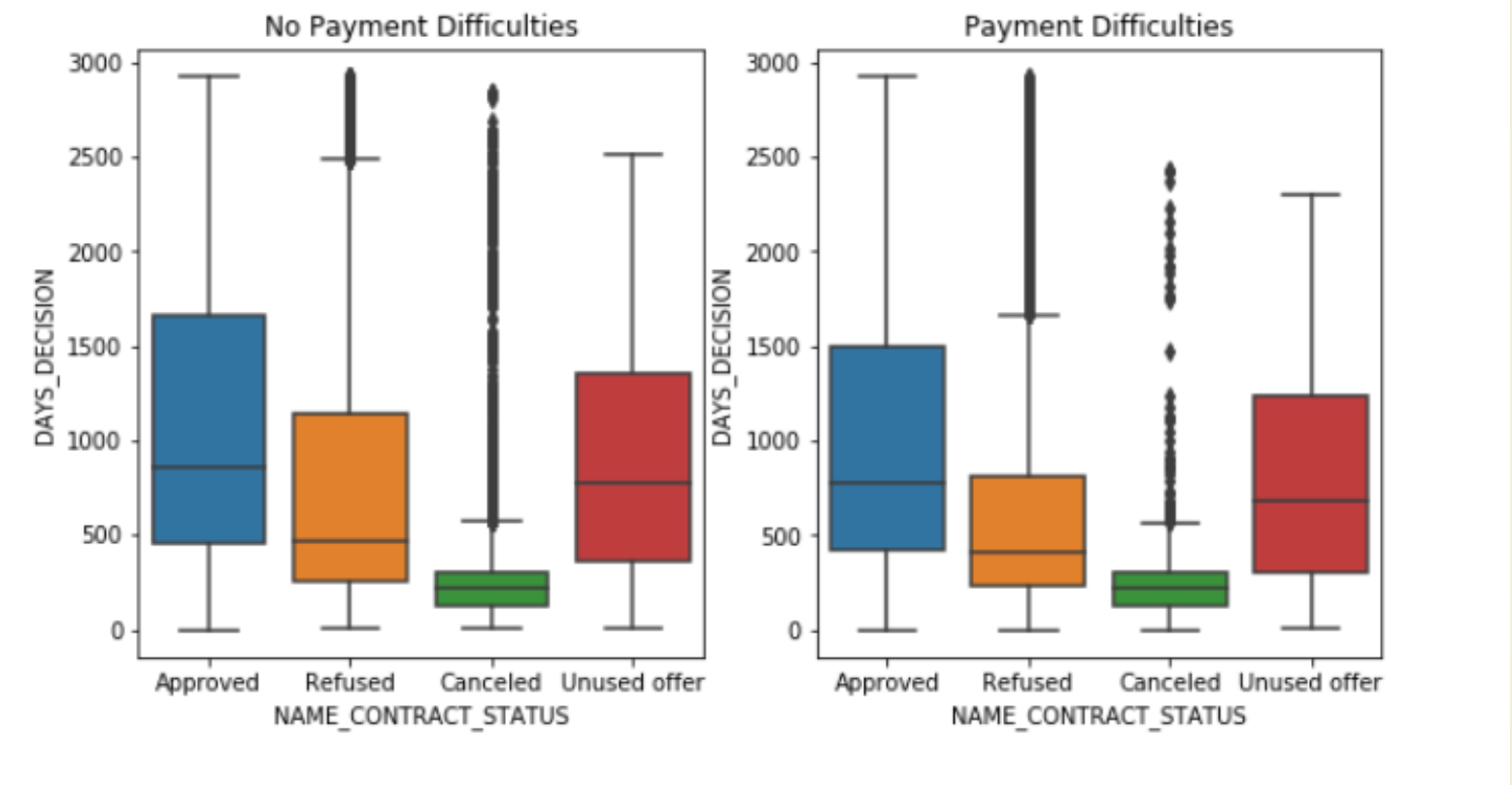
The number of applicants having issues in payment of the loans are the highest for Cash Loans, followed by Consumer Loans and Revolving Loans, however there are higher number of applicants who did not have any issues in payment for the loans taken as Consumer type, thus one may conclude that repayment of Consumer Loans is little more easier than the repayment of Cash Loans.

Findings – Divided Dataframe..(Contd..)



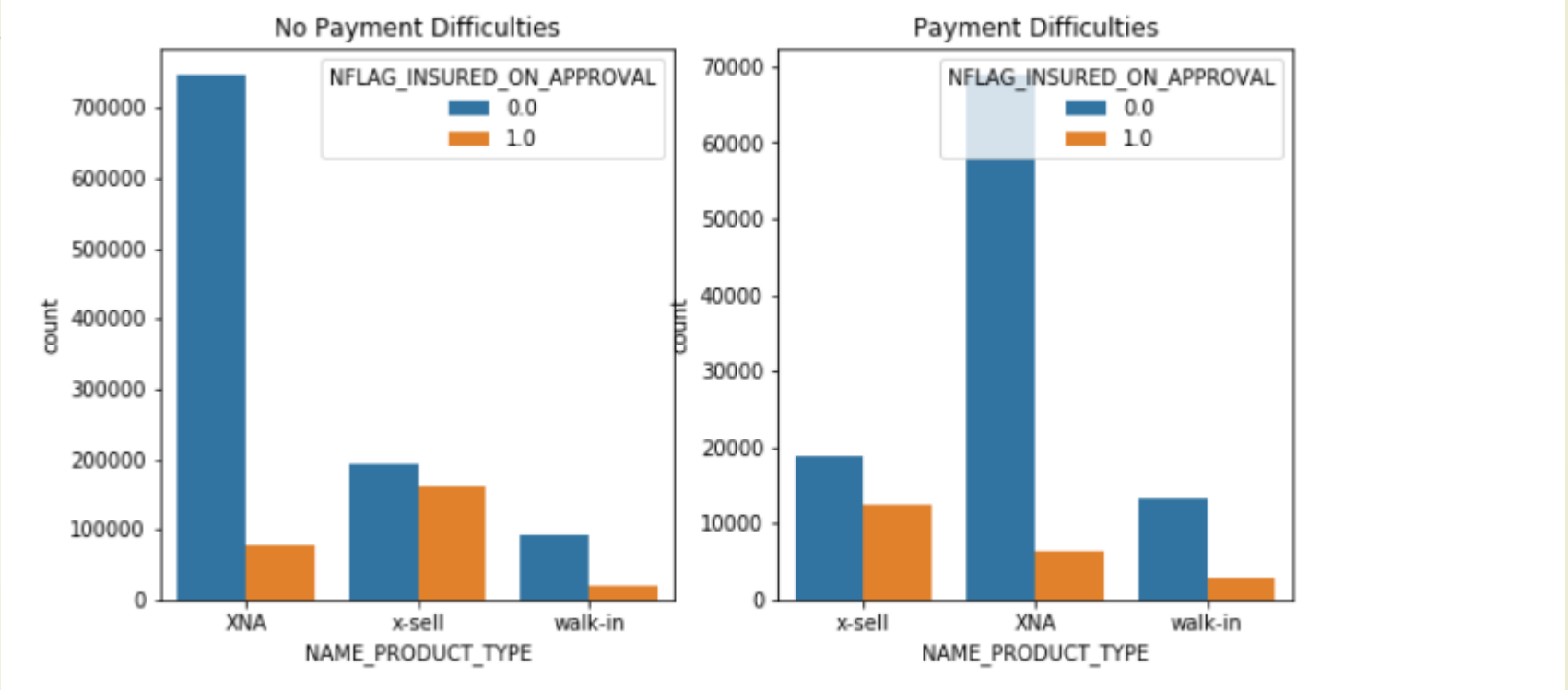
The above graph do not provide much insight into the applicants if they were able to repay without difficulties or not, however we can see the drop in the number of application during the weekend.

Findings – Divided Dataframe..(Contd..)



Comparing count of applications between the above 2 graphs for the ones refused, the counts are more uniformly distributed for "No Payment Difficulties" boxplot than the "Payment Difficulties" boxplot. We can infer that applicants having difficulties in their repayment had got their previous application decision made quite sooner than the applicants who are not having any difficulties.

Findings – Divided Dataframe..(Contd..)



The count of the applicants having "No Payments Difficulties" that had asked for Insurance is a little more than the count of applicants of having "Payment Difficulties"

Conclusion

- **Maximize interest gain**

  We must target applicants who are highly educated, with age group between 38 - 44 years having 10+ working experience, preferably unmarried, whilst earning average and above average income.

- **Minimise the Credit Loss**

  We must try to refrain from applicants having less years of experience, and are married at younger age, it can be seen that males who are businessmen have a lot of issues in repayment, in such case we must also consider their income, which should not be low. We must also consider the status of their previous application, if they had issues earlier in repayment or not and the type of the loan, where we should prefer Consumer loans over Cash Loans, since they involve low chances of risk of not getting timely paid.