Credit EDA Case Study

Satyajit Pattnaik

Business Understanding & Overview

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.

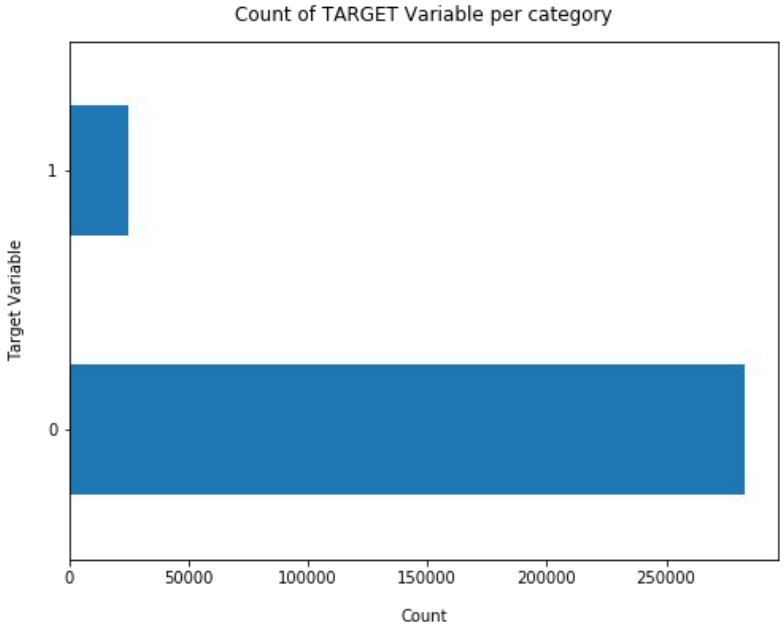
We will perform this Exploratory Data Analysis in order to get actionable insights, and convert them into meaningful stories and present it so that the companies can take necessary actions in order to reject or accept loan application.

Understanding the data

Target Variable

Findings

- Data is highly imbalanced, ratio is almost 92:8.
- Most of the loans were paid back on time (Target: 0)
- We need to analyse the data with other features while taking the target values separately to get some insights.

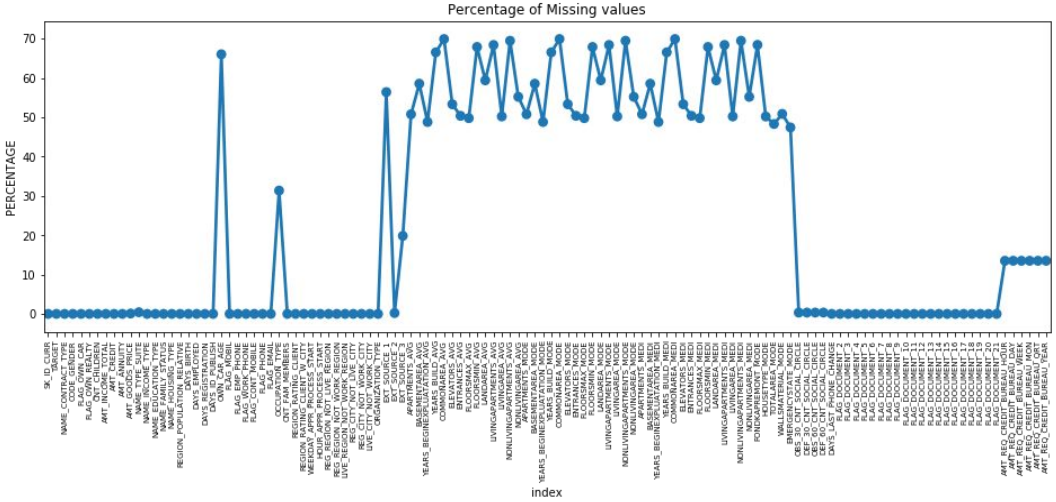


Missing Data

Findings

- Many columns have a lot of missing data (30-70%), some have few missing data (13-19%) and many columns also have no missing data at all.
- For features with less missing values- can use regression to predict the missing values or fill with the mean of the values present, depending on the feature.
- For features with very high number of missing values- it is better to drop those columns as they give very less insight on analysis.
- As there's no thumb rule on what criteria do we delete the columns with high number of missing values, we have done a small analysis and have taken decisions, Analysis to be continued in next slides...

Initial Intuition from the data

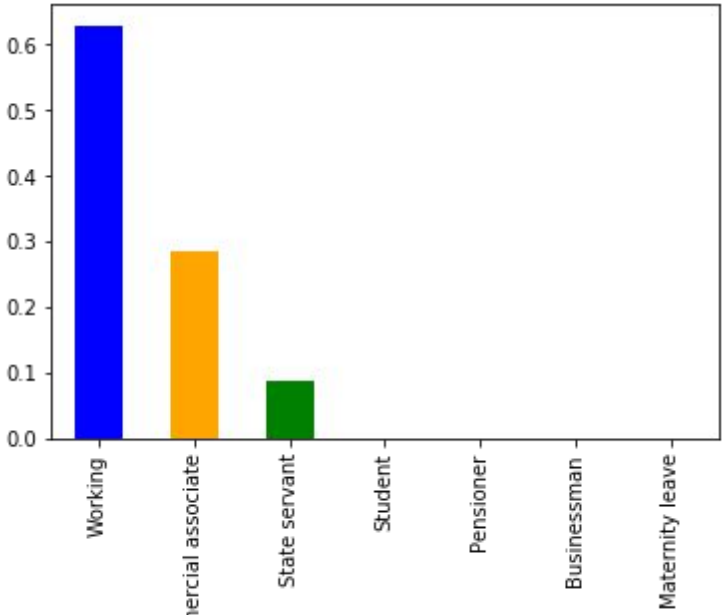


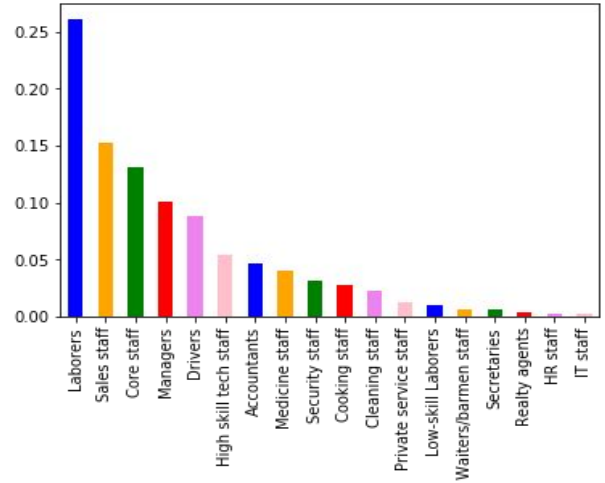
Findings

- Total columns having more than 30% of null values: 64
- As there's no such thumb rule to drop the variables having more than 30% null values, as scenarios might vary from case to case, and the amount of information we think the variable has. For example, most of the columns are Normalized information, and the description is bit unclear, hence I will remove all the Normalized information columns.
- However, we have kept these columns: AMT_GOODS_PRICE, OWN_CAR_AGE, OCCUPATION_TYPE and have analysed further, as these columns seems to carry some important information and two of them are connected with people who have taken Consumer loans, & people having their own car flag as Y, and the occupation type might have an impact on the analysis, hence, once proper analysis has been done, we can think of deleting or keeping these variables.
- In the later stage, we can also analyse the other 61 columns that we are deleting if they carry any important information or not.

Categorical Analysis

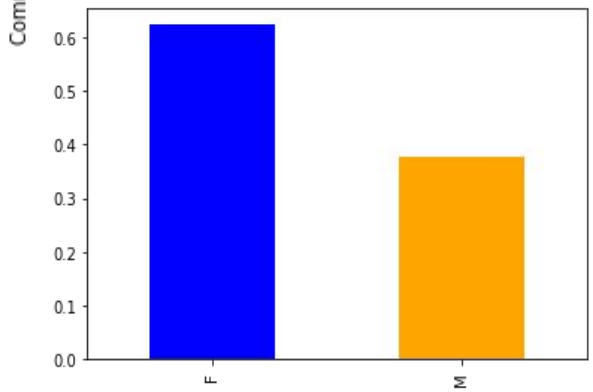
Univariate Analysis

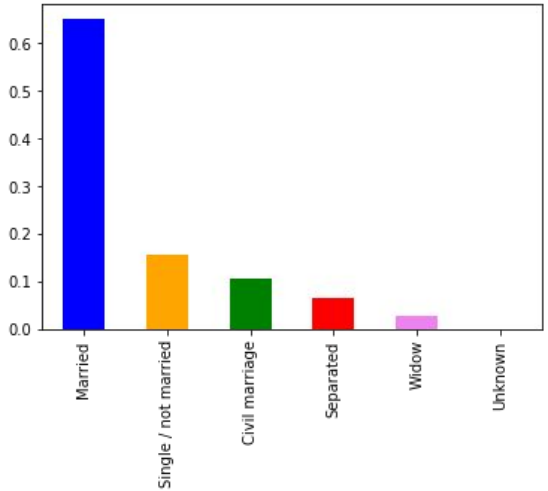




Findings

By just analysing single variables, we won't find much insights related to the defaulters, as here we will just have an idea which category of people are present in abundance, other than that, most of the insights are gathered in analysis of multiple features/variables with target variable.

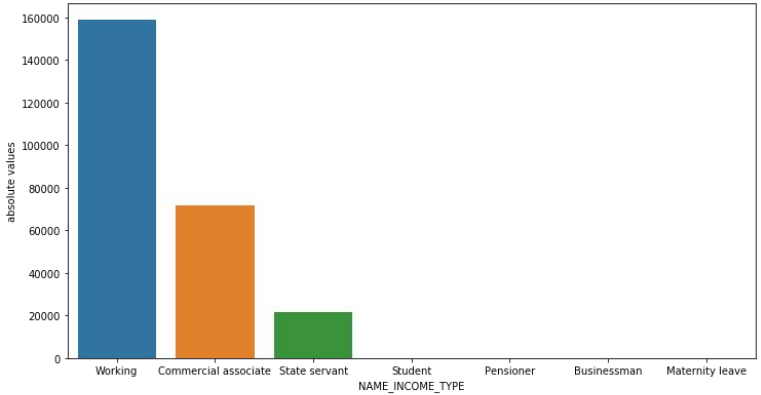


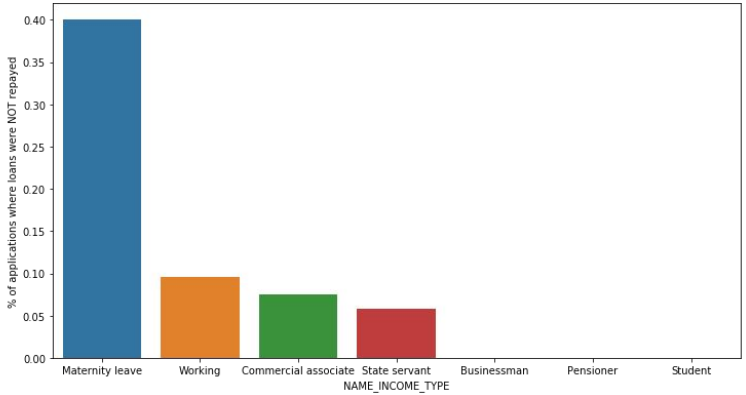


Income Type vs Target Variable

Findings

- The working class applies the most for loans, and have a very low default rate, hence they are reliable.
- Clients who are unemployed or on maternity leave have very high defaults rates even though they are a minority compared to other income types.
- Commercial associates, state servants and pensioners are fairly more reliable.

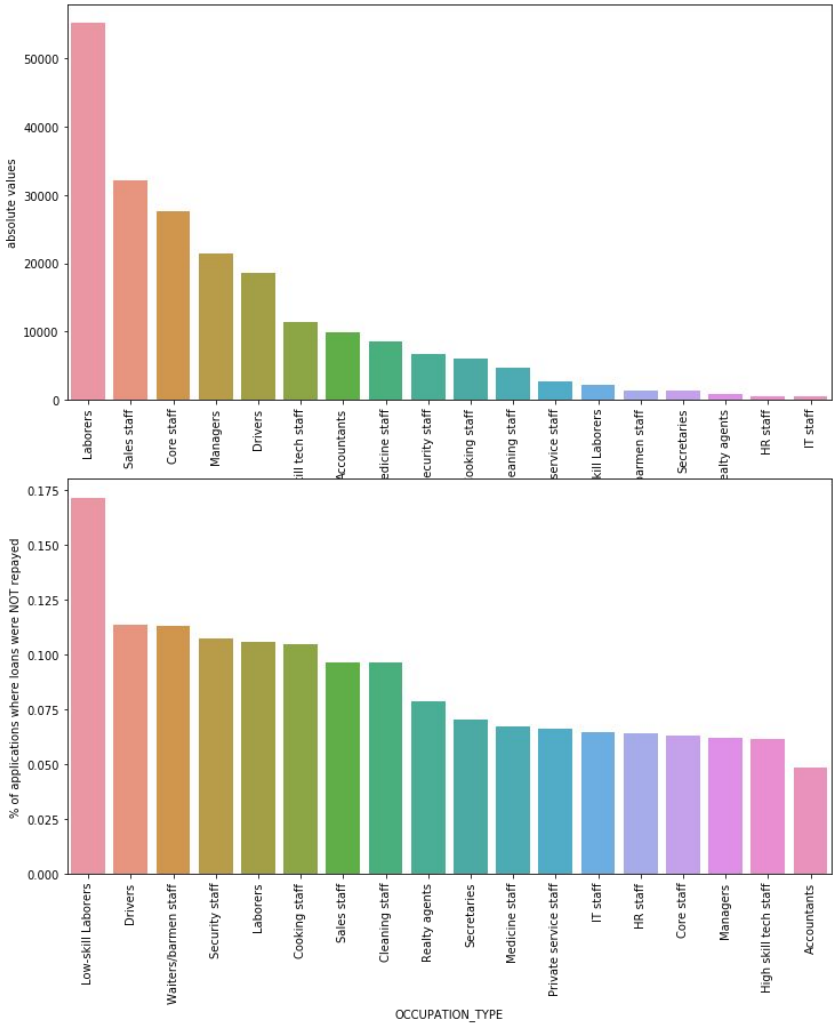




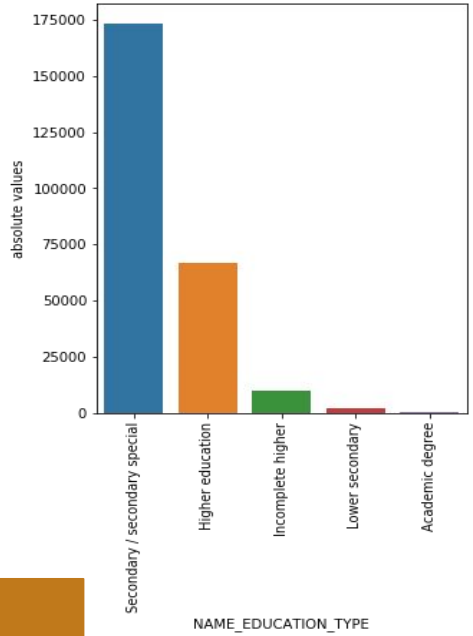
Occupation Type vs Target Variable

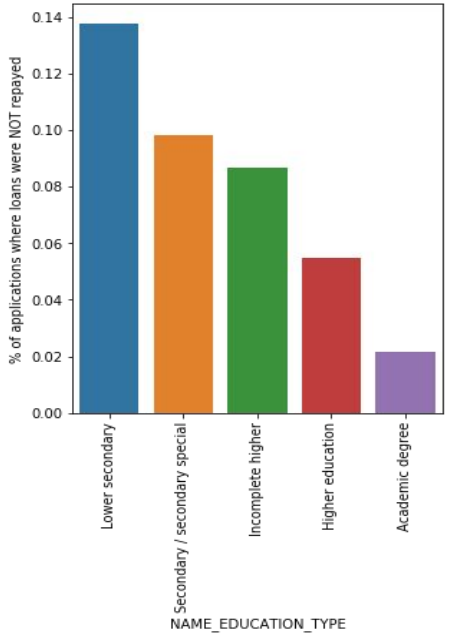
Findings

- Low skilled labourers and lower class staff are most likely to be loan defaulters than high skilled staff and accountants (which is understandable).
- Better the occupation, lesser the chance of defaulting.



Education Type vs Target Variable

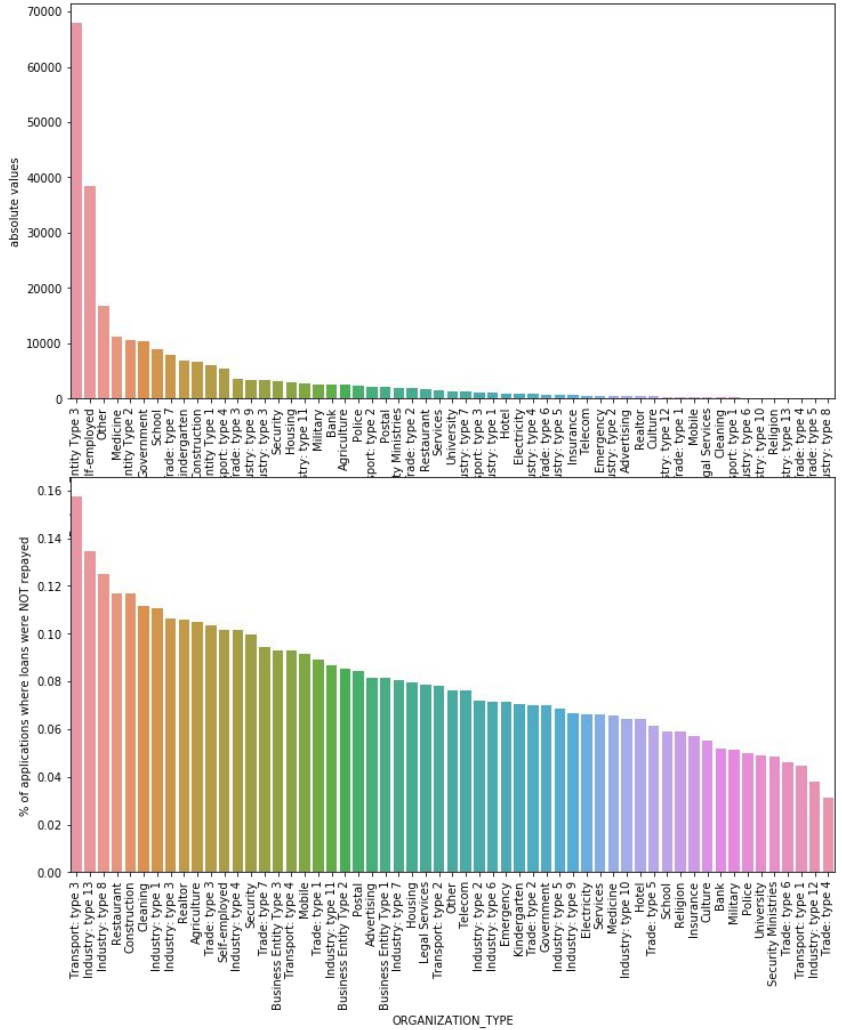




Findings

- Most student loans are for their secondary education or higher education.
- Lower secondary education loans are most risky for the company followed by secondary/secondary special.

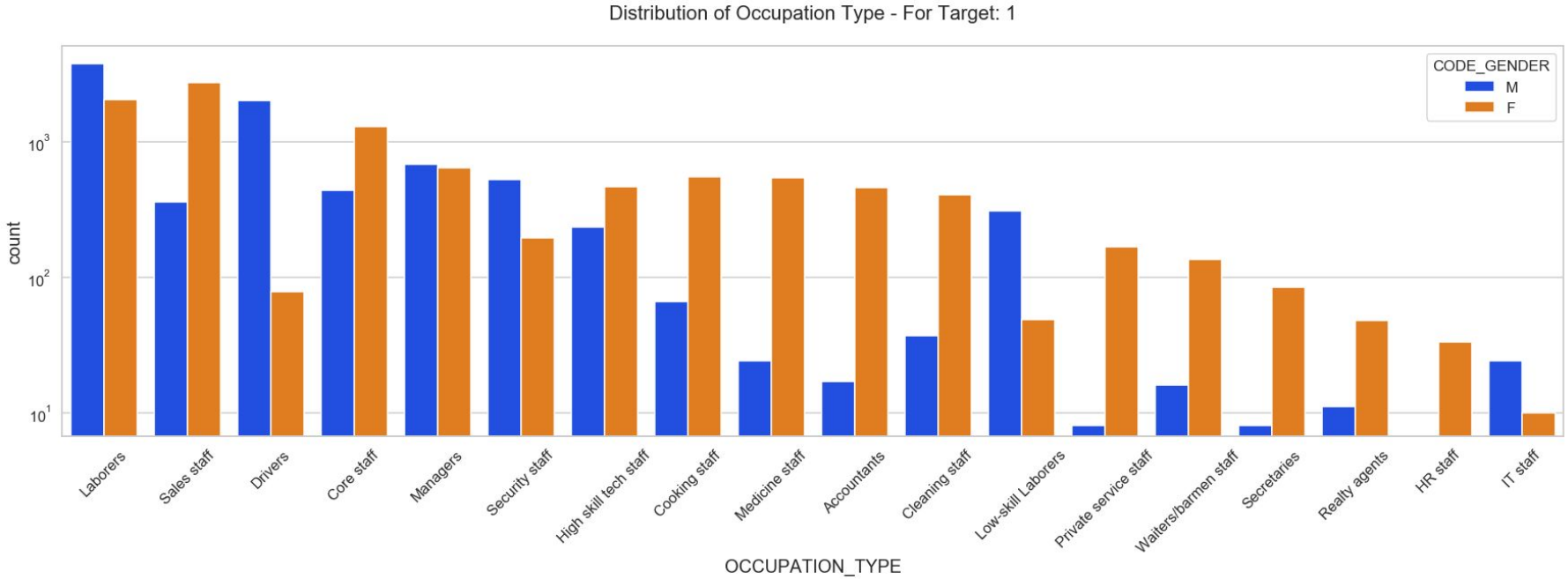
Organization Type vs Target Variable



Findings

- Organizations with highest percent of loans defaulters are Transport: type 3 (16%), Industry: type 13 (13.5%), Industry: type 8 (12.5%) and Restaurant (less than 12%).
- Business Entity type 3 Trade: type 4, Industry: type 12 organizations are most reliable.

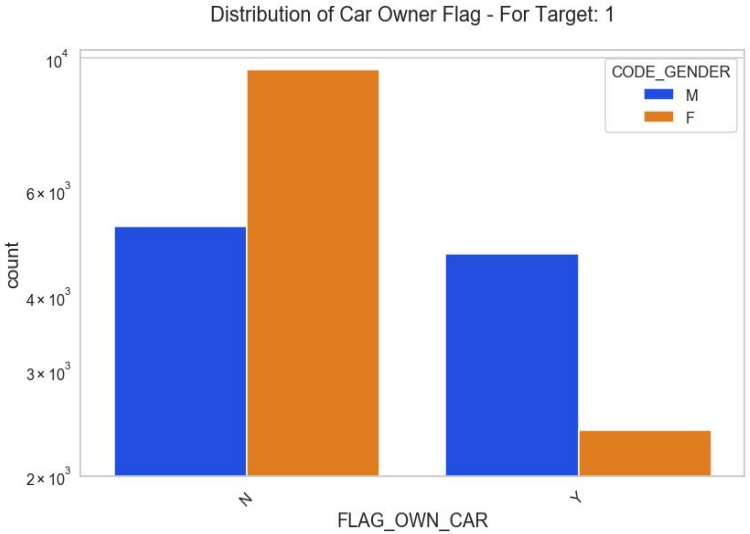
Occupation Type vs Defaulters (Target:1)

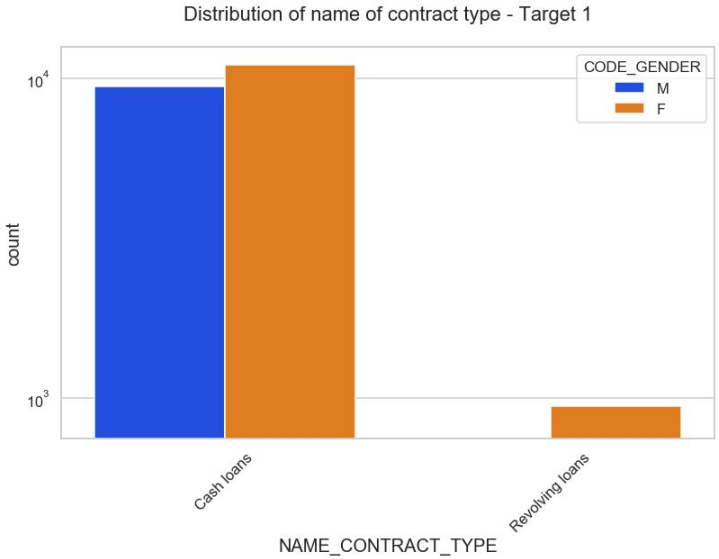


Some good insights

- Female under occupation: Accountants, Private Service Staff, Secretaries, Realty Agents, HR Staff etc are the most defaulted sub categories against their male counter parts.
- Male under occupation: Low Skilled Labourers, Drivers etc are the most defaulted sub categories against their female counterparts.

Flag Own Car, Contract Type vs Defaulters (Target:1)

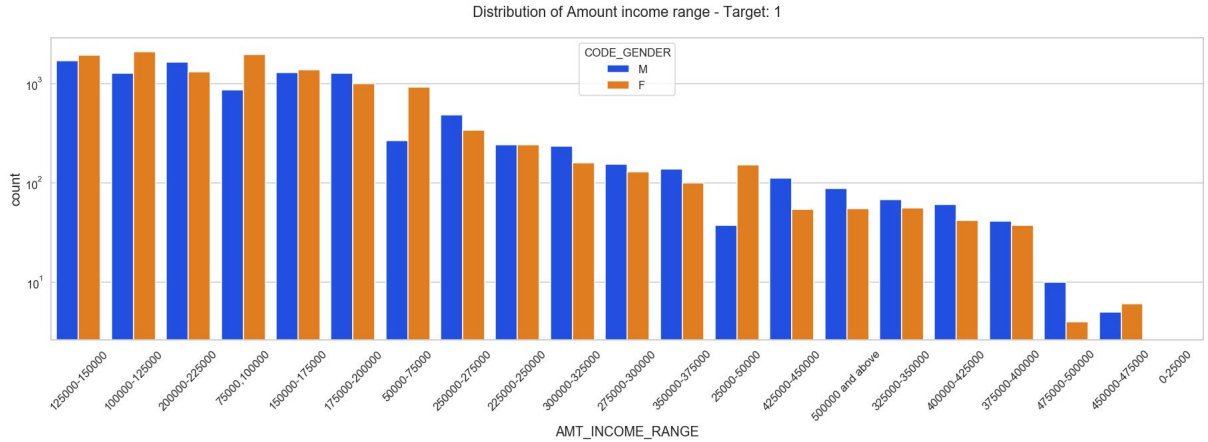


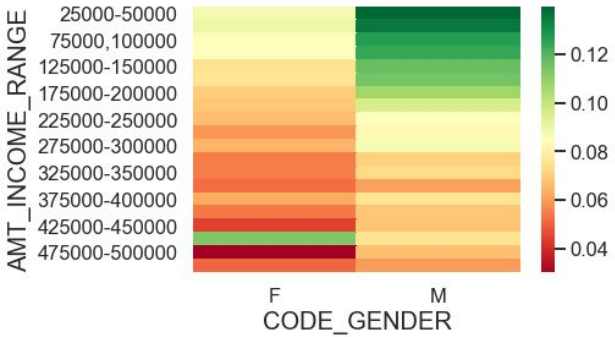


Some good insights

- Overall people without cars are most defaulters, if we further dig in, females without car are the most defaulters as compared to the males without cars.
- Male candidates with cash loans are the ones with highest default rate as compared to other 3 sub categories, but analyzing alongwith the bar plot, it's clear that both males/females having cash loans have almost equal ratio in defaulting, but looking at the revolving loans, even though the numbers are less, but the females who have taken revolving loans are the highest defaulters as compared to their male counter parts.

Income Range vs Defaulters (Target:1)





Some good insights

- For target: 1
  - Male counts are higher than female.
  - Income range from 100000 to 200000 is having more number of credits.
  - This graph show that males are more than female in having credits for that range.
  - Very less count for income range 400000 and above.

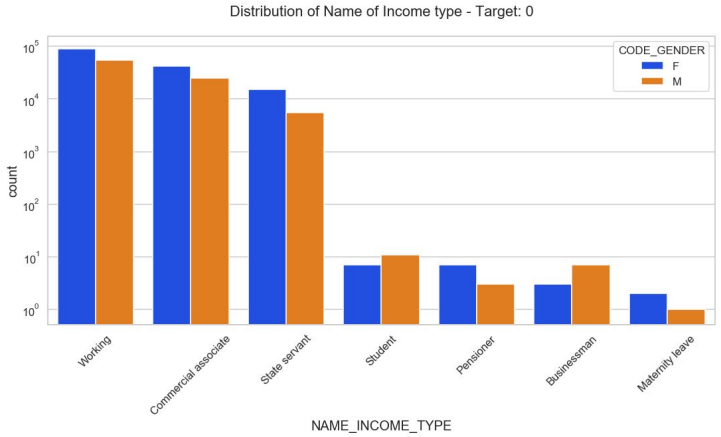
- Male: Female ratio is almost 1:2, but looking at the heatmap, but overall, we can conclude that Male candidates with less income range are the most defaulters, the default rate keeps decreasing as the income range increases.

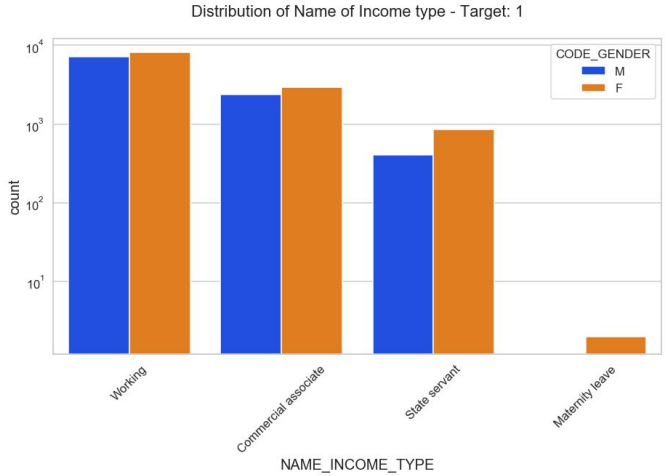
Income Range vs Defaulters (Target:1)

Some good insights

- For income type ‘working’, ’commercial associate’, and ‘State Servant’ the number of credits are higher than other i.e. ‘Maternity leave.
- For this, Females are having more number of credits than male.
- For target value 1: **There is no income type for ‘student’ , ’pensioner’ and ‘Businessman’ which means they don’t do any late payments.**

Females are having marginally higher default rate as compared to males, and most defaulters are working professionals, commercial associates, state servants etc. People with income type: Student, Pensioner, Businessman are not late payers.



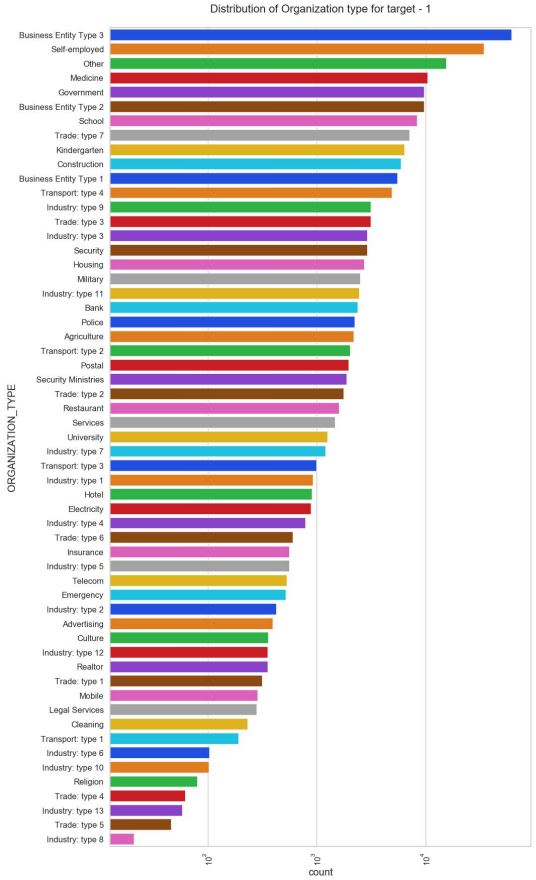


Income Range vs Defaulters (Target:1)

Some good insights

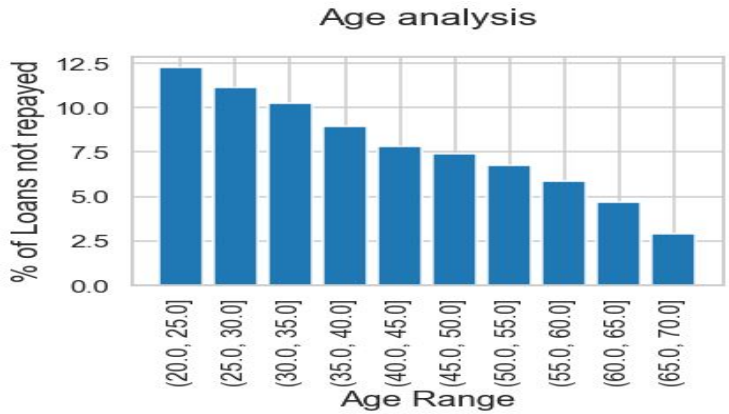
Points to be concluded from this graph.

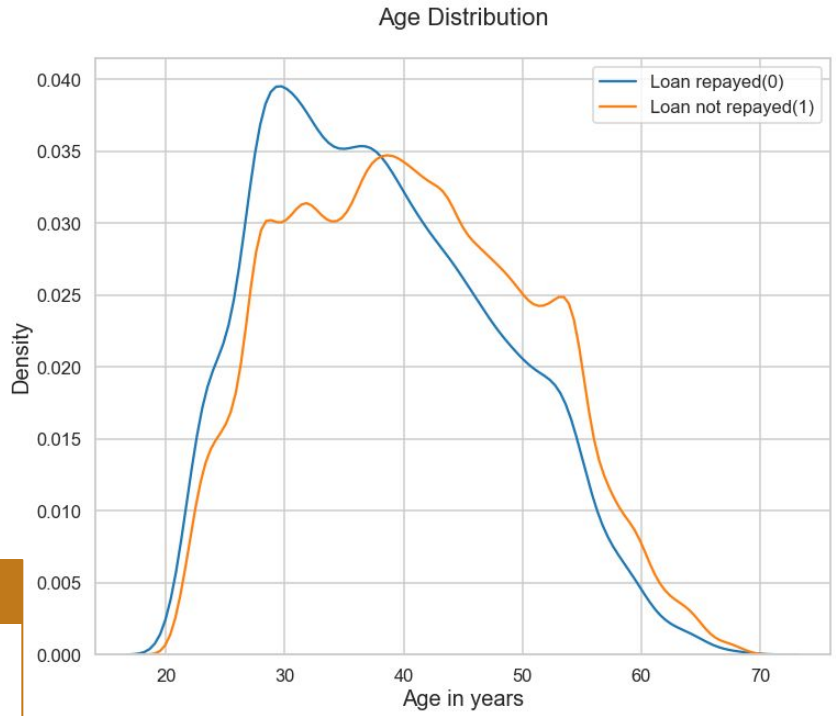
- Clients which have applied for credits are from most of the organization type ‘Business entity Type 3’ , ‘Self employed’ , ‘Other’ , ‘Medicine’ and ‘Government’.
- Less clients are from Industry type 8,type 6, type 10, religion and trade type 5, type 4.
- Same as type 0 in distribution of organization type.



Numerical Analysis

Age Analysis





Findings

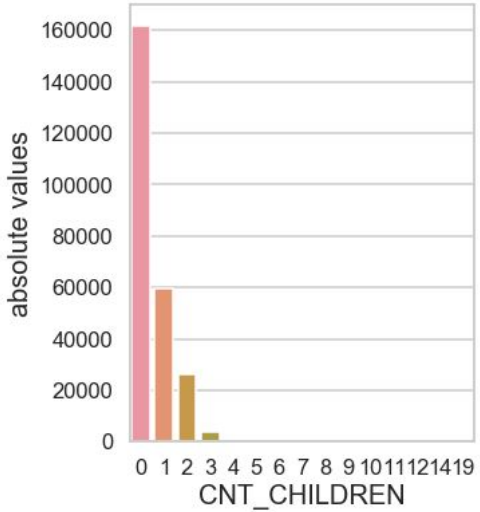
Age Distribution:

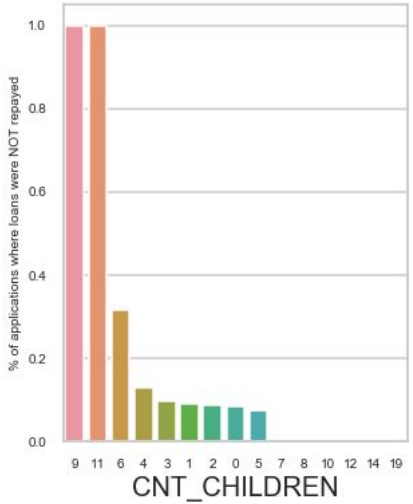
- As clients get older, they tend to repay their loans on time more often.
- Younger clients are less reliable than older clients.
- Even though the correlation (-0.065) is less significant, it does affect the target.

Age Analysis:

- This graph is consistent with our analysis above. Young people have 12% default rate while the oldest have only ~4%.
- Maybe young clients can be given extra guidance for financial planning to help reduce this default rate.

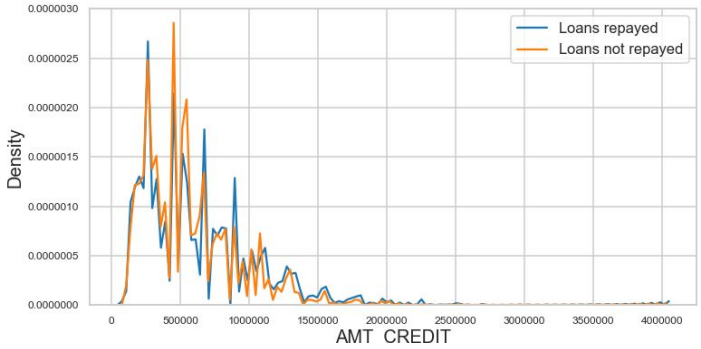
Family Features

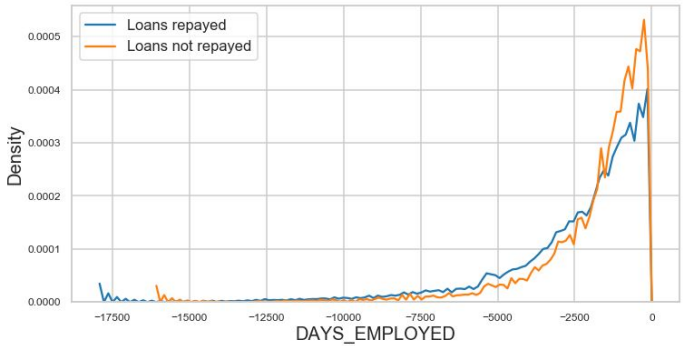


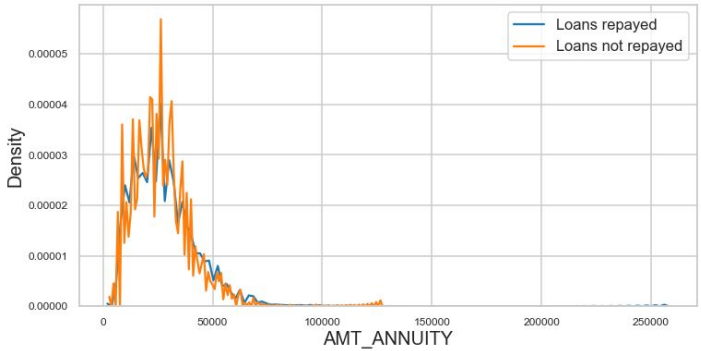


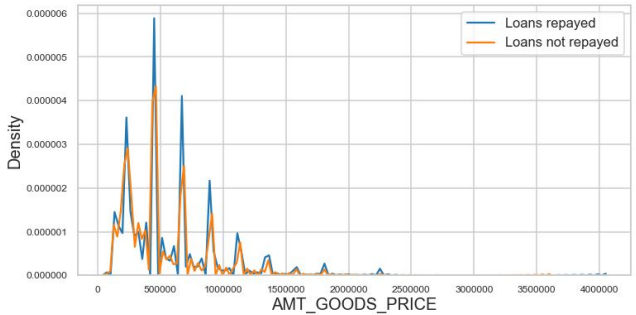
Findings

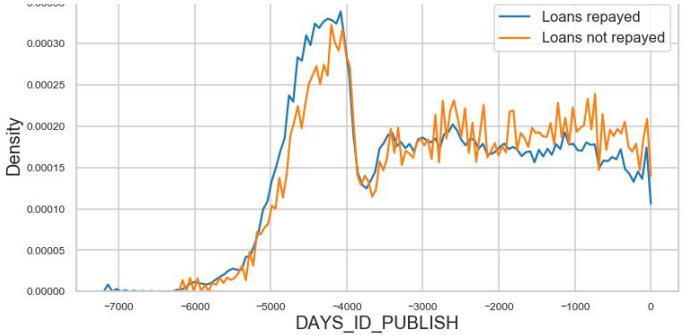
- Most clients have no or very few children and are likely to repay loan on time.
- Clients with very high number of children are risky.

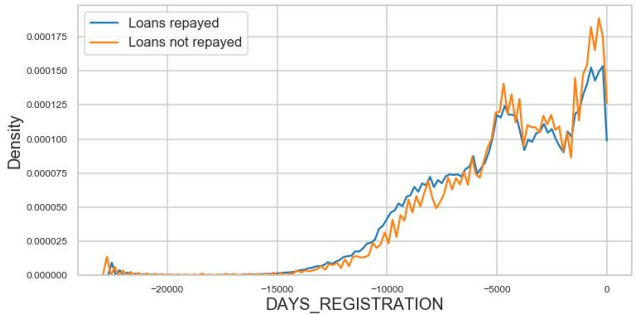












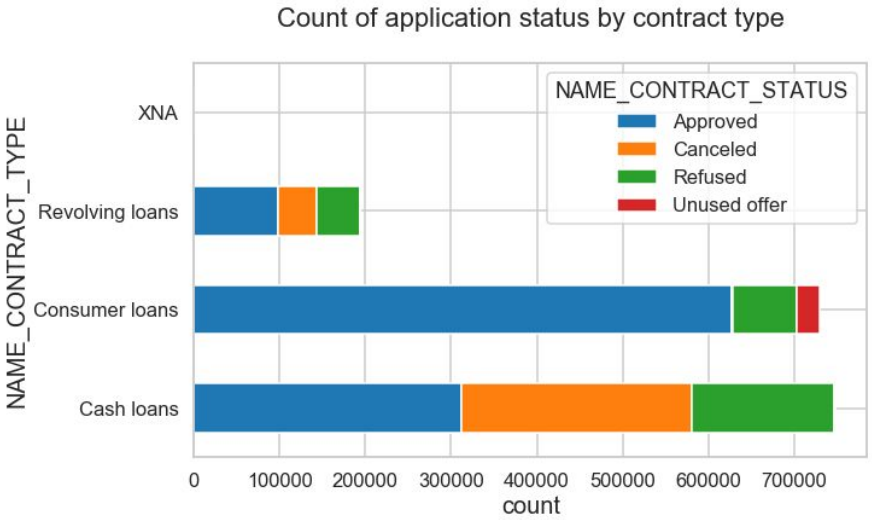
Segmented Analysis

Findings

When clients change their registration closer to application date, they are more likely to default.

Clients who change their identity documents closer to loan application are less reliable than those who change it well in advance.

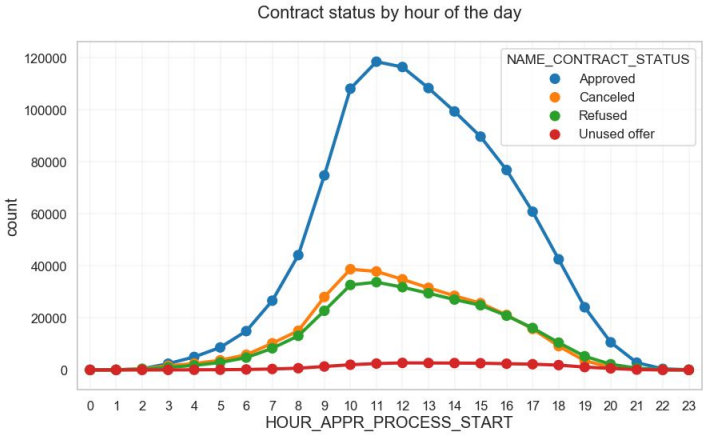
Application Status by Contract Type



Findings

- Consumer loan applications are most approved loans and cash loans are most cancelled and refused loans.
- Consumer loans also rarely cancel, they are the most reliable type.

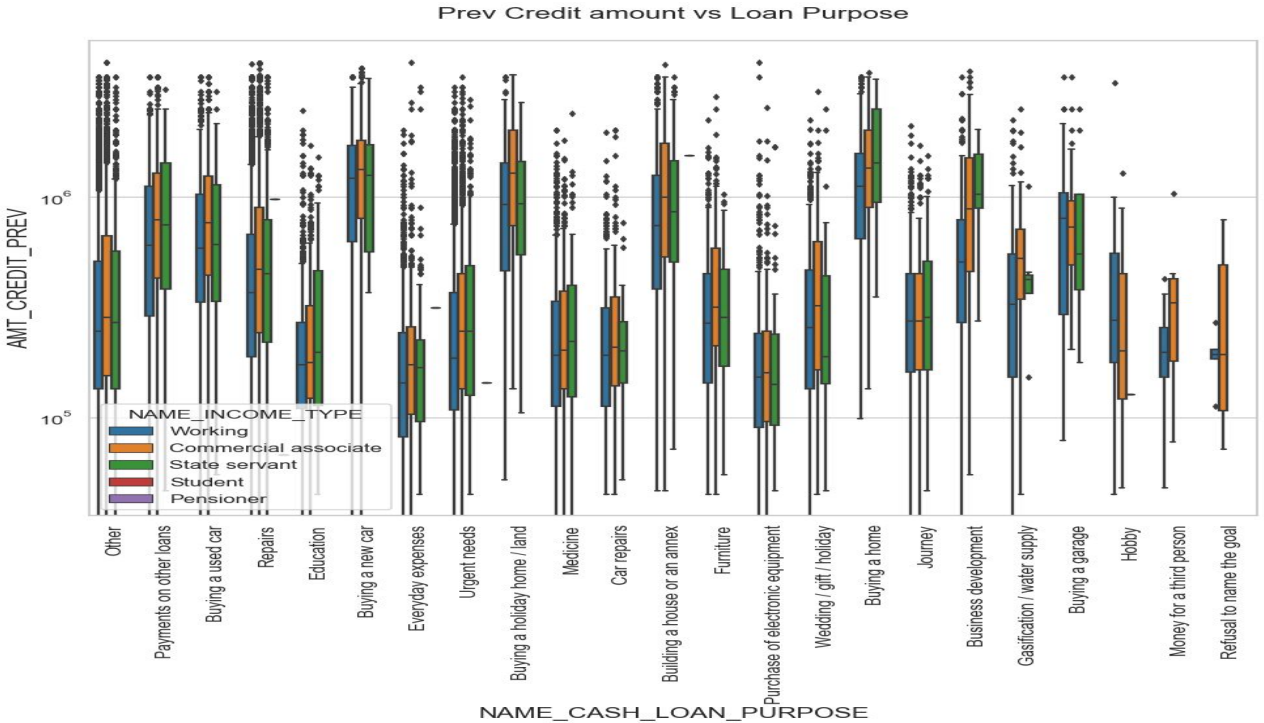
Contract Status by hour of the day



Findings

- Maximum approvals happen around 11 AM.
- Maximum refused and cancelled contracts start application at 10 AM

Previous Credit Amount vs Loan Purpose

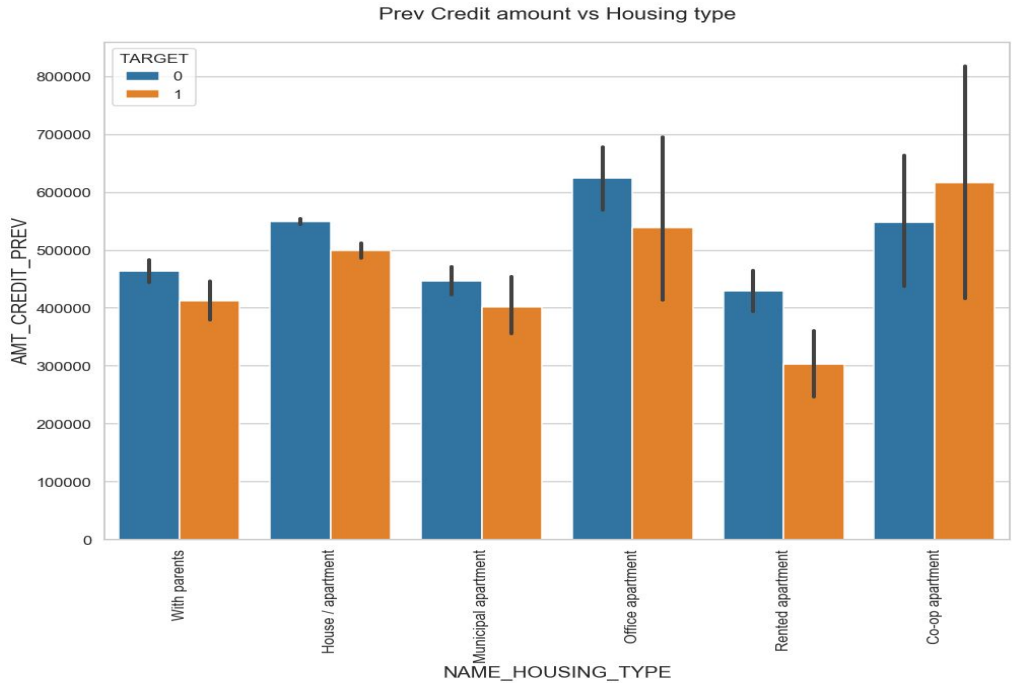


Findings

From the above we can conclude some points-

1. The credit amount of Loan purposes like 'Buying a home','Buying a land','Buying a new car' and Building a house' is higher.
2. Income type of state servants have a significant amount of credit applied
3. Money for third person or a Hobby is having less credits applied for.

Previous Credit Amount vs Housing Type



Findings

From the above we can conclude some points-

Here for Housing type, office apartment is having higher credit of target 0 and co-op apartment is having higher credit of target 1. So, we can conclude that bank should avoid giving loans to the housing type of co-op apartment as they are having difficulties in payment. Bank can focus mostly on housing type with parents or House\apartment or municipal apartment for successful payments.

Final Thoughts:

- Banks should focus more on contract type 'Student' ,'Pensioner' and 'Businessman' with housing 'type other than 'Co-op apartment' for successful payments.
- Banks should focus less on income type 'Working' as they are having most number of unsuccessful payments.
- Also with loan purpose 'Repair' is having higher number of unsuccessful payments on time.
- Get as much as clients from housing type 'With parents' as they are having least number of unsuccessful payments.
- Labourers, Sales Staff, Drivers, seems to be the most defaulters as we concluded earlier, further digging into the female candidates, most of the waiters, private service staff, realty agents, HR staff, IT Staff, Secretaries are the defaulters as compared to their counterparts.
- Female applicants without car are the most defaulters.
- Looking at the current and the previous data, Working, Commercial associates & State servants are mostly prone for being defaulters.
- Most defaulters are in the low income range.

Thanks

The Team

Satyajit Pattnaik