EDA Risk Analysis

Madhusudhan Anand B and Shivaprasad A Metimath

maddymaster@gmail.com
shivaprasadmeti331@gmail.com

Business Understanding

Business Objective:

The finance company is looking for the attributes in a applicant profile which can help them is deciding whether to approve or decline the loan application.

Goal of Analysis:

To find out the relation between the different attributes and their impact on loan defaults. And suggest which attributes contributes a significant difference in Loan Defaults.

The problem or challenge

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
- 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.

The data we had contained 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 Objective

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.

Approach

Data sourcing
- Browse CSV file
- Understand Data
- Understand Business

Data cleaning
- Remove columns
- Manipulate String
- Identify Factors/Numeric
- Convert type

Derived Metrics
- Identify Relevant Numeric Columns
- Define boundaries
- Derive new columns with values based on boundaries

Univariate Analysis
- Identify Relevant Columns
- Identify and treat outliers
- Create plots to visualize data

Bivariate Analysis
- Identify Column pairs
- Observe variations
- Create Plots to visualize data correlation

Summarize Findings
- Highlight trends

My testing method

Each scientist uses different methods of experimentation

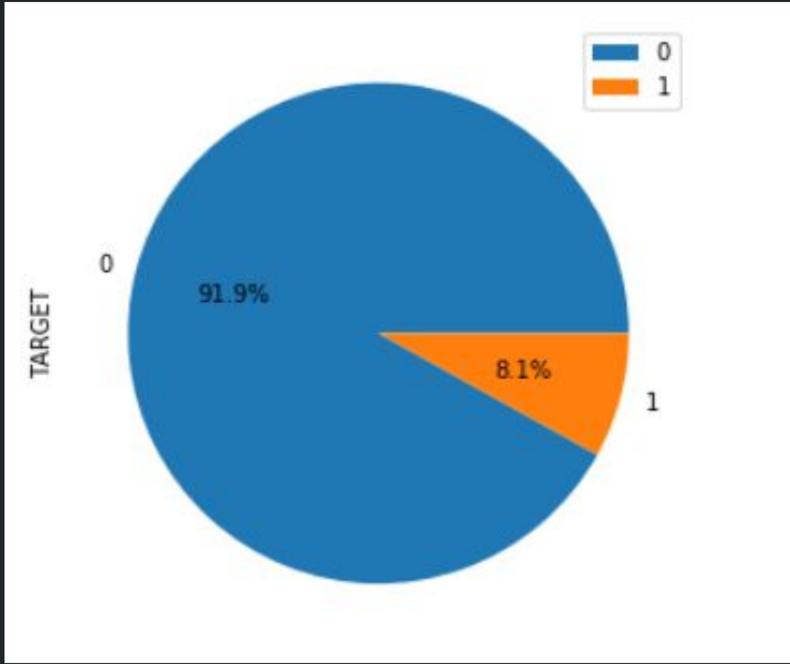
What methods did you use in your experiment?

- Lorem ipsum dolor sit amet, consectetur adipiscing elit
- Incididunt ut labore et dolore
- Consectetur adipiscing elit, sed do eiusmod tempor incididunt

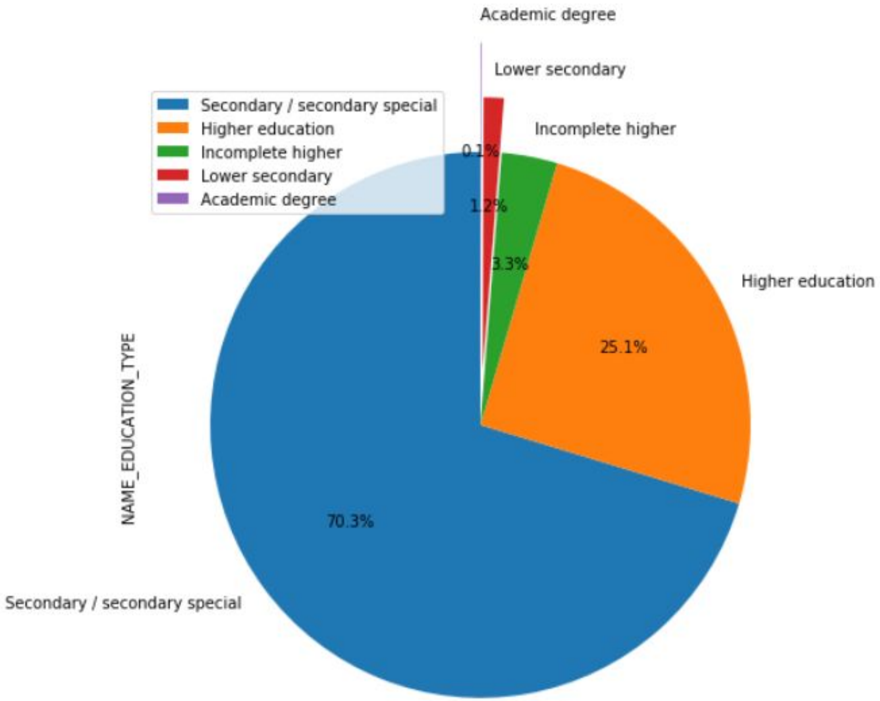
People facing difficulty in paying loan

91.9% face difficulties paying the loan. Only 8.1% pay loan without any difficulty.

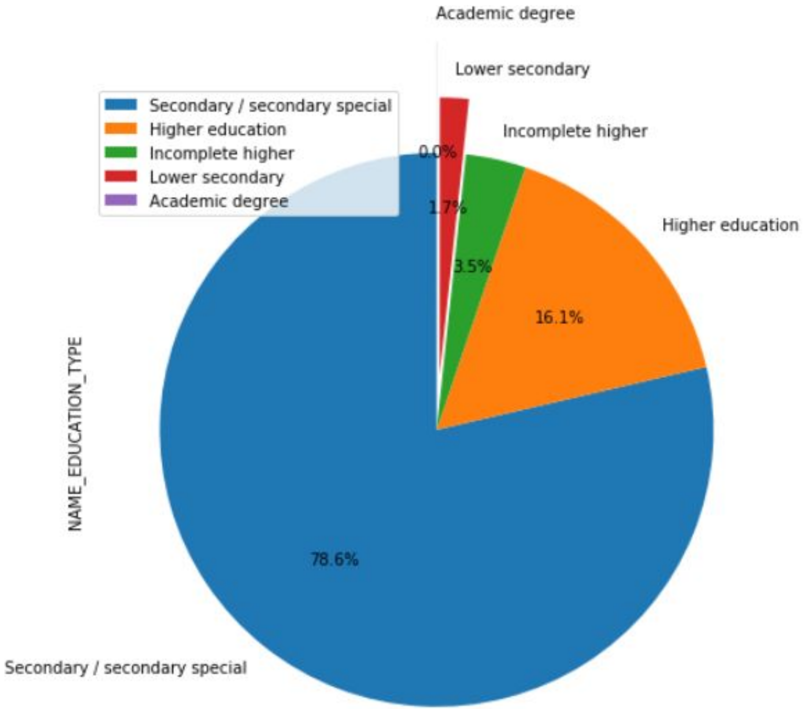
This means that most loan applications present high risk.(92% risk)



Influence of Education Level



Customers who have paid the loan



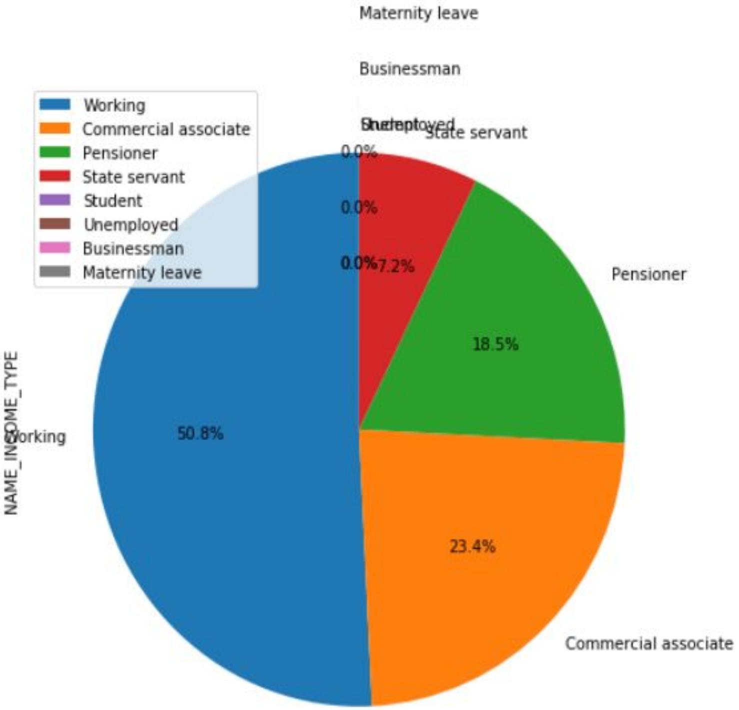
Customers facing difficulties paying the loan

What does this chart mean?

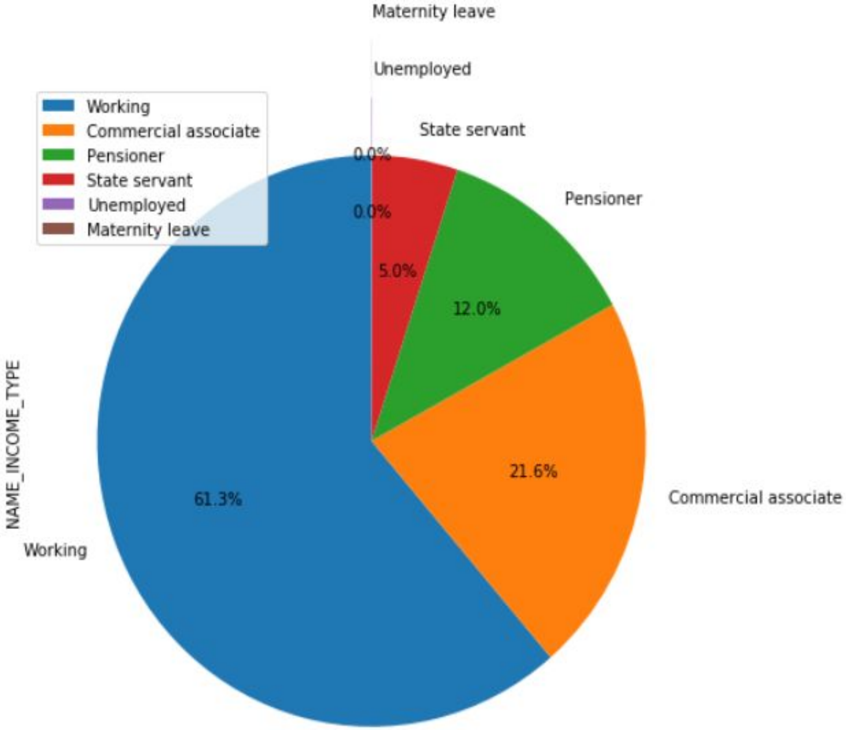
- Education clearly pays a significant role in analysing risk
- Secondary educated sector customer tend to become defaulters at 8.6% more than the others
- While higher education sector customer pay the loan on time are about 19% higher than the rest
- Conclusion: Higher the education, lower the risk and vice versa.

Now let's check Employment Type

IMPACT OF EMPLOYMENT TYPE:



Customer's who have paid the loan

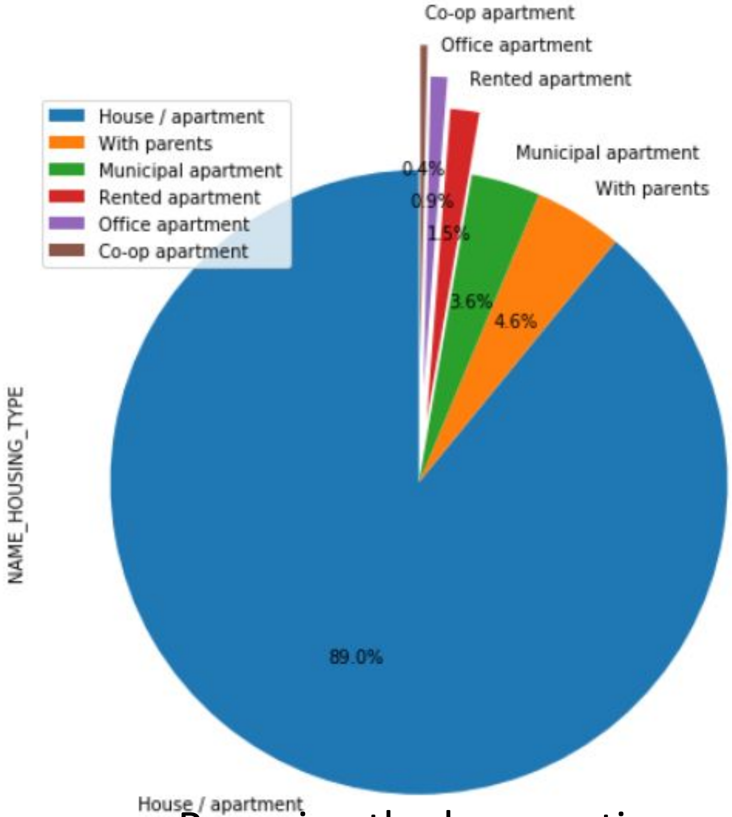


Customer's having difficulties paying the loan

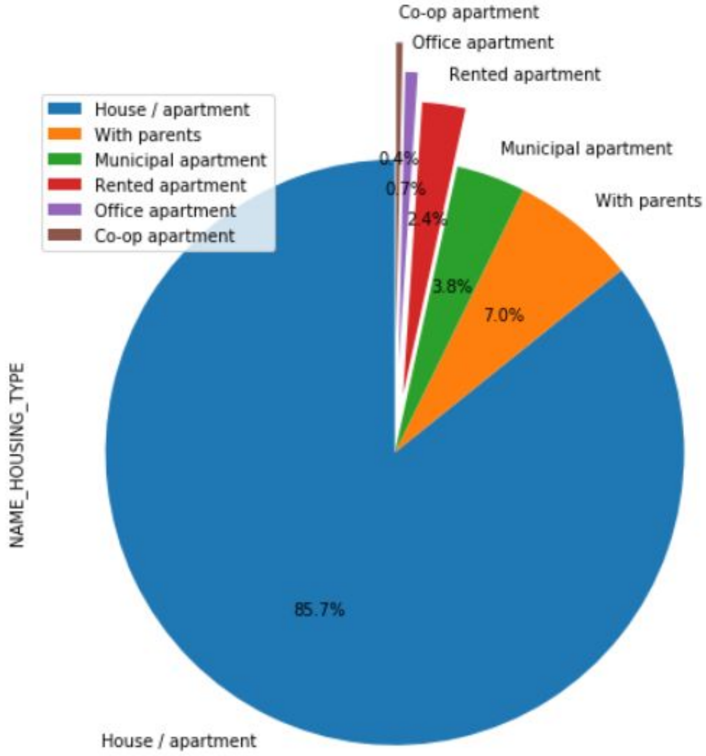
Influence of Employment Type

- **61.3%** of the people who are working tend to face more difficulties in paying back the loan
- While the pensioners, commercial associates and state servants payment details say that they are the top categories paying loan on time

Influence of Owning a House



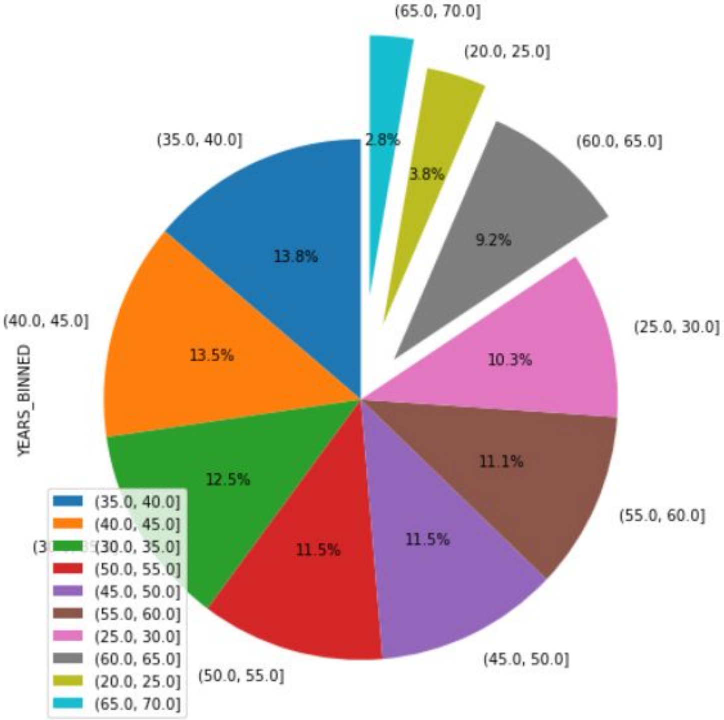
Repaying the loan on time



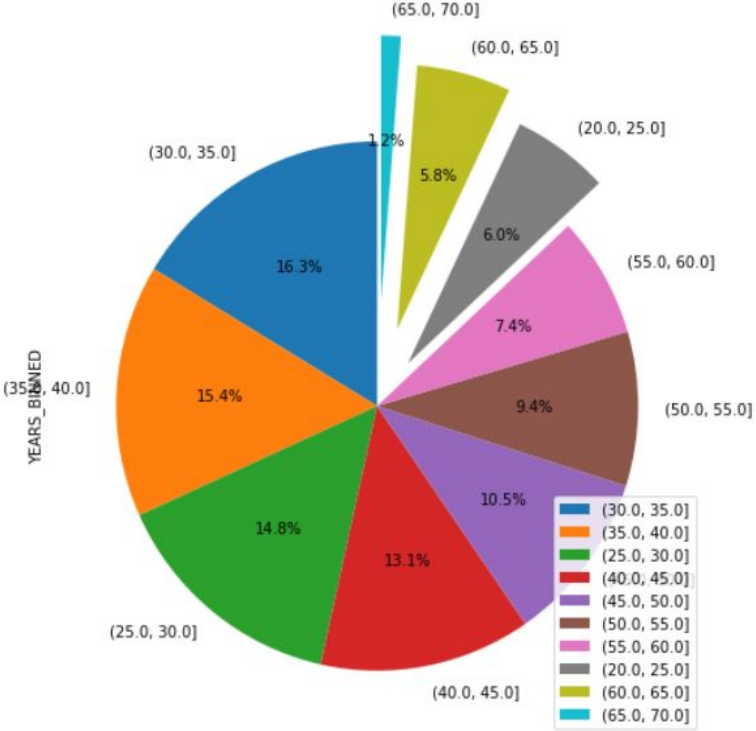
Facing difficulties for paying the loan

- People staying in own house pay the loan on time than others
- And people staying with their parents face more difficulties in paying the loan back
- And rented people also face difficulties

INFLUENCE OF AGE:



Age factor for customers paying on time



Age factor for customers facing difficulties

Observations

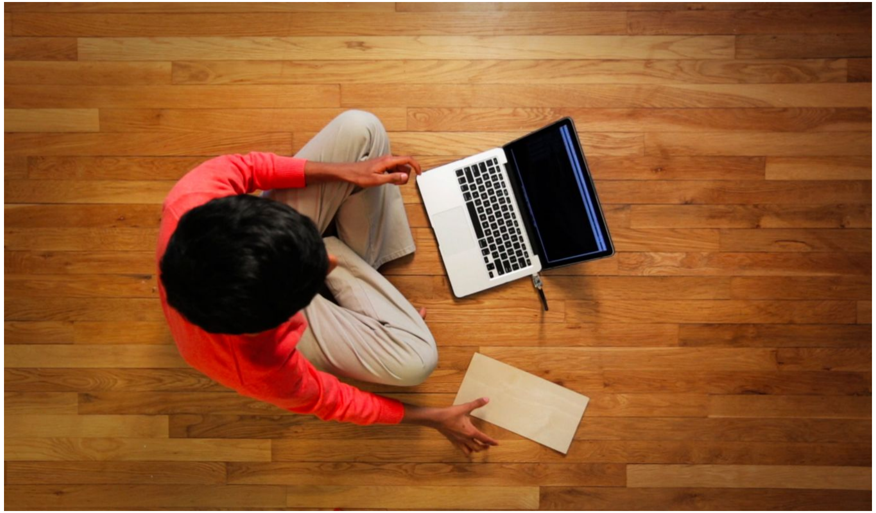
Age group 30 to 40 contribute to over 30% of customers who are not able to payback the loan.

Aha!

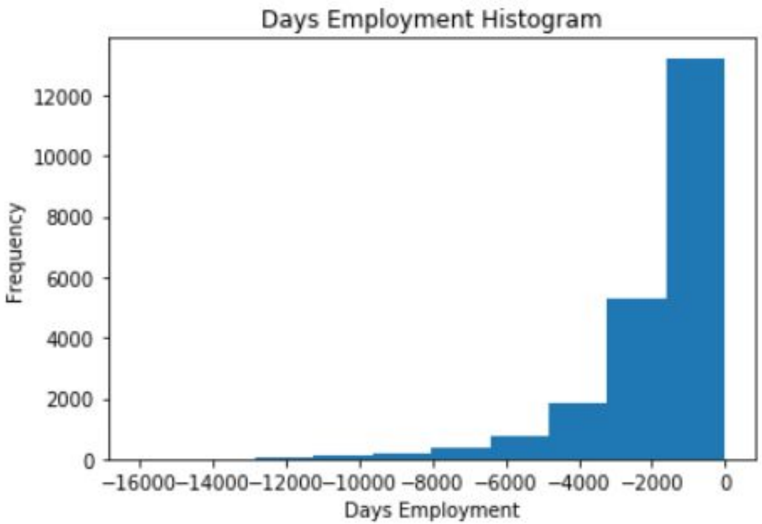
My discoveries

What did you learn after testing?

1. Lorem ipsum dolor sit amet, consectetur adipiscing elit
2. Incididunt ut labore et dolore
3. Consectetur adipiscing elit, sed do eiusmod tempor incididunt



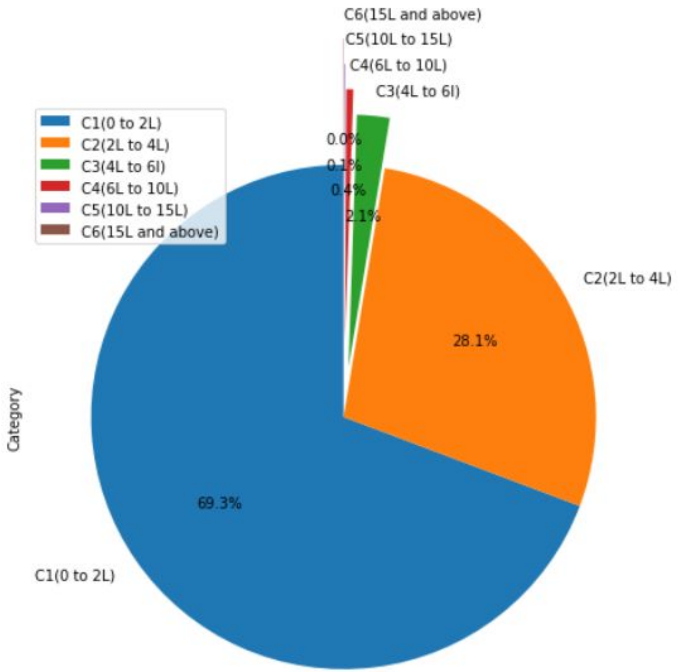
Employment duration of customer



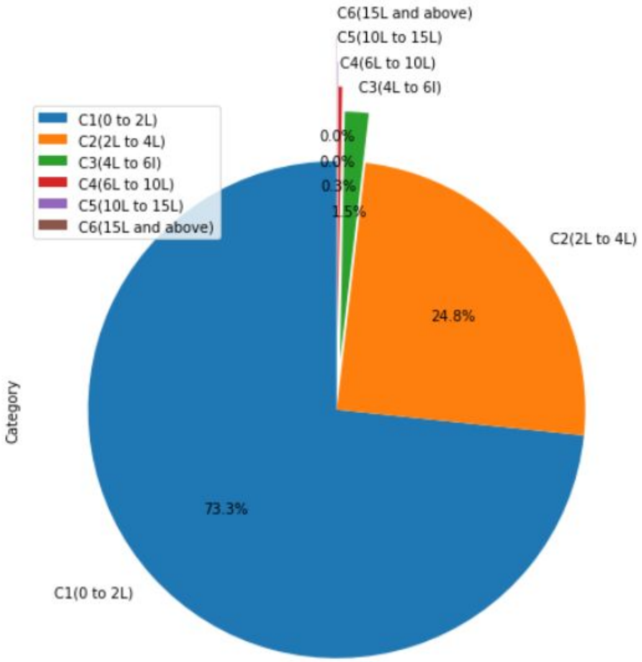
- We can observe that people with less work experience tend to apply for loan early on in their career

The Good, the Bad and the Wealthy

Amount of income:



Total income for the sector paying the loan on time

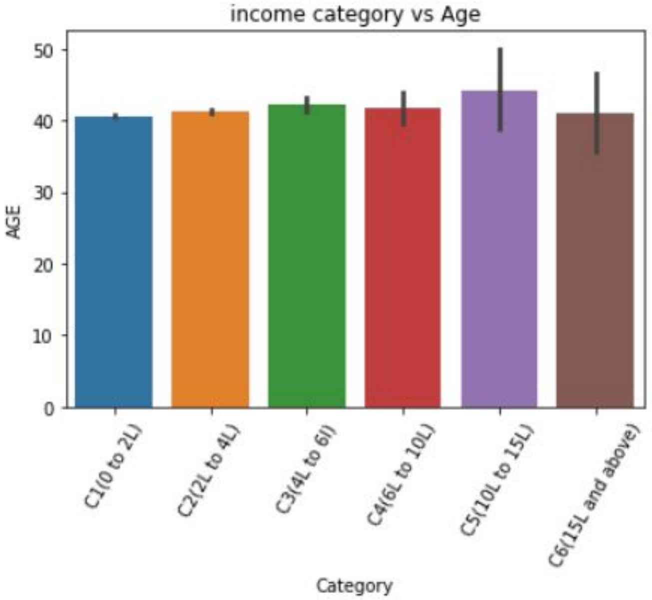


Total income for sector facing difficulties

Influence of amount of income:

- People with income of 0 to 2 lakhs face most difficulties in paying back the loan
- But people with the income range from 2 to 4 lakhs pay the loan o time than the other one

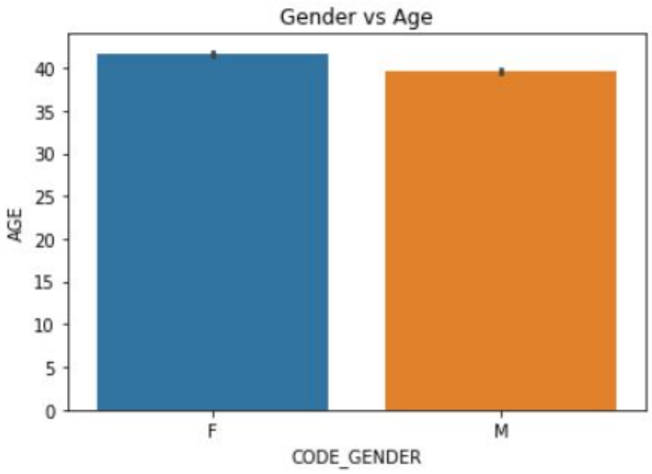
Bivariate analysis of influencing variable



We can see how the age and the total income influence on the number of applications for the loans

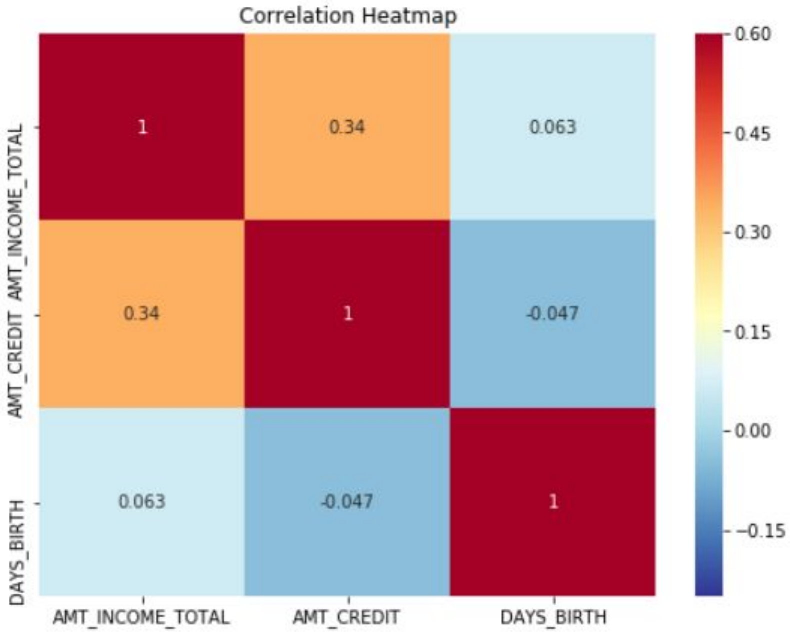
- Category of people with salaries 10L to 15L can be risky. Thats when we need to check other parameters

Gender V/s age:



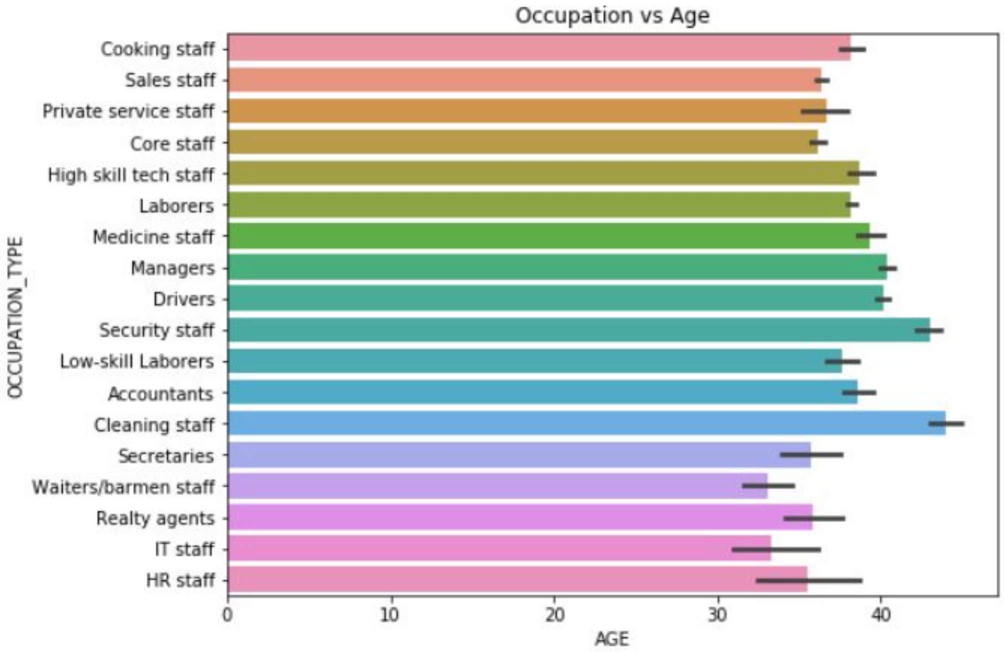
- We can see a slightly higher age of women applying for loan than the men
- And men of younger age tend to apply for the loan
- Different reasons lead to this patterns

CORRELATION OF VARIABLES:



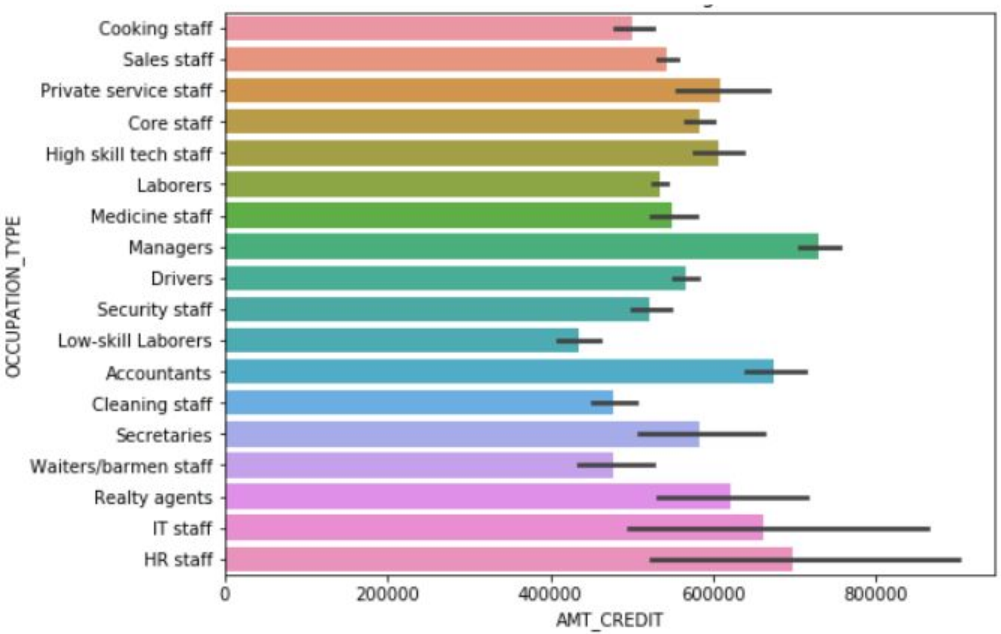
- We can observe that there is a positive correlation btw days birth and total income and amount credit
- But can see that there is a negative correlation between amount credit and days birth

Occupation V/s age:



- Occupation plays major role in one to apply for a loan
- And from above we can observer our set of customers to be aimed at for the business objective

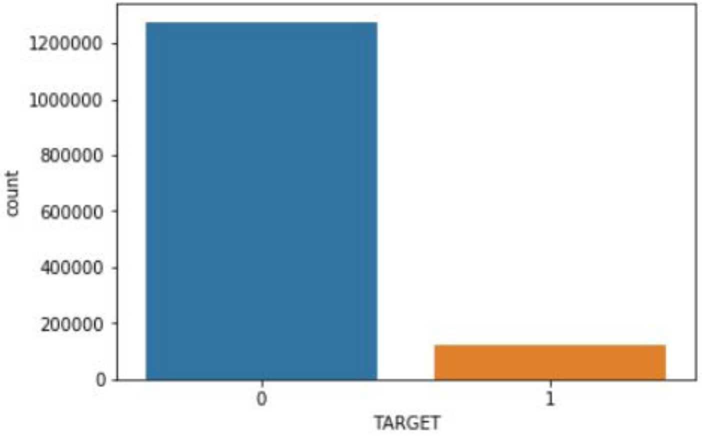
Occupation vs the amount credit:



IT Staff, Realty Agents and HR Staff strangely are high risk as the loan amount is high.

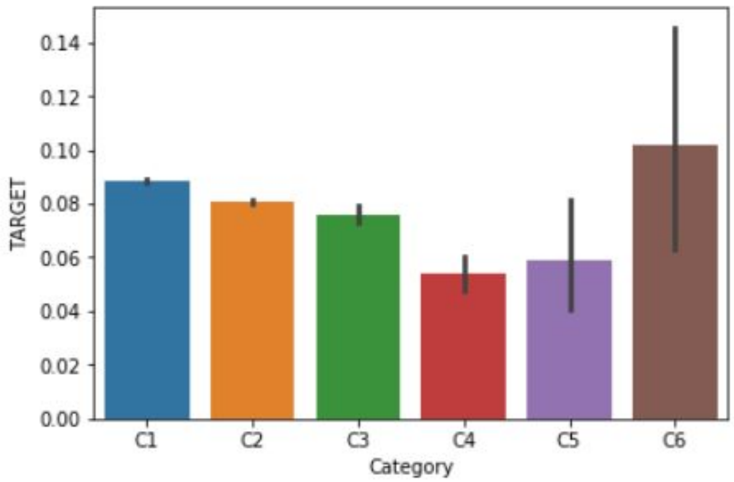
Managers have high appetite for loans, but Drivers tend to be low risky with the loan amount being moderate

Repayers V/s Defaulters:



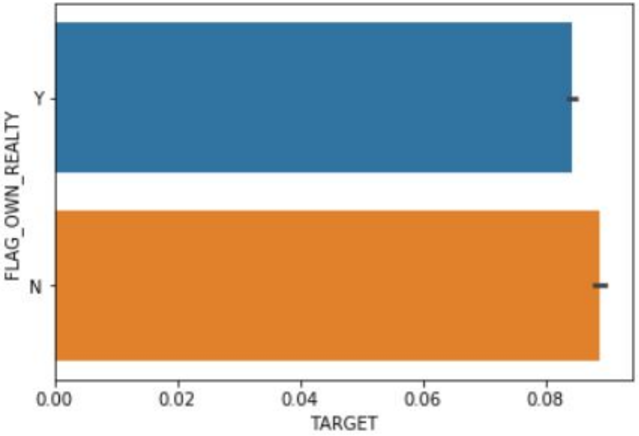
The above plot shows the total numbers customer who are paying the loan and who are having difficulties in paying it back, which is very high

Income categories Vs Target



We can see that the customers with high income tend to become defaulters

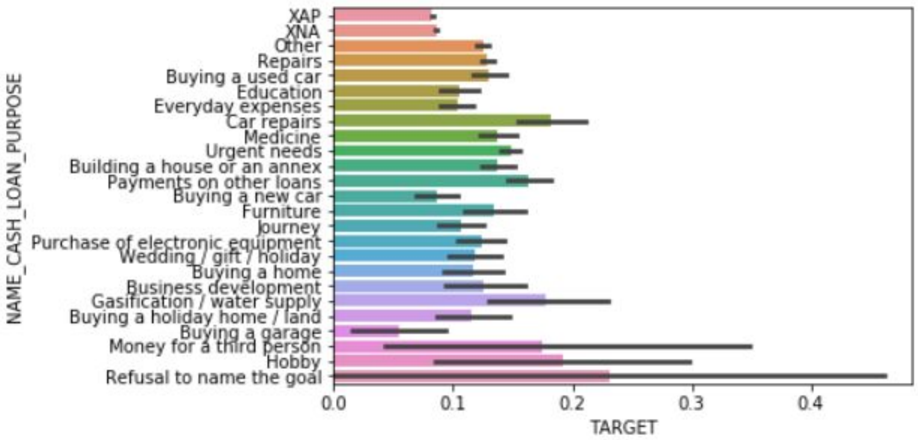
Flag own reality:



The above plot says that people with no own house tend to become defaulters by not paying the loan

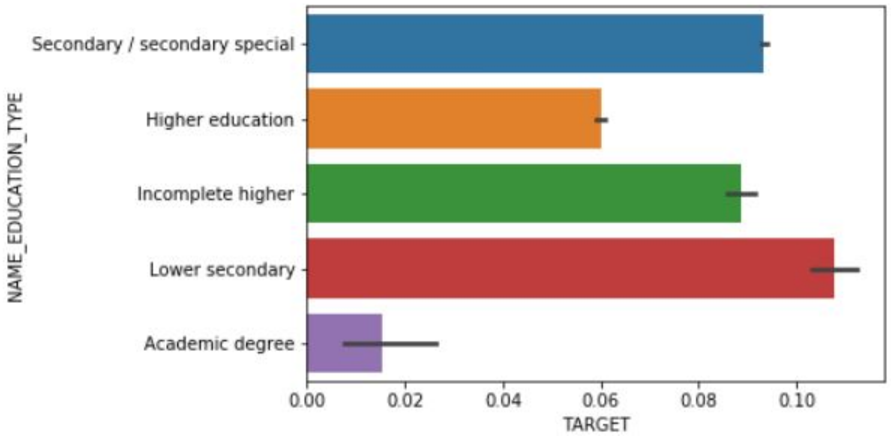
If they have a own house they almost certainly pay-up the loan.

Purpose of loan Vs targets

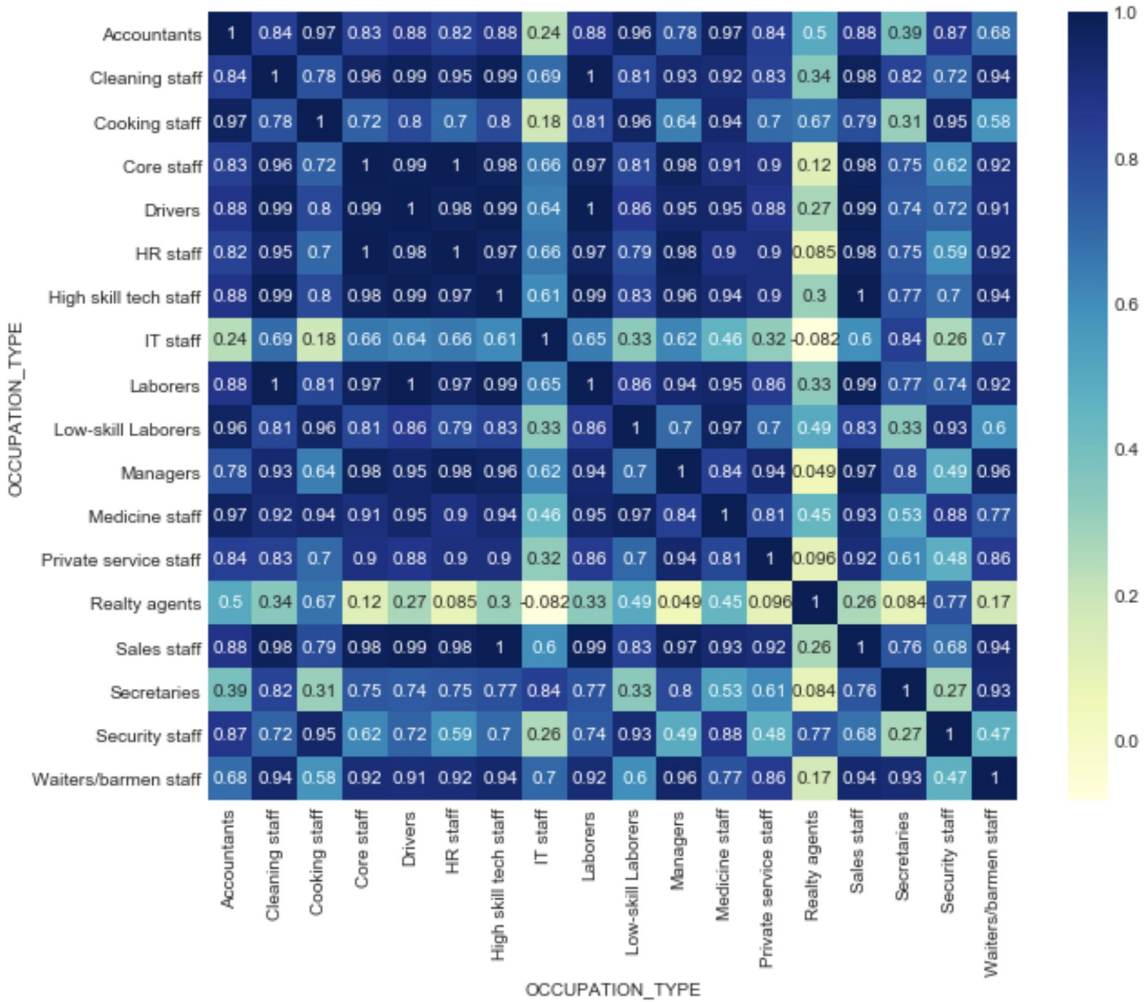


- People with refusal of the goal tend to become defaulters more.
- Hence it is alarming to track these customers carefully before approving the laon

Defaulter Vs education level



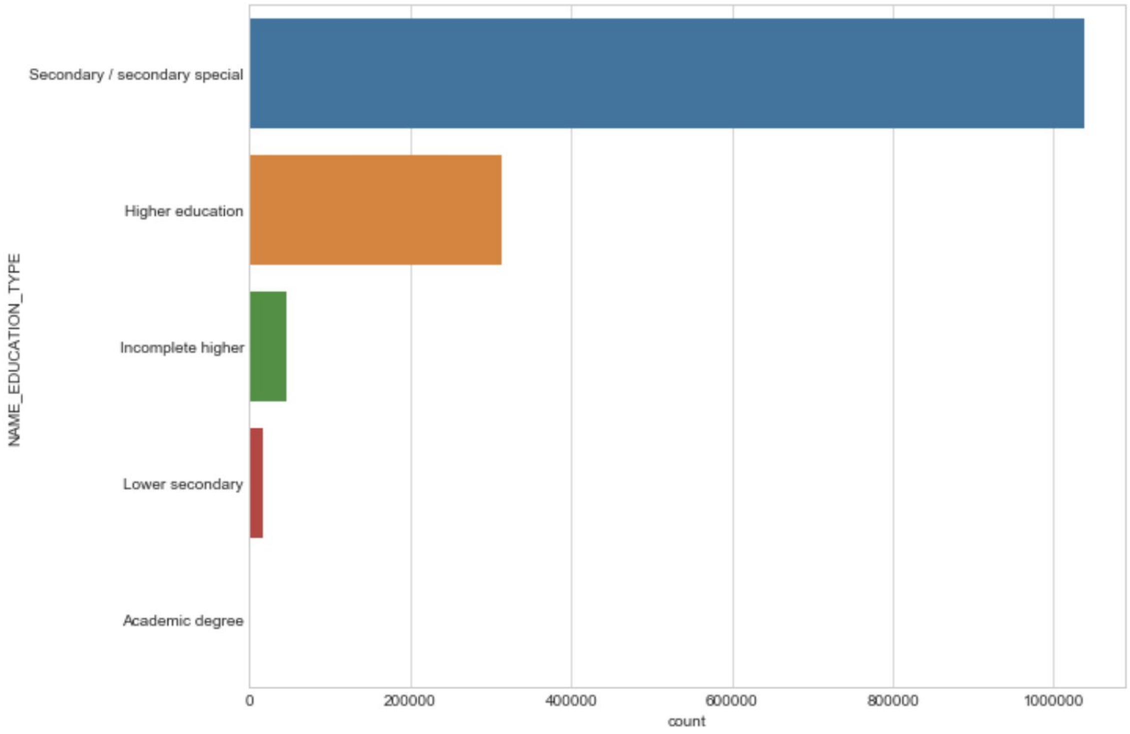
Lower secondary education level customer face difficulties in paying the loan back
And the second one is the secondary educated sector



Drawing correlations, you should not give the loan to a HR Staff and a IT Staff at the same time, both loans can be risky.

This chart will help you understand risk timing

Count of applications educationwise



Percentage Risk by job type and income group

OCCUPATION_TYPE	Accountants	Cleaning staff	Cooking staff	Core staff	Drivers	HR staff	High skill tech staff	IT staff	Laborers	Low-skill Laborers	Managers
NAME_YIELD_GROUP											
XNA	0.065981	0.107125	0.124018	0.079704	0.136392	0.089032	0.083524	0.090323	0.129504	0.212021	0.079684
high	0.062902	0.096698	0.115177	0.073095	0.124964	0.080488	0.077747	0.054054	0.117326	0.191205	0.080136
low_action	0.052066	0.067762	0.103821	0.046683	0.092550	0.041176	0.046346	0.028777	0.084108	0.176871	0.047790
low_normal	0.045243	0.079520	0.089119	0.055716	0.101447	0.055102	0.055242	0.077419	0.094123	0.159798	0.060577
middle	0.054526	0.082504	0.099680	0.067207	0.113370	0.072650	0.062768	0.065445	0.102278	0.182785	0.070251

Risk contd

Medicine staff	Private service staff	Realty agents	Sales staff	Secretaries	Security staff	Waiters/barmen staff
0.090840	0.066715	0.117402	0.111790	0.090960	0.139698	0.122340
0.079665	0.073484	0.106040	0.107099	0.086038	0.117039	0.122555
0.062413	0.050420	0.111111	0.076198	0.038922	0.113699	0.063973
0.058142	0.052799	0.092262	0.084538	0.085286	0.096364	0.097656
0.070276	0.063025	0.072802	0.093548	0.065217	0.102596	0.094488

Driving Factors for Default

92% of risk comes because most applicants are from Secondary and Higher education.

- Lower secondary education level customer face difficulties in paying the loan back
- People with refusal of the goal tend to become defaulters more.
- Hence it is alarming to track these customers carefully before approving the loan
- High income is also high risk as most high income applicants have defaulted.
- Good education, medium salary and academically good is the sweetspot for sanctioning loan.

Conclusion

We can understand risk based on this data. We were able to draw multiple ways of insights based on the given parameters and help the bank know who are high risk versus who are low by income groups, gender, occupation etc.., and helped them reduce risk and give loan to the right people. However the dataset was for a limited time. With more time-series coverage of data, we could come up with better models for more risk mitigation. For now, this is a very good insightful observation

Thank you

Madhusudhan and ShivaPrasad