Analysis on Bank Loan Defaults

A deep dive into the factors affecting credit default

Analysis By – Jayaram Balakrishnan

upGrad

iit-b
ज्ञानमुत्तमम

Problem Statement

- For any commercial bank, the revenue it earns is the difference between the interest it receives from the loans given out to debtors and the interest paid to the depositors.
- More people approaching the bank for loans means more revenue possibilities for the bank.
- However, the bank cannot freely give out a loan to anybody who applies for one. Applicant background checks and credit-worthiness assessments need to be done. Otherwise, there is a high chance of the applicant defaulting on the loan repayment, and the loan turning into a NPA (Non-Performing Asset) for the bank.
- Based on the customer's background checks and credit-worthiness assessment, the bank can arrive at the below decisions on the loan application:
    - Approve the loan.
    - Reject the loan.
    - Provide the loan at a higher interest rate to offset the risk of potential default.
- We need to analyse and identify patterns from the given data to:
    - Identify the parameters affecting the customer's ability to repay the loan on time.
    - Ensure that loan is denied to applicants with payment difficulties, thereby reducing chances of default.
    - Ensure that the loan is not denied to customers with the ability to repay, thereby preventing revenue loss.
- In a nutshell, the objective of this EDA is to help the bank to understand the driving factors behind loan default, i.e. the variables which are strong indicators of default.
- This will lead to a better decisions that are more aligned with revenue maximisation.

What datasets are we working with?

We have been given two datasets in CSV file format, along with the data dictionary:

1. 'application_data.csv' contains all the information of the client at the time of application.
The data is about whether a **client has payment difficulties.**

2. 'previous_application.csv' contains information about the client's previous loan data. It contains the data on whether the previous application had been **Approved, Cancelled, Refused or Unused offer.**

3. 'columns_description.csv' is data dictionary which describes the meaning of the variables.

Tech Stack:

1. Python (IDE: Jupyter Notebook) - Numpy, Pandas, Matplotlib, Seaborn

Assumptions

- The data in the given files has been captured from information provided by the customers, and the genuineness of this information has been verified by the bank to be true.
- There have been no data-entry errors while capturing the information in the bank's system.
- Assumptions made for a few terms whose meanings are not available in the data dictionary:
    - XAP – Not Applicable
    - XNA – Data Not Available
- Currency Unit Assumed: Indian Rupees (INR)
- Region rating where client lives with taking city into account (1,2,3) - Here, the numbers 1, 2, 3 have been interpreted as Tier-1, Tier-2 and Tier-3 cities.
- Reasons quoted for loan default are only probable reasons.

EDA Approach and Methodology

1. Data Loading and Preliminary Inspection
2. Data Cleaning and Processing

2.1 - Handling Missing Data -

2.1.1 - All columns with >40% missing values have been dropped.

2.1.2 - Numerical columns with <40% missing data: Imputation - Median. Mean has not been used to impute as it gets influenced by outlier values.

2.1.3 - Categorical columns with <40% missing data: Imputation - Mode, or isolating into a separate category.

2.1.4 - For occupation type, the missing values have been moved to a new category named 'Unknown'.

2.1.5 - In the gender, contract category and family status columns, 'XNA' values have been replaced with the mode values from these respective columns.

2.2 - Few 'Flag' columns which are not relevant to the analysis have also been dropped.

2.3 - Outlier handling: Outliers are present in numerical columns. Maximum income has been capped to 15 lakhs (1.5 Million) as the order of magnitude of the outlier values are much higher than the rest of the values.

2.4 - Negative values wherever found have been converted to the absolute values.

2.5 - Applicant ages have been derived out of the 'DAYS_BIRTH' column.

2.6 - Critical numerical variables have been segmented into buckets for categorical analysis. New columns have been added for this purpose.

EDA Approach and Methodology - Continuation

3. Analysis and Data Visualisation

   3.1 - Both the current and previous application data sets have been separately analysed to derive insights, before the combined data analysis.

   3.2 - Data Imbalance check with respect to the target variable. Data imbalance is present as 91% of the target variable outcome is 0, i.e. records pertaining to people without payment difficulties. Hence the approach will be to segment the data with respect to the target variable and then analyse.

   3.2 - Univariate analysis has been performed on the numerical and categorical variables to understand the data distribution and customer demographics.

   3.3 - Bivariate analysis has been performed: numeric-numeric, numeric-categorical, categorical-categorical.

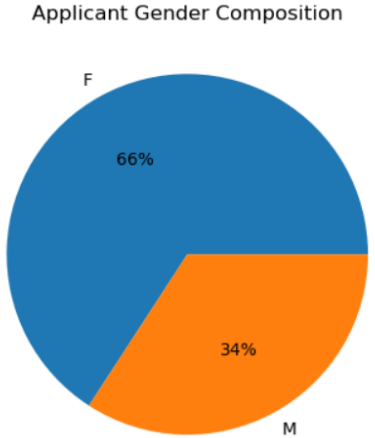
   3.4 - Multivariate analysis of two variables against the target variable.

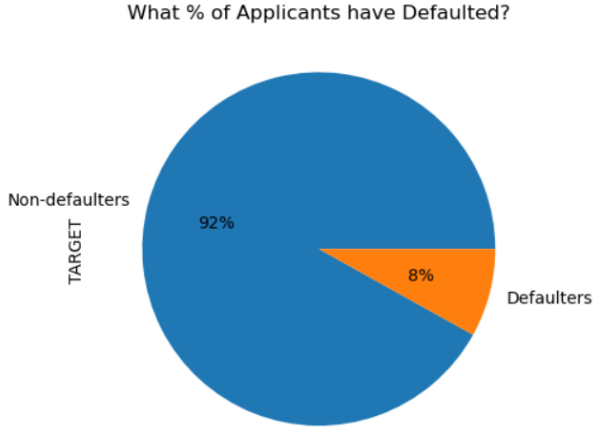
   3.5 - Segmented analysis - Targeted analysis on customers with payment difficulties, by segmenting the data with respect to the target variable and then analysing the correlation between various driver variables.

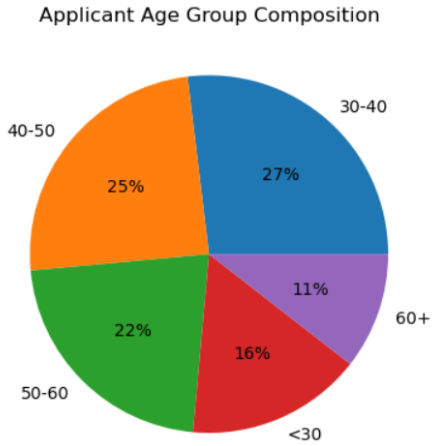
   3.6 - Combining both current and previous application datasets to derive insights.

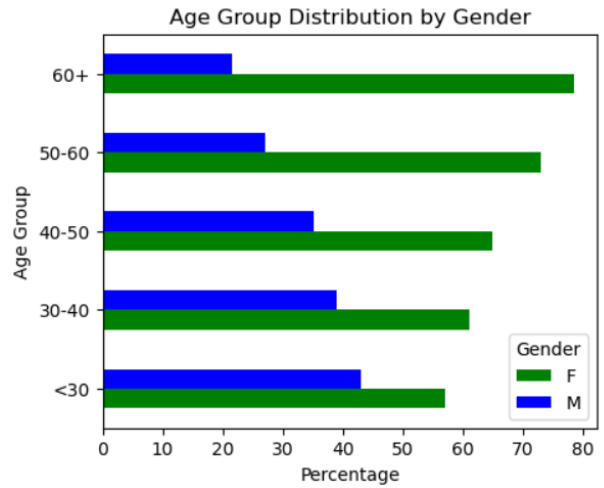
4. Deriving Conclusions and Recommendations

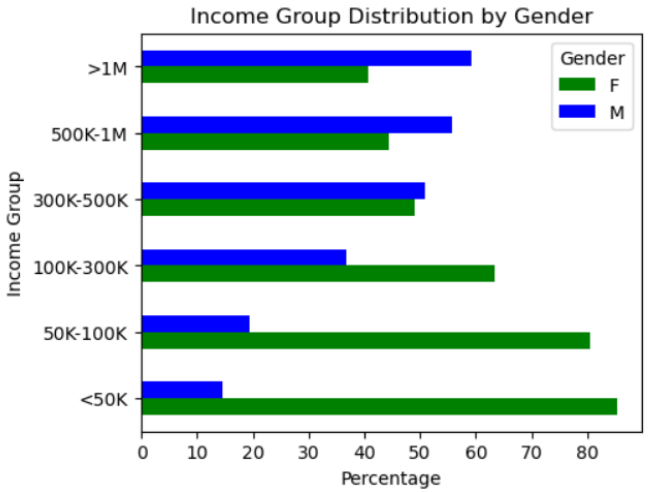
Understanding Applicant Demographics

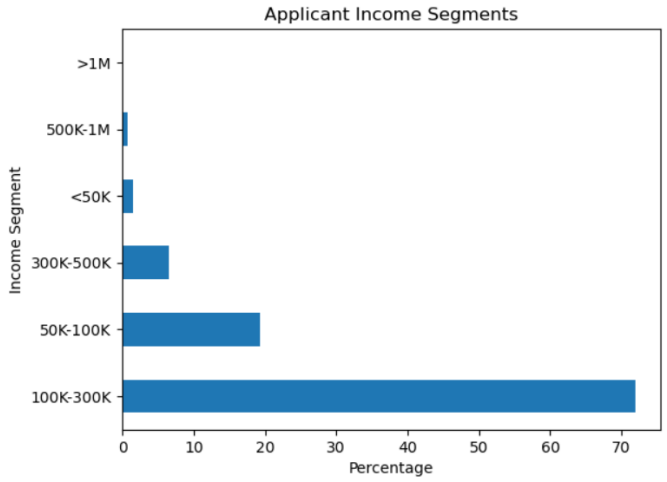


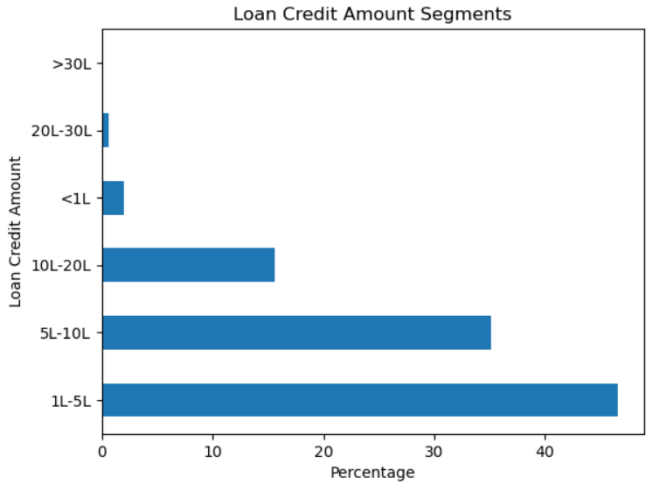


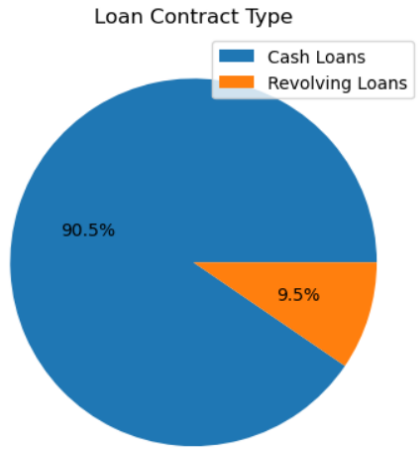








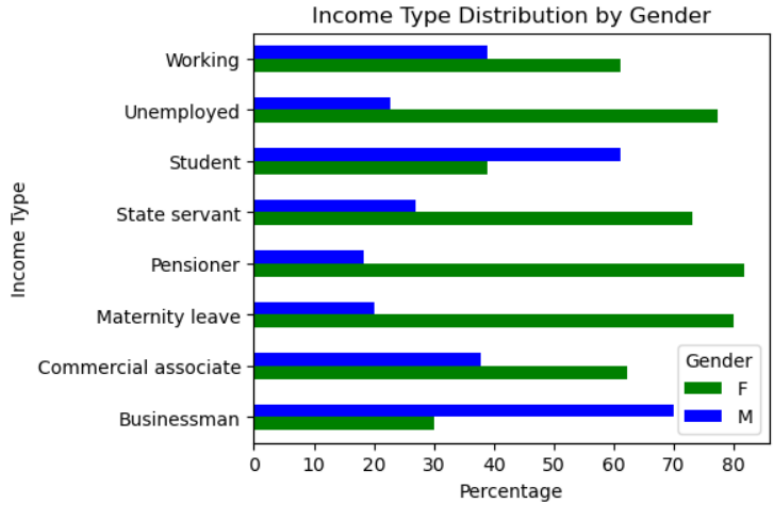




Key Insights:

1. Average applicant age: 44
2. Median income: 1,47,500 INR
3. Around 2/3rd of the applicants are females
4. About 8% of the applicants have payment difficulties
5. Maximum number of applicants are in the 1L-3L income range
6. Average loan amount credited: 5,99,025 INR

Understanding Applicant Demographics - Contd.

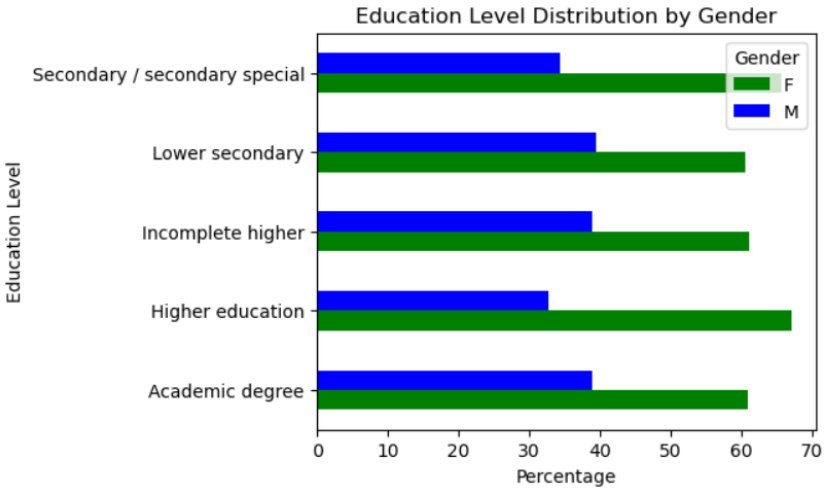


Percentage Distribution of Applicant Income Types

Working	51.632
Commercial associate	23.289
Pensioner	18.003
State servant	7.058
Unemployed	0.007
Student	0.006
Businessman	0.003
Maternity leave	0.002

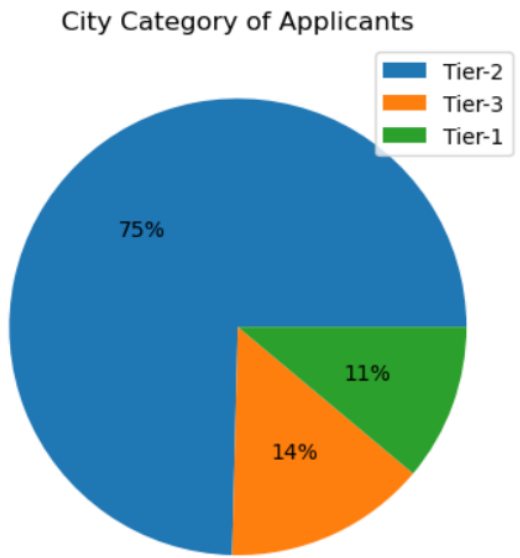
Key Insights:

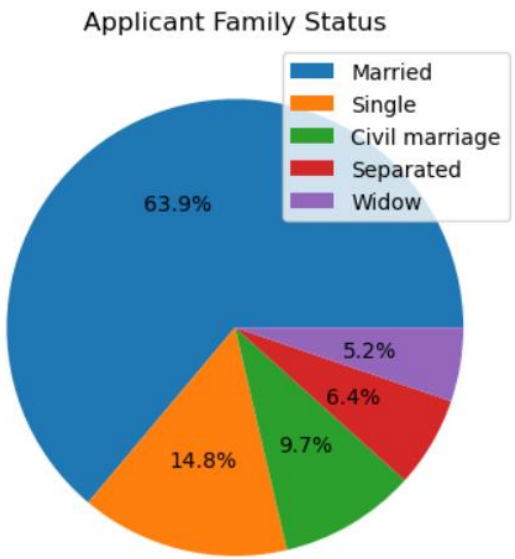
1. Majority of the applicants are from the working class.
2. Majority have studied only upto secondary education.
3. 34% of the applicants who own realty also own a car.
4. 75% of the applicants are from Tier-2 cities.
5. About 64% of the applicants are married.
6. Occupation type is not known for 31% of the applicants.

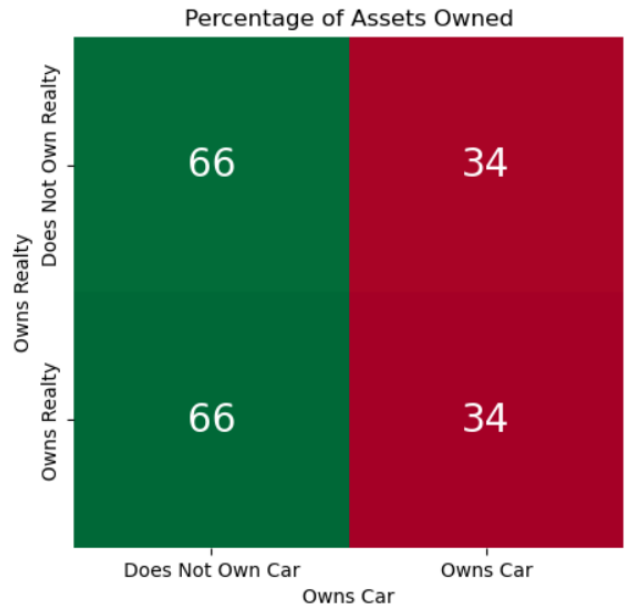


Education Level Distribution of Applicants

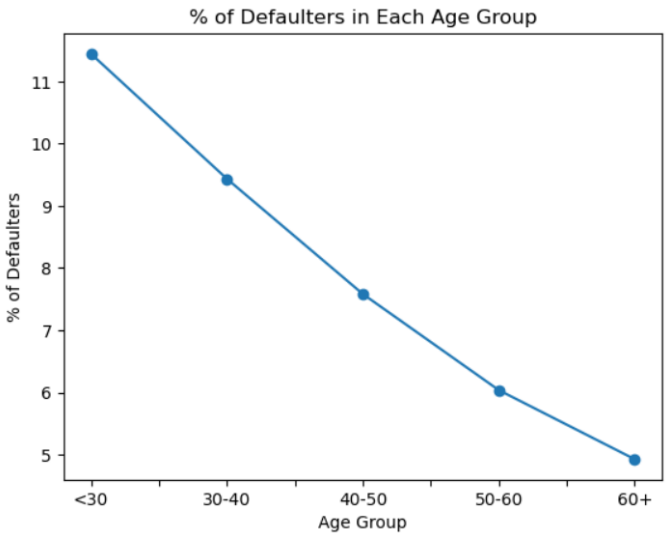
Secondary / secondary special	71.02
Higher education	24.34
Incomplete higher	3.34
Lower secondary	1.24
Academic degree	0.05

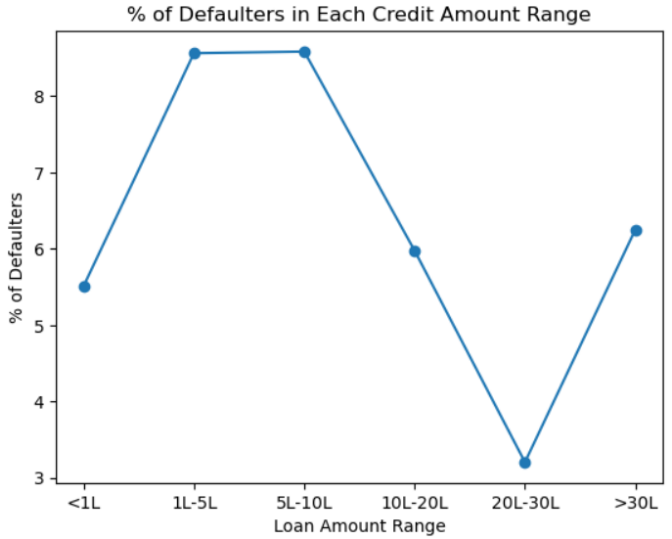


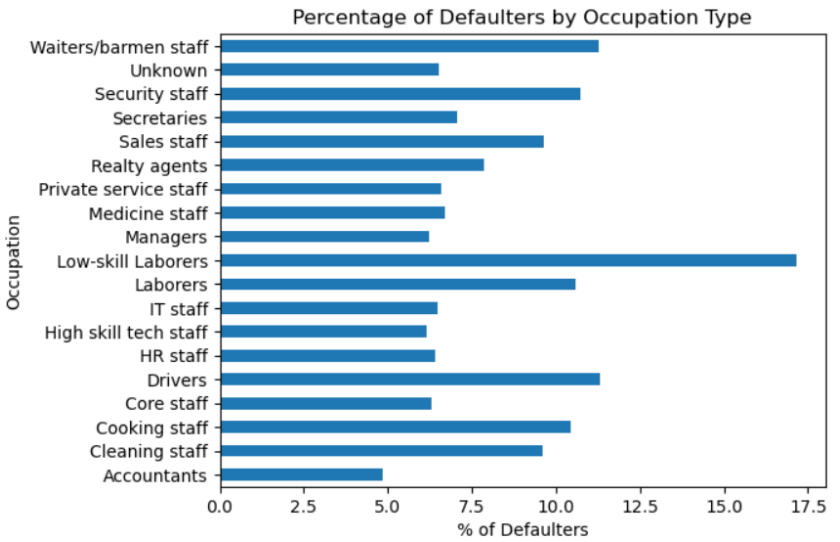


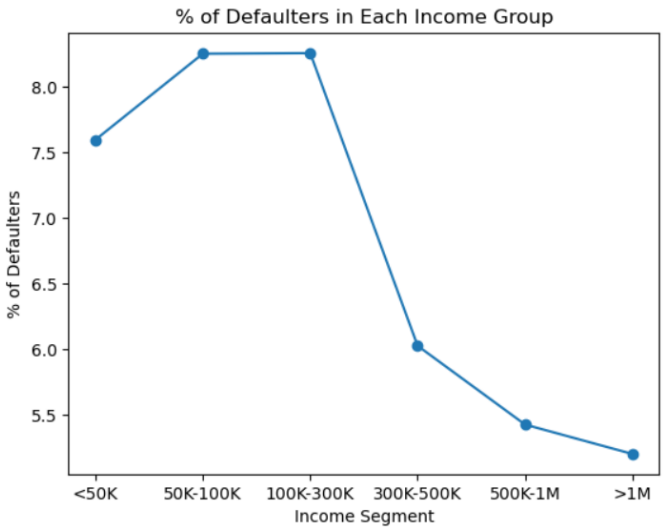


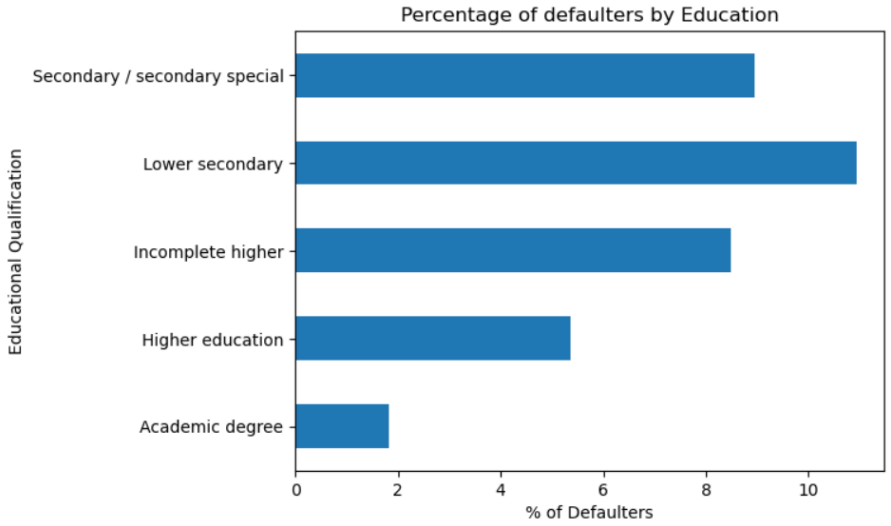
Who are more likely to default?

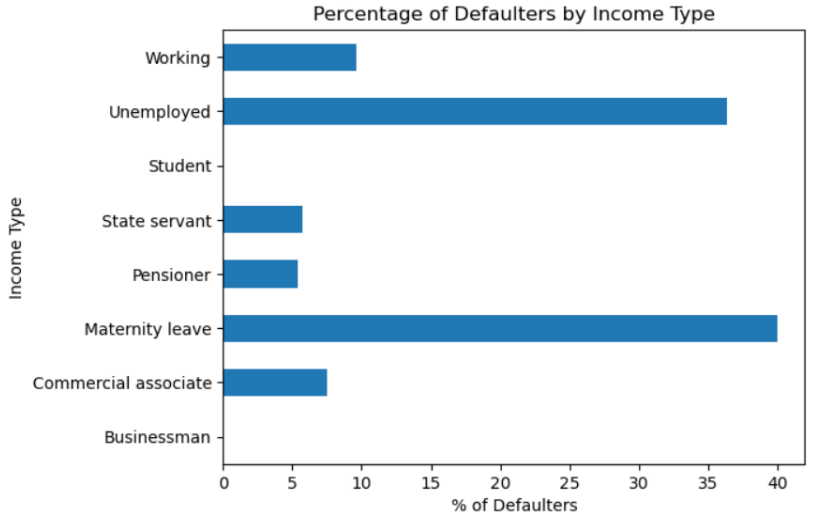




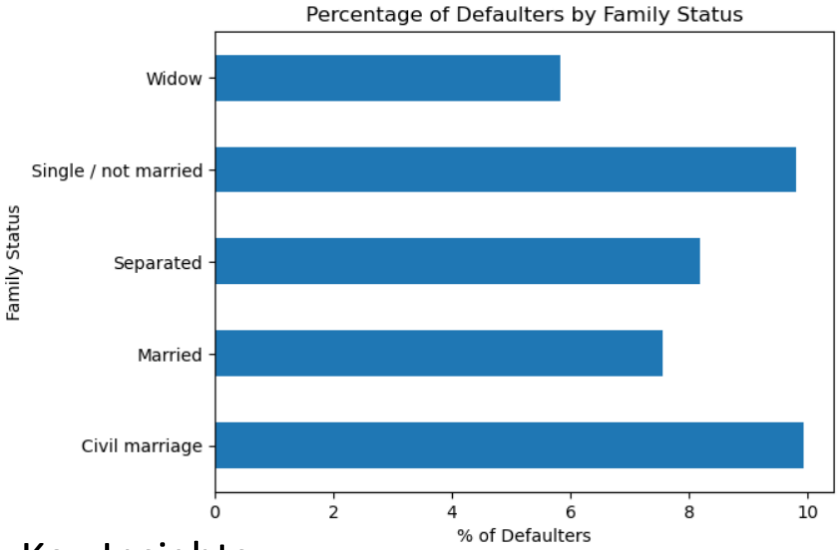


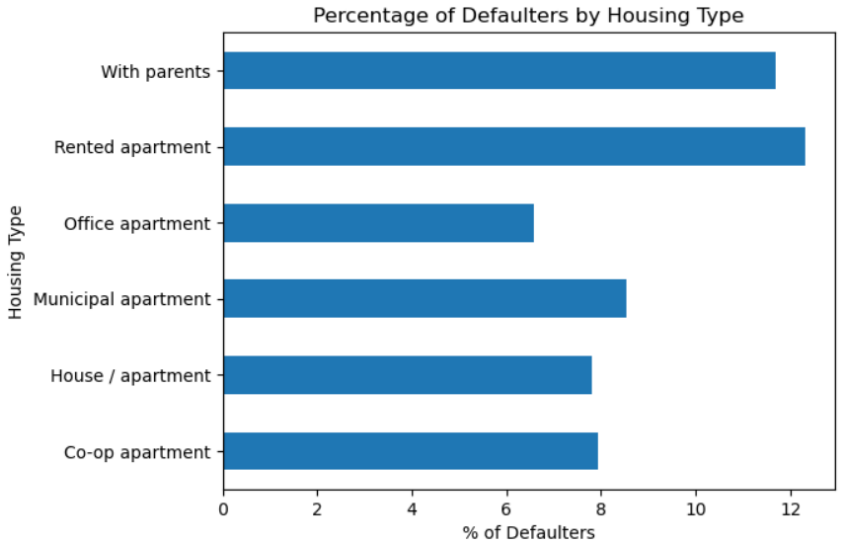


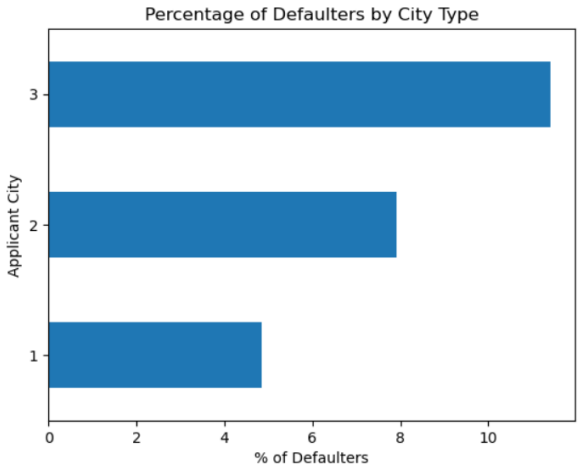




Who are more likely to default? - Contd.



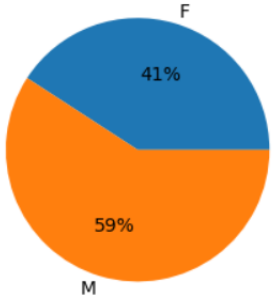




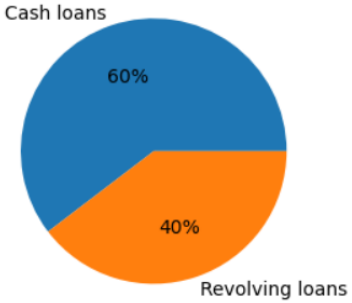
Key Insights:

1. There are more defaulters among the younger age groups (age<40). Possibility of default reduces with increasing applicant age. Reason: Senior people may be better placed in their careers and in a better financial situation. Senior people have lesser tendency to make lesser impulsive purchase decisions.
2. People with income >3 Lakh are much less likely to default. Reason: Better financial situation.
3. People who are better educated are less likely to default. Reason: Educated people are more aware of the consequences of loan default, and refrain from doing so.
4. Default possibility is higher in people taking loans in the 1 Lakh-5 Lakh range, and the 5 Lakh-10 Lakh range.
5. Unemployed applicants have significantly higher chances of default, and businessmen and students are less likely to default. Reason: Lack of regular income among the unemployed.
6. Low skill-labourers, and in general, blue-collar workers, are more likely to default.
7. Applicants who are widows are less likely to default, and unmarried and civil-married people have comparatively higher chances of default. Reason: Lesser financial commitments for widows.
8. Female applicants are less likely to default than male applicants.
9. Applicants living with parents, and the ones living in rented apartments are more likely to default. Reason: High apartment rents, probably high medical expenses of dependent parents.
10. Applicants from Tier-3 cities are more likely to default compared to the ones from Tier-1 cities.
11. Default percentage is higher in cash loans than revolving loans.

Defaulters Gender Composition



Defaulters % by Contract Type



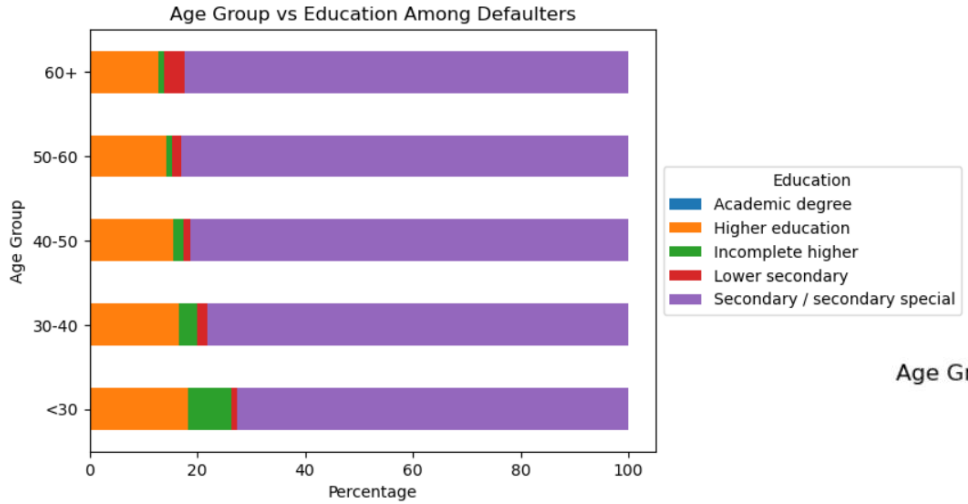
Top Driver Variables for Default:

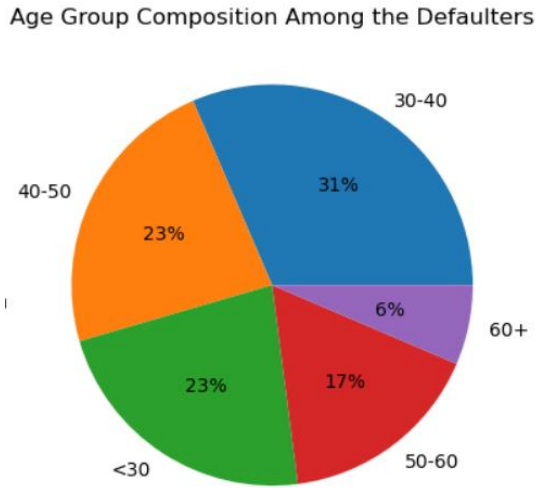
1. Age
2. Income
3. Education
4. Income Type
5. Occupation Type
6. City

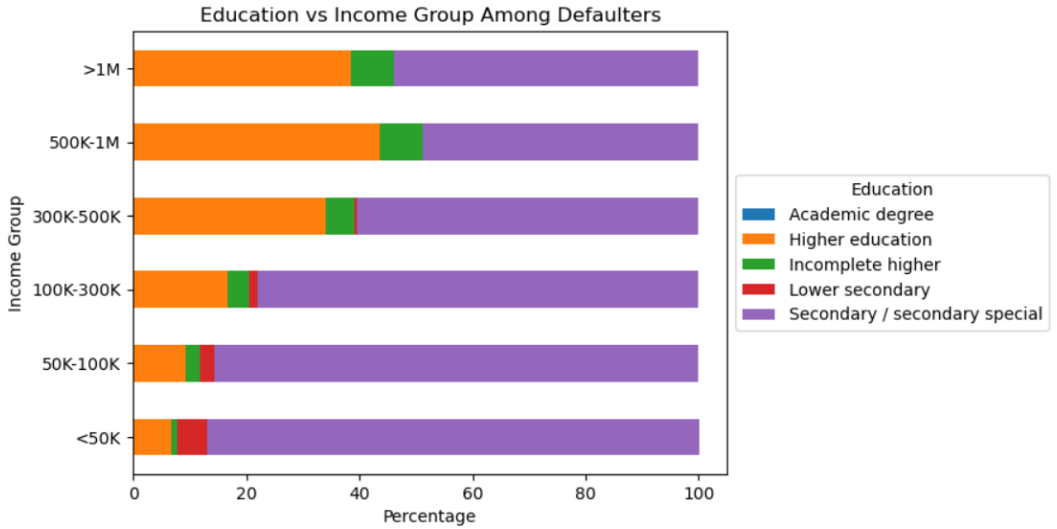
upGrad

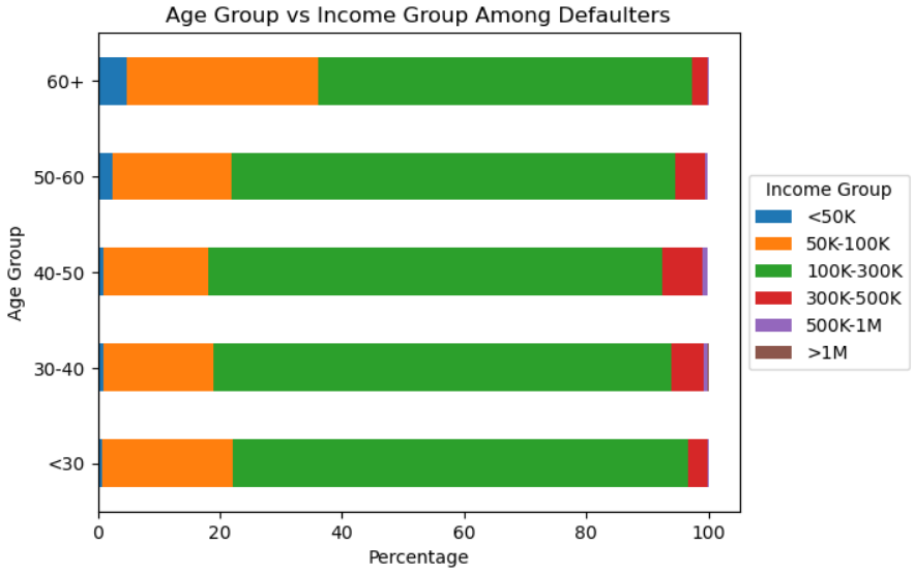
iit-b ज्ञानमुत्तमम

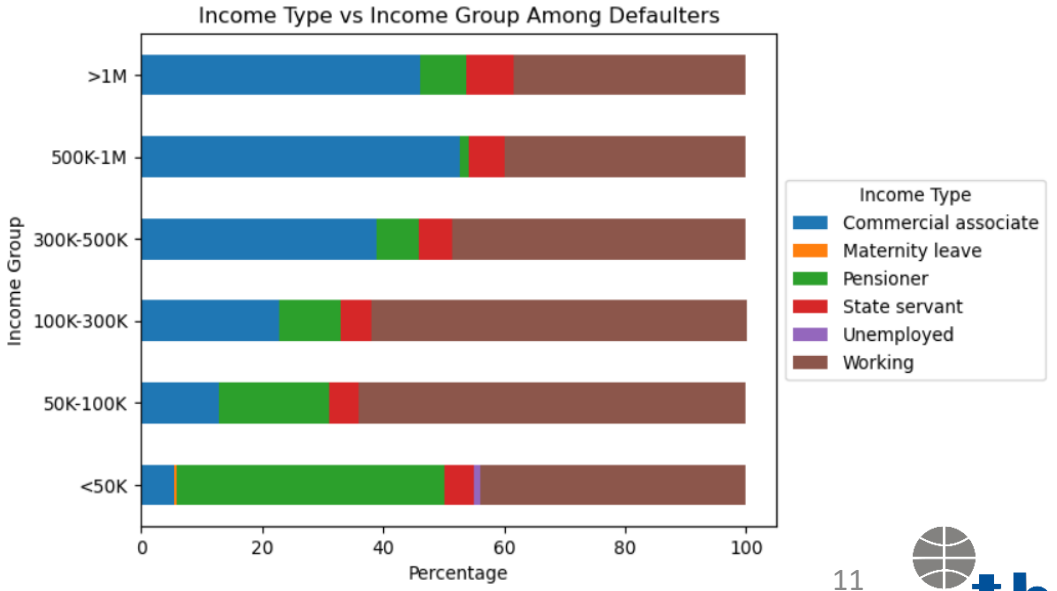
Targeted Analysis on People With Payment Difficulties (Defaulters)



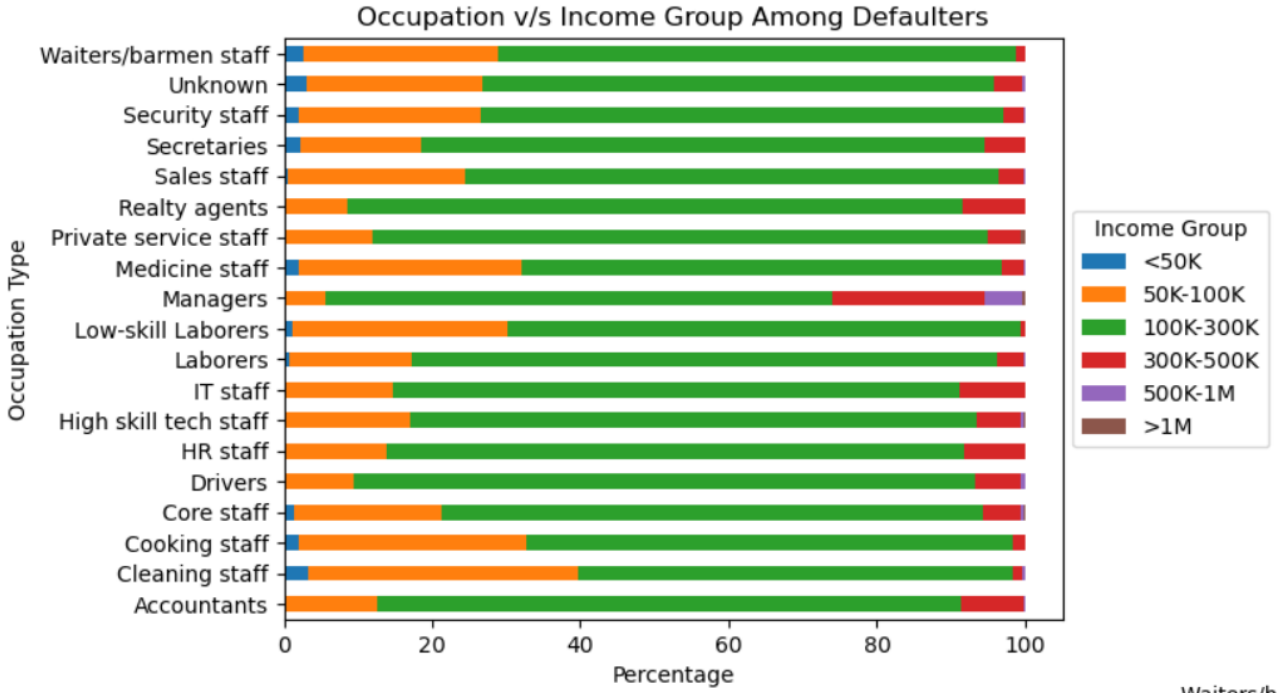


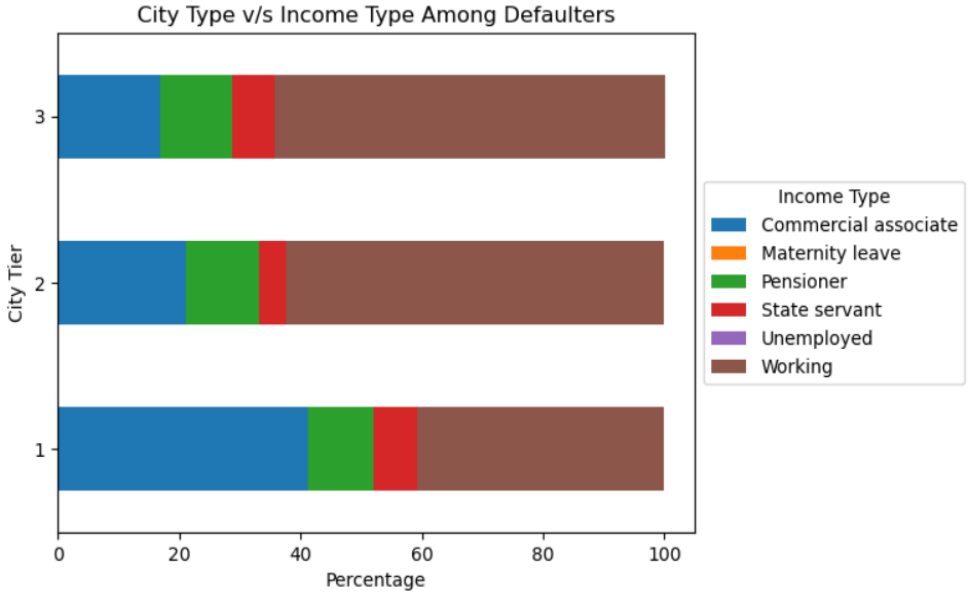


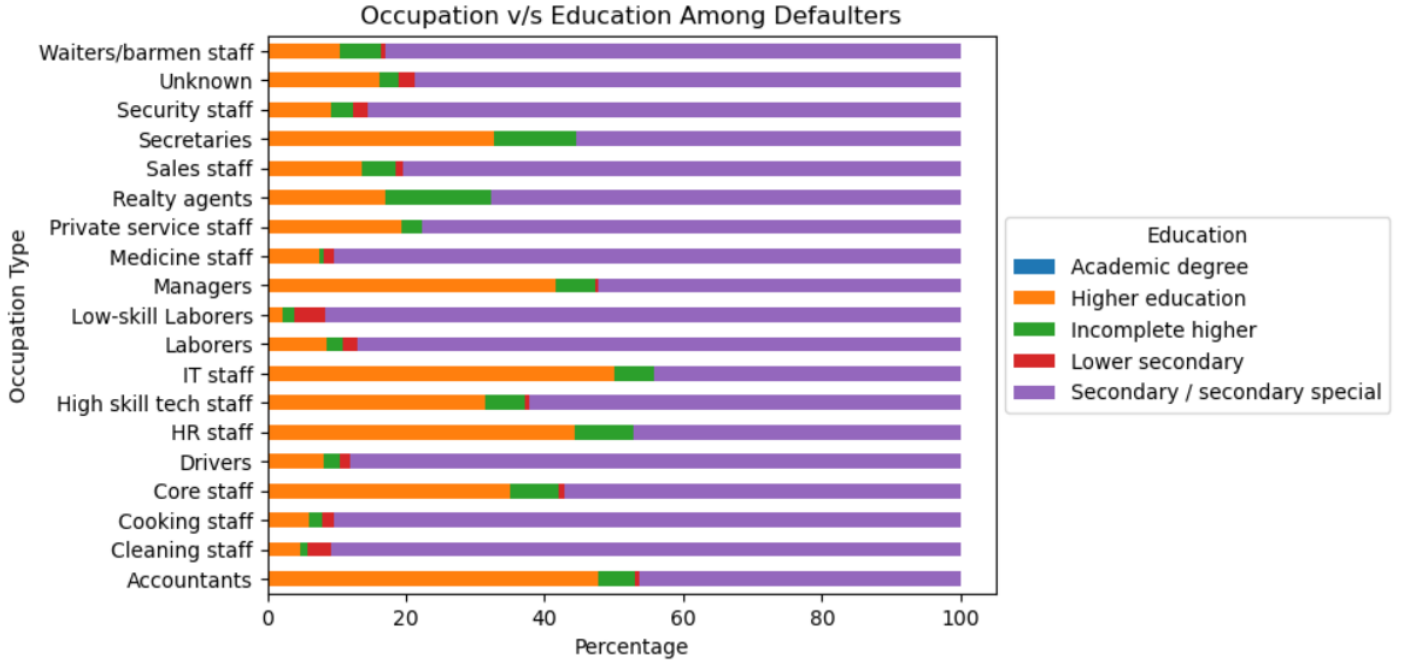




Targeted Analysis on People With Payment Difficulties (Defaulters) - Contd.

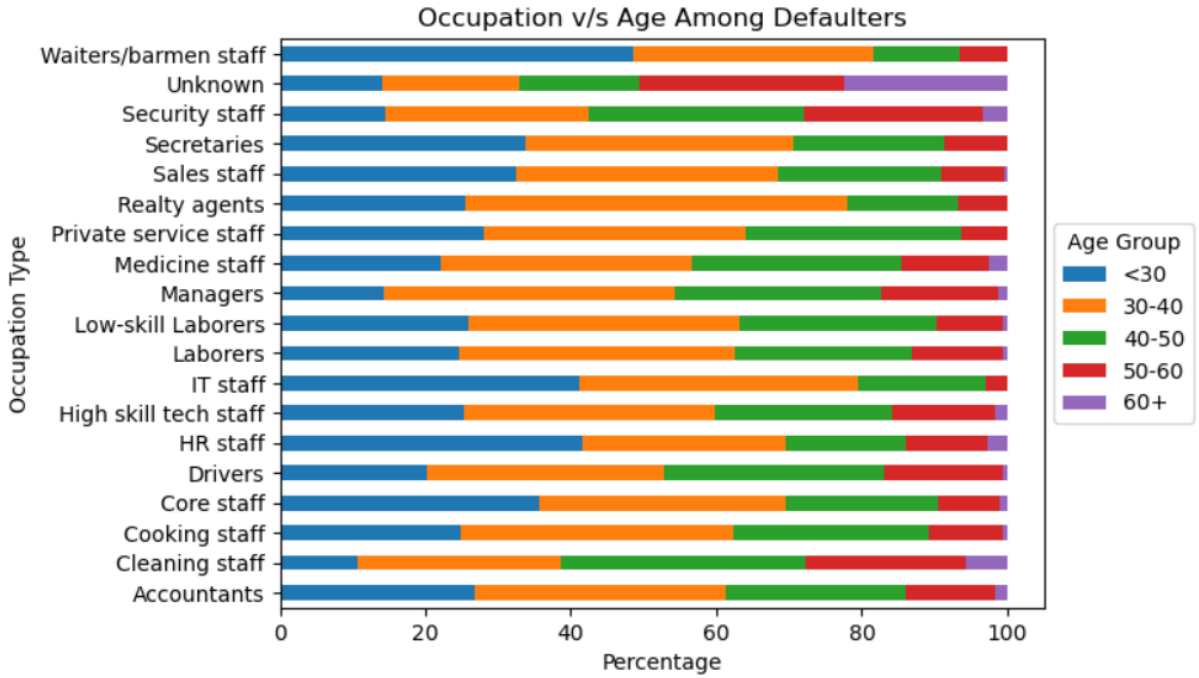


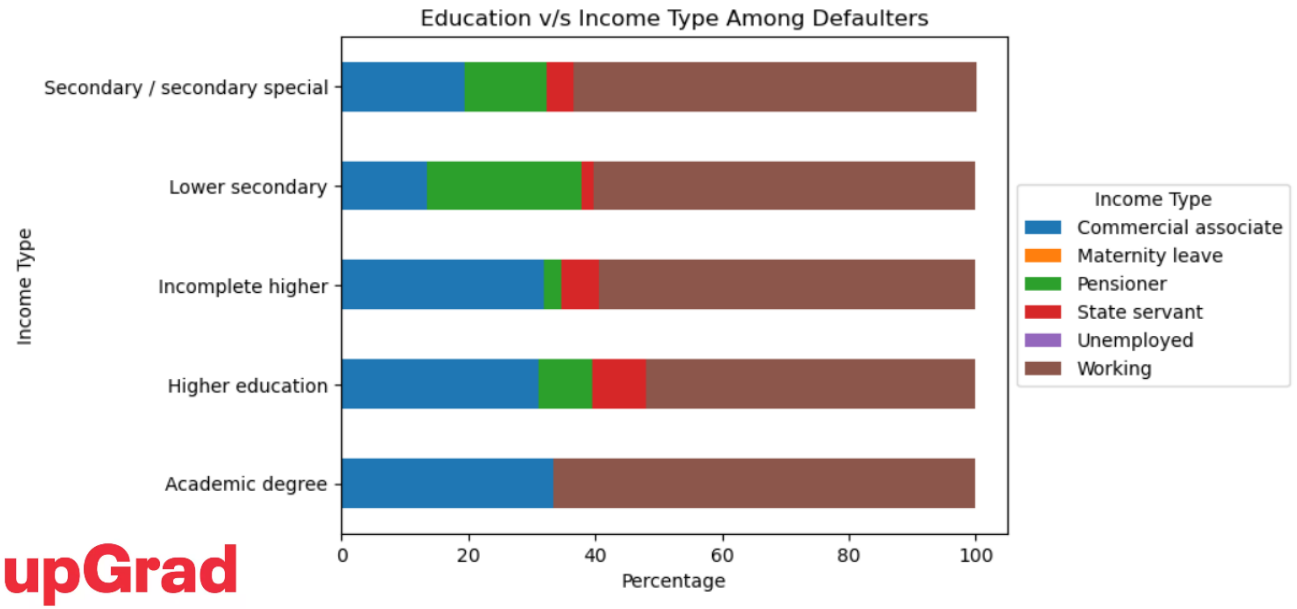




upGrad

Targeted Analysis on People With Payment Difficulties (Defaulters) - Contd.

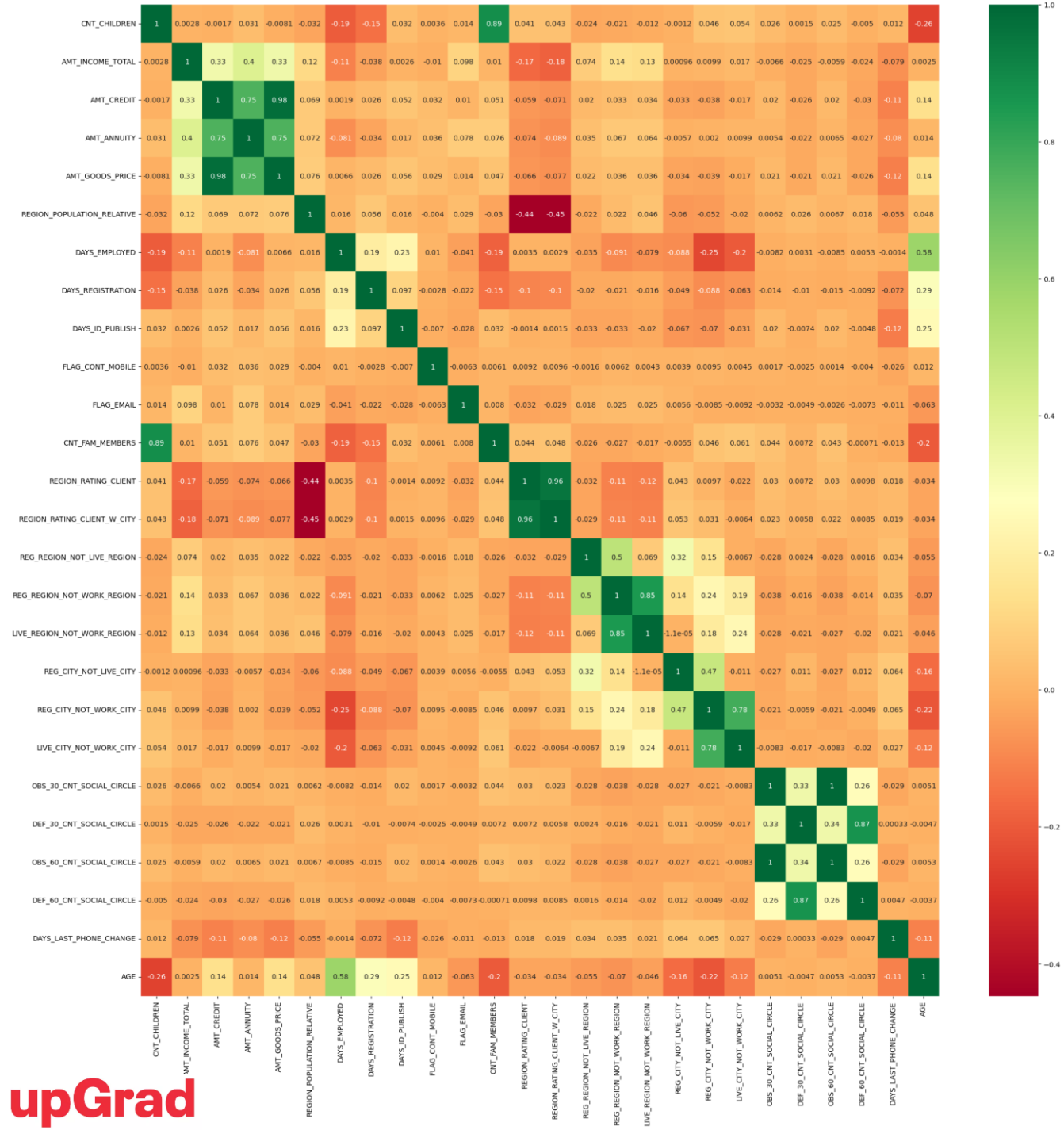




Key Insights - Patterns Among the Defaulters:

1. Dominant Age Group - 31-40.
2. Dominant Income-Age Combination - Age group 30-40 & 1 lakh-3 lakh income bracket.
3. Dominant Age-Education Combination - Age group 50-60 with Secondary / secondary special education.
4. Dominant Income Group-Education Combination - Income <50k having Secondary / secondary special education.
5. Dominant Income Group-Income Type Combination - Working professionals with income range 50k-100k.
6. Dominant Education-Income Type Combination - Working professionals with academic degree.
7. Dominant Income Type-City Combination - Working professionals from Tier-3 cities.
8. Dominant Income Group-Occupation Type Combination - Drivers with income in the range 1 lakh-3 lakh.
9. Dominant Education-Occupation Type Combination - Low-skill labourers with secondary education.
10. Dominant Age Group-Occupation Type Combination - Realty Agents in the 30-40 age group.

Applications from above combinations must be thoroughly scrutinized as these have the highest risk of loan default.

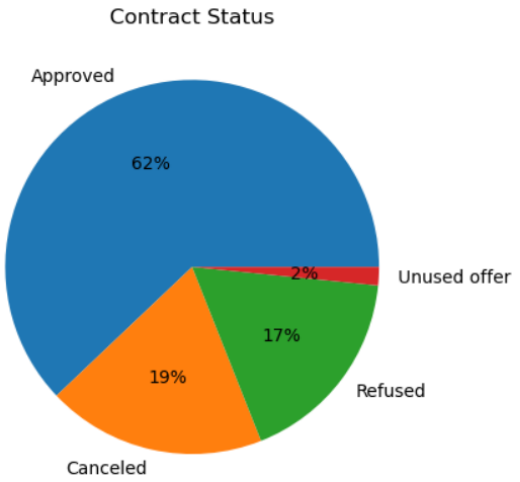


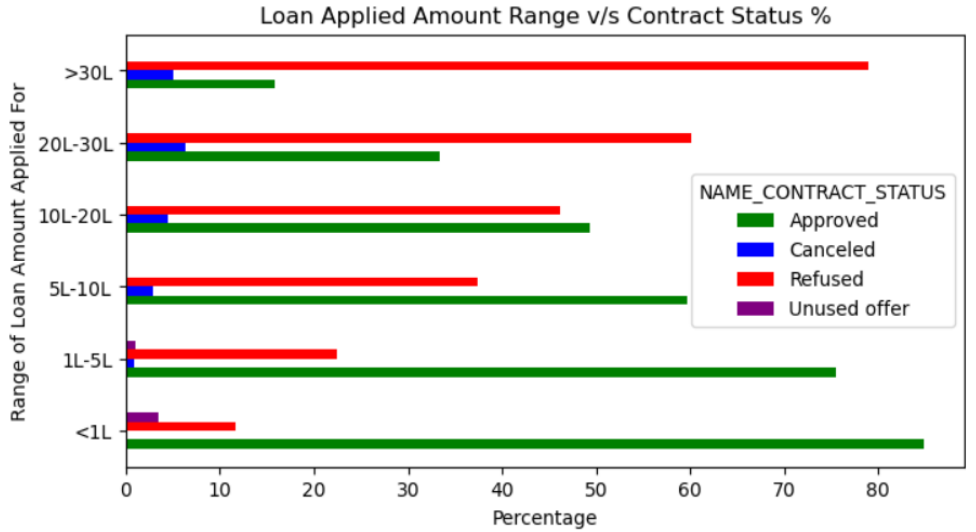
Correlation Heatmap for the Defaulters

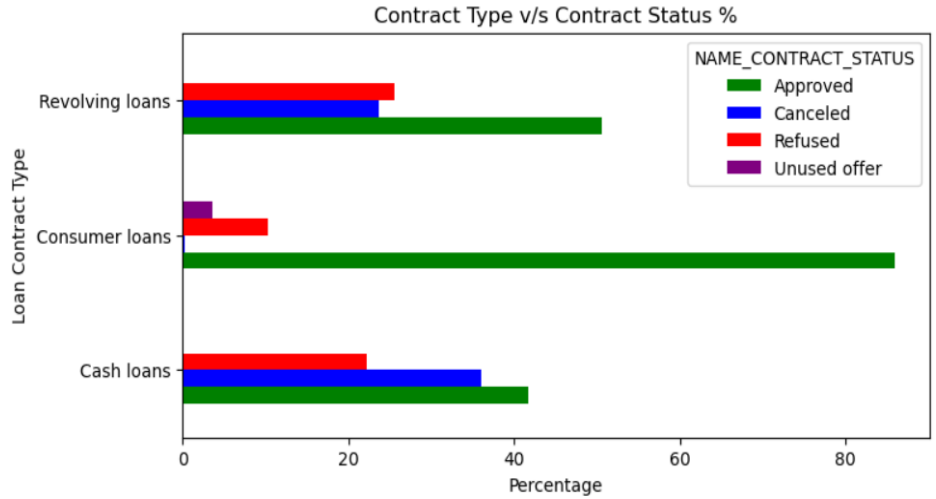
Top 10 Correlations:

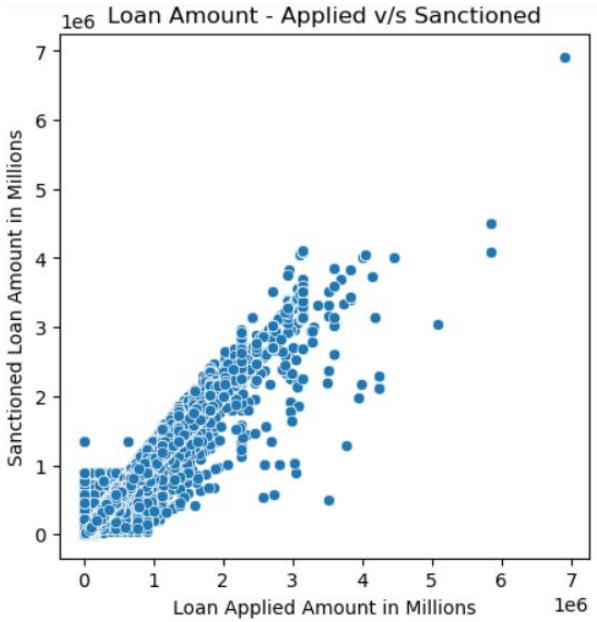
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998270
AMT_CREDIT	AMT_GOODS_PRICE	0.982783
REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.956637
CNT_FAM_MEMBERS	CNT_CHILDREN	0.885484
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.869016
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.847885
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.778540
AMT_GOODS_PRICE	AMT_ANNUITY	0.752295
AMT_CREDIT	AMT_ANNUITY	0.752195
AGE	DAYS_EMPLOYED	0.581769

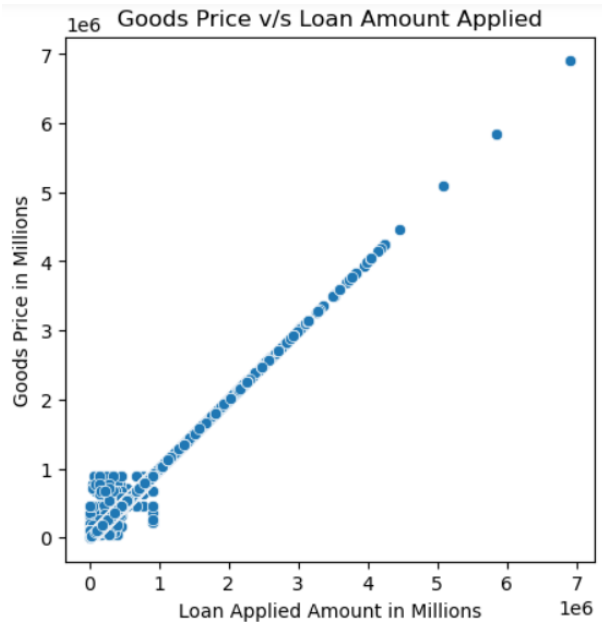
Insights From Previous Applications







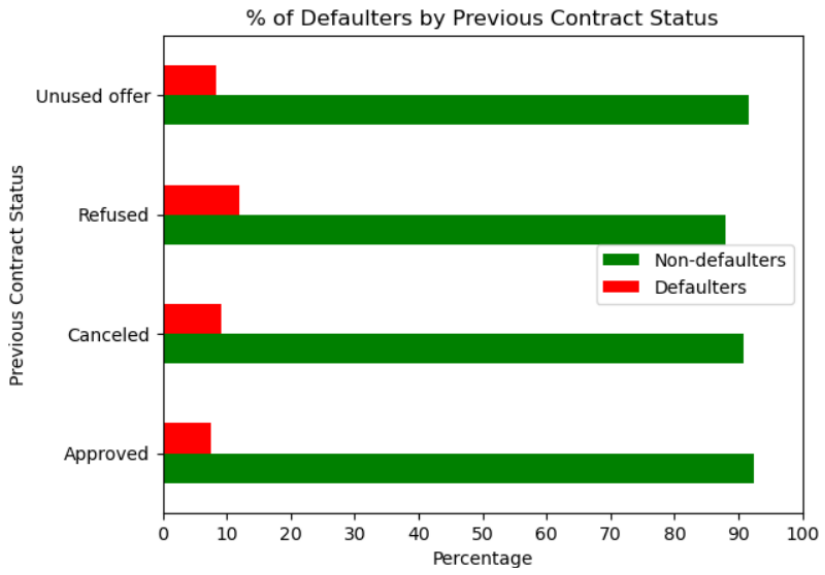


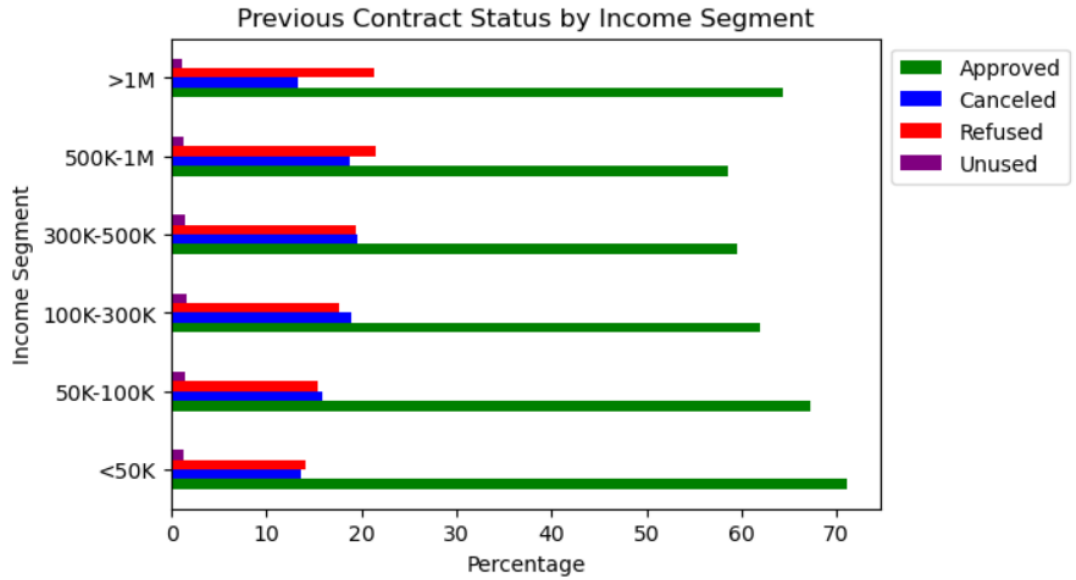


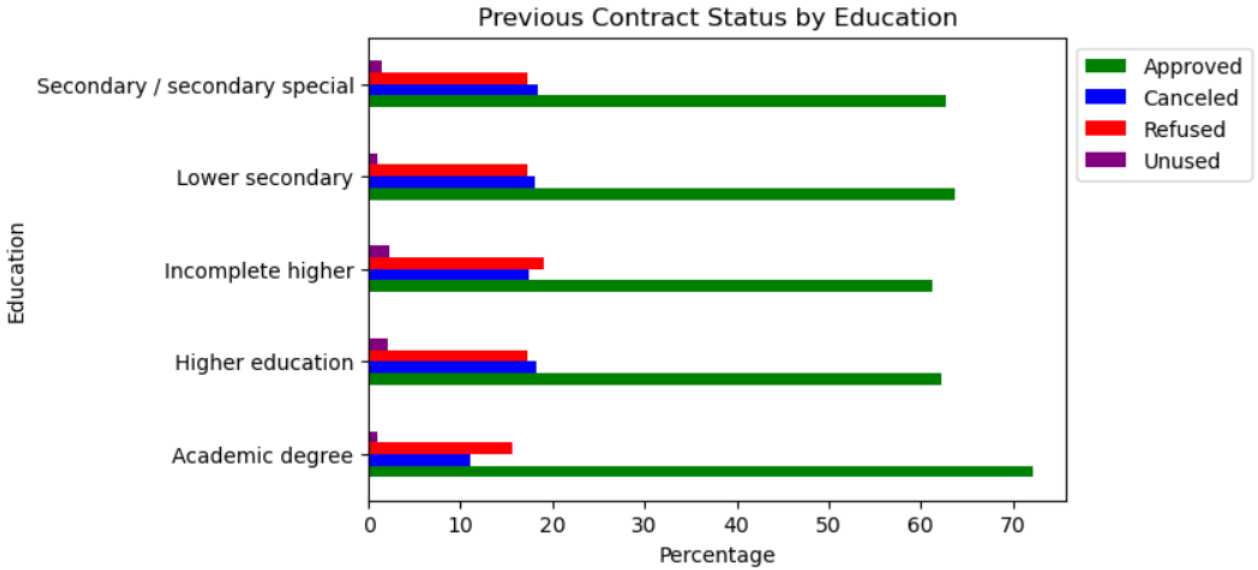
Key Insights:

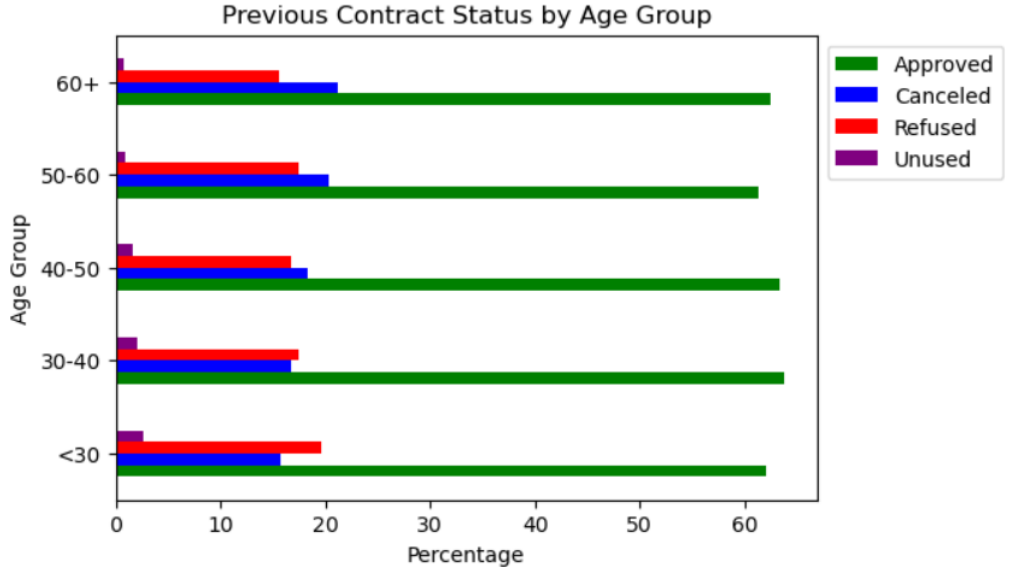
1. We can see a linear relationship between the goods price and loan amount applied for goods prices >1 million, i.e. This suggests that most big ticket items have been purchased fully on loan without any down payment.
2. Mostly linear relationship between applied and sanctioned loan amounts suggests that bank has disbursed the applied loan amount in most cases.
3. Loan applications for amounts less than 10 lakhs have mostly been approved, and we can observe more rejections for the higher loan amounts.
4. Consumer loan applications are mostly approved, whereas cash loans and revolving loans have a much lower approval percentage.

Insights from Combined Analysis of Current and Previous Applications

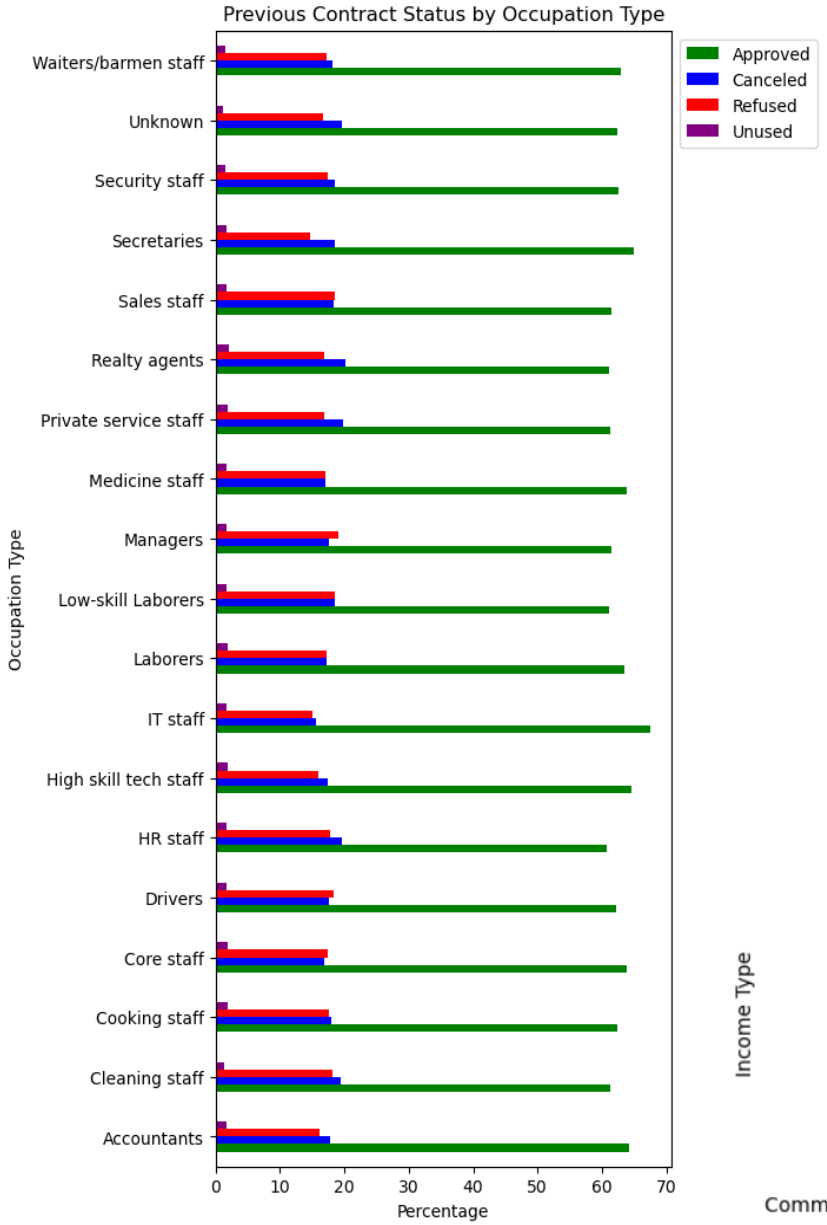






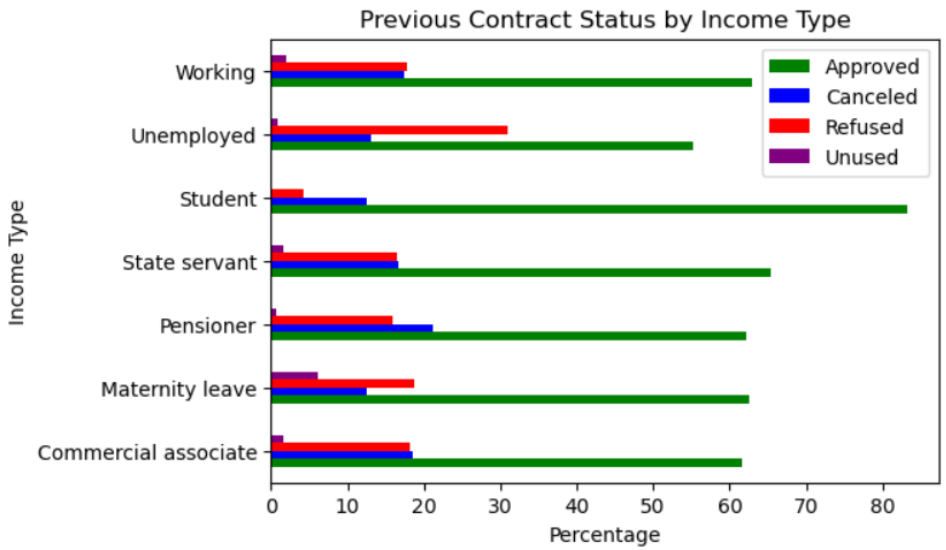


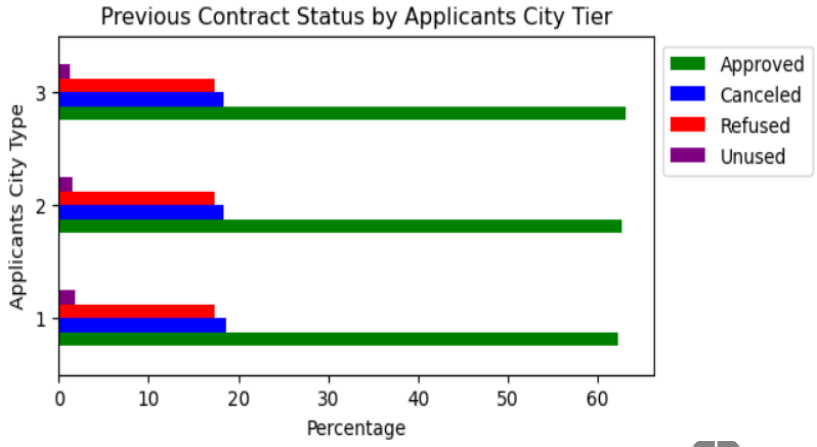
Insights from Combined Analysis of Current and Past Applications - Contd.



Key Insights:

1. 7.6% of the applicants, whose previous applications were approved, have defaulted. This is a **risk** for the bank.
2. 88% of the applicants, for whom the previous applications were refused/rejected, have not defaulted. This is a huge **revenue loss** for the bank.
3. Age does not seem to have been a criteria for accepting or rejecting loans previously, as the acceptance percentage is mostly uniform across all age groups.
4. Applicants' city does not seem to have been a criteria for accepting or rejecting loans previously, as the acceptance percentage is mostly uniform across all city types.
5. Loan applications from applicants with academic degrees have high acceptance rate.
6. Loan applications from students (possibly Education Loans) have high acceptance rate and those from unemployed applicants have the highest rejection rate.
7. IT Staff are less likely to default, and their applications have the highest acceptance percentage.





Conclusions

1. Possibility of loan default decreases with increasing applicant age.
2. Possibility of loan default is low with the higher income groups, i.e. people with income greater than 3 lakh INR.
3. Chances of loan default are low with the well-educated applicants.
4. Applicants doing Blue-collar jobs like labourers, drivers, waiters, cleaning staff, etc are more likely to default.
5. Unemployed applicants have higher chances of default.
6. Applicants from Tier-3 cities are more likely to default.
7. Applicants living with parents, and the ones living in rented apartments are more likely to default.
8. Female applicants have lower chances of defaulting on the loan than male applicants.

Recommendations

1. Give more importance to applications from higher age groups, and ensure that the win rate is high for these groups.
2. Invite more loan applications from high-income groups by providing attractive interest rates.
3. Invite more loan applications from well-educated applicants by providing attractive interest rates. Thoroughly scrutinize applicants who possess only secondary / secondary special education.
4. Focus more on applications from state servants, pensioners, businessmen. Ensure high win rate for these applicants by offering loans at competitive interest rates.
5. Thoroughly scrutinize loan applications from blue-collar workers and ensure that they have repayment capability before accepting. If required insist for collateral mortgage to offset risk.
6. Do not prefer applications from unemployed applicants, particularly for unsecured loans. Secured loans against a collateral may possibly be offered as the risk gets lowered.
7. Give more preference to applicants from Tier-1 cities by offering attractive interest rates. For applicants from Tier-3 cities, the interest rate can probably be hiked to offset the risk of default.
8. Even though it has been found in this analysis that male applicants are more likely to default than females, gender has deliberately not been projected as a key driver variable as it is not practically possible to implement. Because, in this day and age, no bank or any corporate entity would want to deliberately introduce gender bias into their customer assessment process. It may lead to bad press and possible legal hassles.

Thank You!