Bank Credit EDA

Analysis of driver variables of payment difficulty

Nguyen Bao Khang

Problem statement

Context

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.

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.

Problem statement

Business Objective

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.

Analysis approach

Target 1 rate/Payment difficulty rate

Payment difficulty rate will be used as the primary metrics for measuring driving factors (or driver variables) for the chance of having payment difficulty.

$$Payment\ difficulty\ rate\ of\ a\ subset = \frac{Amount\ of\ target\ 1\ in\ the\ subset}{Amount\ of\ total\ sample\ in\ subset}$$

Note: the analysis below will often refer to payment difficulty rate as PD rate for short.

Assumption about the data

- The data contains the complete record of the bank in a recent timeframe.

- If this is just a subset of the complete data, it should be a stratified random sample that can reasonably represent the population (ie no particular segment or part of which was intentionally left out).

- While not mentioned by the problem statement, based on the information in the data, we can assume that the bank is based in India. The income data fits Indian annual median income in rupee (source: Statista Research Department) and Education variable describes Indian education system.

Data preprocessing

Summary of data preprocessing

Missing data

- Missing entries were investigated and imputed with appropriate values.
- Columns with too many missing value were removed

Outliers

- Investigate if the extreme values are valid or incorrect, and handle accordingly.
- Create bins for highly skewed numerical data

Data standardization

- Convert data to appropriate unit where applicable (age from days to years, etc)
- Reduce redundant columns

Summary of data preprocessing

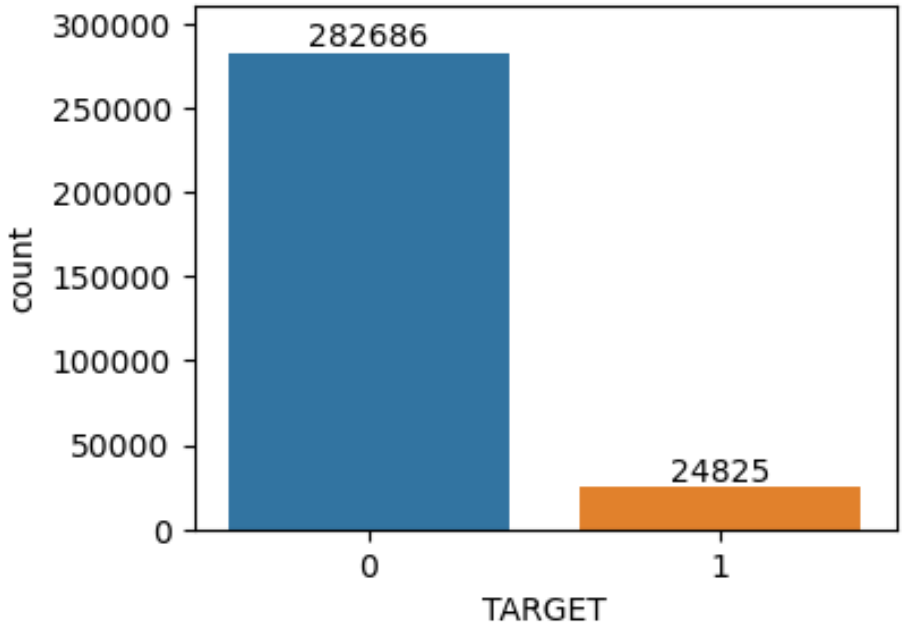
New data were created if applicable:

- REG_WORK_LIVE_CITY: information about clients' location regarding to reg/work/live address
- Data about previous application, regarding the status of the application and their amount

Univariate analysis

Target variable

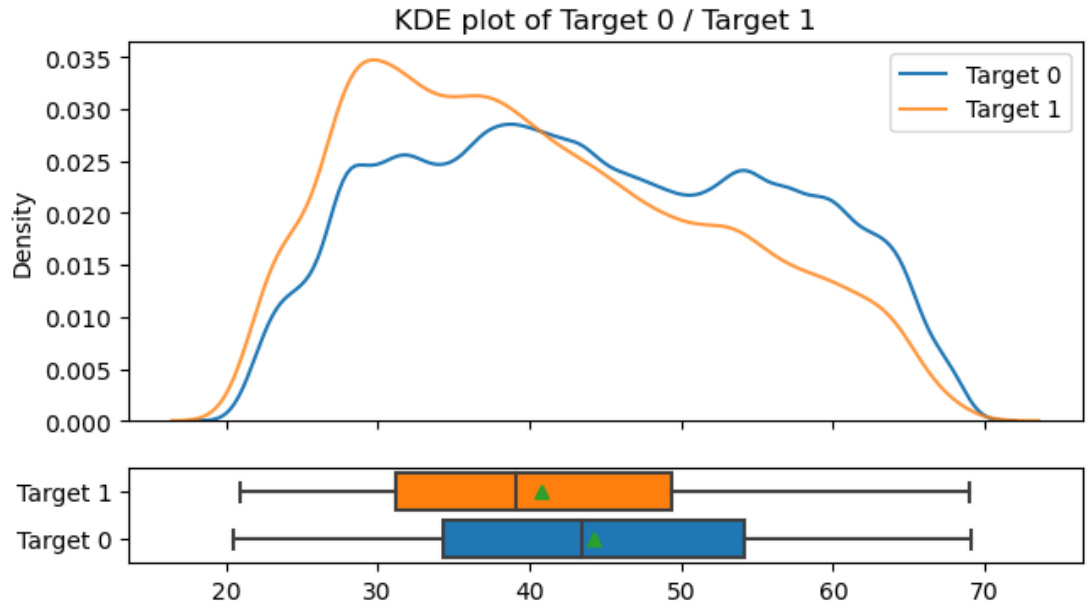
- The dataset has an average **payment difficulty rate of 8.07%**.
- Since the target variable is imbalanced, we will use this rate as benchmark for analyzing segmented / categorical variables
- A segment (for numerical features) or a categorical class (for categorical features) has higher chance of payment difficulty than average if its rate is significantly higher than 8.07% and vice versa.

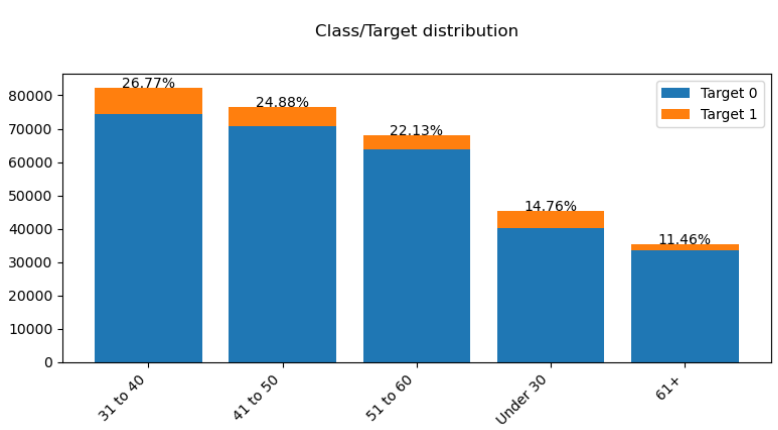


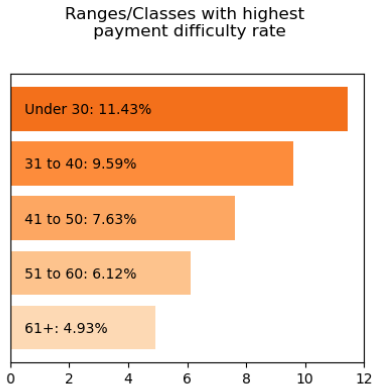
AGE

- Age seems to have a good correlation to payment difficulty rate, as we can see that the age groups with the highest rate are Under 30% and gradually decrease as the age goes up.
- Under 30 and 31-40 has significantly higher chance of payment difficulty (11.4% and 9.6%) than average (8.07%), while the rest have lower payment difficulty rate. 61+'s PD rate is the lowest, at 4.93%.
- Nearly 1/3 of clients with PD the bank have had are from 31-40 group - the group with 2nd highest risk. The bank may need to apply stricter requirement in the future on this age group to limit the clients with PD from said group.

AGE

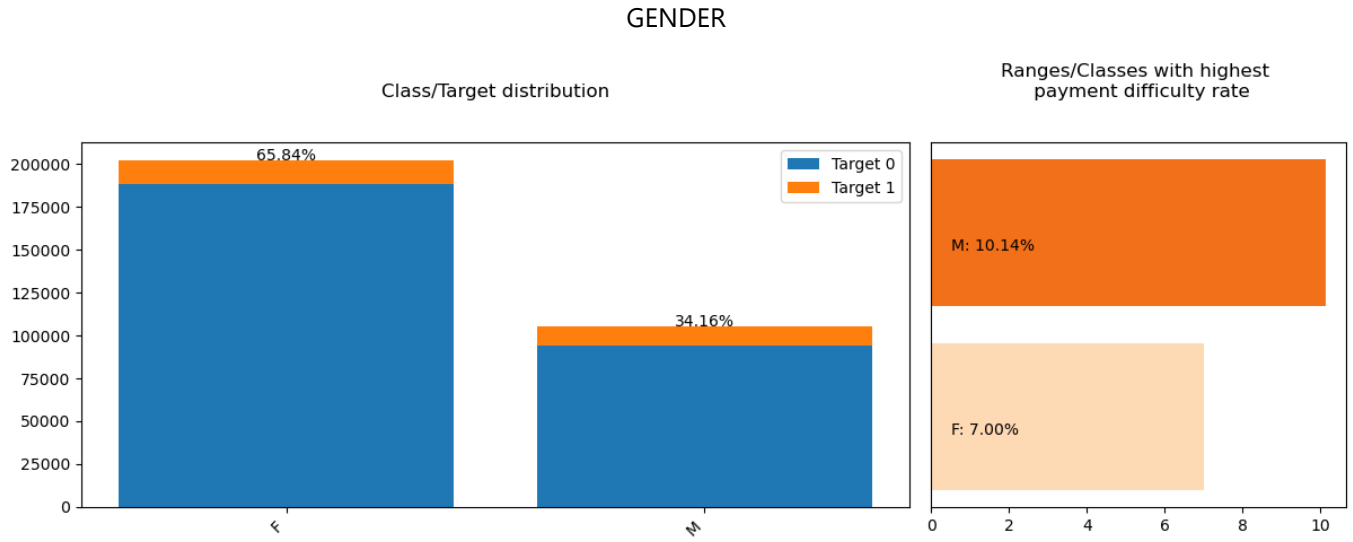






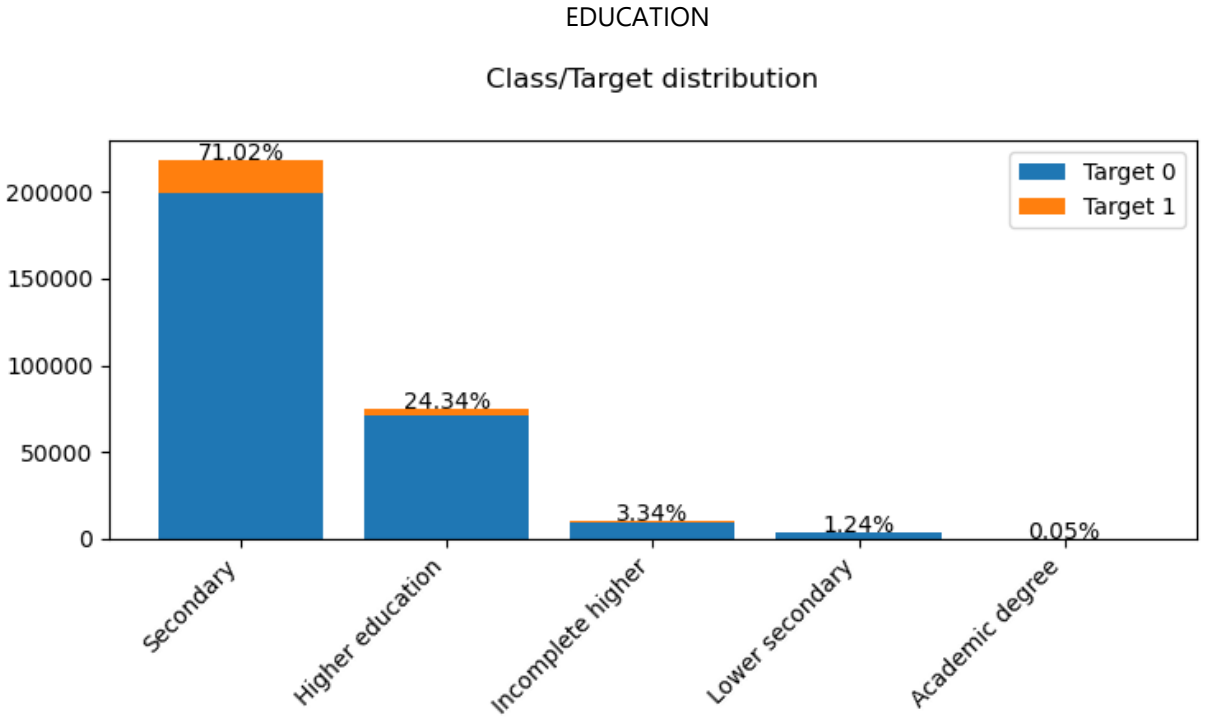
GENDER

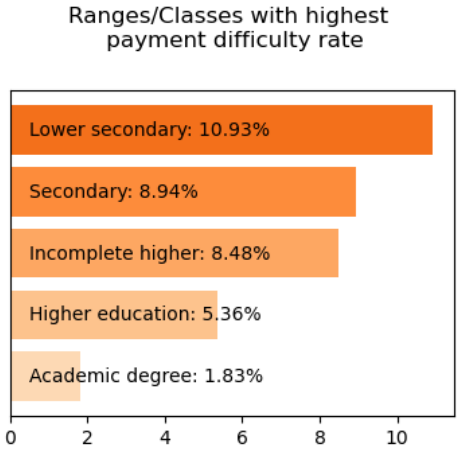
- Male has significantly higher chance of payment difficulty than Female. More than 2/3 of applicants are female.



EDUCATION

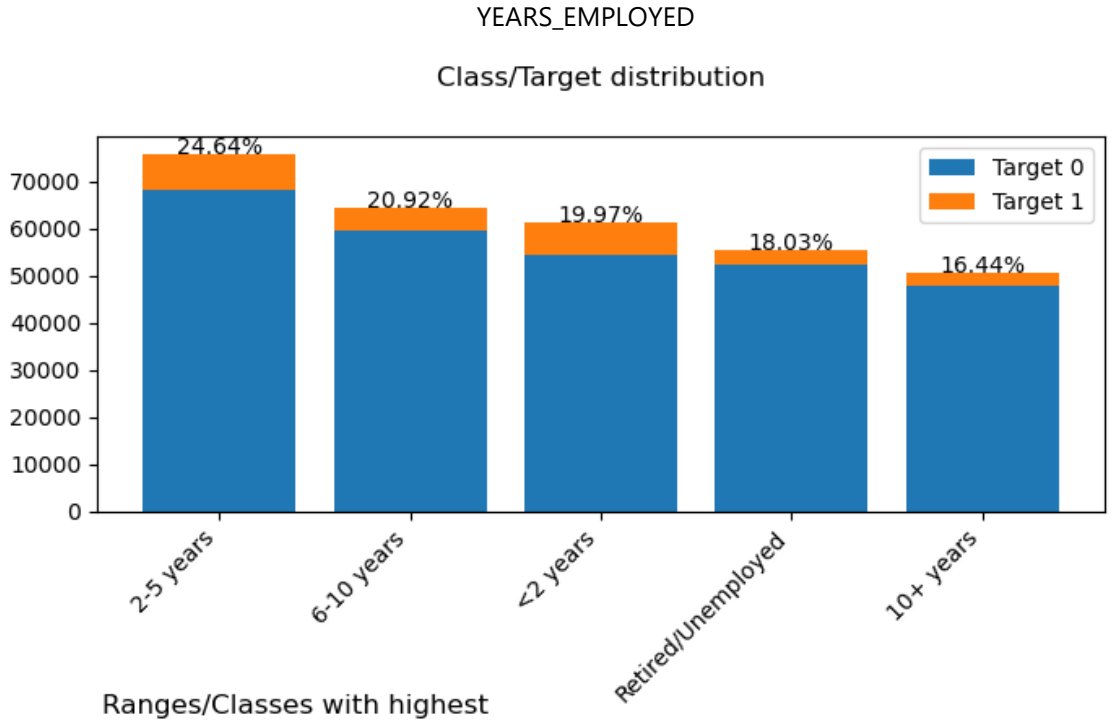
- Lower secondary and Secondary customer has the highest PD rate, while higher education and academic degree customer has the lowest risk.
- More than 70% of customers are Secondary/Lower Secondary, 27% are Higher/Incomplete higher and only a tiny fraction has an academic degree.

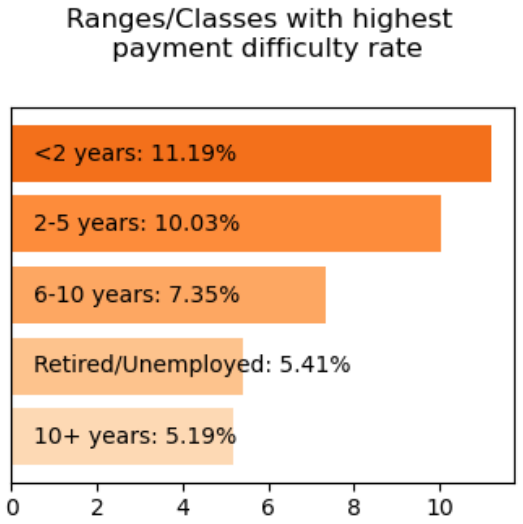




YEARS_EMPLOYED

- Customers with less than 5 years of employment have significantly higher risk of payment difficulty than average, while customers with 10+ years or retired has the lowest payment difficulty rate.
- 58% of clients with PD the bank has had are from under 5 years of employment - the groups with highest risk. The bank may need to apply stricter requirement in the future on this group to limit the clients with PD from said group.



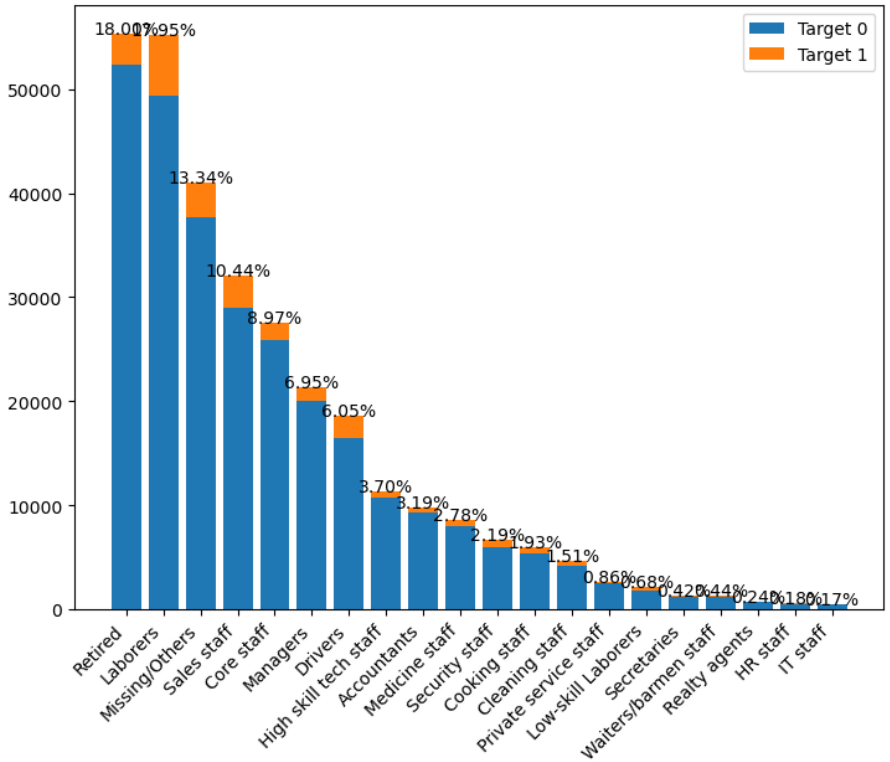


OCCUPATION_TYPE

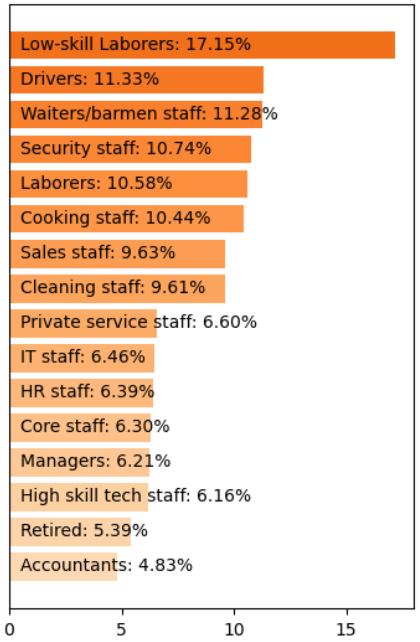
- Low skill labor, driver, waiter, security staff, cleaning staff (ie blue collar, heavy labor jobs) has the highest PD rate
- Accountant/high skill tech, manager,... have lowest rate.

OCCUPATION_TYPE

Class/Target distribution

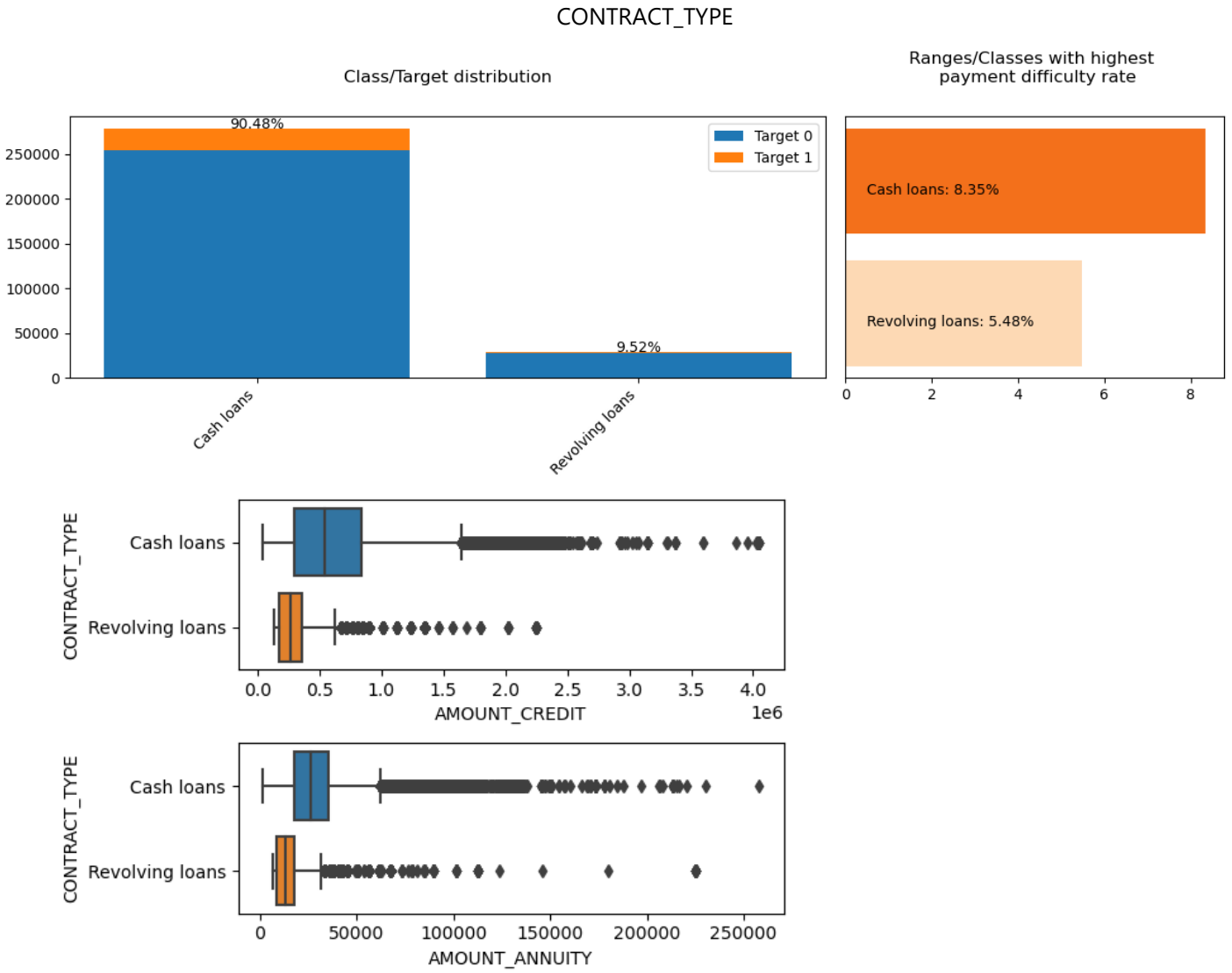


Ranges/Classes with highest payment difficulty rate (Top 8 and bottom 8 only).



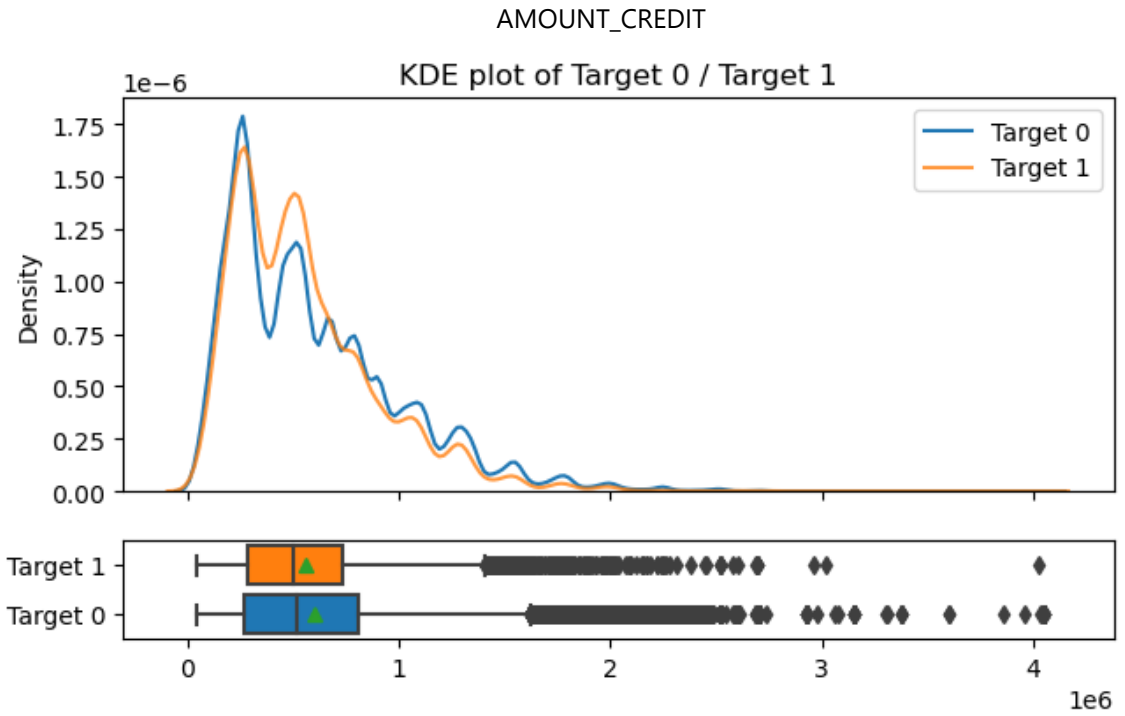
CONTRACT_TYPE

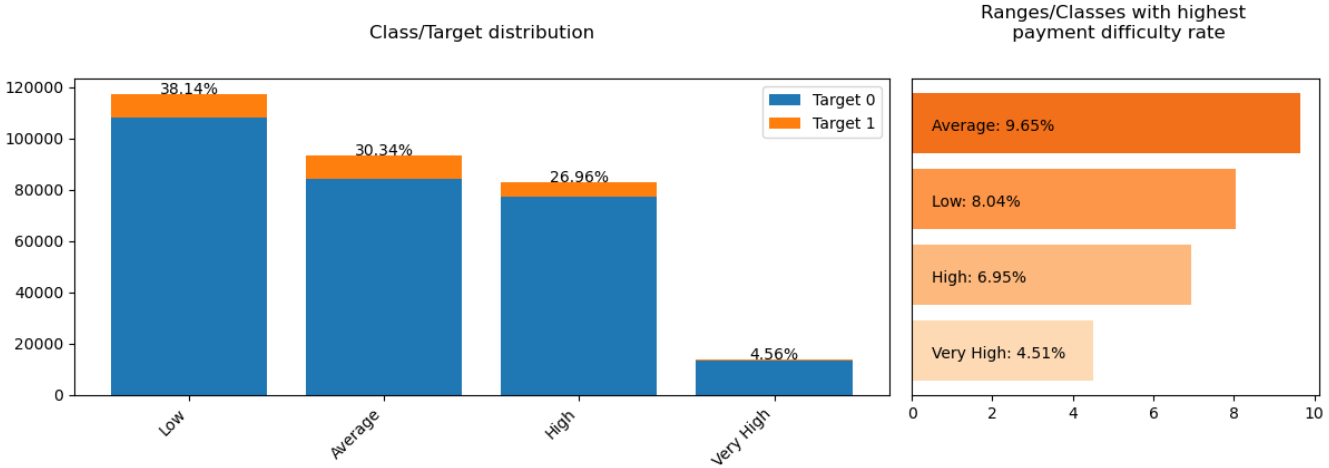
- Cash loan customers has higher chance of payment difficulty than Revolving loan, due to the fact that Cash loan tends to have much higher annuity amount.
- About 90% of contracts are Cash loan and 10% are Revolving loan.



AMOUNT_CREDIT

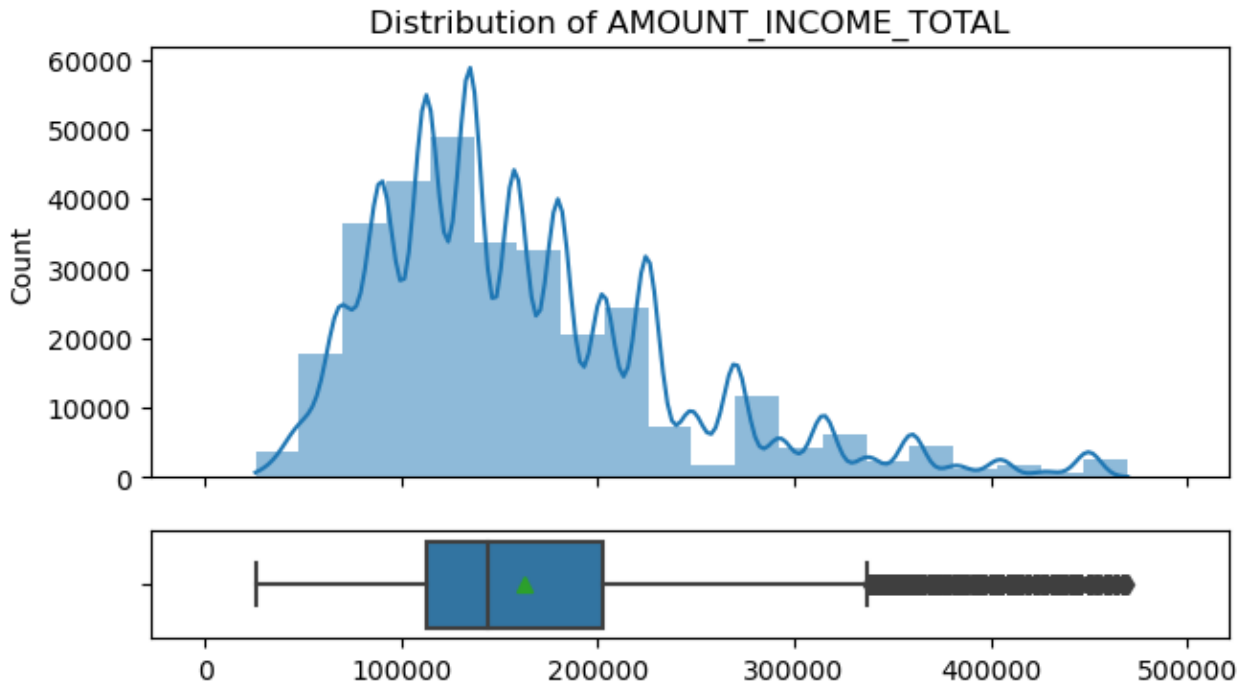
- The 'Average' range has higher than average payment difficulty rate, while the high and very high range has significantly lower payment difficulty rate.
- There seems to be correlation between the AMOUNT_CREDIT and payment difficulty rate, where the customers who was approved low/average amount have higher chance of payment difficulty than those with high/very high amount

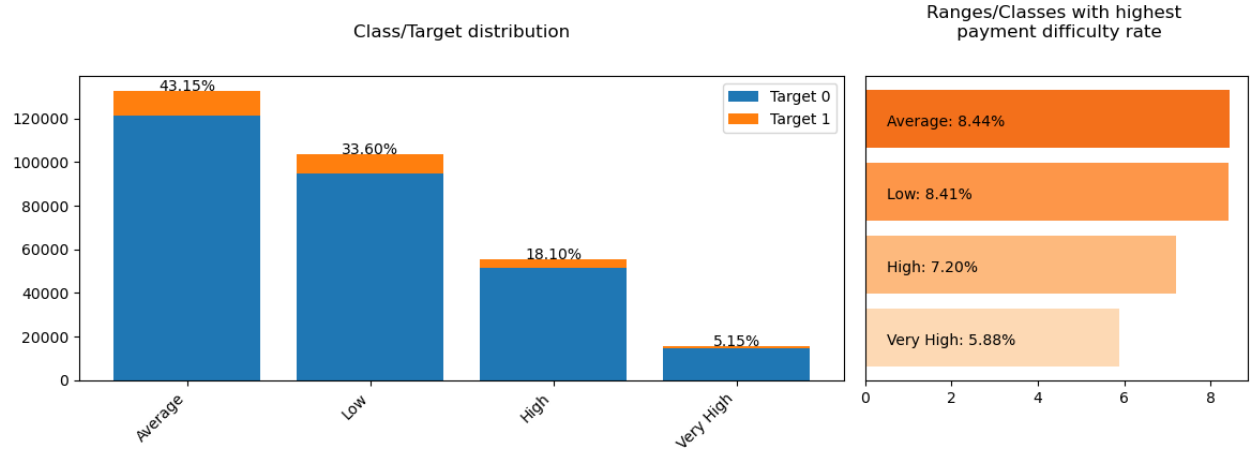




AMOUNT_INCOME_TOTAL

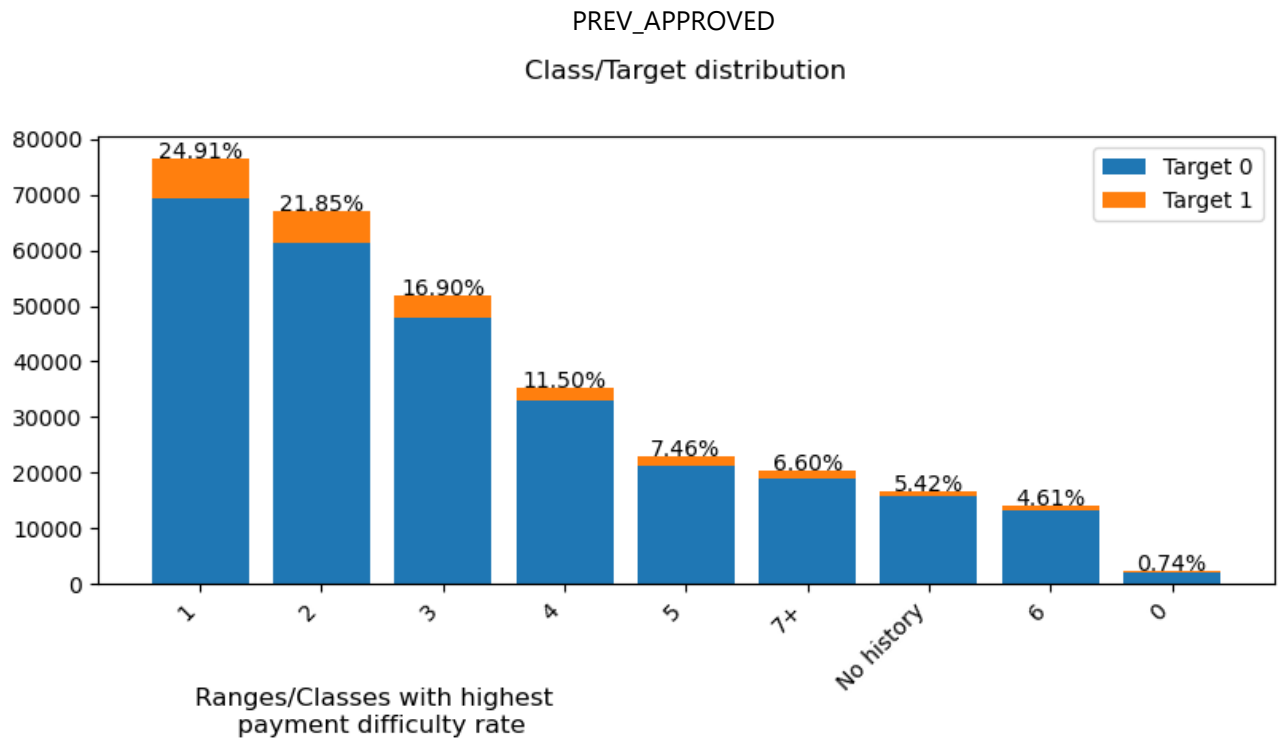
- AMOUNT_INCOME_TOTAL has a very right skewed distribution, with many extreme values in the high range.
- The low and mid ranges has higher than average PD rate, while the high and very high range has significantly lower PD rate.

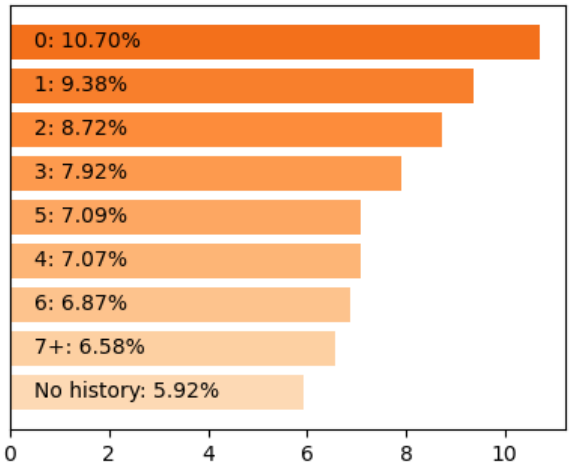




PREV_APPROVED

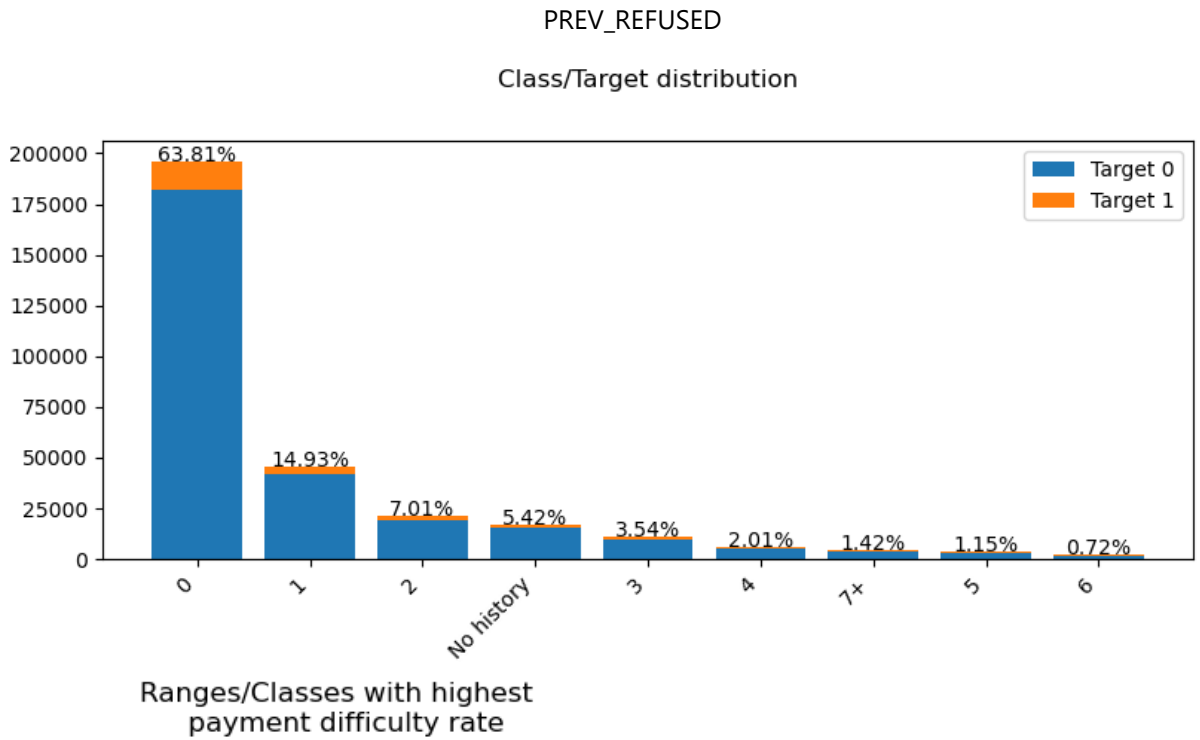
- The number of past approved application has a negative correlation to PD rate. But the group with the lowest rate are those with no history at all.
- In other words, while high number of previously approved applications can be a proven track record of how well a client can manage to repay, their chance of payment difficulty are still higher than who have never applied for a loan.
- Conversely, people with very few previous approved loans (ie did apply but got rejected) have significantly higher risk of payment difficulty.

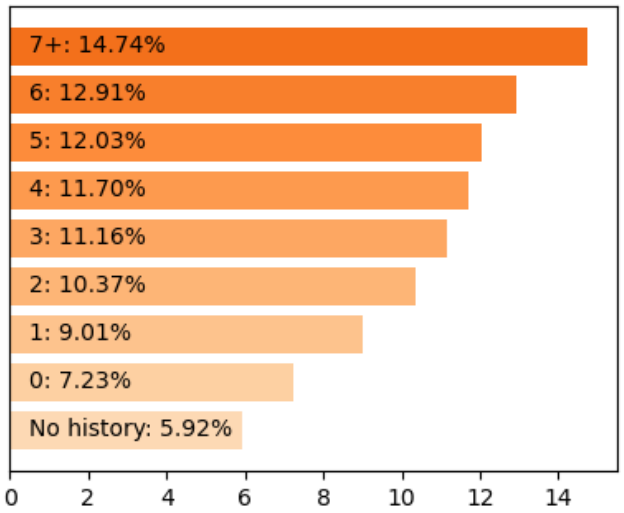




PREV_REFUSED

- In an opposite manner to PREV_APPROVED, the more times a client have been refused the loan, the higher their chance of payment difficulty are.
- Note that for clients who had even only 1 refused loan, their chance of payment difficulty is already higher than the average rate (9.01% vs 8.07%)

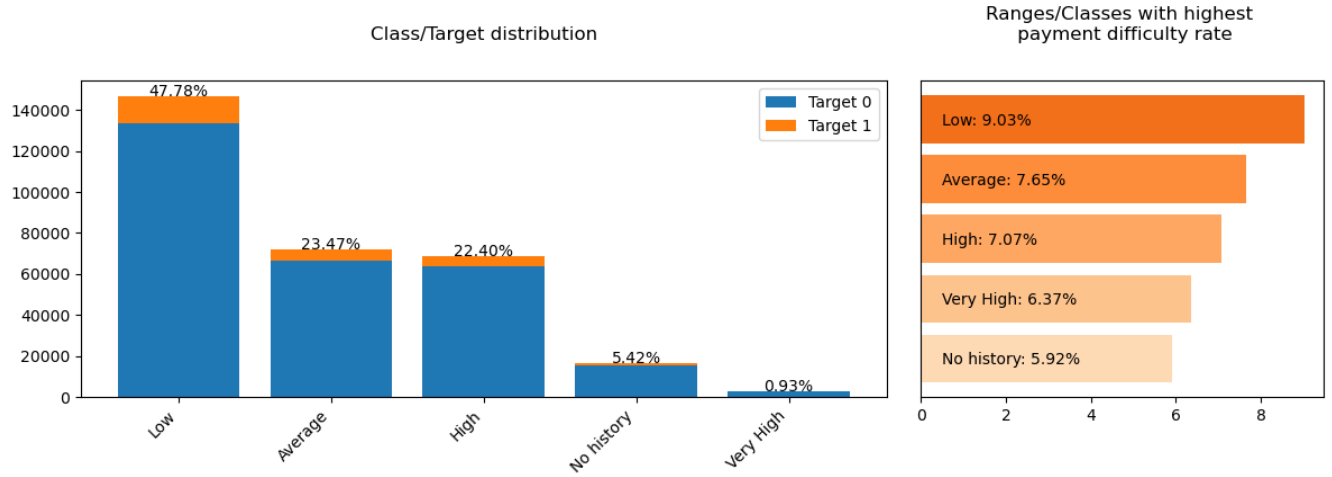




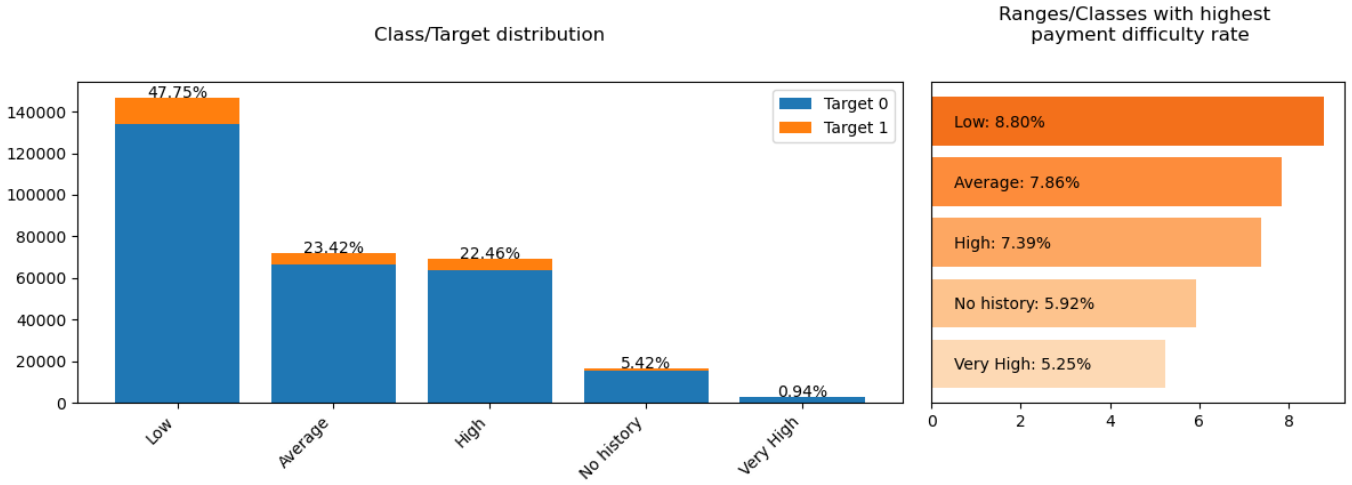
TOTAL_CREDIT_APPROVED

- Different from previous TOTAL/AVERAGE CREDIT_APPLIED, TOTAL/AVERAGE CREDIT_APPROVED are the sum amount of credit actually approved by the bank.
- The people with low total approved amount and low approved amount per application has the highest PD rate, while the ones with very high approved amount/average approve amount per application has lowest chance of PD.
- They seem to be better indicators for detecting clients with payment difficulty, as it reflects past evaluation from the banks.

TOTAL_CREDIT_APPROVED

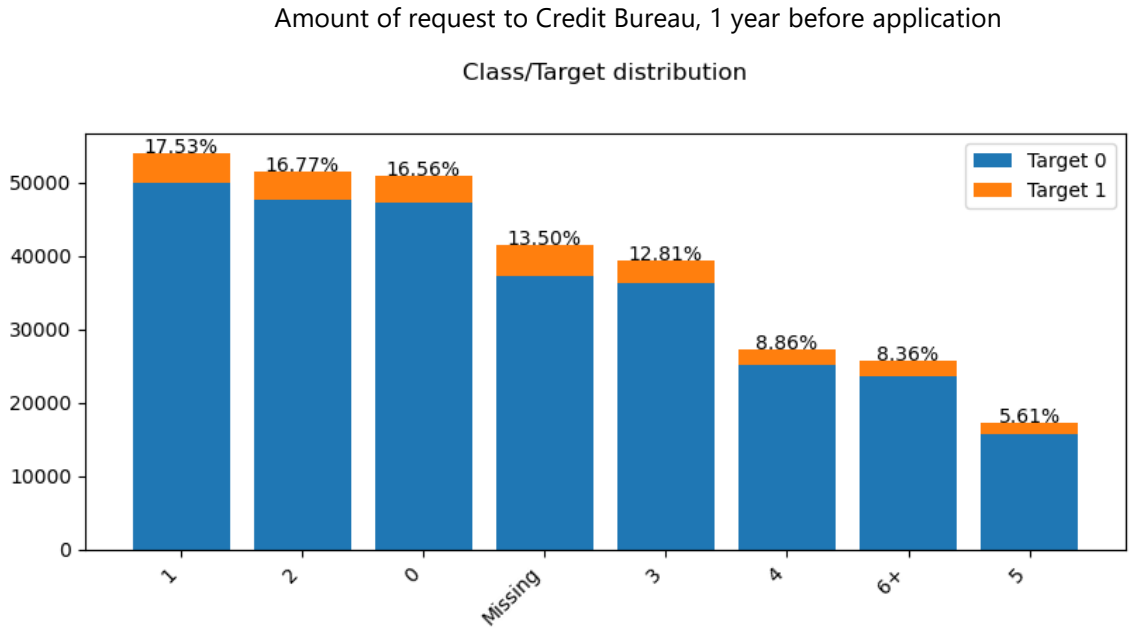


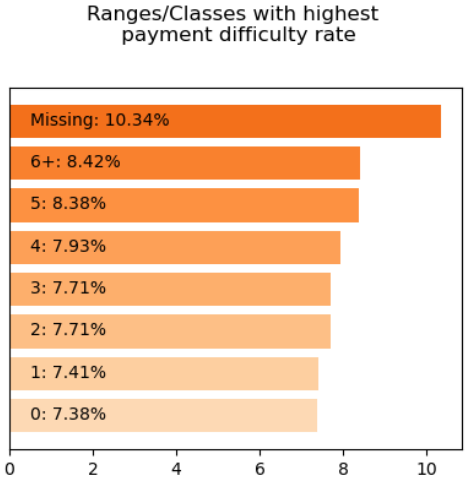
AVERAGE_CREDIT_APPROVED



AMOUNT_REQ_CB_YEAR

- Having many requests, meaning have applied for loans many time, can be a sign of financially struggling.
- People who had many requests in the year prior to the loan application have significantly higher payment difficulty rate than those with fewer request.
- The clients that have missing value for amount request have the highest PD rate (10.34%). The bank needs to investigate the source of this segment and apply stricter procedure/requirements for such clients.
- Overall, this seems to be a good indicator of payment difficulty chance. The number of requests has a strong correlation to PD rate.

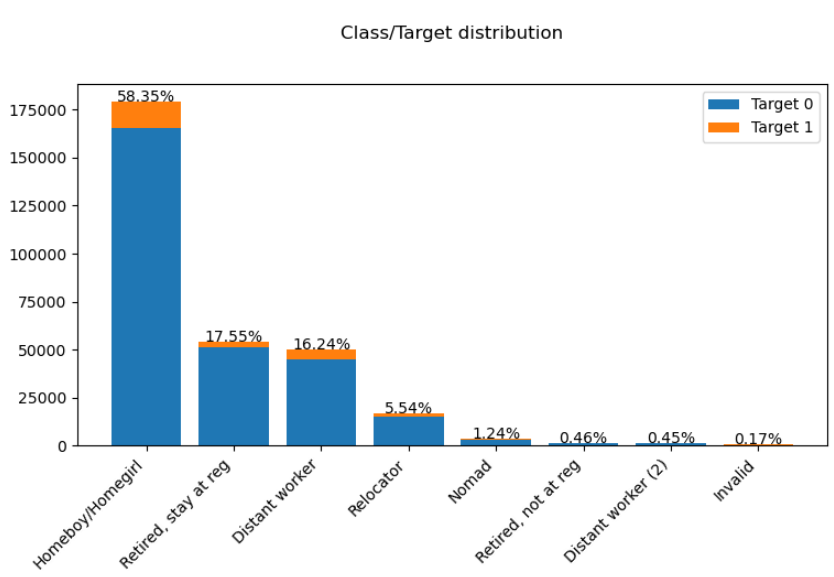


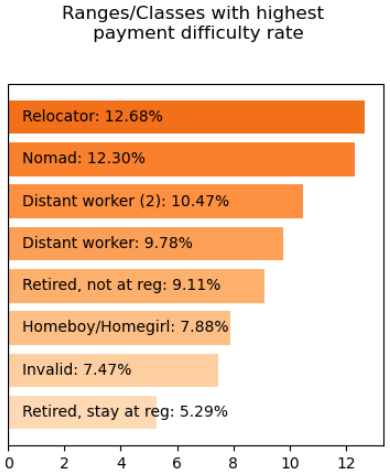


REG_WORK_LIVE_CITY

- We can observe that those who don't live at their registered city and/or associated with 2 cities make the top 5 highest PD rate:
  - Relocator: 12.68%
  - Nomad: 12.3%
  - Distant worker (2): 10.47%
  - Distant worker: 9.78%
  - Retired, not at reg: 9.11%
- Meanwhile, those who associated with only 1 city has the lowest PD rate:
  - Homeboy/Homegirl
  - Retired, at reg
- - Living in a different city than registered address most likely means having to pay for rent, which can take up a large portion of monthly salary. Not to mention the cost associated with moving, commuting, etc...

REG_WORK_LIVE_CITY





1. REG = LIVE = WORK: The **Homeboy/Homegirl**.
2. REG = LIVE ≠ WORK: The **Distant worker**
3. REG = WORK ≠ LIVE: The **Distant worker (2)**
4. REG ≠ LIVE = WORK: The **Relocator**
5. REG ≠ LIVE ≠ WORK: The **Nomad**
6. REG = LIVE (for Retiree): Retired, at reg
7. REG ≠ LIVE (for Retiree): Retired, not at reg
8. The rest: **Invalid**

Summary of univariate analysis

Numerical features that shows a clear relationship with payment difficulty rate:

1. **AGE**: younger age has higher PD rate and vice versa.
2. **YEARS_EMPLOYED**: low years of employment has higher PD
3. **AMOUNT_INCOME_TOTAL**: Low/average range has higher rate than high/very high range
4. **PREV_APPROVED**: More previous approved loan means lower PD rate. The lowest is no history.
5. **PREV_REFUSED**: Less previous refused loan means lower PD rate. The lowest is no history.
6. **TOTAL/AVERAGE_CREDIT_APPLIED**: Low total approved credit and low average approved credit per loan has higher rate.
7. **AMOUNT_CREDIT**: clients with low/average credit has higher PD rate than those in high/very high range.
8. **AMOUNT_ANNUITY**: average and high annuity has higher rate than either low or very high range.
9. **AMOUNT_REQ_CB_YEAR**: More requests in the year prior to the loan application have significantly higher payment difficulty rate than those with fewer request.
10. **CAR_AGE**: Client who owns new car (car age less than 5 years) has lower PD rate than those with old car (10+ years) or no car.

Summary of univariate analysis

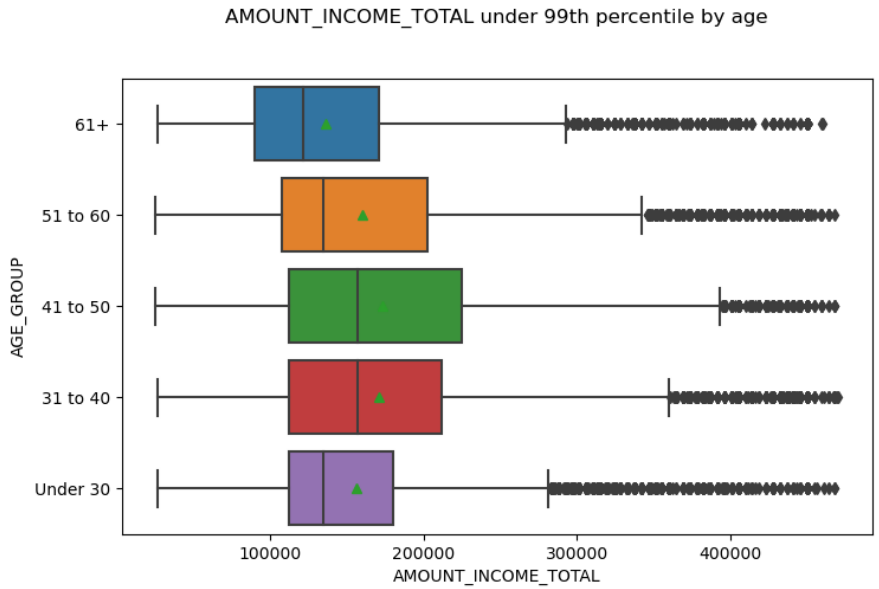
Insights and notable categorical features:

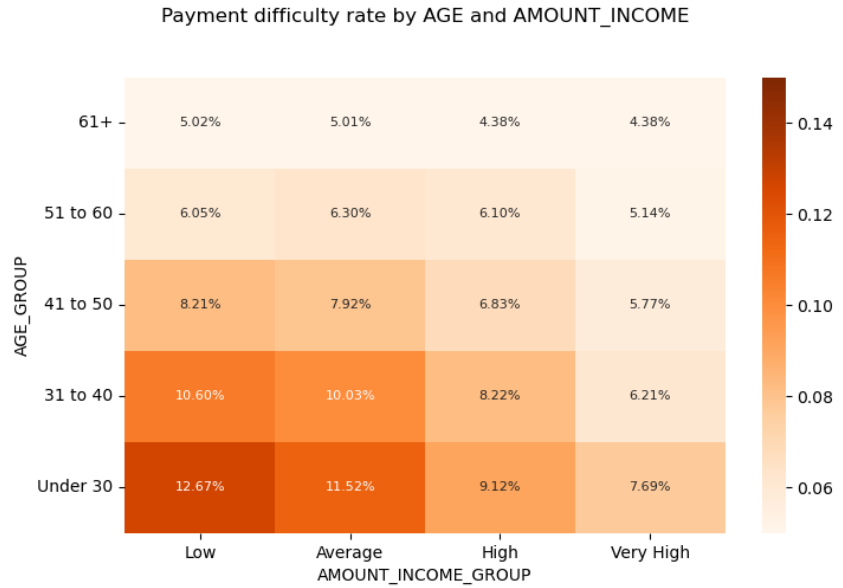
- **OCCUPATION_TYPE**: Heavy labor jobs has the highest PD rate while accountant, manager, high skill tech, ... have lowest rate.
- **EDUCATION**: the level of education also correlate to payment difficulty rate. Lower levels tends to have higher rate and vice versa.
- **COUNT_CHILDREN**: People who have children has higher PD rate than those who do not, and the rate increase with the number of children
- **COUNT_FAMILY_MEMBER/MARITAL_STATUS**: Single and people with big family have higher PD rate than married couples.
- **HOUSING_TYPE**: People who live with parents or rent has significantly highest PD rate than those live in other housing type.
- **GENDER**: Male has significantly higher chance of payment difficulty than Female.
- **REG_WORK_LIVE_CITY/CITY_COUNT**: People associated with 2 cities or more have significantly higher PD rate than those who live and work in the same city as their registered city.

Multivariate analysis

AMOUNT_INCOME_TOTAL x AGE

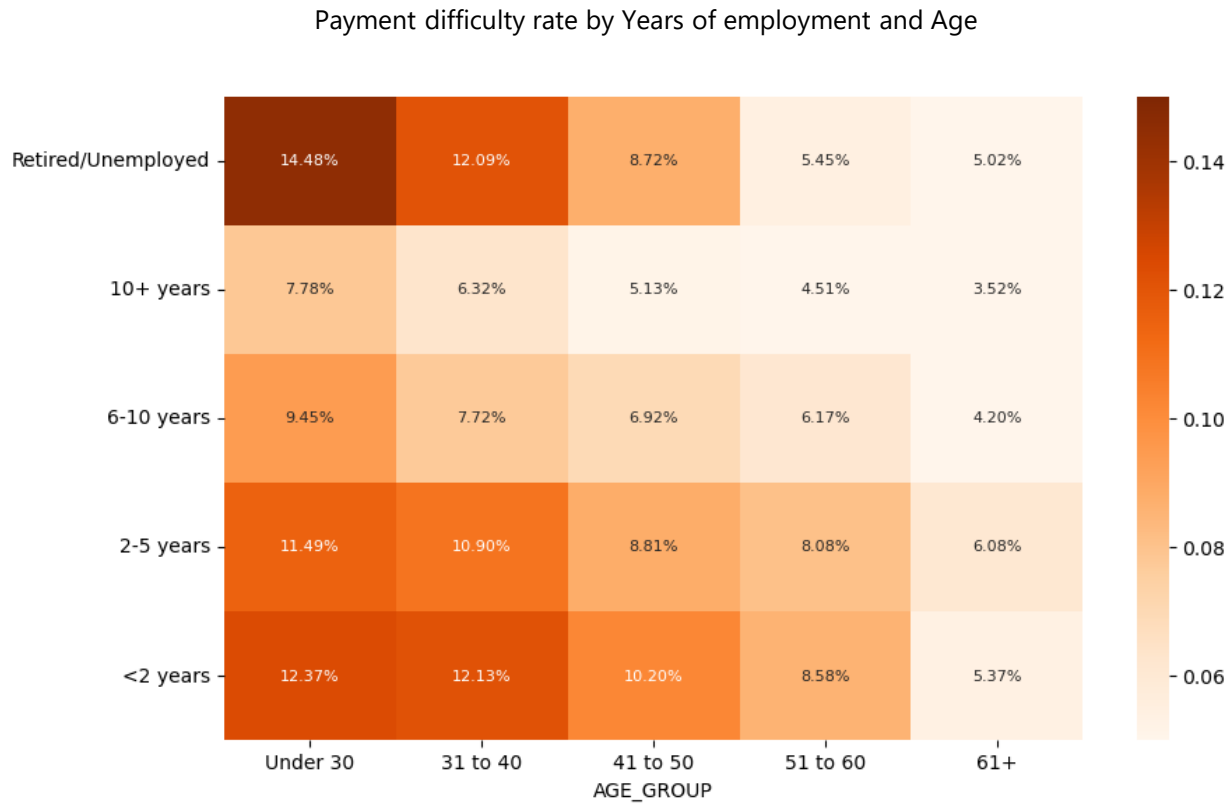
- AGE and AMOUNT_INCOME_TOTAL together has a correlation with payment difficulty rate.
- The older a client is and the higher their income is tend to lead to lower payment difficulty rate.
- Age group 41-50 has highest mean and 75th percentile of income. The reason can be that people between 41-50 are still working and have reached higher management positions. Older than that and people start to retire (at around 55~60), which leads to lower income than the former.





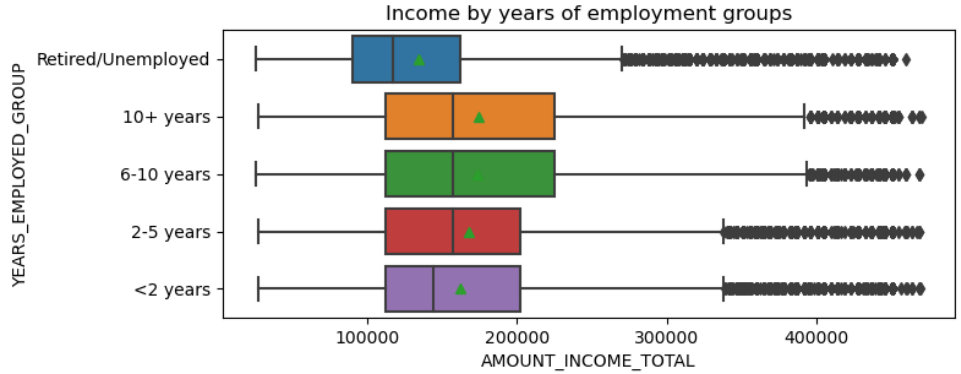
YEARS_EMPLOYED x AGE

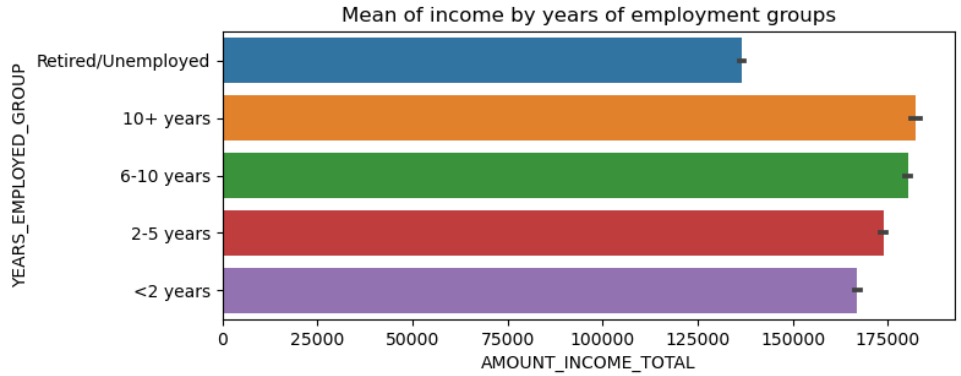
- YEARS_EMPLOYED and AGE together has a correlation with payment difficulty rate.
- The more experienced a client is and the higher their income is tend to lead to lower payment difficulty rate.

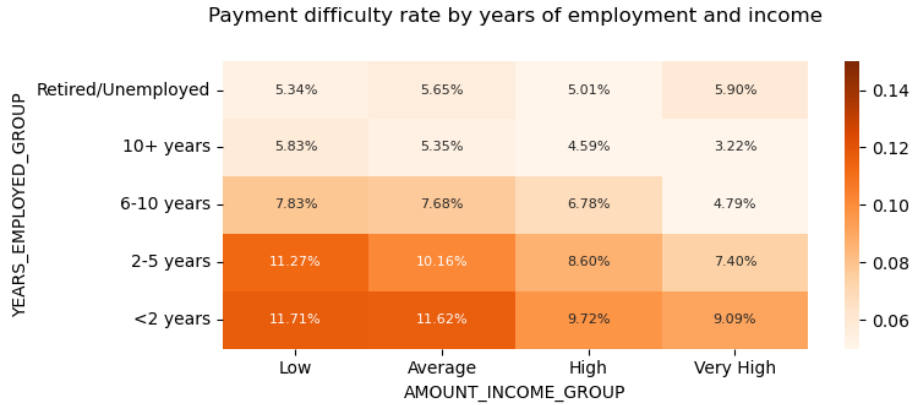


YEARS_EMPLOYED x AMOUNT_INCOME_TOTAL

- YEARS_EMPLOYED and AMOUNT_INCOME_TOTAL together has a correlation with payment difficulty rate.
- The more experienced a client is and the higher their income is tend to lead to lower payment difficulty rate, and vice versa.
- The mean income tends to increase as the years of employment increase, but the difference is not very noticeable.
- The 75th percentile income of 6-10 years and 10+ years of employment is also higher than that of under 5 years of employment.
- That being said, these 2 features have no correlation (Correlation coefficient: 0.032). This is because the actual income can be varied vastly between profession, organization, etc,...

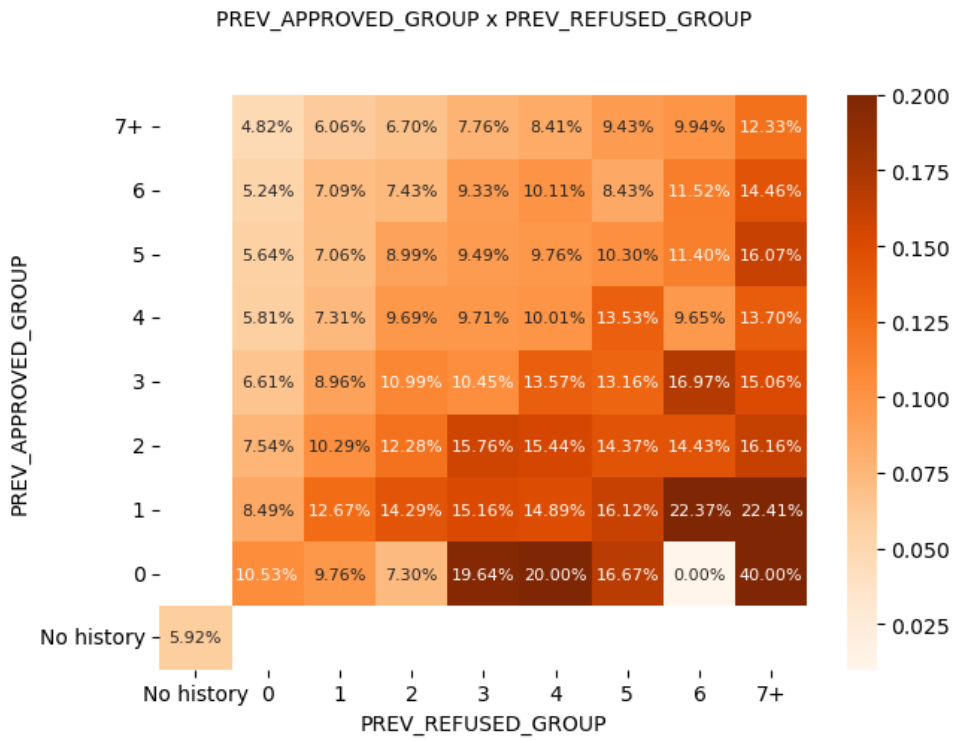


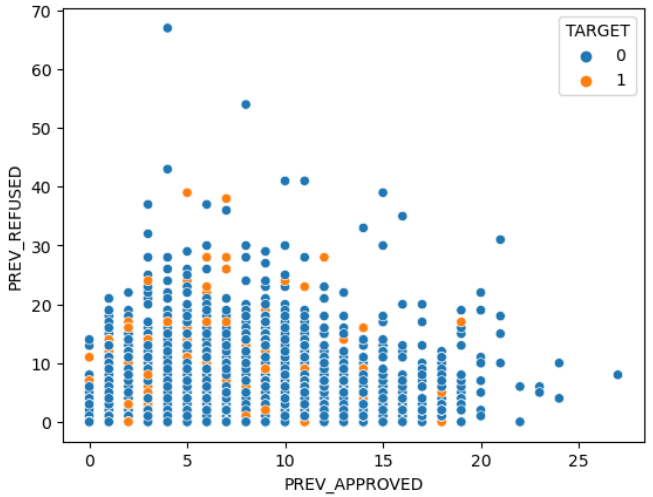




PREV_APPROVED x PREV_REFUSED

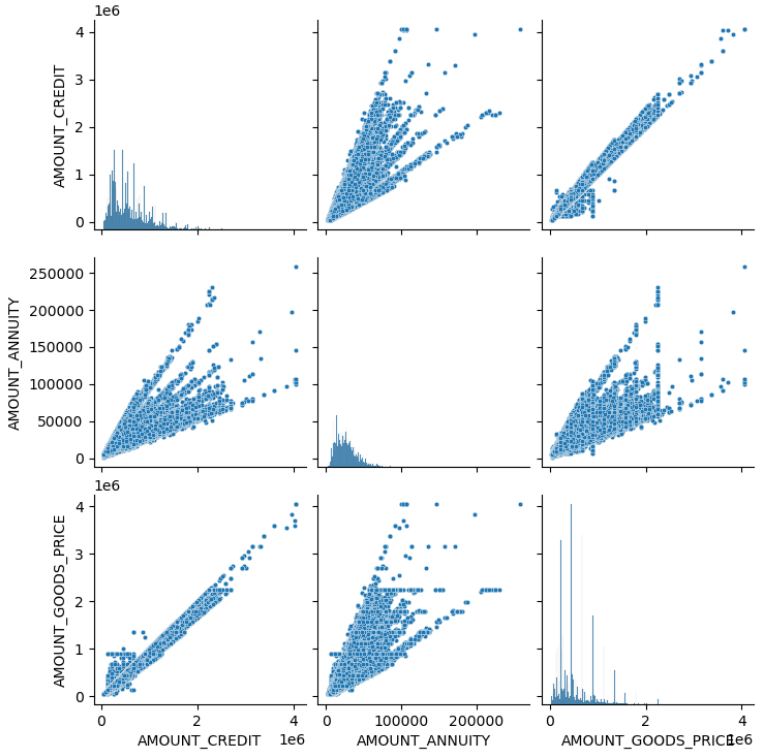
- - PREV_APPROVED and PREV_REFUSED has a weak positive correlation, ie they tend to increase together, though it is hard to see the trend on the scatterplot.
- - On the heatmap of payment difficulty rate, we can see that PRE_APPROVED and PRE_REFUSED affect the rate together.
  - A person with very low approved/high refused pose an extremely high risk of payment difficulty (can go up to 20% ~ 40%)
  - Conversely, a person with very high approved/low refused has significantly lower chance of payment difficulty (as low as 4.8%)



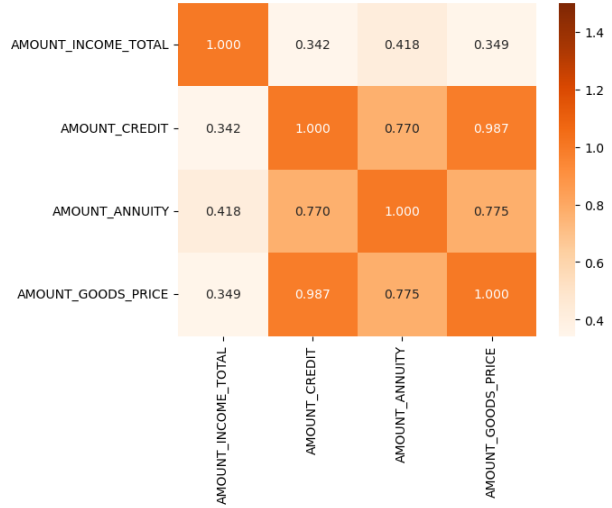


AMOUNT_INCOME_TOTAL, AMOUNT_CREDIT, AMOUNT_ANNUITY and AMOUNT_GOODS_PRICE

- - AMOUNT_CREDIT, AMOUNT_ANNUITY, AMOUNT_GOODS_PRICE has very strong correlation to each other.
- - There's a moderate/weak correlation between Income and AMOUNT_CREDIT/AMOUNT_ANNUITY/AMOUNT_GOODS_PRICE

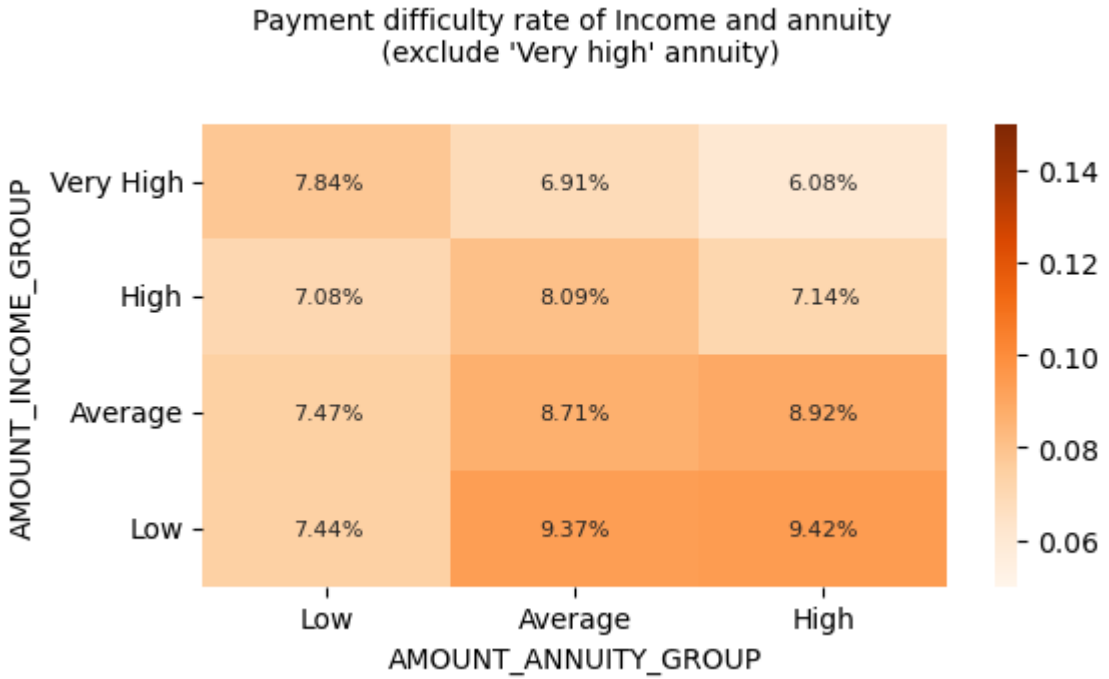


Correlation coefficient between variables



AMOUNT_INCOME_TOTAL x AMOUNT_ANNUITY

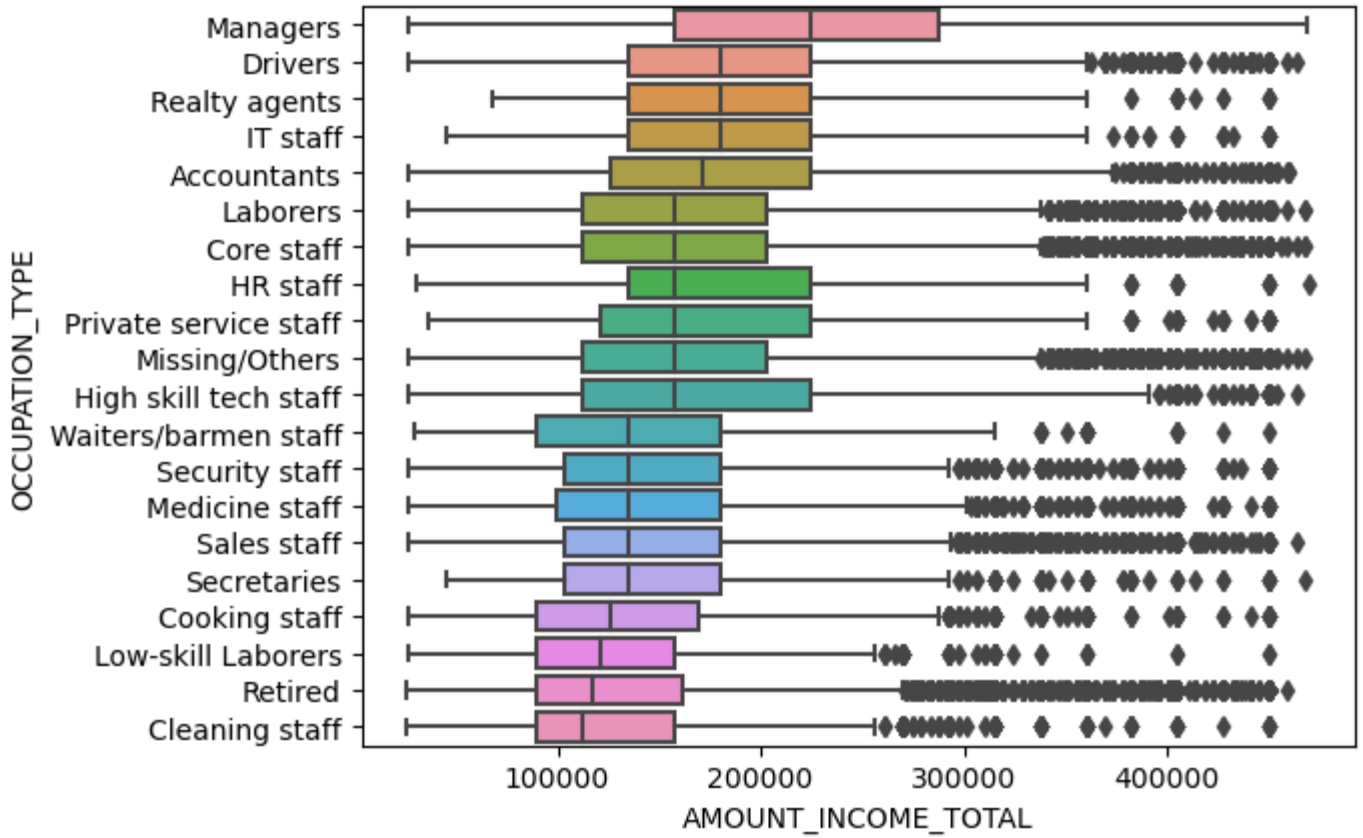
- Income and annuity together has a correlation with payment difficulty rate.
- Client in low/average income range but average/high annuity has the highest risk of payment difficulty.



OCCUPATION_TYPE x AMOUNT_INCOME

- Managers position has the highest mean income and overall low payment difficulty rate.
- Drivers have the second highest mean income, but they have very high payment difficulty rate across all income range.
- Cooking staff, low-skill labor, retired and cleaning staff has the lowest mean income. They also have high payment difficulty risk except for Retire.

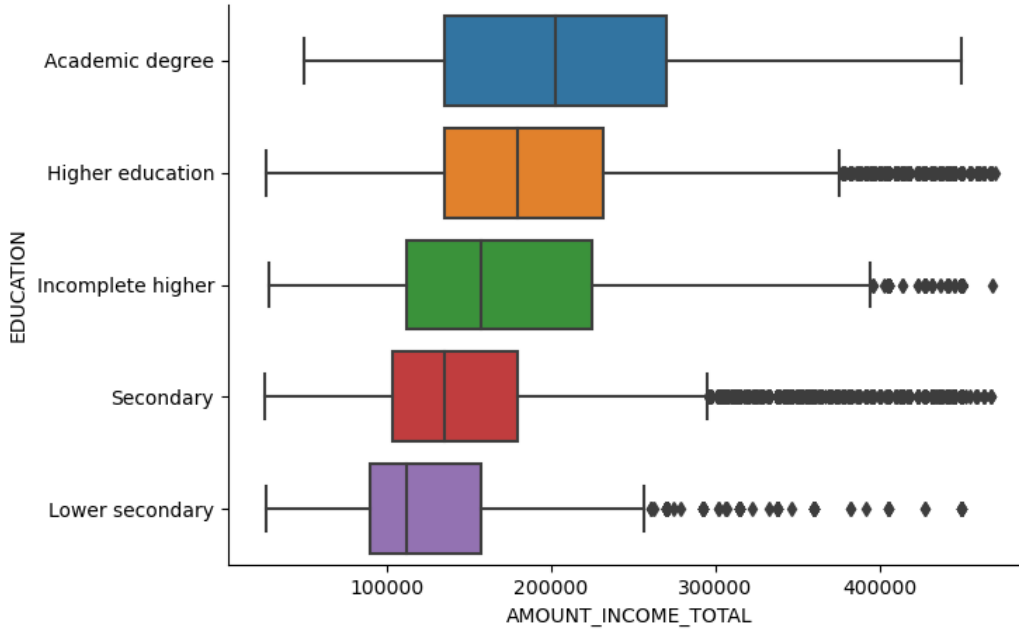
Income by occupation, sorted by median



EDUCATION x AMOUNT_INCOME

- We can see that education level greatly affects the payment difficulty rate.
- Clients with Secondary/Lower secondary education tend to have higher PD rate than average, even in the High income range.
- On the other hand, client with higher education have significantly lower PD rate than average, regardless of income range.
- Overall, there is a correlation between education and income, as seen in the boxplot. Higher education customers have higher income range.
- While education and income can both affect the payment difficulty rate, education is the more deciding factor of the two.

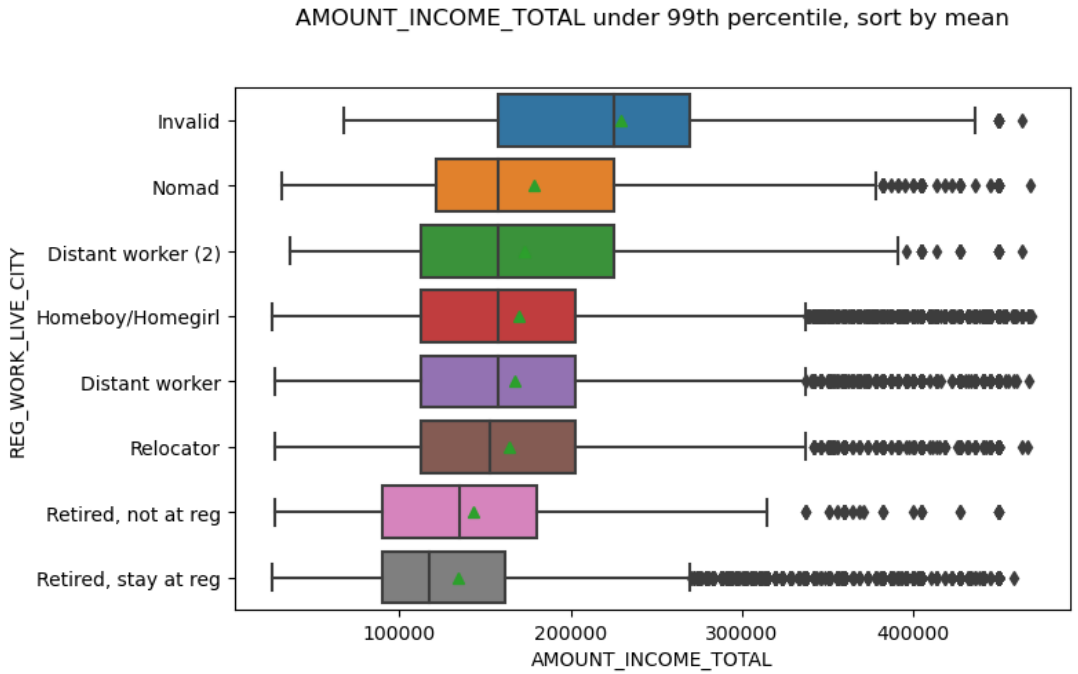
Income by Education level, sorted by median

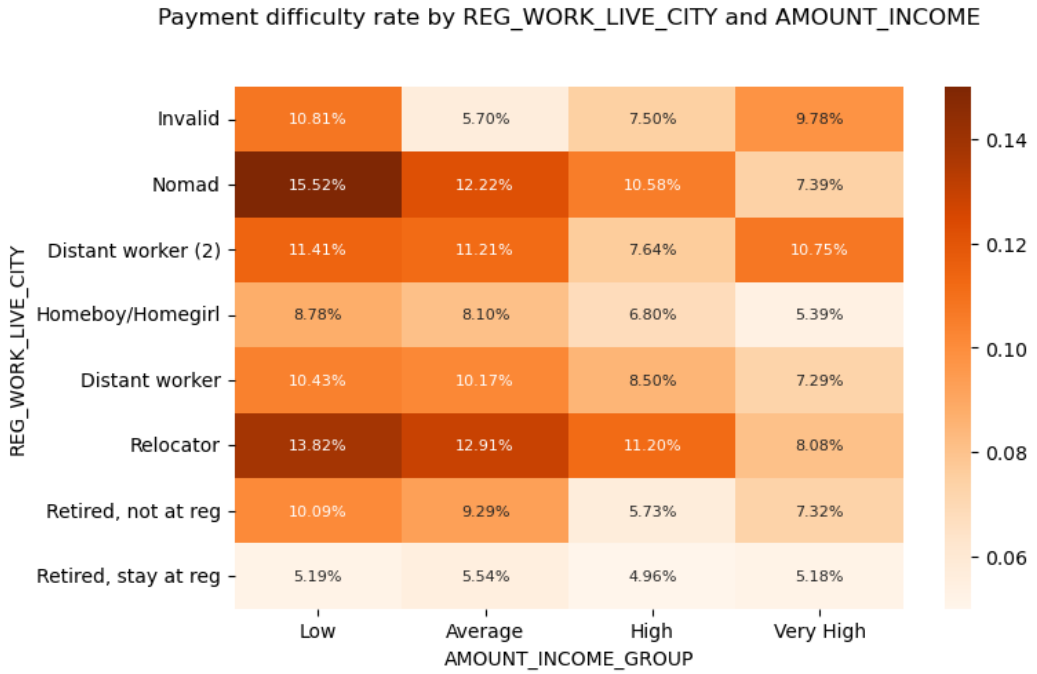


Payment difficulty rate by Education level and Income range

AMOUNT_INCOME_TOTAL x reg/work/live city

- Nomad (reg≠work≠live) and Relocator(reg≠work=live) have the highest PD rate, even for high income customers.
- During our analysis so far, Retiree tend to have very low payment difficulty rate. Only now we can see a segment where Retiree has higher PD rate than average: those who are living in a different city than their registered address and have low/average income.





Summary of multivariate analysis

- **AGE** and **AMOUNT_INCOME_TOTAL** together has a correlation with payment difficulty rate, although they do not have a correlation to each other.

- **YEARS_EMPLOYED** and **AGE** together has a correlation with payment difficulty rate. The more experienced a client is and the higher their income is tend to lead to lower payment difficulty rate.

- **YEARS_EMPLOYED** and **AMOUNT_INCOME_TOTAL** together has a correlation with payment difficulty rate.

- **PRE_APPROVED** and **PRE_REFUSED** also affect the rate together

- **AMOUNT_CREDIT, AMOUNT_ANNUITY, AMOUNT_GOODS_PRICE** has very strong correlation to each other. There's a moderate/weak correlation between Income and AMOUNT_CREDIT/AMOUNT_ANNUITY/AMOUNT_GOODS_PRICE. Income and annuity together has a correlation with payment difficulty rate. Specifically, clients in low/average income but average/high annuity have the highest payment difficulty rate.

- In term of **Occupation** type, Managers position has the highest mean income and overall low payment difficulty rate. Cooking staff, low-skill labor, retired and cleaning staff has the lowest mean income. They also have high payment difficulty risk except for Retire.

- We can see that **EDUCATION** level greatly affects the payment difficulty rate across all income ranges. While education and income can both affect the payment difficulty rate, education is the more deciding factor of the two.

Business Recommendation

Most important features

1. **AGE**: younger age has higher PD rate and vice versa.
2. **YEARS_EMPLOYED**: low years of employment has higher PD
3. **AMOUNT_INCOME_TOTAL**: Low/average range has higher rate than high/very high range
4. **PREV_APPROVED**: More previous approved loan means lower PD rate. The lowest is no history.
5. **PREV_REFUSED**: Less previous refused loan means lower PD rate. The lowest is no history.
6. **TOTAL/AVERAGE_CREDIT_APPLIED**: Low total approved credit and low average approved credit per loan has higher rate.
7. **AMOUNT_CREDIT**: clients with low/average credit has higher PD rate than those in high/very high range.
8. **AMOUNT_ANNUITY**: average and high annuity has higher rate than either low or very high range.
9. **AMOUNT_REQ_CB_YEAR**: More requests in the year prior to the loan application have significantly higher payment difficulty rate than those with fewer request.
10. **OCCUPATION_TYPE**: Heavy labor jobs has the highest PD rate while accountant, manager, high skill tech, ... have lowest rate.
11. **EDUCATION**: the level of education also correlate to payment difficulty rate.
12. **REG_WORK_LIVE_CITY/CITY_COUNT**: People associated with 2 cities or more have significantly higher PD rate than those who live and work in the same city as their registered city.

Most important combinations

Combination of features:

1. **AGE** and **AMOUNT_INCOME_TOTAL**
2. **YEARS_EMPLOYED** and **AGE**
3. **YEARS_EMPLOYED** and **AMOUNT_INCOME_TOTAL**
4. **PREV_APPROVED** and **PREV_REFUSED**
5. **REG_WORK_LIVE_CITY** and **AMOUNT_INCOME**
6. **EDUCATION** and **AMOUNT_INCOME**
7. **AMOUNT_INCOME** and **AMOUNT_ANNUITY**

Green flags

General traits of an ideal customers that the bank should focus on:

- Age 41 or older, more than 5 years of employment
- Retiree living at their registered address
- Income in high range
- Has high amount of previous approved contracts and low amount of refused.
- Low amount of AMOUNT_REQ_CB in the last year
- Finished higher education
- Holding management position or high skilled staff occupation
- Live and work in the same city of their registered address
- Married, have no children
- Own new car
- Have low annuity amount relatively to their income
- Have no defaulted individual in their recent social circle

Red flags

General traits of customer with higher risk of payment difficulty:

- Young age (Under 30)
- Older age (30-50) but have low years of employment
- Older age (30-50) but only finished Secondary education
- Retiree living in a different city than their registered address
- Income in low range
- Income in high range but have low education level
- Renting apartment or living with their parents
- Has low amount of previous approved contracts and high amount of refused
- High amount of AMOUNT_REQ_CB in the last year
- Working as low-skilled labor or heavy-manual labor
- Relocated to or work in a different city than their registered address
- Single, or married but have many children
- Own old car (>10 years) or do not own car at all
- High annuity amount relative to their income
- Have defaulted individual in their recent social circle

Other Recommendations

Issues that need considering stricter requirements/approval process in the future

- The clients that have missing value for amount request to Credit bureau have the highest PD rate (10.34%). The bank needs to investigate the source of this segment and apply stricter procedure/requirements for such clients.

- Nearly 1/3 of clients with PD the bank have had are from 31-40 group - the group with 2nd highest risk. The bank may need to apply stricter requirement in the future on this age group to limit the clients with PD from said group.

- 58% of clients with PD the bank has had are from under 5 years of employment - the groups with highest risk. The bank may need to apply stricter requirement in the future on this group to limit the clients with PD from said group.

- In term of required documents, the bank might need to review the procedure associated with Document_3. Could it be that DOC_3 is the easiest to apply/obtain thus allow people with higher PD rate exploit this?

- Secondary/Lower Secondary has the highest PD rate, which makes up more than 70% of the current clients. The bank might need to apply stricter requirements on this group.

Thank you.