Assessment

Predict how capable each applicant is of repaying a loan.

NGUYEN THI KIM NHU

1. EXPLORE DATA

TARGET DEFINITION | DATA REVIEW | **FACILITATE NEW VARIABLES** | MISSING/OUTLIER VALUES

TARGET (binary variable) indicates whether a customer has payment difficulties or not.

Definition of Bad

Target = 1, defined bad. That presents a client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan.

Definition of Good

Target = 0, all other cases.

1. EXPLORE DATA

TARGET DEFINITION | DATA REVIEW | **FACILITATE NEW VARIABLES** | MISSING/OUTLIER VALUES

Selected variable are based on two standards as follows:

- ❑ First, the variables have been **suggested from the book "Intelligent Credit Scoring"**
- ❑ Second, the resource provided must be **sufficient** to facilitate the calculation of variables

- Removing
  - ❑ Rows with 'CODE_GENDER' equal to 'XNA'
  - ❑ Rows with 'NAME_FAMILY_STATUS' equal to 'Unknown'

1. EXPLORE DATA

TARGET DEFINITION | DATA REVIEW | **FACILITATE NEW VARIABLES** | MISSING/OUTLIER VALUES

No	Table	New Variables	Description
0	application_{train|test}.csv	BUREAU_SCORE	Mean of 'EXT_SOURCE_1', 'EXT_SOURCE_2', and 'EXT_SOURCE_3' for each applicant.
1	application_{train|test}.csv	credit_annuity_ratio	Ratio of 'AMT_CREDIT' to 'AMT_ANNUITY' for each applicant.
2	application_{train|test}.csv	credit_goods_price_ratio	Ratio of 'AMT_GOODS_PRICE' to 'AMT_CREDIT', indicating the degree of financing for each loan.
3	application_{train|test}.csv	credit_downpayment	Difference between 'AMT_GOODS_PRICE' and 'AMT_CREDIT', serving as collateral for the loan.
4	application_{train|test}.csv	DEBT_TO_INCOME_RATIO	Ratio of 'AMT_CREDIT' to 'AMT_INCOME_TOTAL', representing the debt-to-income ratio for each applicant.
5	application_{train|test}.csv	Age	Age of each applicant calculated from 'DAYS_BIRTH', divided (-365)
6	application_{train|test}.csv	NEW_EMPLOY_TO_BIRTH_RATIO	Ratio of the days employed relative to the client's age
7	bureau.csv	Number of public records	Count of public records for each applicant in the 'bureau' dataset.

1. EXPLORE DATA

TARGET DEFINITION

DATA REVIEW

FACILITATE NEW VARIABLES

MISSING/OUTLIER VALUES

No	Table	New Variables	Description
8	bureau.csv	Time at bureau	Longest time at the bureau (in years) for each applicant calculated from 'DAYS_CREDIT'.
9	bureau.csv	COUNT_ACTIVE	Count of 'Active' values in the 'CREDIT_ACTIVE' column for each applicant in the bureau table
10	credit_card_balance.csv	TIME_AS_CUSTOMER	Time as a customer (in years) calculated from the 'MONTHS_BALANCE' column in the 'credit_card' dataset.
11	previous_application.csv	PREV_IR	Average of interest rate calculated from the 'AMT_ANNUITY', 'CNT_PAYMENT', and 'AMT_CREDIT' about all the previous loans for each application.
12	previous_application.csv	PREV_APPROVED_RATIO	Ratio of approved contracts to total previous contract for each application.
13	installments_payments.csv	avg_past_due_prev	Average of past due installments for each application

After the computation of new variables, all features will be merged into the application and uniquely identified by the 'SK_ID_CURR' (unique ID for each loan in the sample).

1. EXPLORE DATA

TARGET DEFINITION

DATA REVIEW

FACILITATE NEW VARIABLES

MISSING/OUTLIER VALUES

307 505 obs

The missing value rate is 6.16%. Despite various ways to fill missing values, dropping them using the `dropna` method is the optimal approach in the shortest time without removing too much data. After analysis, the missing values comprise approximately 5% of the "1" target values and 6.26% of the "0" target values.

Variables	Missing	Variables	Missing
SK_ID_CURR	0	BUREAU_SCORE	172
TARGET	0	credit_annuity_ratio	12
CODE_GENDER	0	credit_goods_price_ratio	276
FLAG_OWN_CAR	0	credit_downpayment	276
FLAG_OWN_REALTY	0	DEBT_TO_INCOME_RATIO	0
CNT_CHILDREN	0	AMT_public_records	0
AMT_INCOME_TOTAL	0	Age	0
AMT_CREDIT	0	NEW_EMPLOY_TO_BIRTH_RATIO	55374
AMT_ANNUITY	12	Time_at_bureau	0
AMT_GOODS_PRICE	276	COUNT_ACTIVE	0
NAME_EDUCATION_TYPE	0	TIME_AS_CUSTOMER	0
NAME_FAMILY_STATUS	0	PREV_IR	18945
NAME_HOUSING_TYPE	0	avg_past_due_prev	15866
REGION_POPULATION_RELATIVE	0	PREV_APPROVED_RATIO	18945
OCCUPATION_TYPE	0		

1. EXPLORE DATA

TARGET DEFINITION

DATA REVIEW

FACILITATE NEW VARIABLES

MISSING/OUTLIER VALUES

307 505 obs

dropna

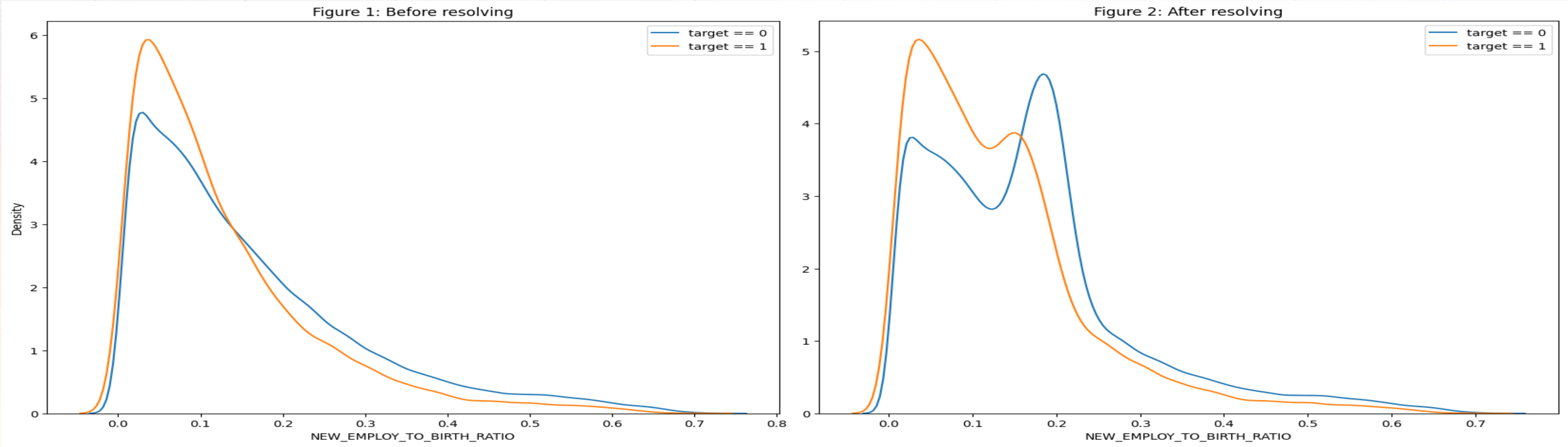
287 642 obs

I dropped the missing values (NaNs) for all variables, retaining only the specific NaNs in the 'NEW_EMPLOY_TO_BIRTH_RATIO' variable. Handling missing values in this selective manner significantly alters the distribution of the remaining variables.

1. EXPLORE DATA

TARGET DEFINITION | DATA REVIEW | FACILITATE NEW VARIABLES | **MISSING/OUTLIER VALUES**

Distribution of NEW_EMPLOY_TO_BIRTH_RATIO by Target



During the analysis, I observed that the processing of the variable related to "DAYS_EMPLOYED" involves addressing outliers in the form of extreme values, as some values represent periods exceeding 100 years. After replacing these outlier values with NaN, the distribution of the variable was examined (Figure 1). At the cutoff point of 0.15, the distribution of the target variable (0 or 1) showed that the occurrences with target = 0 outnumbered those with target = 1.

To impute missing values, linear regression was employed, and the resulting graph is showed in Figure 2. The cutoff point at the equivalent value of 0.15 was also observed, indicating that the distribution of target = 0 remains more prevalent than that of target = 1. Even though the graph shows two distinct peaks, I consider this compromise acceptable given the time constraints.

1. EXPLORE DATA

TARGET DEFINITION

DATA REVIEW

FACILITATE NEW VARIABLES

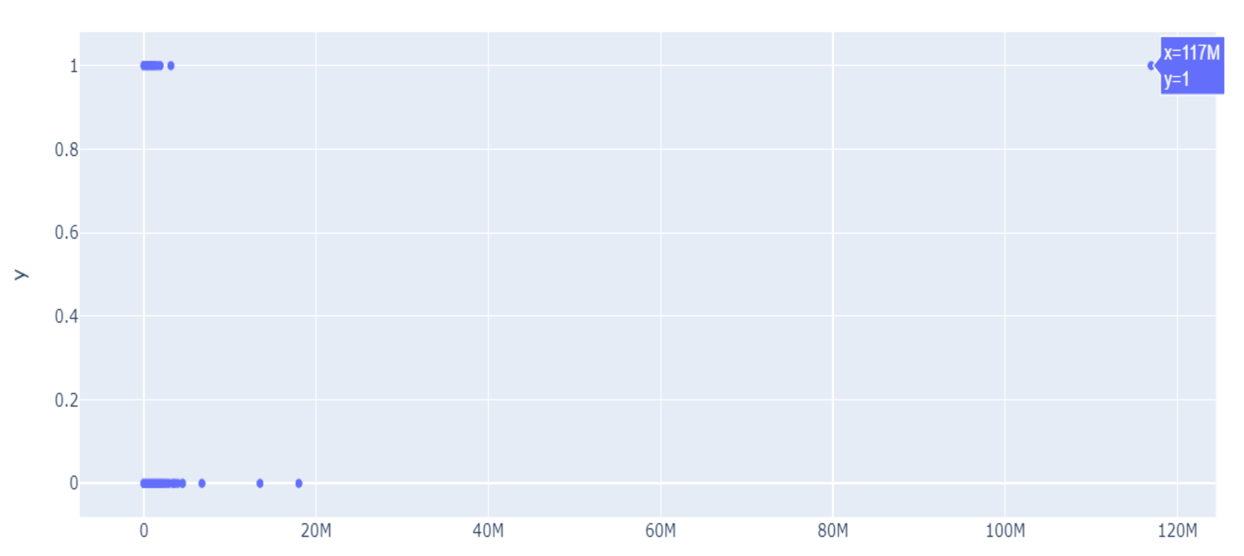
MISSING/OUTLIER VALUES

FIND OUTLIERS

	AMT_INCOME_TOTAL
count	287642.000
mean	167138.157
std	241223.643
min	25650.000
25%	112500.000
50%	144000.000
75%	202500.000
max	117000000.000

Summary statistics is helpful to determine whether or not the dataset has outliers. As we can see, the AMT_INCOME_TOTAL columns have outliers. The max is 117 000 000 while its mean is 167 138. The mean is sensitive to outliers, but the fact the mean is so small compared to the max value indicates the max value is an outlier.

I defined one observation outlier in AMT_INCOME_TOTAL → drop this outliner



1. EXPLORE DATA

TARGET DEFINITION

DATA REVIEW

FACILITATE NEW VARIABLES

MISSING/OUTLIER VALUES

WORKING WITH OUTLIERS

For integer variables with a limited number of outliers, scaling by capping the maximum value is a common approach. This helps mitigate the impact of outliers and allows for better scaling of the data. Adjust the threshold values as needed based on your data characteristics and requirements.

```
data2.loc[data2['AMT_public_records'] > 30, 'AMT_public_records'] = 30
data2.loc[data2['COUNT_ACTIVE'] > 13, 'COUNT_ACTIVE'] = 13
data2.loc[data2['CNT_CHILDREN'] > 10, 'CNT_CHILDREN'] = 10
data2.loc[data2['avg_past_due_prev'] > 40, 'avg_past_due_prev'] = 40
```

307 505 obs

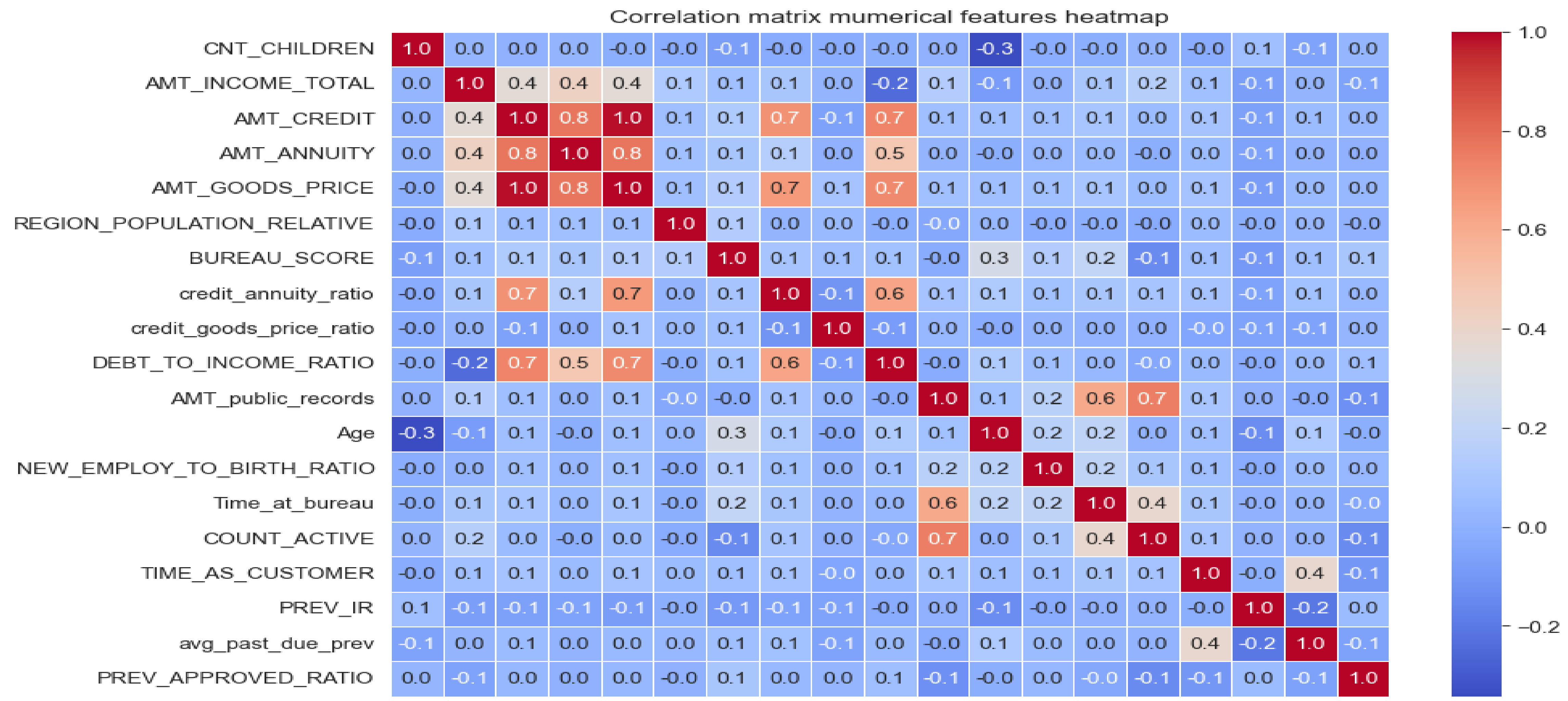
dropna

287 642 obs

Working missing/ outliers

267 366 obs

2. VARIABLE ANALYSIS



The variables related to loan amounts and financial information, such as income, exhibit a relatively high degree of correlation. This is understandable as they involve straightforward mathematical formulas. Similarly, variables computed from the bureau table also show a high level of correlation. Further selection will be performed later. For the remaining variables, the correlation is within an acceptable range.

2.1. SCALING MAX VALUE

In the process of working with the data, I observed the presence of outliers, which needs to be addressed. I have redrawn the distribution plots of the variables and scaled the values accordingly. This is essential for the next step in the process, which involves binning the data to apply the Weight of Evidence (WOE) analysis method.

2.1. SCALING MAX VALUE

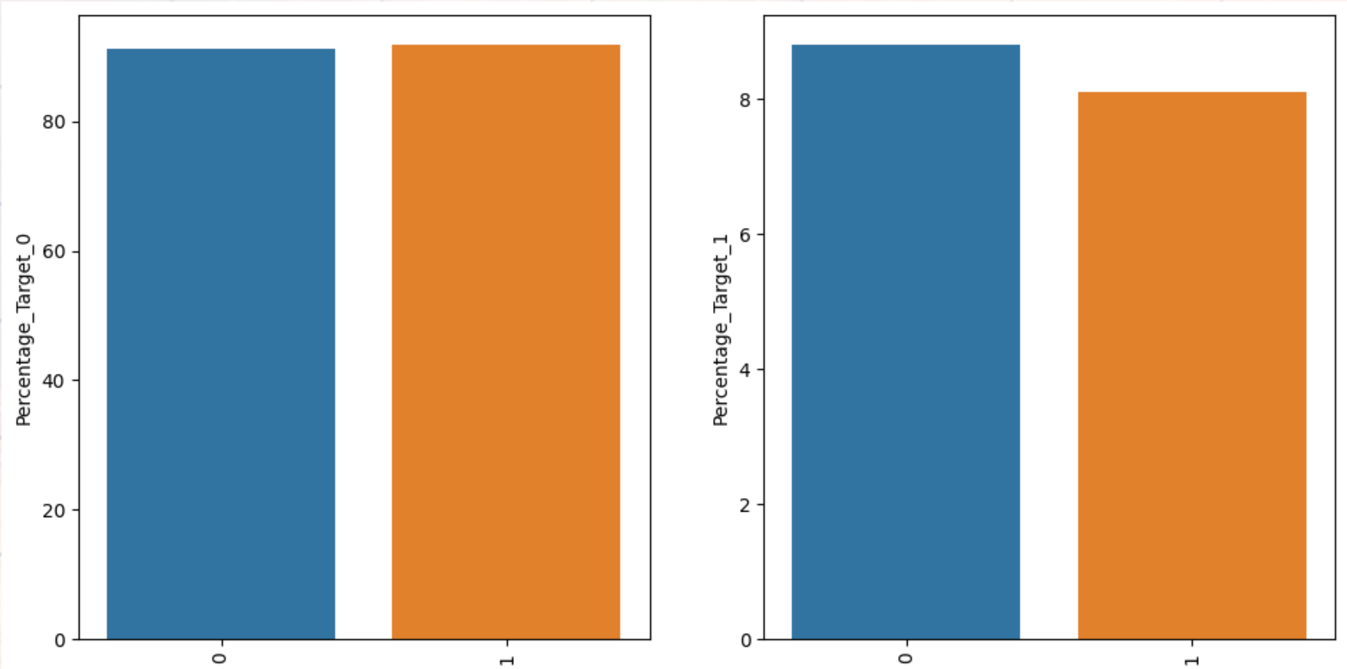


Figure: **Distribution of FLAG_NEW_CUSTOMER BY TARGET**

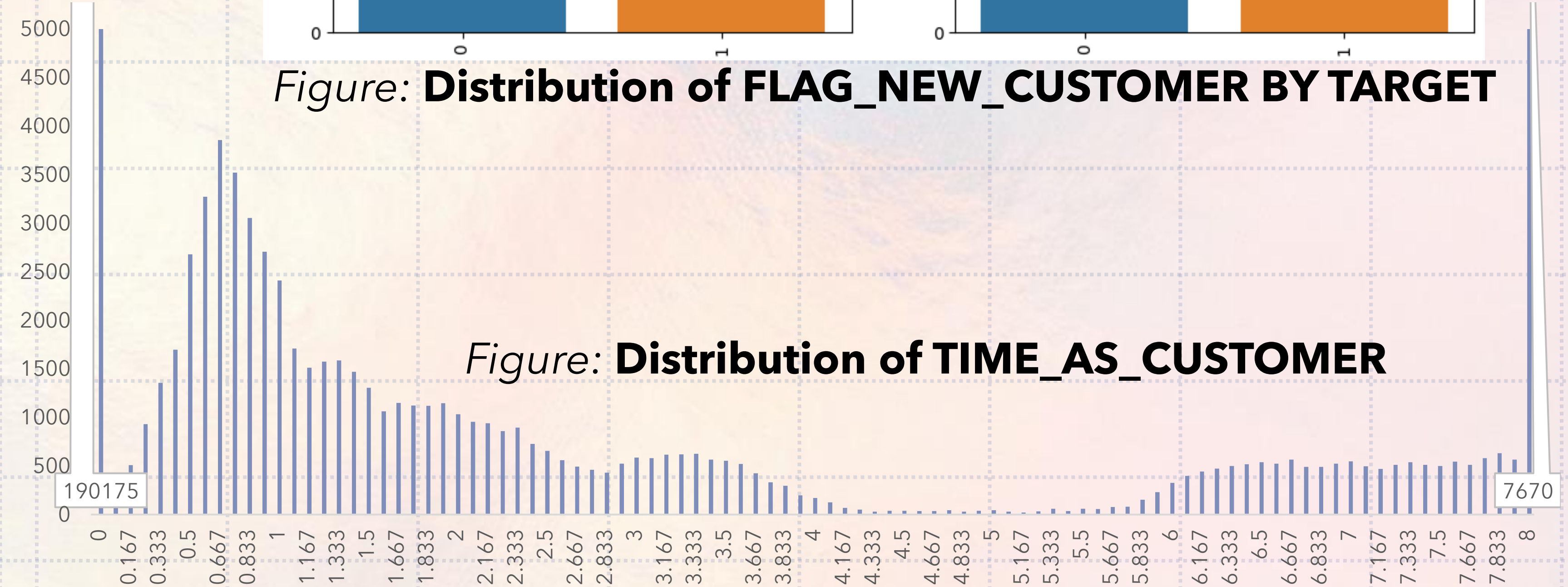
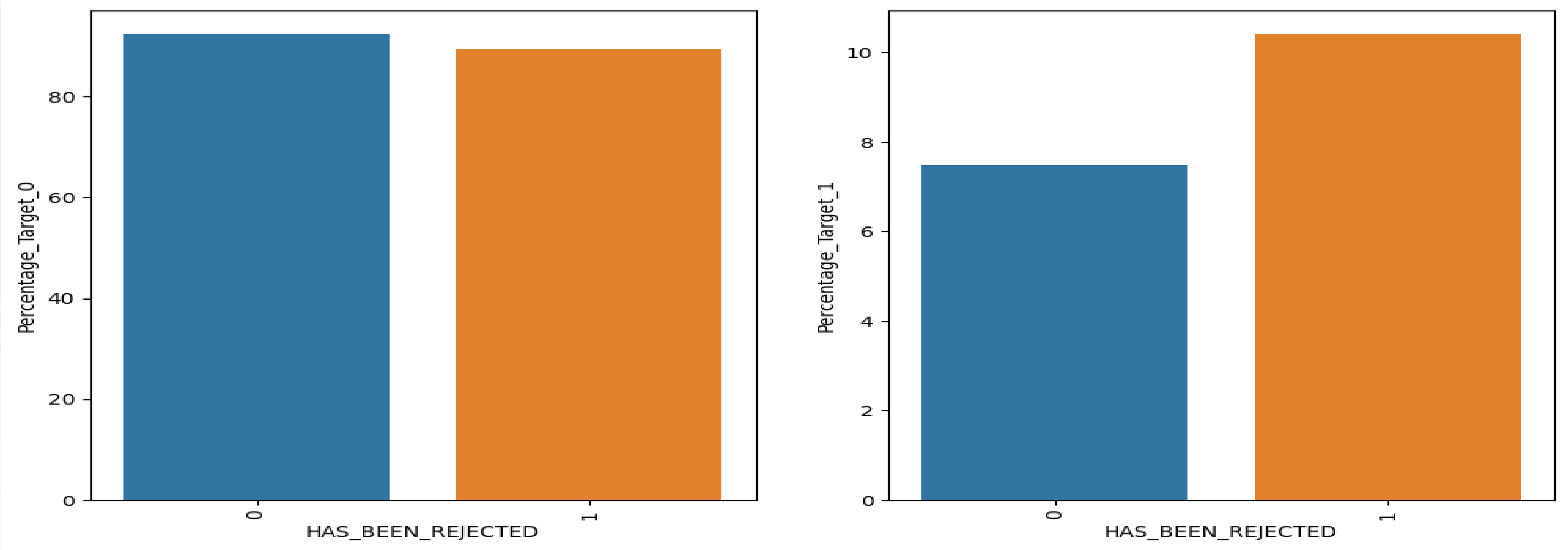


Figure: **Distribution of TIME_AS_CUSTOMER**

I also noticed that the variable TIME_AS_CUSTOMER exhibits unusual variability, with a significant concentration of values at 0, indicating that most loans are from new customers. Therefore, the TIME_AS_CUSTOMER variable will be transformed into a FLAG_NEW_CUSTOMER variable based on the condition that if TIME_AS_CUSTOMER = 0, the customer is considered new, and FLAG_NEW_CUSTOMER will take the value of 1. If TIME_AS_CUSTOMER is greater than 0, FLAG_NEW_CUSTOMER will take the value of 0 (indicating not a new customer).

2.1. SCALING MAX VALUE



In a similar manner, for the variable PREV_APPROVED_RATIO, I observed that more than 70% of the values are concentrated at 1 (indicating that previous loans of applicants were approved 100%). Therefore, I will transform it into a binary variable, HAS_BEEN_REJECTED, where 0 represents customers who have never been rejected for a loan, and 1 represents customers who have been rejected in previous loan applications. Thus, by assigning the value HAS_BEEN_REJECTED = 0 when PREV_APPROVED_RATIO = 1 and HAS_BEEN_REJECTED = 1 when PREV_APPROVED_RATIO < 1.

2.1. SCALING MAX VALUE

After processing the above procedures, the remaining dataset contains 267366 observations, and the variables will be further analyzed using the Weight of Evidence (WOE) technique as shown:

```
<class 'pandas.core.frame.DataFrame'>
Index: 267366 entries, 0 to 307504
Data columns (total 24 columns):
 #   Column                       Non-Null Count   Dtype  
---  ------                       --------------   -----  
 0   SK_ID_CURR                   267366 non-null  int64  
 1   TARGET                       267366 non-null  int64  
 2   CODE_GENDER                  267366 non-null  object 
 3   FLAG_OWN_CAR                 267366 non-null  object 
 4   FLAG_OWN_REALTY              267366 non-null  object 
 5   CNT_CHILDREN                 267366 non-null  int64  
 6   NAME_EDUCATION_TYPE          267366 non-null  object 
 7   NAME_FAMILY_STATUS           267366 non-null  object 
 8   NAME_HOUSING_TYPE            267366 non-null  object 
 9   REGION_POPULATION_RELATIVE   267366 non-null  float64
 10  OCCUPATION_TYPE              267366 non-null  object 
 11  BUREAU_SCORE                 267366 non-null  float64
 12  credit_annuity_ratio         267366 non-null  float64
 13  credit_goods_price_ratio     267366 non-null  float64
 14  DEBT_TO_INCOME_RATIO         267366 non-null  float64
 15  AMT_public_records           267366 non-null  float64
 16  Age                          267366 non-null  float64
 17  NEW_EMPLOY_TO_BIRTH_RATIO    267366 non-null  float64
 18  Time_at_bureau               267366 non-null  float64
 19  COUNT_ACTIVE                 267366 non-null  float64
 20  PREV_IR                      267366 non-null  float64
 21  avg_past_due_prev            267366 non-null  int32  
 22  FLAG_NEW_CUSTOMER            267366 non-null  int64  
 23  HAS_BEEN_REJECTED            267366 non-null  int64  
dtypes: float64(11), int32(1), int64(5), object(7)
memory usage: 50.0+ MB
```

2.2. GROUPED VARIABLES

I also assign ordinal group numbers based on the WOE in a way that corresponds to the order for the model to follow. For example, for specific attributes within a characteristic, if the WOE is lower, indicating a higher proportion of bads for that attribute when labeled, I will assign a higher order label.

Extract the results for object variables as follows:
To better understand, please open the attached code to view the details

```
IV of  CODE_GENDER : 0.04597551745907483
  CODE_GENDER   Count   Bads   Goods  Tot Distr  Distr Good  Distr Bad  Bad Rate     WOE    IV
0           0  180534  12906  167628     67.523       0.684      0.581     0.071  16.349 0.017
1           1   86832   9320   77512     32.477       0.316      0.419     0.107 -28.230 0.029
2       Total  267366  22226  245140      1.000       1.000      1.000     0.083   0.000   NaN
IV of  FLAG_OWN_CAR : 0.003882449576440671
  FLAG_OWN_CAR   Count   Bads   Goods  Tot Distr  Distr Good  Distr Bad  Bad Rate    WOE    IV
0            0   86182   6577   79605     32.234       0.325      0.296     0.076  9.293 0.003
1            1  181184  15649  165535     67.766       0.675      0.704     0.086 -4.179 0.001
2        Total  267366  22226  245140      1.000       1.000      1.000     0.083  0.000   NaN
IV of  FLAG_OWN_REALTY : 0.001417089632750136
  FLAG_OWN_REALTY   Count   Bads   Goods  Tot Distr  Distr Good  Distr Bad  Bad Rate    WOE    IV
0               0  188966  15357  173609     70.677       0.708      0.691     0.081  2.467 0.000
1               1   78400   6869   71531     29.323       0.292      0.309     0.088 -5.745 0.001
2           Total  267366  22226  245140      1.000       1.000      1.000     0.083  0.000   NaN
IV of  NAME_EDUCATION_TYPE : 0.03976430796593128
  NAME_EDUCATION_TYPE   Count   Bads   Goods  Tot Distr  Distr Good  Distr Bad  Bad Rate     WOE    IV
0                   0     115      1     114      0.043       0.000      0.000     0.009 233.563 0.001
1                   1   57456   3252   54204     21.490       0.221      0.146     0.057  41.292 0.031
2                   2    8776    760    8016      3.282       0.033      0.034     0.087  -4.469 0.000
3                   3  197462  17814  179648     73.855       0.733      0.801     0.090  -8.955 0.006
4                   4    3557    399    3158      1.330       0.013      0.018     0.112 -33.183 0.002
5               Total  267366  22226  245140      1.000       1.000      1.000     0.083   0.000   NaN
```

2.2. GROUPED VARIABLES

Extract the results for object variables as follows:
To better understand, please open the attached code to view the details

```
IV of  CODE_GENDER : 0.04597551745907483
  CODE_GENDER   Count   Bads   Goods  Tot Distr  Distr Good  Distr Bad  Bad Rate     WOE    IV
0           0  180534  12906  167628     67.523       0.684      0.581     0.071  16.349 0.017
1           1   86832   9320   77512     32.477       0.316      0.419     0.107 -28.230 0.029
2       Total  267366  22226  245140      1.000       1.000      1.000     0.083   0.000   NaN
IV of  FLAG_OWN_CAR : 0.003882449576440671
  FLAG_OWN_CAR   Count   Bads   Goods  Tot Distr  Distr Good  Distr Bad  Bad Rate    WOE    IV
0            0   86182   6577   79605     32.234       0.325      0.296     0.076  9.293 0.003
1            1  181184  15649  165535     67.766       0.675      0.704     0.086 -4.179 0.001
2        Total  267366  22226  245140      1.000       1.000      1.000     0.083  0.000   NaN
IV of  FLAG_OWN_REALTY : 0.001417089632750136
  FLAG_OWN_REALTY   Count   Bads   Goods  Tot Distr  Distr Good  Distr Bad  Bad Rate    WOE    IV
0               0  188966  15357  173609     70.677       0.708      0.691     0.081  2.467 0.000
1               1   78400   6869   71531     29.323       0.292      0.309     0.088 -5.745 0.001
2           Total  267366  22226  245140      1.000       1.000      1.000     0.083  0.000   NaN
IV of  NAME_EDUCATION_TYPE : 0.03976430796593128
  NAME_EDUCATION_TYPE   Count   Bads   Goods  Tot Distr  Distr Good  Distr Bad  Bad Rate     WOE    IV
0                   0     115      1     114      0.043       0.000      0.000     0.009 233.563 0.001
1                   1   57456   3252   54204     21.490       0.221      0.146     0.057  41.292 0.031
2                   2    8776    760    8016      3.282       0.033      0.034     0.087  -4.469 0.000
3                   3  197462  17814  179648     73.855       0.733      0.801     0.090  -8.955 0.006
4                   4    3557    399    3158      1.330       0.013      0.018     0.112 -33.183 0.002
5               Total  267366  22226  245140      1.000       1.000      1.000     0.083   0.000   NaN
```

2.2. GROUPED VARIABLES

I have written the `WOE_numerical_iv` function to calculate Weight of Evidence (WOE) and Information Value (IV) for numerical variables, but due to even distribution, some values result in inf (infinity) values. I rely on the results of the `WOE_numerical_iv` function to choose suitable `custom_bin_edges` in the `WOE_inf` function and recalculate the WOE for attributes. Additionally, I am considering the most reasonable approach, whether to group or categorize characteristics. Here is example for BUREAU_SCORE:

0.5861238683817757

	BUREAU_SCORE_BINNED	Count	Bads	Goods	Tot Distr	Distr Good	Distr Bad	Bad Rate	WOE	IV
0	(0.0, 0.1]	2409	789	1620	0.901	0.007	0.035	0.328	-168.115	0.049
1	(0.1, 0.2]	7118	1919	5199	2.662	0.021	0.086	0.270	-140.390	0.091
2	(0.2, 0.3]	17387	3549	13838	6.503	0.056	0.160	0.204	-103.981	0.107
3	(0.3, 0.4]	35419	5030	30389	13.247	0.124	0.226	0.142	-60.191	0.062
4	(0.4, 0.5]	57406	4936	52470	21.471	0.214	0.222	0.086	-3.688	0.000
5	(0.5, 0.6]	67887	3575	64312	25.391	0.262	0.161	0.053	48.921	0.050
6	(0.6, 0.7]	57705	1956	55749	21.583	0.227	0.088	0.034	94.939	0.132
7	(0.7, 0.8]	21536	463	21073	8.055	0.086	0.021	0.021	141.745	0.092
8	(0.8, 0.9]	499	9	490	0.187	0.002	0.000	0.018	159.661	0.003
9	(0.9, 1.0]	0	0	0	0.000	0.000	0.000	NaN	NaN	NaN
10	Total	267366	22226	245140	1.000	1.000	1.000	0.083	0.000	NaN

C:\Users\kateo\AppData\Local\Temp\ipykernel_8048\1599241667.py:14: SettingWithCopyWarning:

Using WOE_numerical_iv

0.5860874772962553

	BUREAU_SCORE_BINNED	Count	Bads	Goods	Tot Distr	Distr Good	Distr Bad	Bad Rate	WOE	IV
0	(0.0, 0.1]	2409	789	1620	0.901	0.007	0.035	0.328	-168.115	0.049
1	(0.1, 0.2]	7118	1919	5199	2.662	0.021	0.086	0.270	-140.390	0.091
2	(0.2, 0.3]	17387	3549	13838	6.503	0.056	0.160	0.204	-103.981	0.107
3	(0.3, 0.4]	35419	5030	30389	13.247	0.124	0.226	0.142	-60.191	0.062
4	(0.4, 0.5]	57406	4936	52470	21.471	0.214	0.222	0.086	-3.688	0.000
5	(0.5, 0.6]	67887	3575	64312	25.391	0.262	0.161	0.053	48.921	0.050
6	(0.6, 0.7]	57705	1956	55749	21.583	0.227	0.088	0.034	94.939	0.132
7	(0.7, inf]	22035	472	21563	8.242	0.088	0.021	0.021	142.119	0.095
8	Total	267366	22226	245140	1.000	1.000	1.000	0.083	0.000	NaN

C:\Users\kateo\AppData\Local\Temp\ipykernel_8048\411095794.py:8: SettingWithCopyWarning:

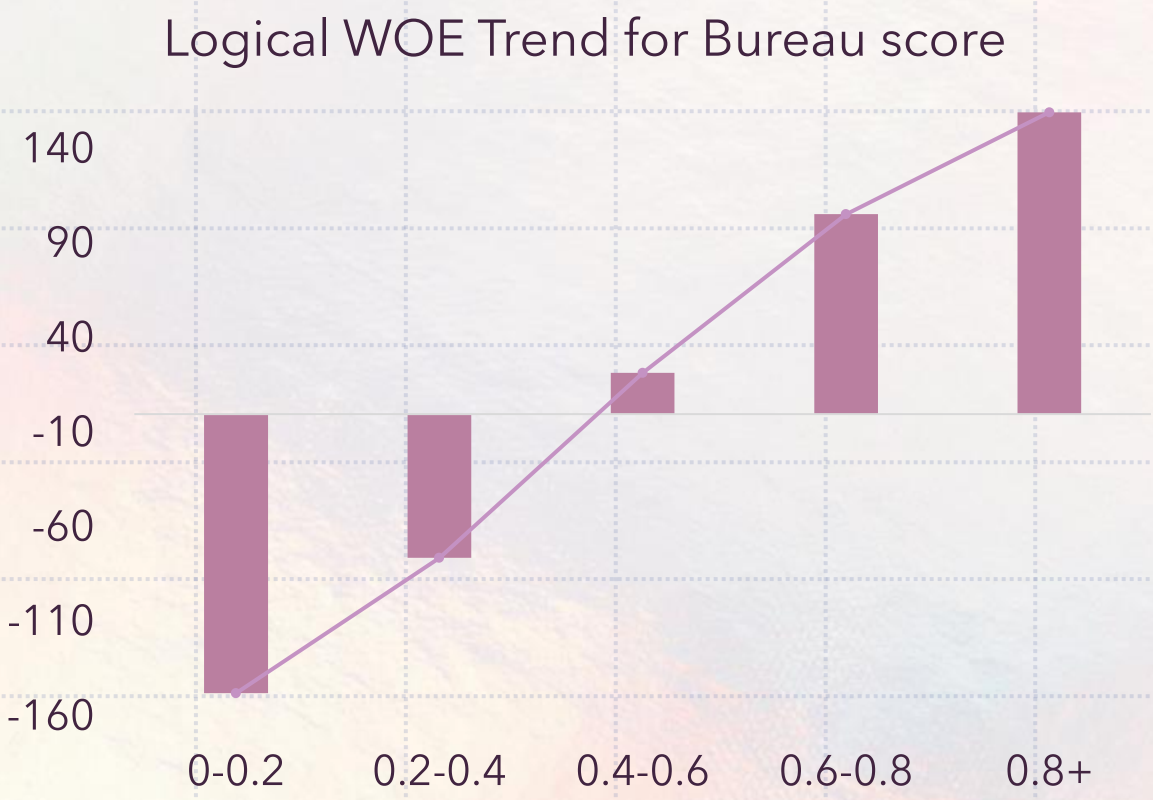
Using WOE_inf

2.2. GROUPED VARIABLES

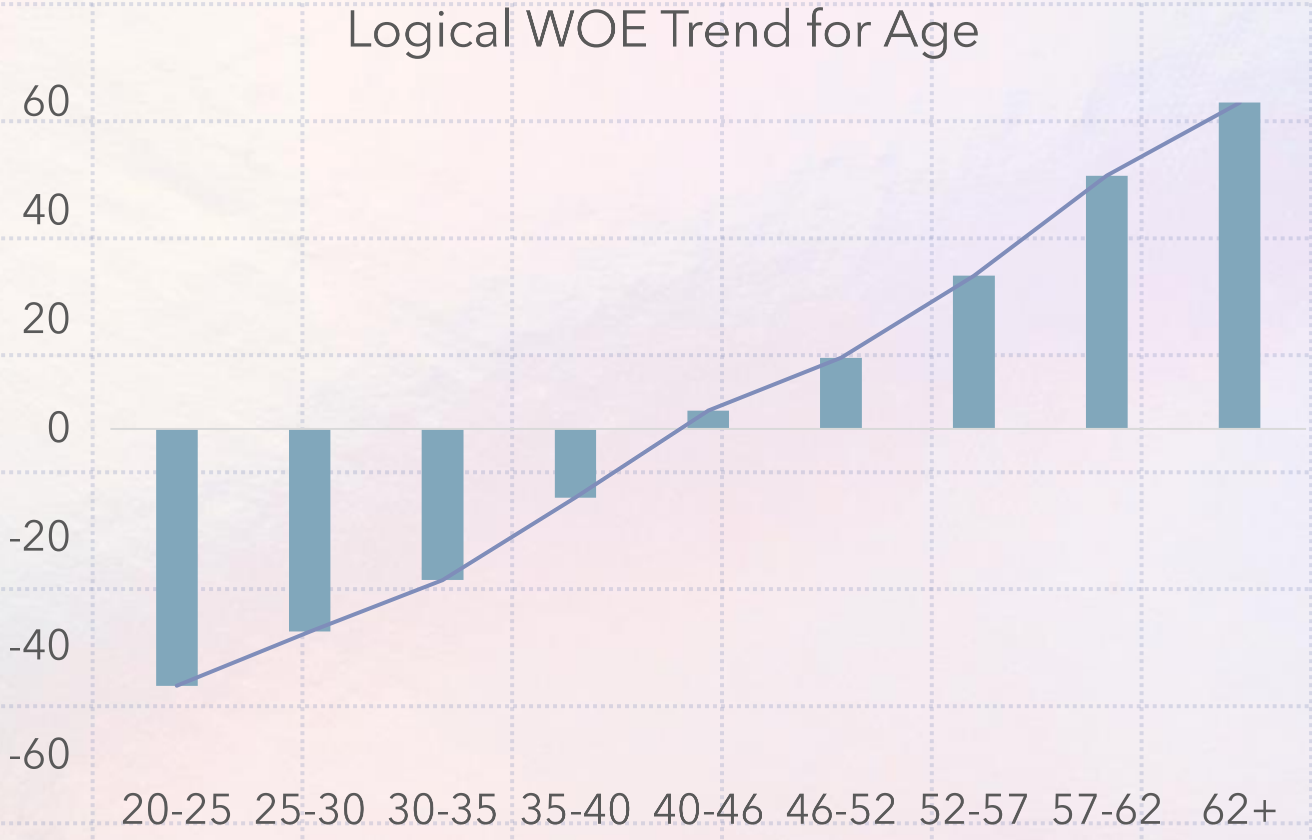
According to the calculated results and data segmentation, the Information Value (IV) is arranged in descending order as follows (the WOE calculations are presented in the next slides, and details are provided in the accompanying code section).

Variables	IV	Variables	IV
BUREAU_SCORE	0.53450	HAS_BEEN_REJECTED	0.02819
Age	0.09627	NAME_FAMILY_STATUS	0.02354
NEW_EMPLOY_TO_BIRTH_RATIO	0.08663	NAME_HOUSING_TYPE	0.01635
Time_at_bureau	0.08322	REGION_POPULATION_RELATIVE	0.01223
OCCUPATION_TYPE	0.08122	DEBT_TO_INCOME_RATIO	0.01043
credit_annuity_ratio	0.07063	avg_past_due_prev	0.00725
credit_goods_price_ratio	0.05870	AMT_public_records	0.00557
CODE_GENDER	0.04598	FLAG_OWN_CAR	0.00388
NAME_EDUCATION_TYPE	0.03976	FLAG_NEW_CUSTOMER	0.00172
PREV_IR	0.02903	CNT_CHILDREN	0.00161
COUNT_ACTIVE	0.02843	FLAG_OWN_REALTY	0.00142

2.3. VARIABLES ANALYSIS



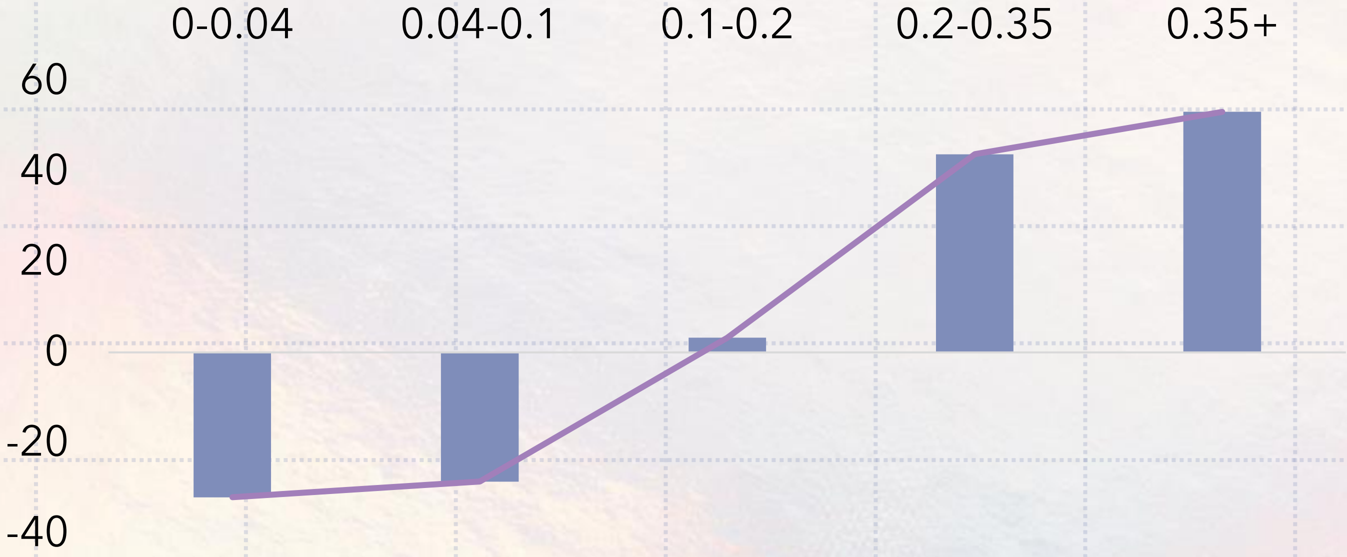
As can be seen clearly, groupings in this characteristic have a linear relationship with WOE; that is, they denote a linear and logical relationship between attributes in bureau score and proportion of bads. This confirms that people who have lower bureau score tend to be of a higher risk than the higher population.



Similarly to the bureau score, age also exhibits a logical relationship; however, the extent of its impact on the proportion of bads is not as significant as the bureau score.

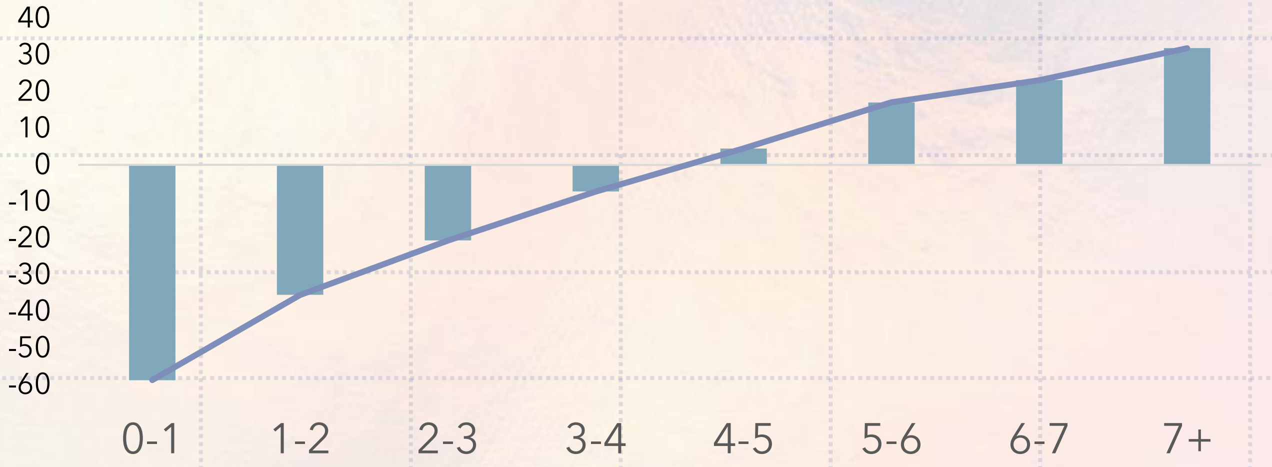
2.3. VARIABLES ANALYSIS

Logical WOE Trend for NEW_EMPLOY_TO_BIRTH_RATIO



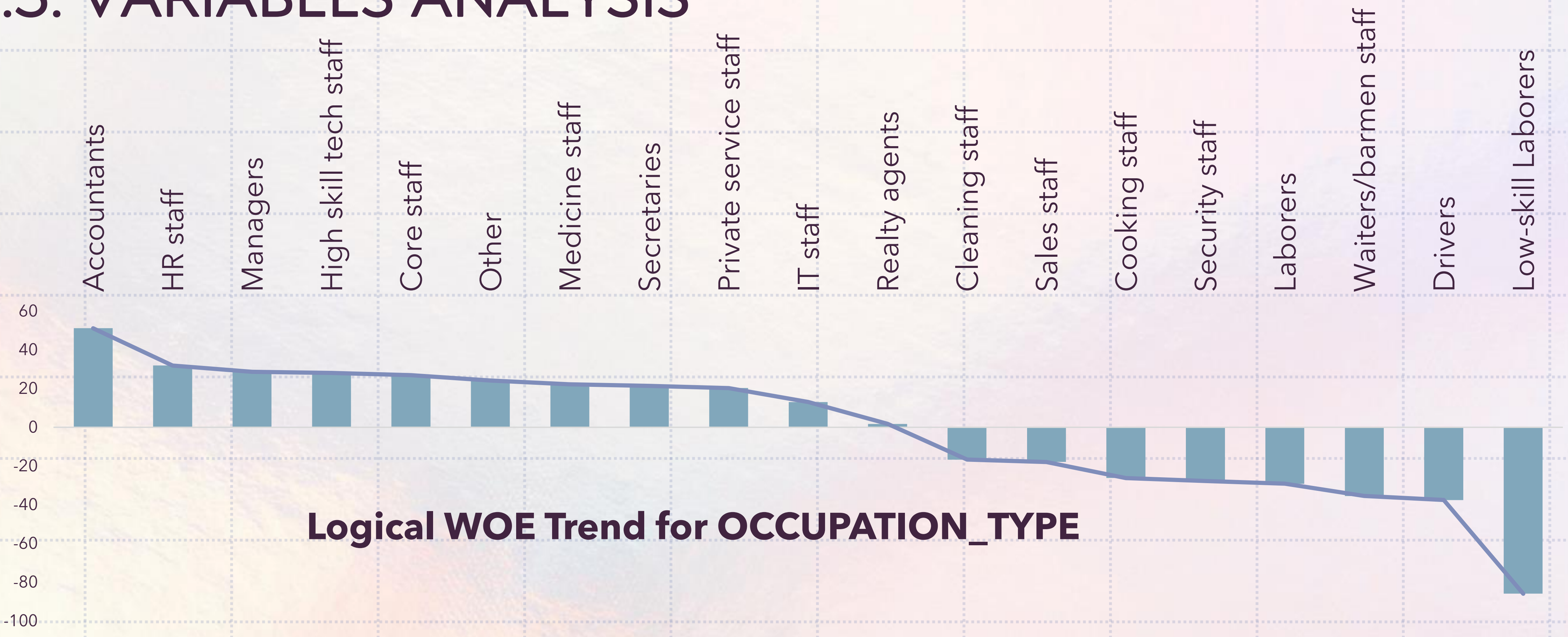
For the variable NEW_EMPLOY_TO_BIRTH_RATIO, it's worth noting that previously, I used a univariate regression to predict error values in the dataset. This may have influenced the unclear logic between the two ends of the variable. However, we can still observe that the ratio of working time to age is lower, indicating that a person with a shorter working time is at a higher risk of falling into a bad state

Logical WOE Trend for Time at bureau



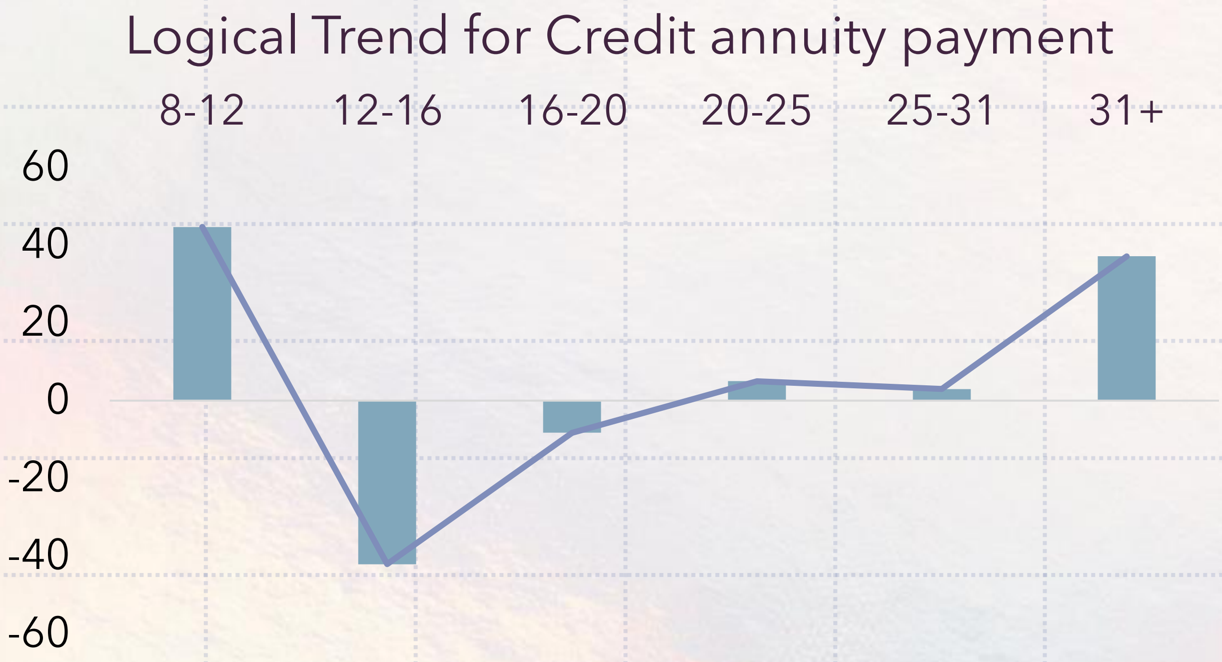
The longer the customer's tenure at the bureau, the better the customer's likelihood of performing well when taking out new loans. A clear linear relationship indicates this trend.

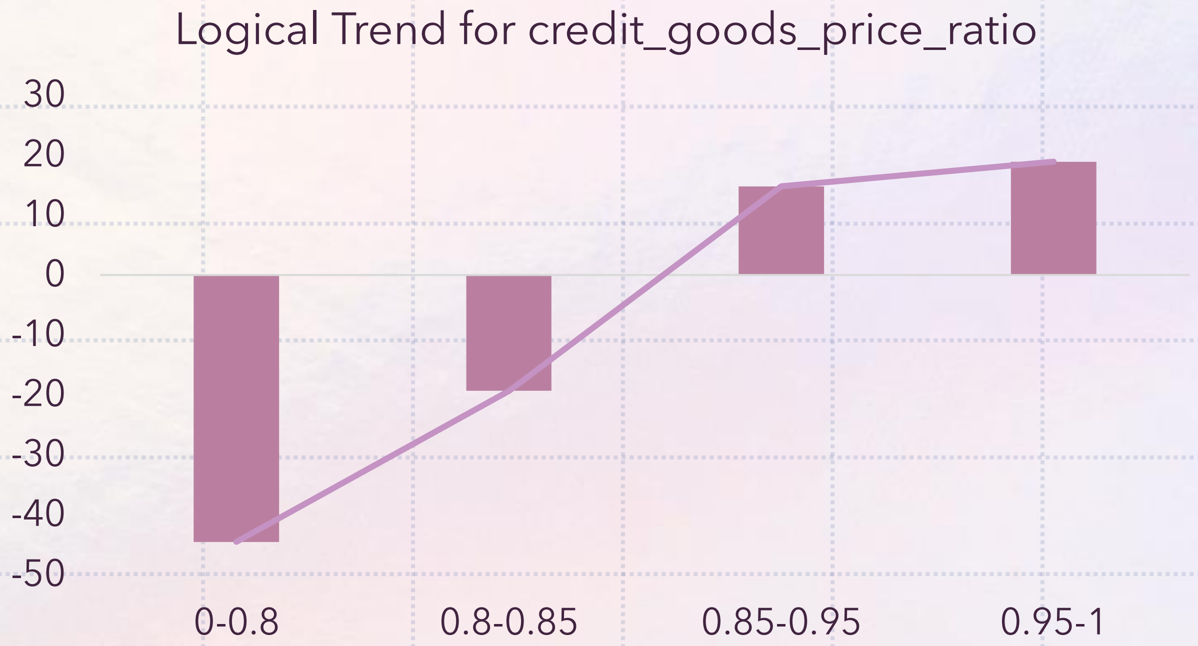
2.3. VARIABLES ANALYSIS



Occupation types play an important role in risk profile analysis. For professions requiring high expertise and offering high salaries, the likelihood of customers falling into bad debt is significantly lower. On the other hand, occupations with inherent instability, such as cleaning staff, sales staff, and waiters, pose higher challenges in debt repayment. Jobs that involve physical labor, like security staff and drivers, may not provide substantial income for comfortable debt repayment. Particularly noteworthy is the 'Low-skill Laborers' category, highlighted by a significantly negative WOE, indicating a higher risk of loan default. This strongly supports the notion that high-income professions tend to mitigate the risk of falling into bad debt for customers.

2.3. VARIABLES ANALYSIS





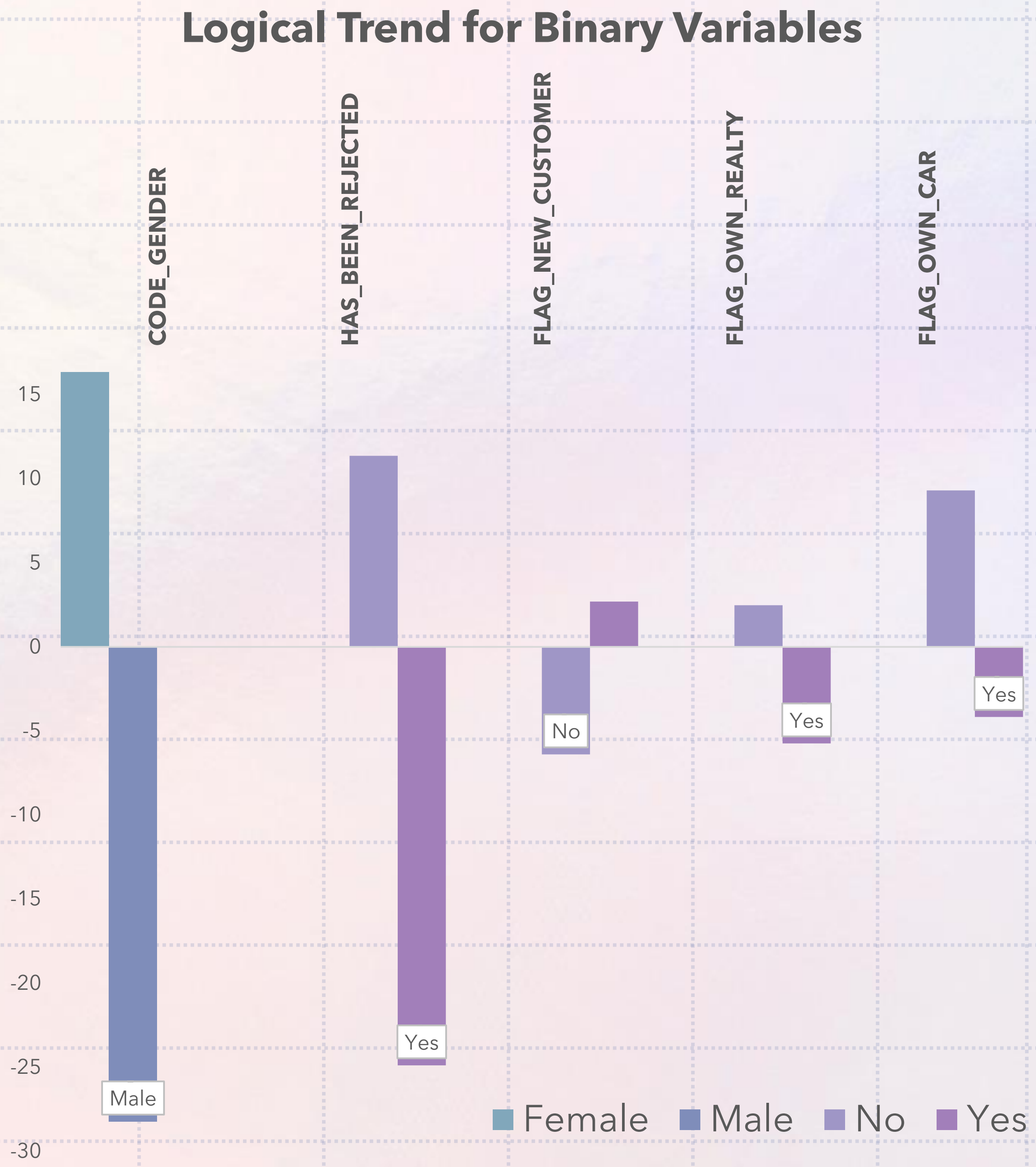
The number of repayments for previous loans ranging from 8 to 12 months indicates a more positive trend compared to other terms. Starting from December onwards, a higher number of repayment periods is associated with a lower risk of customers falling into bad debt.

If loans are for purchasing goods, a higher proportion of the loan amount being covered tends to signal a more positive outcome. On the other hand, when loans cover less than 80% of the value of the goods, there is a significant proportion of bads.

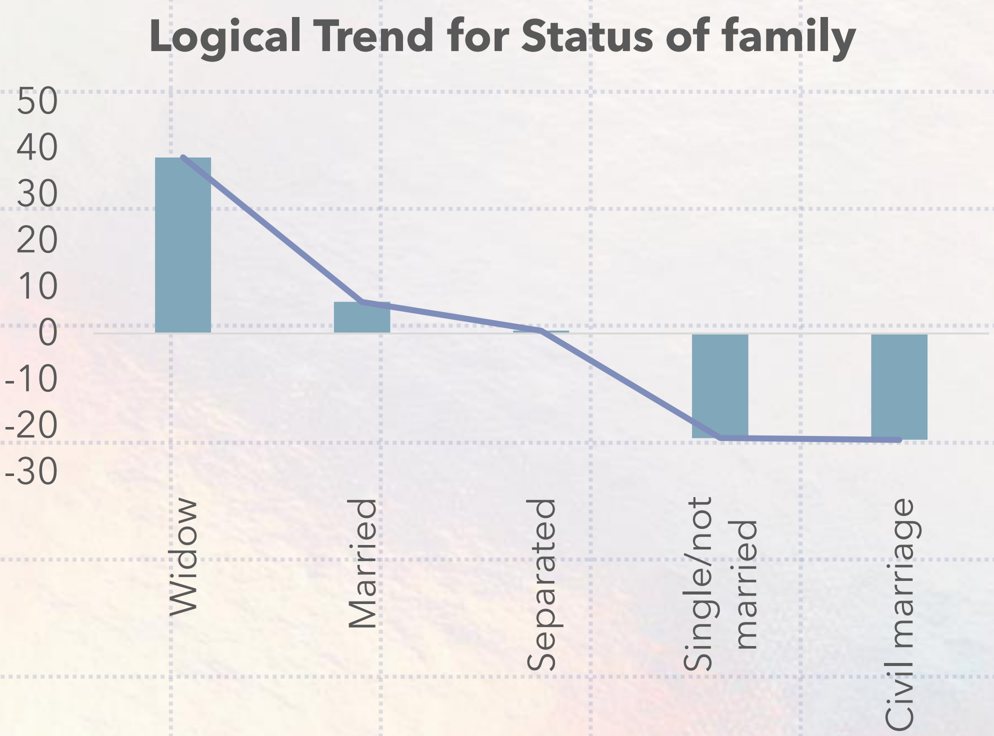
2.3. VARIABLES ANALYSIS

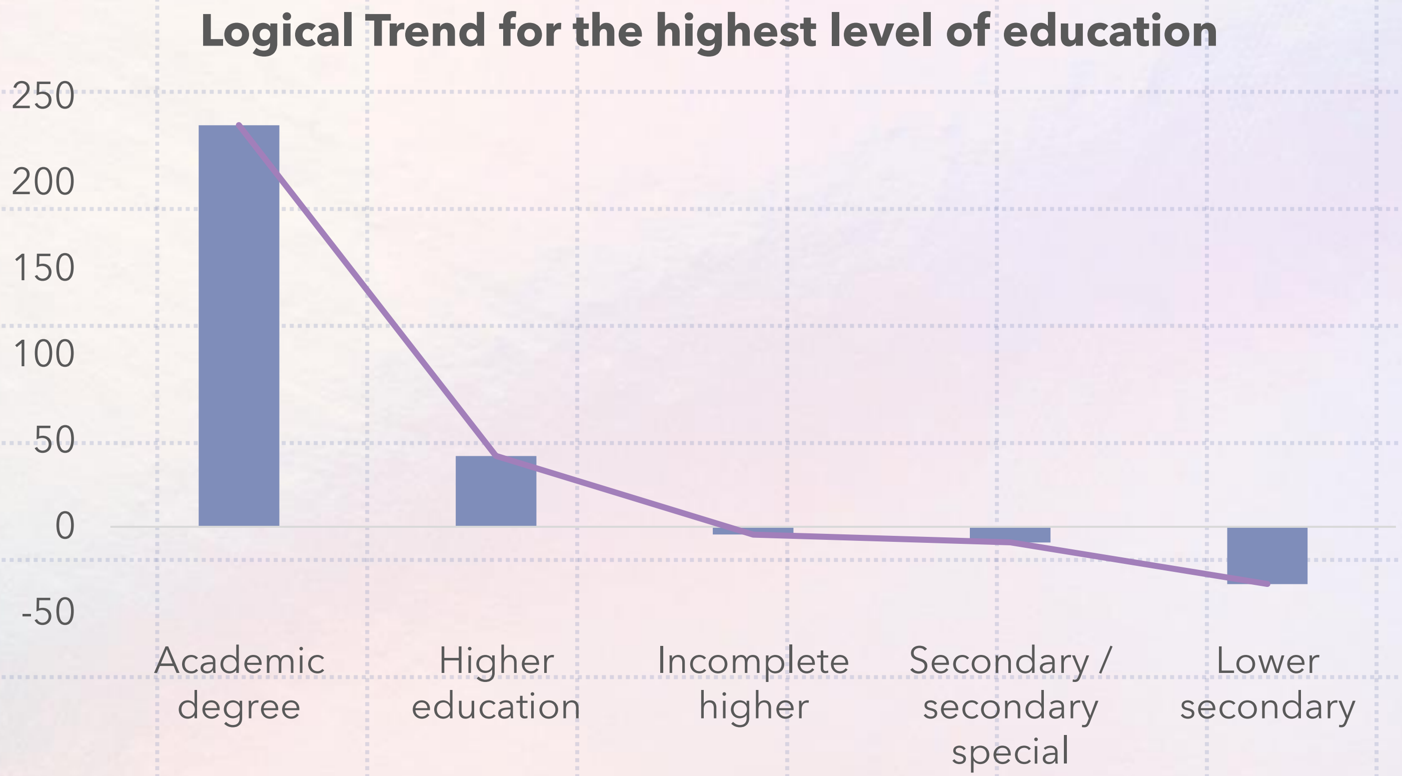
For binary variables, the observed trends are as follows: The trend of bad debt tends to concentrate on the following characteristics: male applicants and applicants who have been previously denied credit. This trend aligns well with business experience and is relatively easy to accept. However, a new finding is that customers who are not former clients of HomeCre tend to have a higher incidence of bad debt than new customers. This could be attributed to HomeCre's leniency in approving loans based solely on the factor of having been a previous customer, leading to potential negative outcomes (ethical risks).

Regarding two variables, whether the client owns real estate (FLAG_OWN_REALTY) and whether the client owns a car (FLAG_OWN_CAR), an unusual observation is that if a client owns both a house and a car, the proportion of bad outcomes is higher. Since the provided data only indicates whether the client owns a car (1 = yes, 0 = no) and owns a house or flat, this should be considered for regression analysis to properly model the relationships.



2.3. VARIABLES ANALYSIS

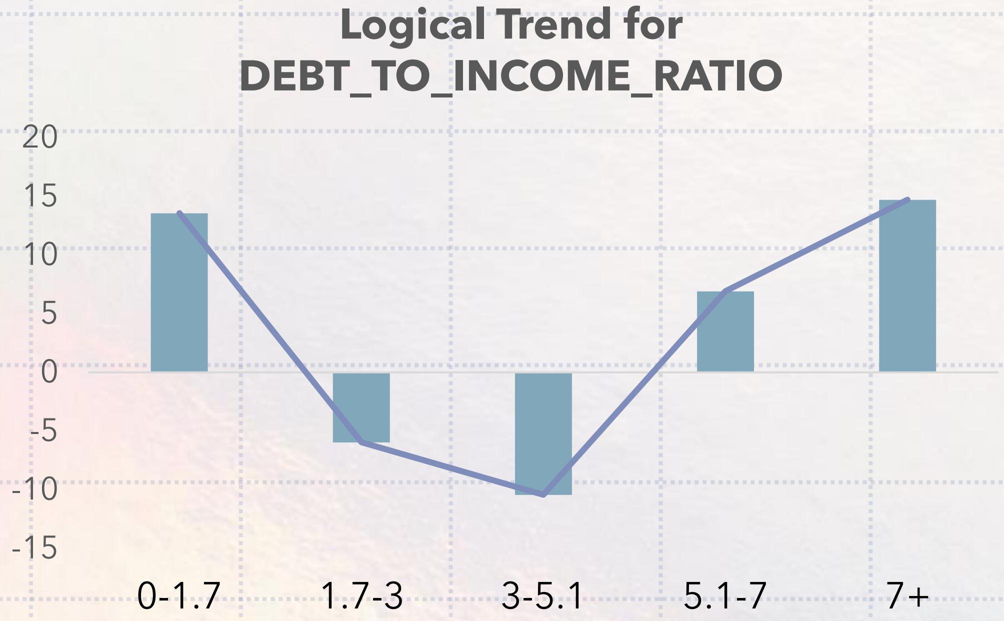




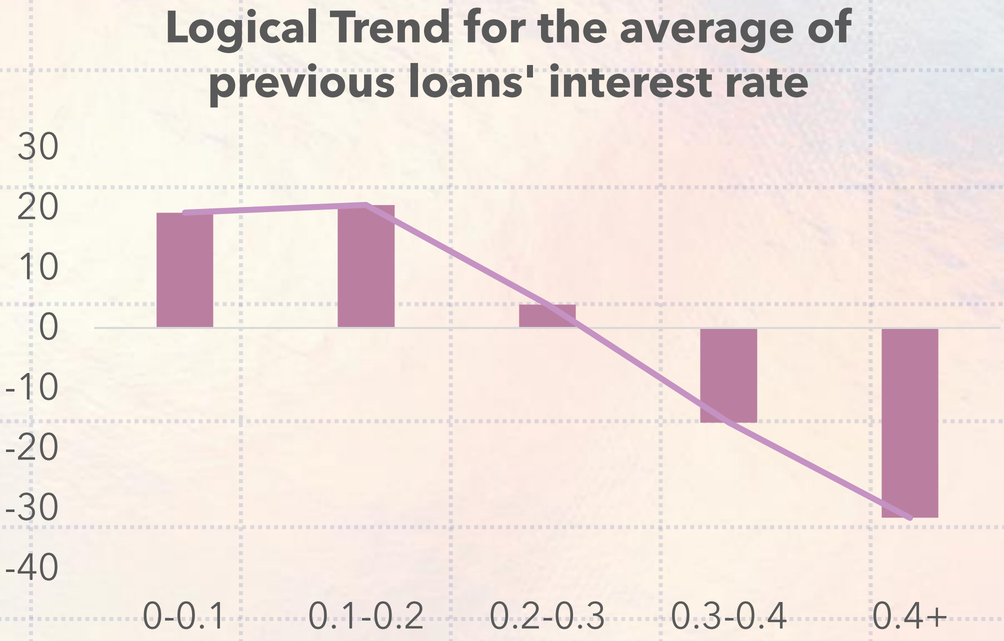
Customers who are single or divorced tend to pose an increased risk for loans, with divorced individuals being slightly less risky. Those who are married exhibit a higher level of safety. A noteworthy point is that widows represent the lowest risk among these categories.

Clearly, individuals with higher educational qualifications provide a sense of security for credit institutions when lending money. There is only one single case among the 115 instances of loans extended to individuals with an academic degree (nearly 100%). Following this, the safety level of loans gradually decreases as the educational qualification decreases from high to low.

2.3. VARIABLES ANALYSIS

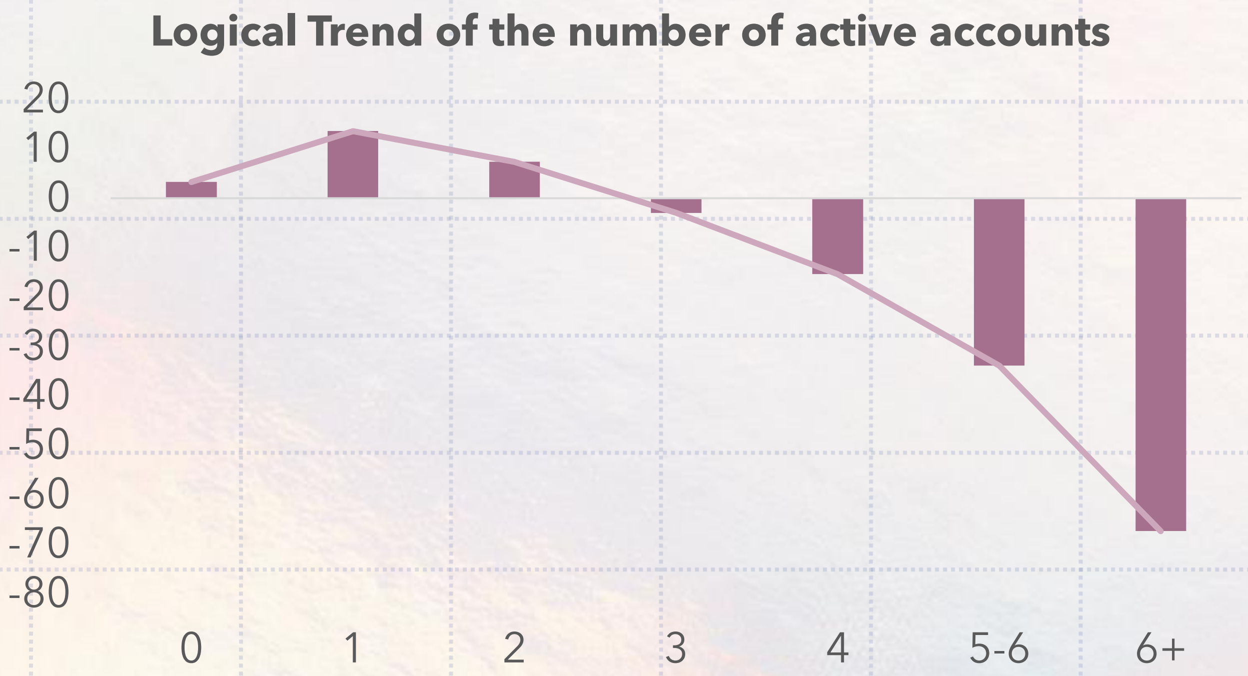


DEBT_TO_INCOME_RATIO is the ratio between AMT_CREDIT and INCOME. The definition does not specify whether INCOME is monthly or yearly and does not provide the unit of measurement, making it challenging to precisely interpret the relationship from DEBT_TO_INCOME_RATIO. Therefore, it is difficult to provide a clear explanation in this case.

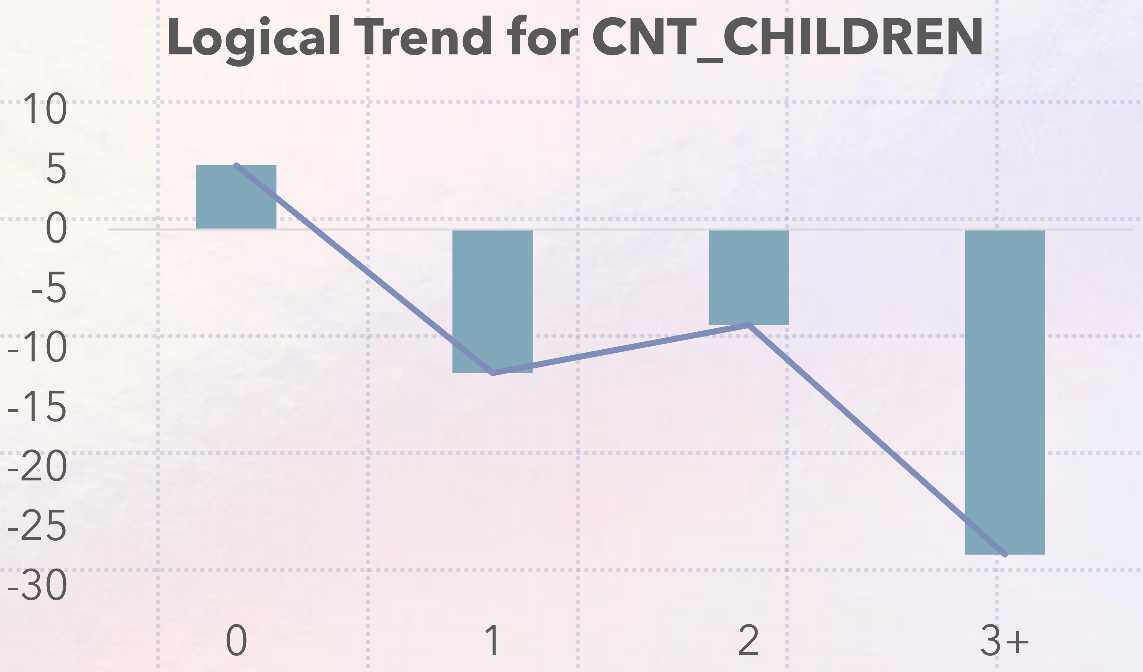


I have calculated the average interest rate for previous loans of applicants. It can be observed that at interest rates lower than 0.2, the proportion of bad outcomes is very low. As the interest rate increases, the proportion of bad outcomes also rises. The peak is reached when the interest rate is higher than 0.4, indicating the highest risk for customers.

2.3. VARIABLES ANALYSIS

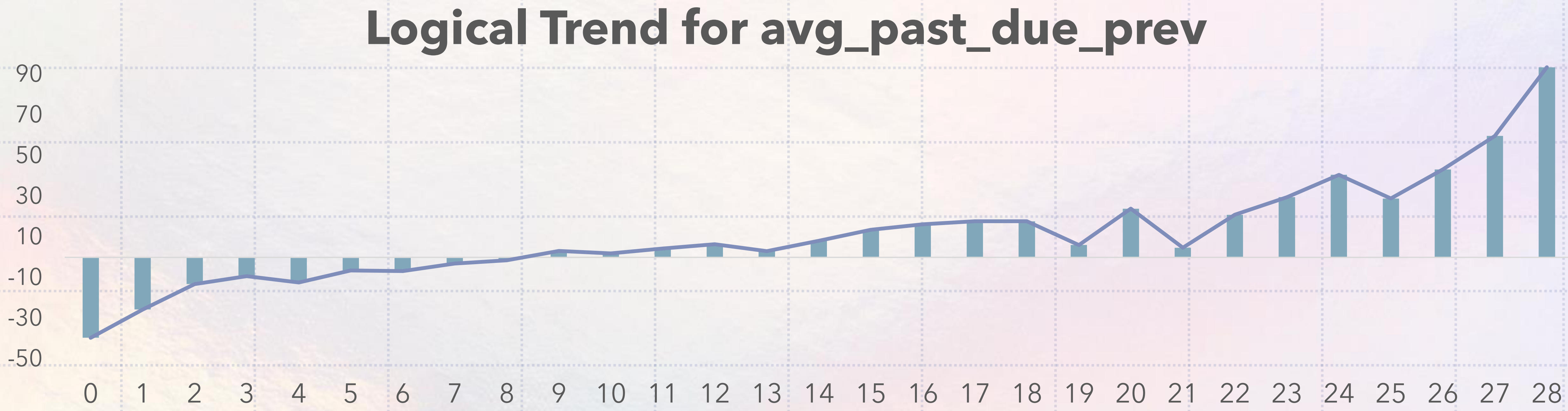


Customers with only one active account tend to have the highest ability to repay debt. As the number of accounts increases, it tends to reduce their transparency in the credit bureau. However, individuals with no active accounts also pose an increased risk when engaging in loans. This is because there is limited information about their spending habits without active accounts.



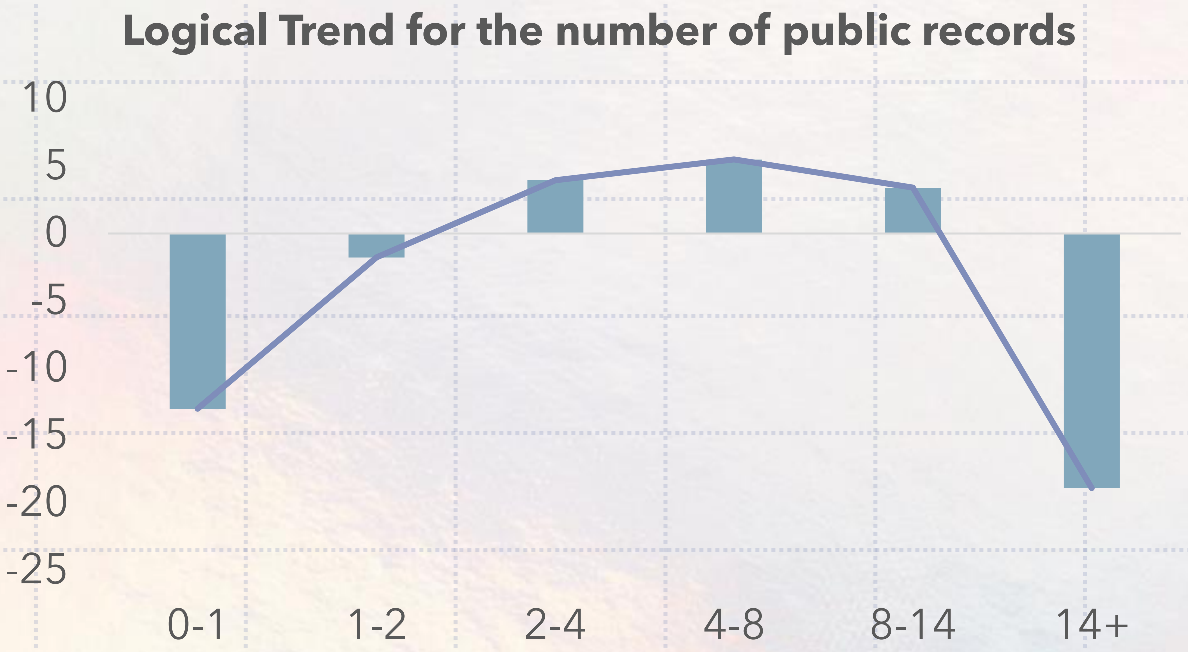
The more children a customer has, the more financial responsibilities they carry. Therefore, it is not difficult to observe that as the number of children increases, the customer's ability to repay debt tends to decrease, leading to an increased risk for the loan.

2.3. VARIABLES ANALYSIS

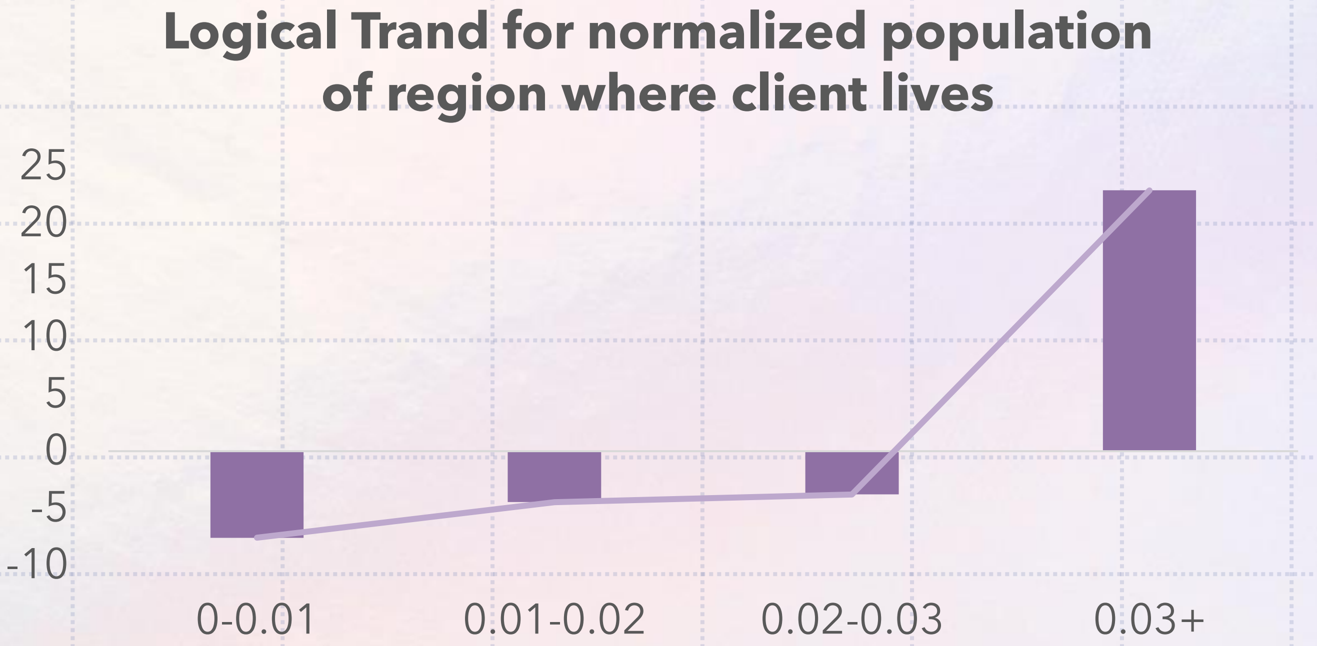


This is a variable I calculated by counting the installments that are past due. However, there is a paradox where the more installments are past due, the better the customer's ability to repay the debt. This seems inconsistent with business experience.

2.3. VARIABLES ANALYSIS

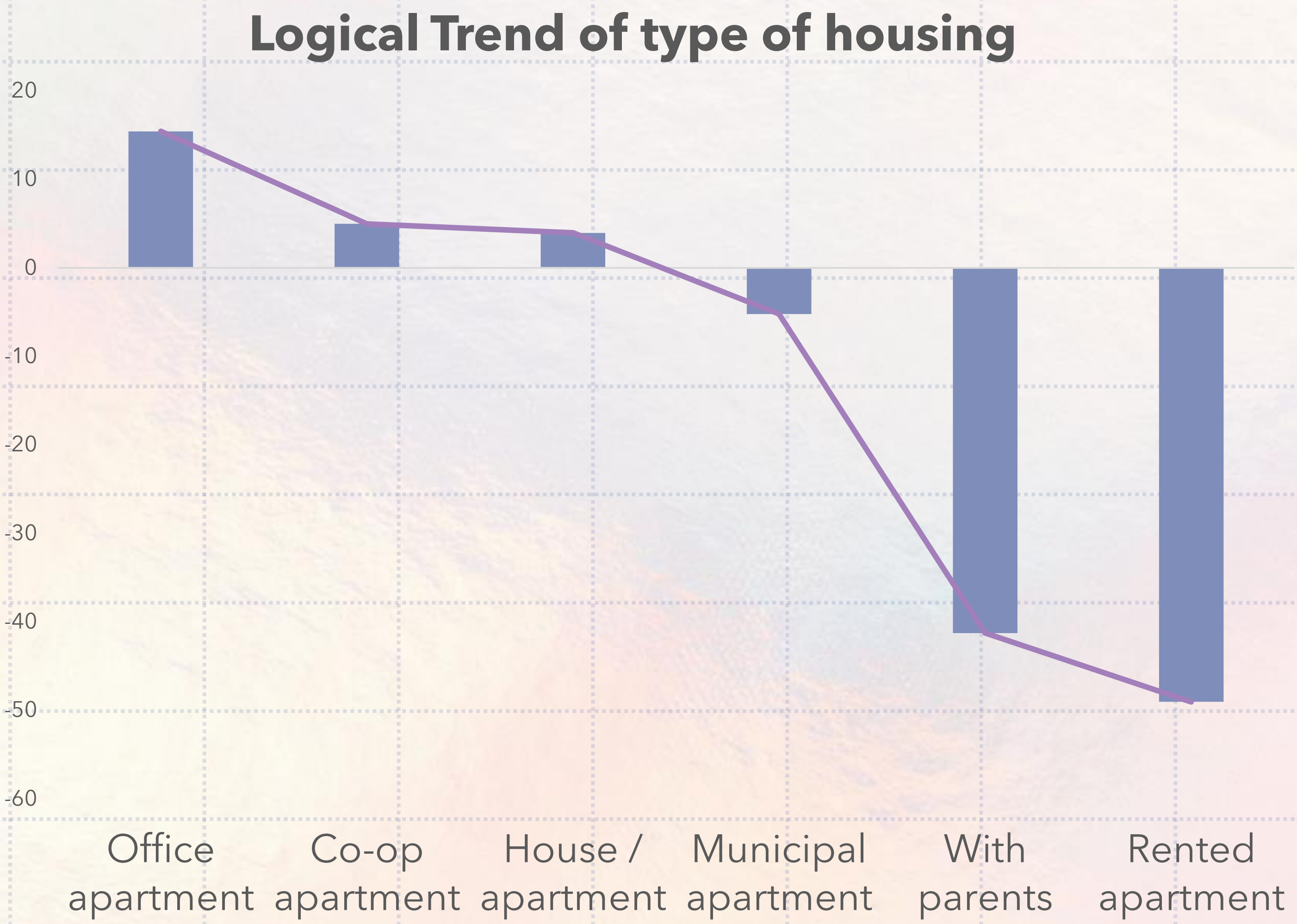


The more records are public, the lower the risk. However, having an excessive number of records also raises many questionable issues, which is evident in the increase in the proportion of bad outcomes.



According to business experience, it also suggests that areas with a higher population density can be assumed to be urban areas, where income levels are likely to be higher, and there are more opportunities for employment and business. This, in turn, helps increase the likelihood of customers repaying their debts. Conversely, in sparsely populated areas, which could be rural regions, the majority of the population may have lower incomes, potentially increasing the risk associated with loans.

2.3. VARIABLES ANALYSIS



Certainly, renters typically incur additional costs for meals and daily expenses, leading to higher overall expenditures. An even more notable discovery is that individuals living with their parents rank second in terms of debt repayment risk. This finding could be an interesting contribution that needs further observation, suggesting an assumption that these individuals may still be financially dependent or lack the ability to live independently.

2.3. VARIABLES ANALYSIS

Variables	Trend	Variables	Trend
BUREAU_SCORE	-	HAS_BEEN_REJECTED	+
Age	-	NAME_FAMILY_STATUS	+
NEW_EMPLOY_TO_BIRTH_RATIO	-	NAME_HOUSING_TYPE	+
Time_at_bureau	-	REGION_POPULATION_RELATIVE	-
OCCUPATION_TYPE	+	DEBT_TO_INCOME_RATIO	Unknown
credit_annuity_ratio	-	avg_past_due_prev	Unknown
credit_goods_price_ratio	-	AMT_public_records	-
CODE_GENDER	+	FLAG_OWN_CAR	Unknown
NAME_EDUCATION_TYPE	+	FLAG_NEW_CUSTOMER	-
PREV_IR	+	CNT_CHILDREN	+
COUNT_ACTIVE	+	FLAG_OWN_REALTY	Unknown

2.4. VARIABLES SELECTION

Based on Information Value (IV) to select suitable variables for our model. Additionally, I use the decision tree algorithm to determine feature importance. Therefore, I have two sets of features: 1) IV_features and 2) decision_tree_features.

Variables	IV
BUREAU_SCORE	0.53450
Age	0.09627
NEW_EMPLOY_TO_BIRTH_RATIO	0.08663
Time_at_bureau	0.08322
OCCUPATION_TYPE	0.08122
credit_annuity_ratio	0.07063
credit_goods_price_ratio	0.05870
CODE_GENDER	0.04598
NAME_EDUCATION_TYPE	0.03976
PREV_IR	0.02903
COUNT_ACTIVE	0.02843
HAS_BEEN_REJECTED	0.02819
NAME_FAMILY_STATUS	0.02354

Feature	Importance
BUREAU_SCORE	0.143989
NEW_EMPLOY_TO_BIRTH_RATIO	0.10304
Age	0.101757
PREV_IR	0.092996
DEBT_TO_INCOME_RATIO	0.086268
Time_at_bureau	0.070561
REGION_POPULATION_RELATIVE	0.060315
avg_past_due_prev	0.051768
credit_goods_price_ratio	0.045549
credit_annuity_ratio	0.044068
OCCUPATION_TYPE	0.035467
AMT_public_records	0.034781
COUNT_ACTIVE	0.028768
NAME_FAMILY_STATUS	0.020027

3. MODEL

We observe that the model achieves an accuracy of 91%. However, examining the confusion matrix reveals that only 35 out of 4495 observations with the target equal to 1 are correctly predicted. This highlights the impact of computing on an imbalanced dataset, where there are numerous instances of target = 0, while target = 1 comprises just over 9% (22226 obs) of the entire dataset.

```
Accuracy: 0.9161648651681191
Classification Report:
              precision    recall  f1-score   support

           0       0.92      1.00      0.96     48980
           1       0.59      0.01      0.02      4494

    accuracy                           0.92     53474
   macro avg       0.75      0.50      0.49     53474
weighted avg       0.89      0.92      0.88     53474

Confusion Matrix:
 [[48956    24]
 [ 4459    35]]
```

Consequently, I will implement a straightforward approach to bring the number of target = 0 closer to target = 1. I removed target = 0.

In practice, data imbalance is entirely normal as adverse cases are generally less frequent than positive cases. Our task is to address this issue in machine learning to achieve the best results. This also emphasizes that the accuracy of the model is just a reference number; more attention should be given to the confusion matrix to understand the essence of the dataset.

3. MODEL

After balancing the dataset, we observe a significant decrease in accuracy, now standing at only 67%. However, it's essential to note that the dataset used has shown improved performance in predicting adverse cases, which aligns with the project's objective of risk alerting. Accuracy, as a metric, may not be the sole indicator of model performance, especially in imbalanced datasets. The emphasis should be on the model's ability to effectively identify and predict the minority class, which, in this case, pertains to the instances of risk.

```
Accuracy: 0.6788887639185693
Classification Report:
              precision    recall  f1-score   support

           0       0.67      0.69      0.68      4421
           1       0.69      0.67      0.68      4470

    accuracy                           0.68      8891
   macro avg       0.68      0.68      0.68      8891
weighted avg       0.68      0.68      0.68      8891

Confusion Matrix:
 [[3051 1370]
 [1485 2985]]
```

IV_features

3. MODEL

Afterward, I ran an additional logistic regression model using the feature set selected by the decision tree's feature importance. The accuracy remained similar; however, the variables chosen by the decision tree exhibited better predictive signals for risk when forecasting a higher number of observations with a value of 1. This suggests that the feature set derived from the decision tree's feature importance might be more effective in capturing the characteristics associated with instances of risk in the dataset. It underscores the importance of considering feature relevance and selection techniques tailored to the specific nature of the problem at hand.

```
Accuracy: 0.6761894050163086
Classification Report:
              precision    recall  f1-score   support

           0       0.67      0.68      0.68      4421
           1       0.68      0.67      0.68      4470

    accuracy                           0.68      8891
   macro avg       0.68      0.68      0.68      8891
weighted avg       0.68      0.68      0.68      8891

Confusion Matrix:
 [[3009 1412]
 [1467 3003]]
```

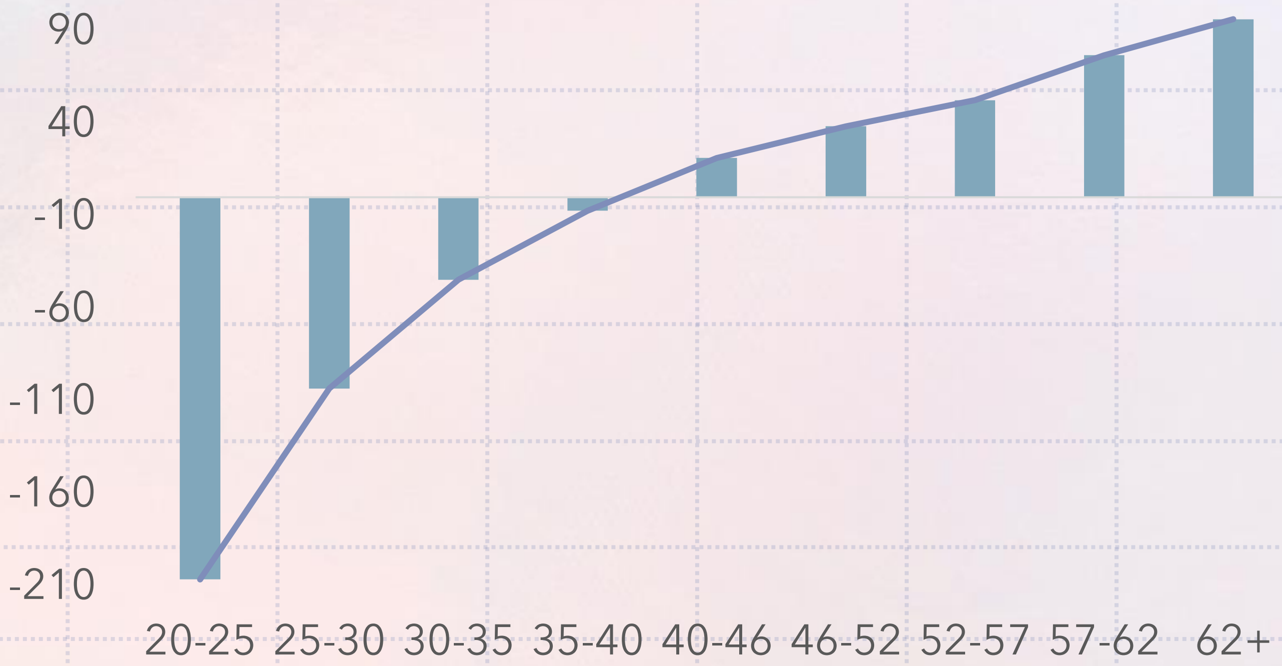
Feature importance decision tree

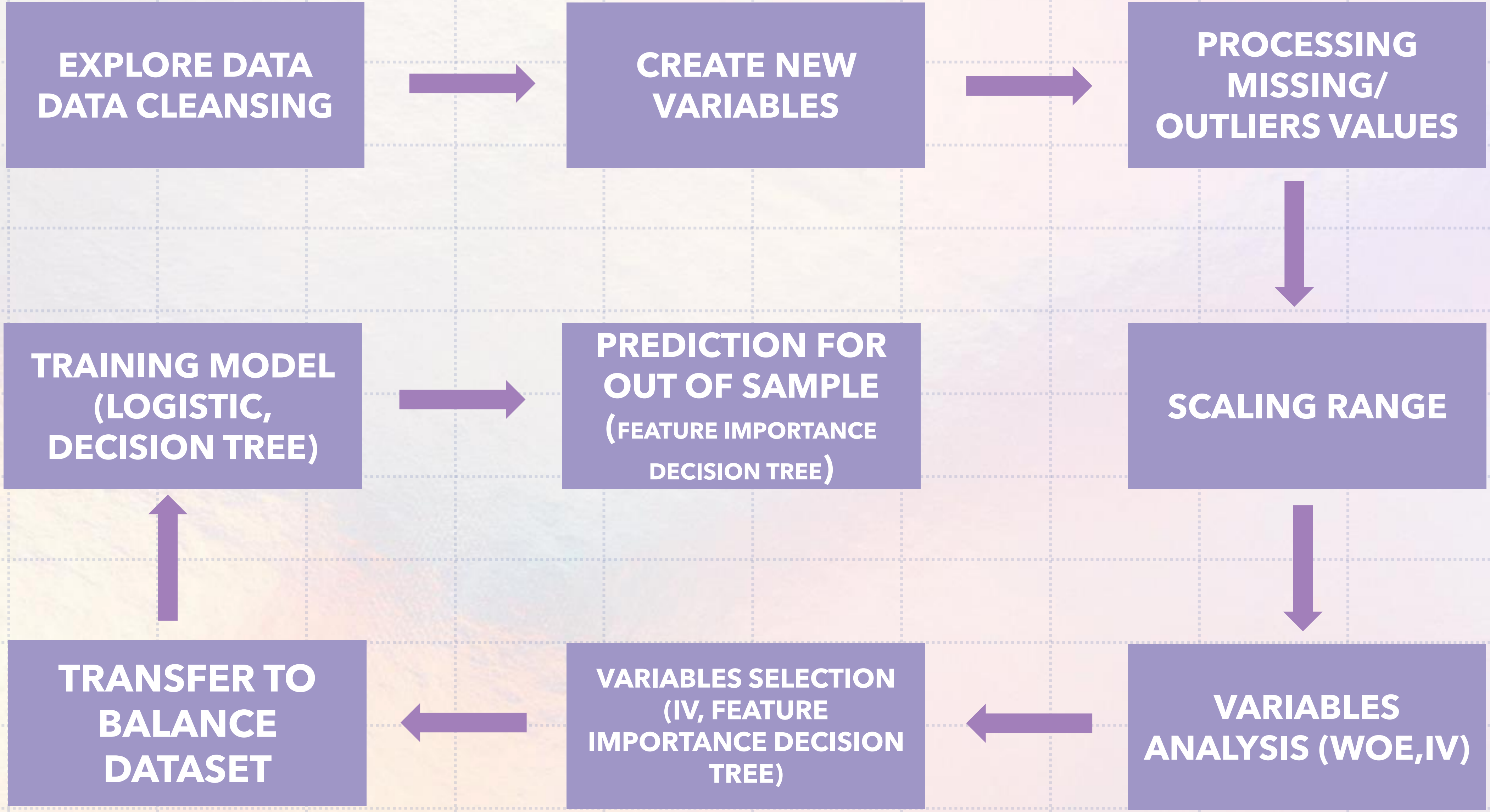
4. PREDICTION FOR OUT OF SAMPLE

BUREAU_SCORE_BINNED	Count	Bads	Goods	Tot Distr	Distr Good	Distr Bad	Bad Rate	WOE	IV
(0.0, 0.3]	3732	3703	29	8.623	0.001	0.246	0.992	-547.675	1.340
(0.3, 0.4]	5946	5250	696	13.738	0.025	0.348	0.883	-264.778	0.857
(0.4, 0.5]	10076	4830	5246	23.280	0.186	0.321	0.479	-54.452	0.073
(0.5, 0.6]	11911	1239	10672	27.520	0.378	0.082	0.104	152.618	0.452
(0.6, inf]	11617	47	11570	26.840	0.410	0.003	0.004	487.888	1.986
Total	43282	15069	28213	1.000	1.000	1.000	0.348	0.000	NaN

Model Logistic with feature importance decision tree was used for prediction beyond the sample because it can forecast risks better, with fewer Type 1 errors. The prediction results are in the Excel file named "final_destination." Below are some charts illustrating the relationship between the predicted target variable and various features:

Logical WOE Trend for Age out of sample




EXPLORE DATA
DATA CLEANSING
CREATE NEW VARIABLES
PROCESSING MISSING/ OUTLIERS VALUES
SCALING RANGE
VARIABLES ANALYSIS (WOE,IV)
VARIABLES SELECTION (IV, FEATURE IMPORTANCE DECISION TREE)
TRANSFER TO BALANCE DATASET
TRAINING MODEL (LOGISTIC, DECISION TREE)
PREDICTION FOR OUT OF SAMPLE
(FEATURE IMPORTANCE DECISION TREE)