Home Credit Default Risk

Problem Statement:

Home Credit is an organization that serves the unbanked population with access to loans. Such individuals that do not have a built-up credit score have a challenging time securing loans from financial institutions.

Home Credit strives to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience. In order to make sure this underserved population has a positive loan experience; Home Credit makes use of a variety of alternative data--including telco and transactional information--to predict their clients' repayment abilities

So, problem statement would be predicting how likely each applicant is of repaying a loan?

Dataset Description:

There are seven data sets that are at the disposal of Home Credit:

Application_train.csv - this is the principal table and presents all the application information. There is a single row per application, which has a unique identifier.

previous_application.csv - this file presents previous applications for people in the sample through Home Credit. There is a row per each application.

installments_payments.csv - this is the repayment history on loans given out through Home Credit for people in the sample. Each row is a made or missed payment.

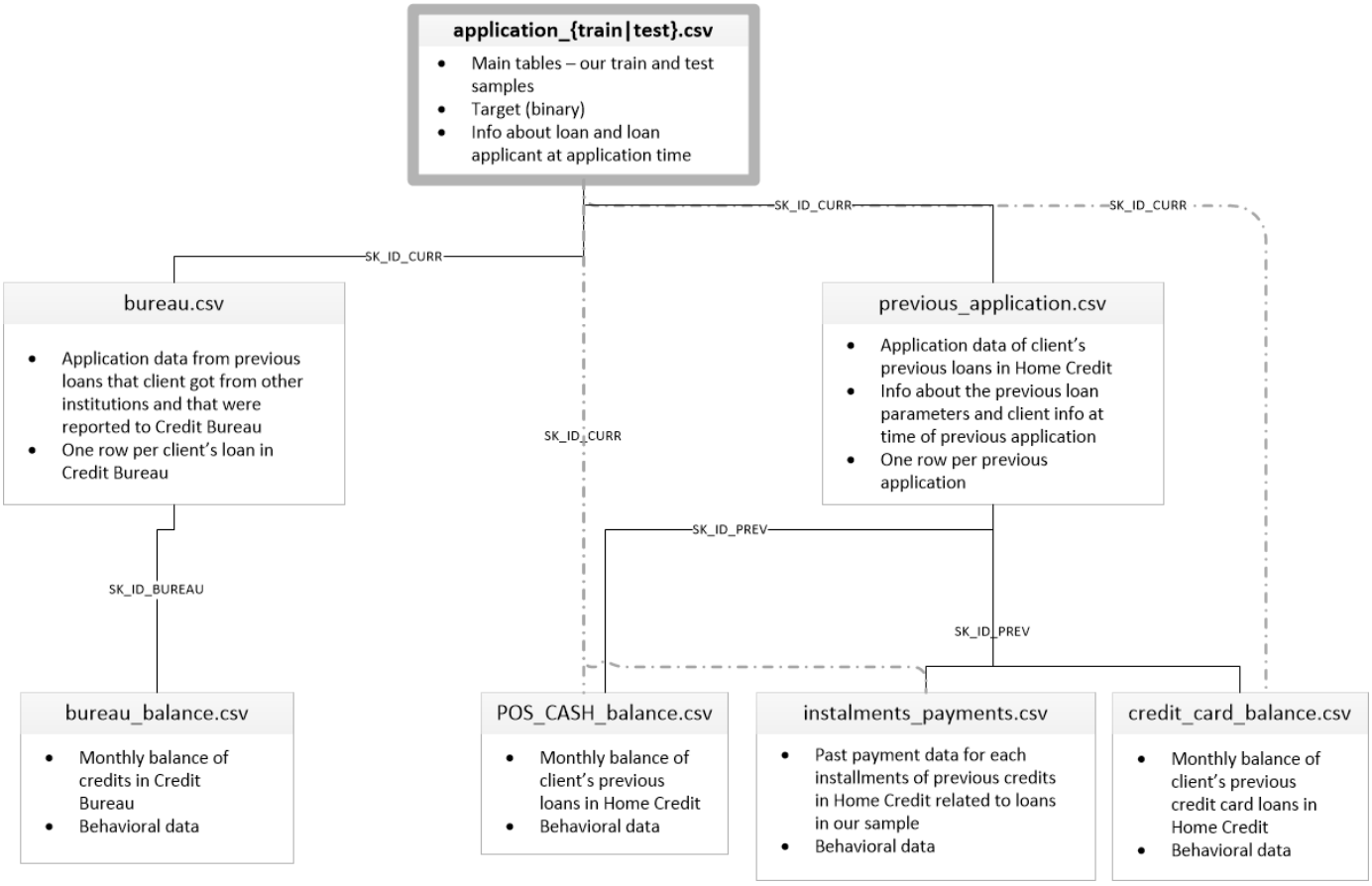
bureau.csv - credit information from other financial institutions that were reported to Home Credit. Each row represents a credit that was given to an individual in the sample.

bureau_balance.csv - monthly information per credit, per loan for users in the sample..

POS_CASH_balance.csv - like bureau_balance.csv, this data set is the internal version of the previous monthly breakdown of balances for consumer credit and cash loans that were taken out through Home Credit.

credit_card_calance.csv - each row in this data set represents a monthly balance of credit cards that were issued to applicants in the sample through Home Credit.

Links between the Data sets:



Data Wrangling and Cleaning:

A) **bureau_balance.csv :**
   - Analyzed dataset with few rows & identified categorical column
   - "STATUS " column being used to denote "**Status of Credit Bureau loan during the month (active, closed, DPD0-30,… [C means closed, X means status unknown, 0 means no DPD, 1 means maximal did during month between 1-30, 2 means DPD 31-60,… 5 means DPD 120+ or sold or written off])**" which typically denotes on which month Paid , Not paid , closed . This column dropped & instead dummy variables created for Numerical Conversion.
   - Then it summing based on '**SK_ID_BUREAU**' , which will be representing '**Month_Balance_Count**' . Also, we are dropping '**MONTHS_BALANCE**' original column
   - Grouped Data set is ready for Merge

B) **bureau.csv**
   - Transforming Categorical variables to Numerical variables & Dropping actual columns
   - Merging with **bureau_balance** dataset based on '**SK_ID_BUREAU'**
   - Now Merged Data set is ready

C) credit_card_balance.csv

- It has Each month credit record
- Transforming Categorical variables to Numerical variables & Dropping actual columns
- Step 2: Creating new unique DF for **'SK_ID_PREV'** & **'SK_ID_CURR'**
- Dropping **''SK_ID_CURR''** from **credit_card_balance** data frame**.**
- Grouping DF based on **'SK_ID_PREV' –** Summing past Installement payment information and grouping by previous ID.
- Merging with Step 2 DF, based on **'SK_ID_PREV'**
- So it has unique **'SK_ID_PREV'** & **'SK_ID_CURR'** with all data's Merged together

D) previous_application.csv

- Transforming Categorical variables to Numerical variables & Dropping actual columns

E) POS_CASH_BALANCE.csv

- Transforming Categorical variables to Numerical variables & Dropping actual columns

F) installments_payments.csv

- Transforming Categorical variables to Numerical variables & Dropping actual columns
- Dropping **''SK_ID_CURR''** from **installments_payments** data frame**.**
- New Variable Created to know whether Payment Made on date or Late
- Grouping DF based on **'SK_ID_PREV' –** Summing past Installement payment information and grouping by previous ID.
- Merging with Step 2 DF, based on **'SK_ID_PREV'** from **credit_card_balance.csv**
- So, it has unique **'SK_ID_PREV'** & **'SK_ID_CURR'** with all data's Merged together

General Merging Strategy:

- Dropping **''SK_ID_CURR''** from **installments_payments, pos_cash_balance, cc_balance** data frame**.**
- **previous_application** & **installments_payments** Merged based on **'SK_ID_PREV'**
- Above Merged to **cc_balance** based on **'SK_ID_PREV'**
- 1-1 Mapping enabled for **'SK_ID_PREV'** & **'SK_ID_CURR'.** Then Merged with above 4 data set
- So Right side of above Image is completely Merged.
- All Previous Data is Grouped By **'SK_ID_CURR'** then dropped **'SK_ID_PREV',** to merge with training data set
- Merged Previous Data sets & bureau grouped datasets with training data set through '**SK_ID_CURR'**
- All data sets Merged with Training Data set. Merging Process completed.

Missing Variables Handling:

- There are 283 columns with missing variables out of 339 columns in the data frame.
- Above 35 % Missing variables columns dropped as it is not going to impact predictions.

- np.isfinite() Method being used to drop few rows from the data set which depends on 'late' & 'closed' Which is being referred from Missing variable % table . which shares same % of missing values.
- **'OCCUPATION_TYPE'** is important feature though it has 31 % of missing values, so NAN marked as Unemployment
- Rest of below 10 % missing variables being filled with wither 0 or Mean, based on feature reference from excel.

After Handling Missing variables, data set stored to CSV as a single Merged data set .

Outlier Handling:

- Have not gone through complete data set columns, so looked at few important columns
- DOB has not any outliers
- Days Employed has outlier, it's been handling after divided by -365 then > 0's will be marked & masked as 0.
- Could see few outliers in '**AMT_INCOME_TOTAL'** it' been handled by 3 standard deviations of the mean.

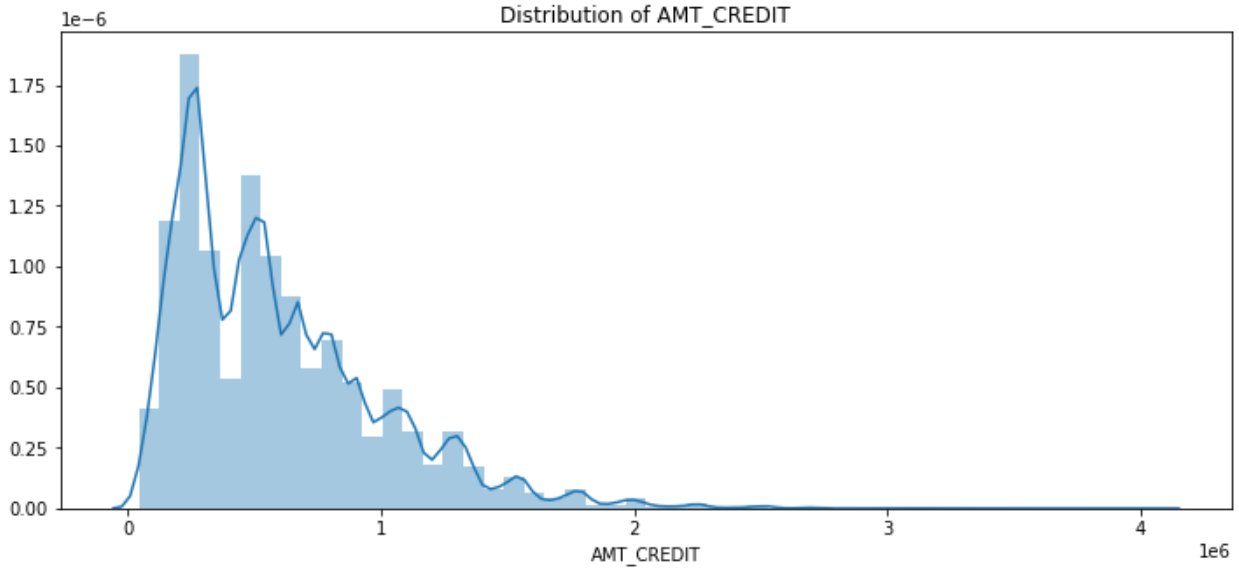
Application_train.csv – Preparation:

- Identified categorial variables
- Transforming Categorical variables to Numerical variables & Dropping actual columns.
- Replaced all negative values to 0
- X_train & y_train identified for feature selection
- Training set and validation set are split in following percentages: 66.66% : 33.33%.
- Top 10 features identified based on '**mutual_info_classif'**

EDA With Application train data set:

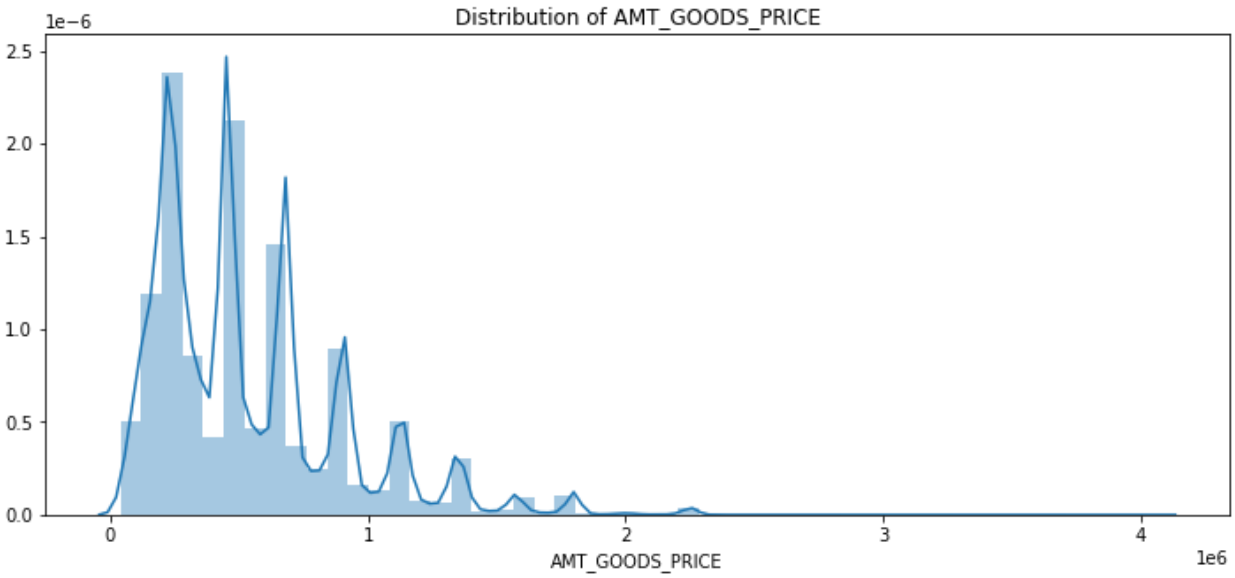
Data Exploration:

1. Distribution of Amount Credit



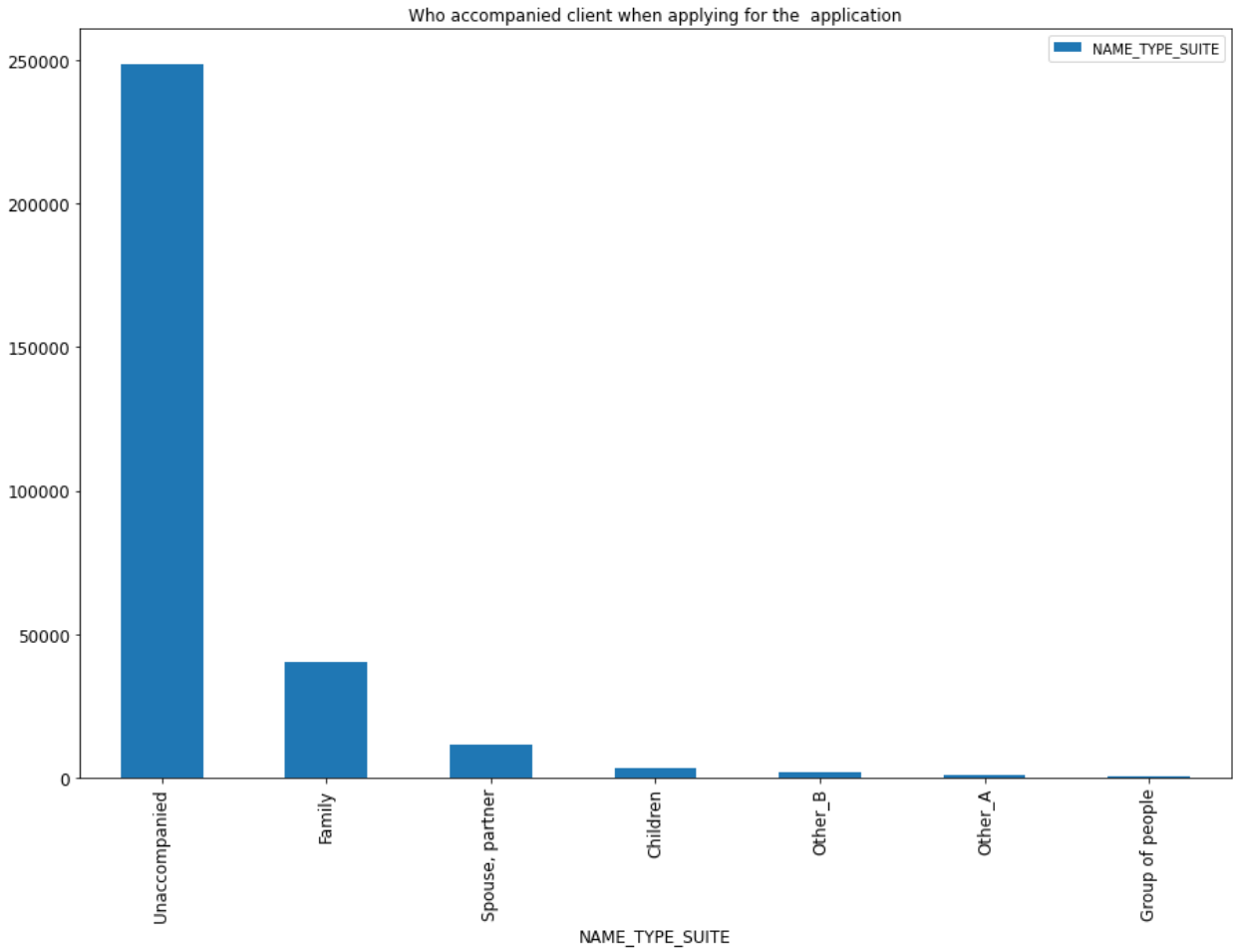
Distribution is right side skewed, between 0 . 1,50000 has more entries

2. Distribution of Amount Goods Price



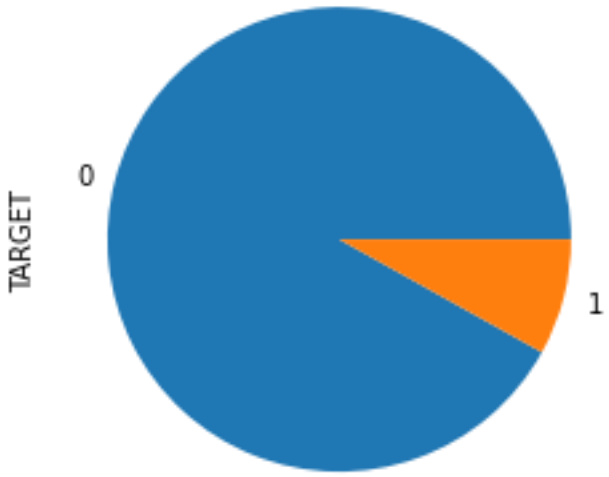
Majority of the amount of goods price spreaded between 0-1.5

3. Who accompanied client when applying loan ?



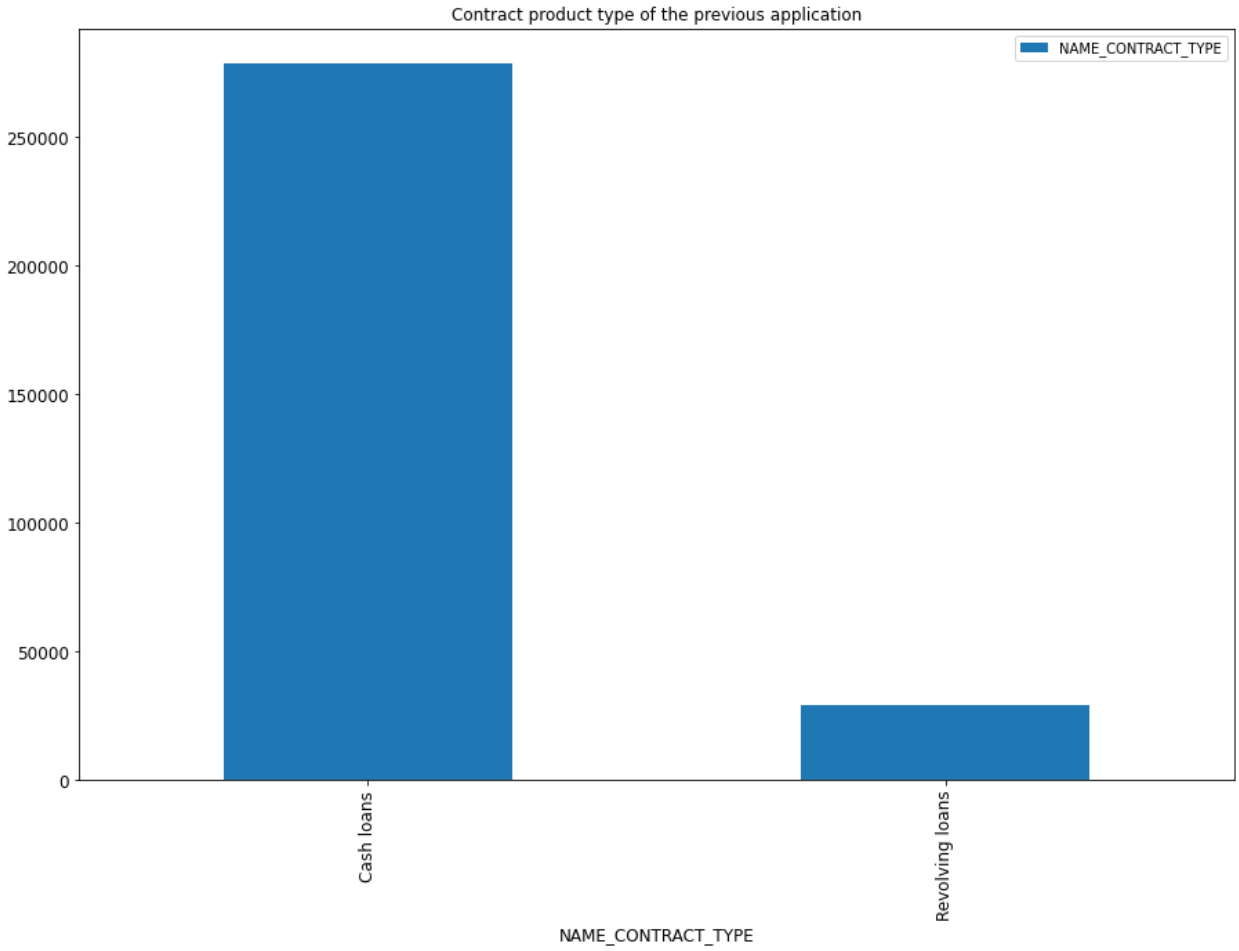
Majority of the applicants are unaccompanied

4. Highly imbalanced data!



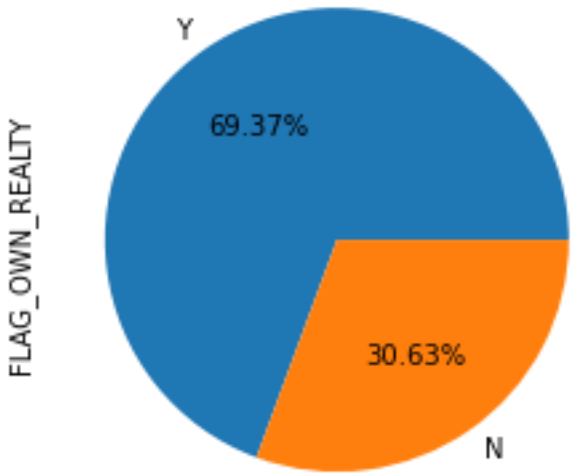
As we can see data is highly imbalanced.

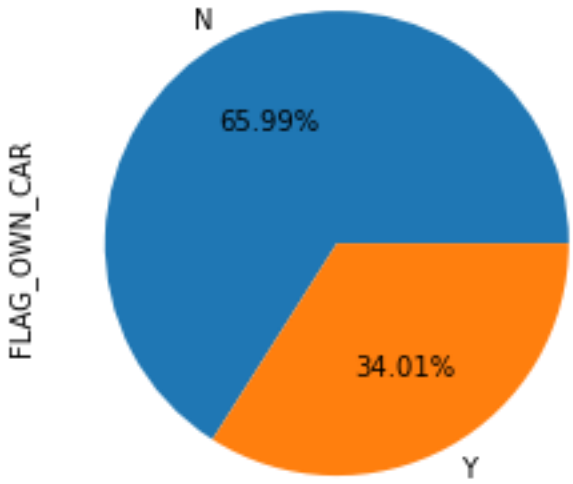
5. Contract Type of Previous Loan app :



Most of the loans are Cash loans which were taken by applicants.

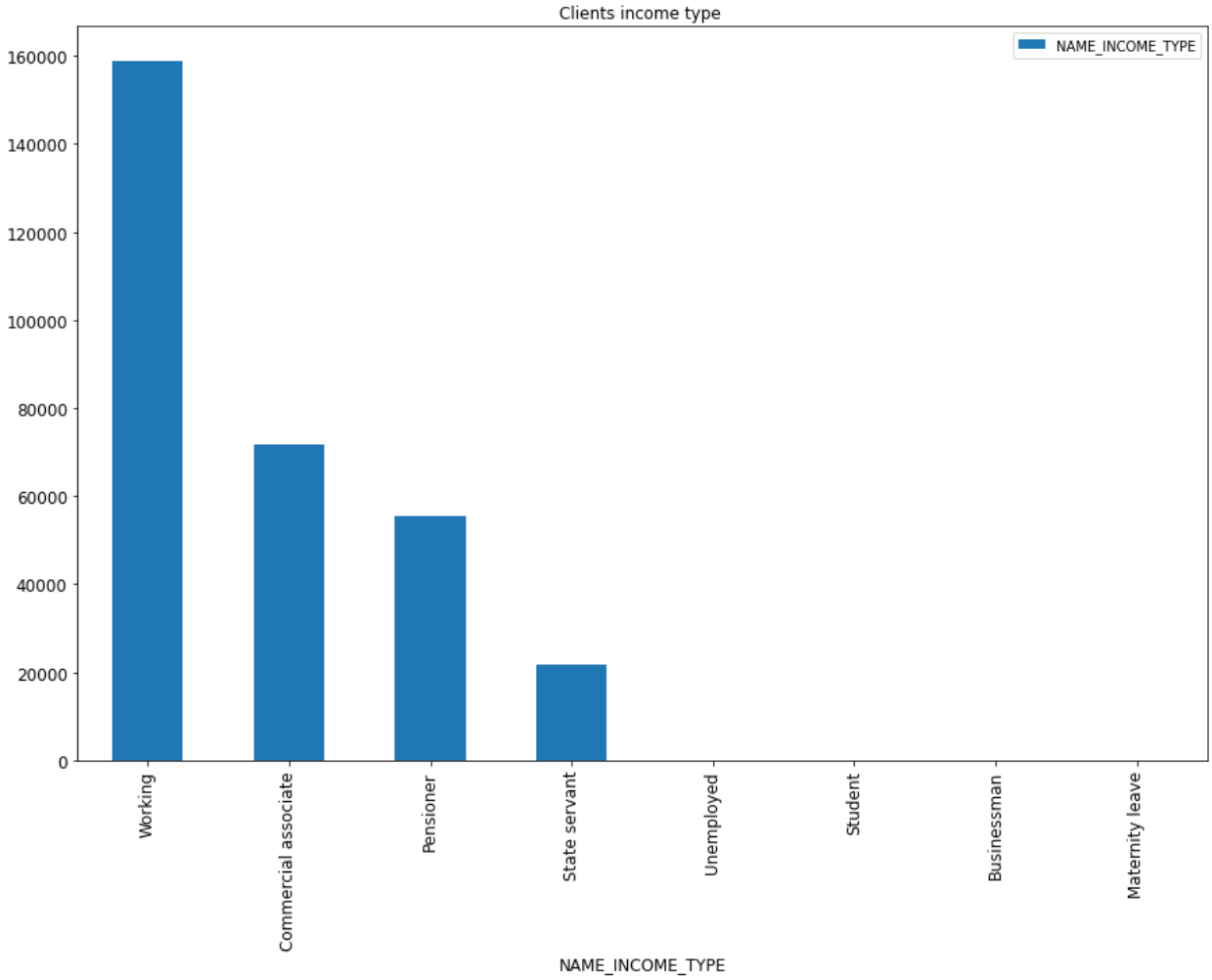
6. Own Relaty & Own Car





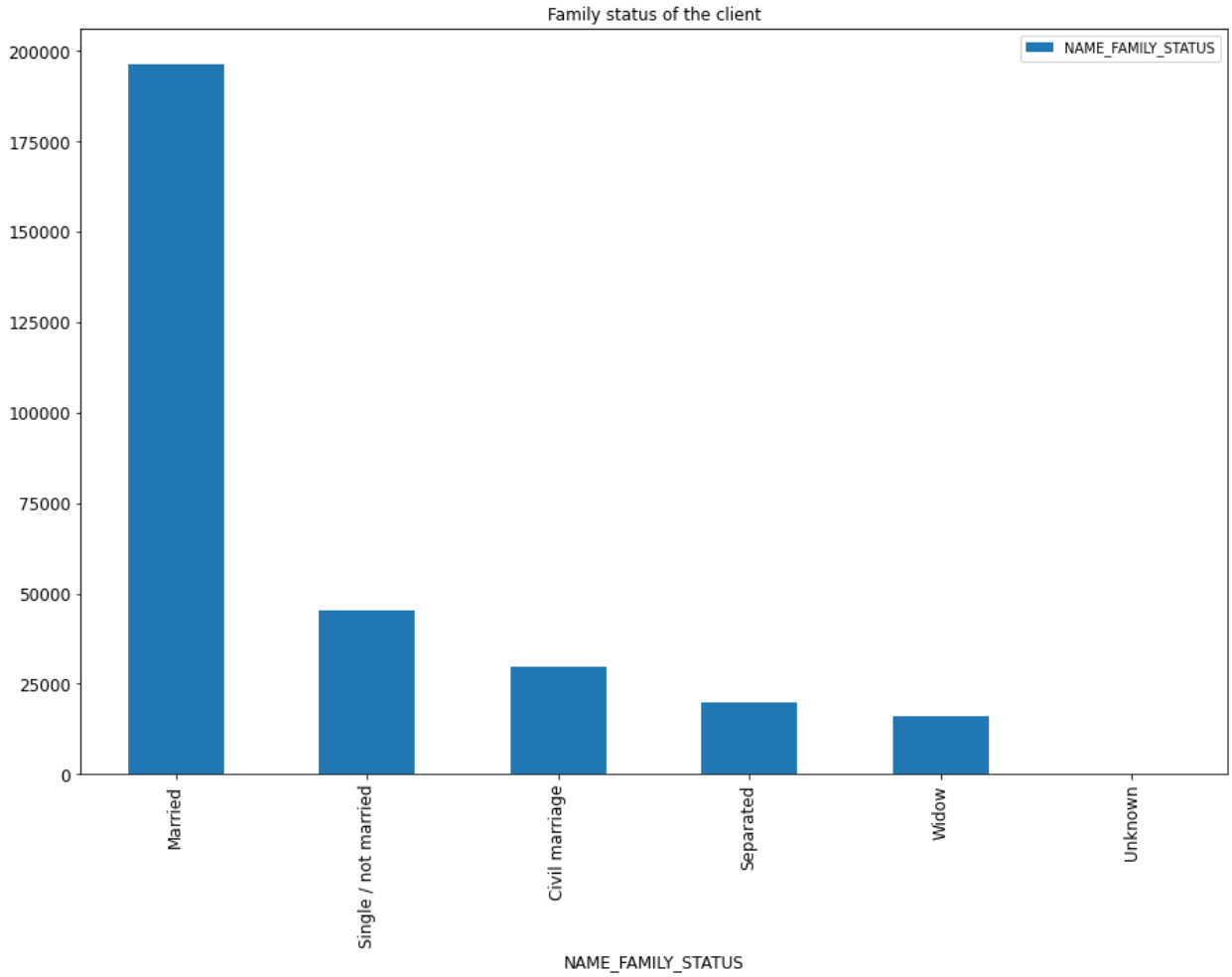
70 % applicants has own realty & 65 % has own car

7. Client's Income Type :



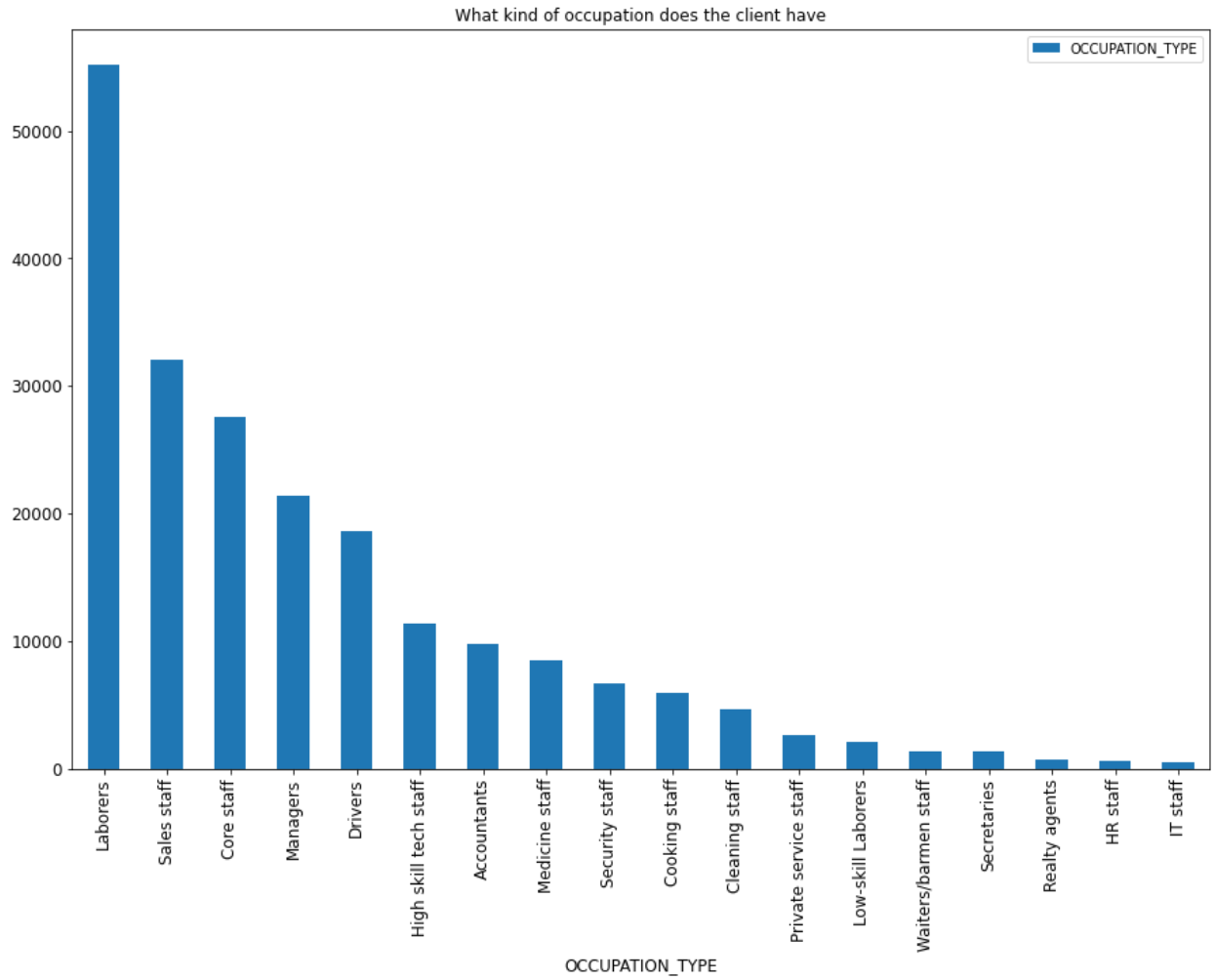
Top 3 ratios go like Working , Commercial & Pesnioner

8. Family Status of the Client :



#Majority of the applicants are Married

9. Client's Occupational type



Top Applicant's who applied for loan :

Laborers - Apprx. 55 K

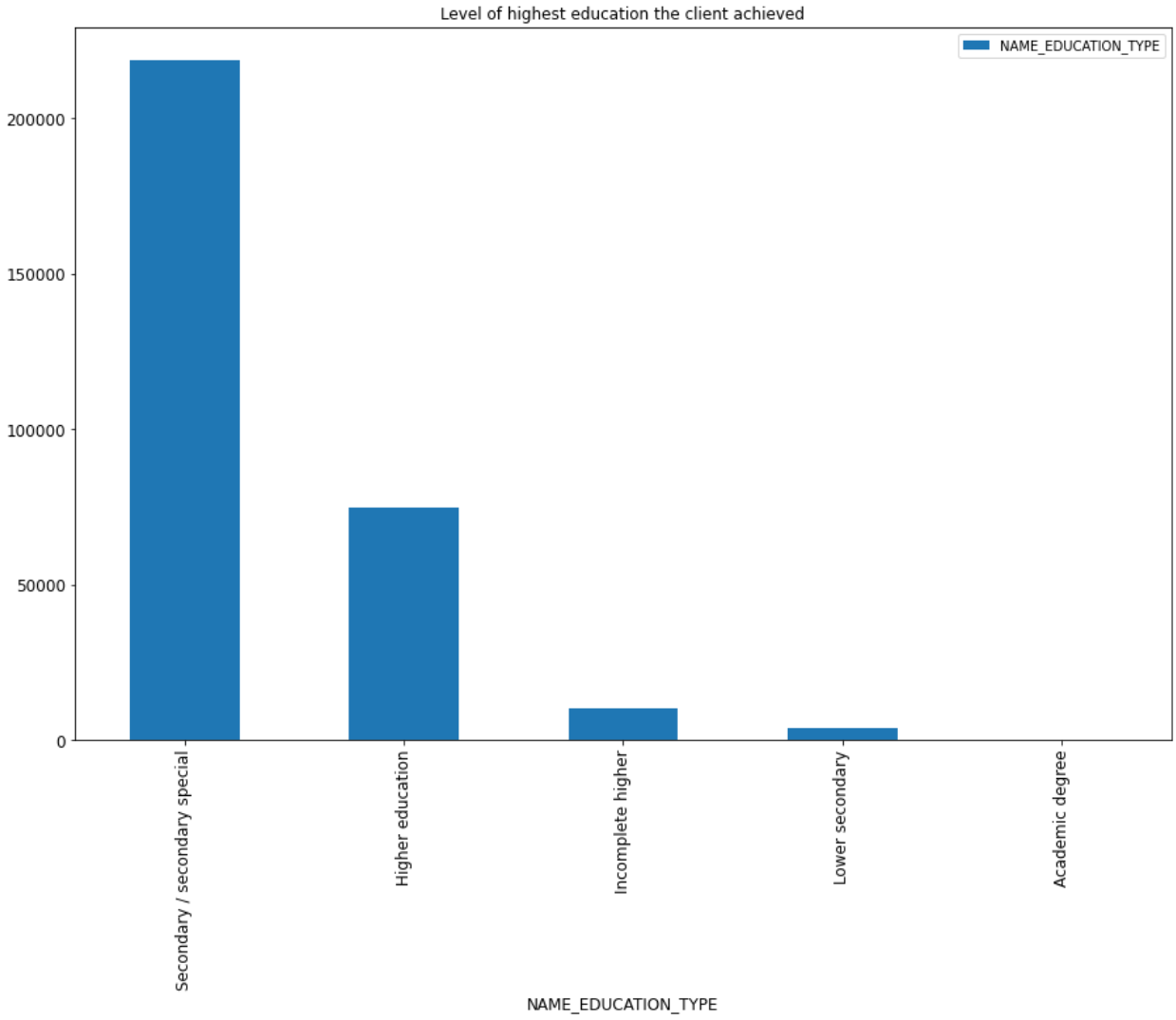
Sales Staff - Approx. 32 K

Core staff - Approx. 28 K

Managers - Approx. 21 K

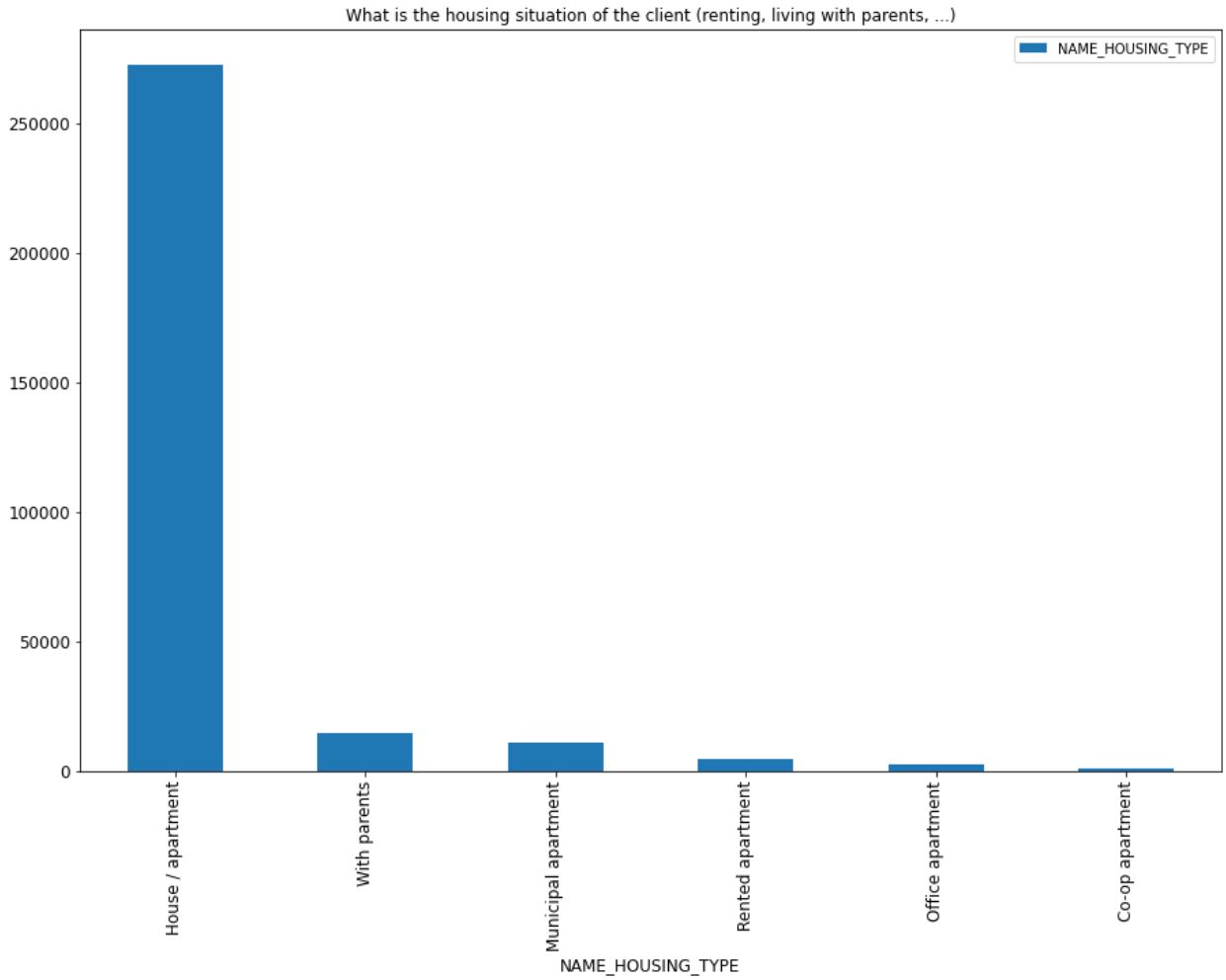
Drivers - Approx. 19 K

10. Client Educational Type



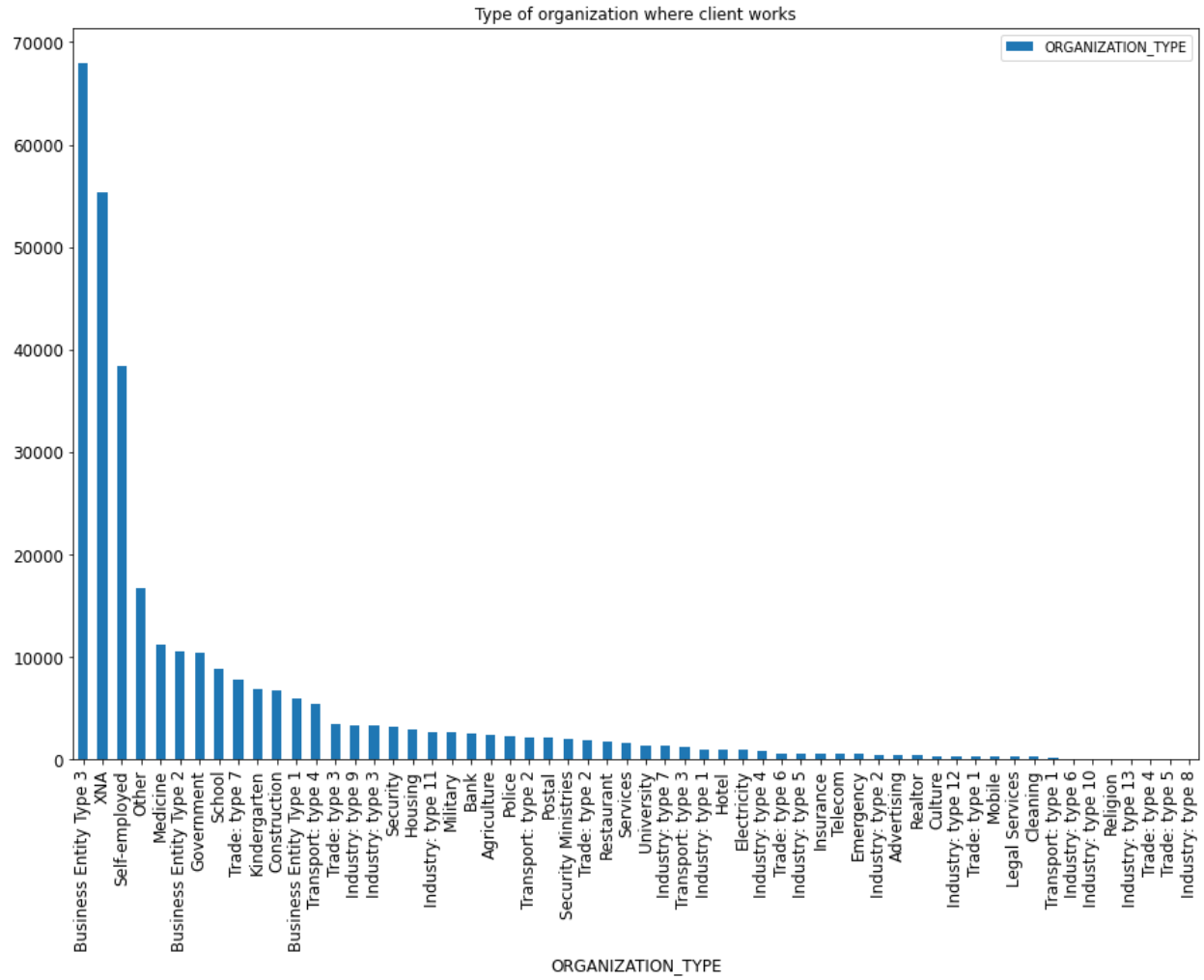
Majority of applicants have secondary and 2nd most having higher education.

11. Client Housing Type:



Approx. 90 % peoples applied for loan, they mentioned type of house is House / Appartment

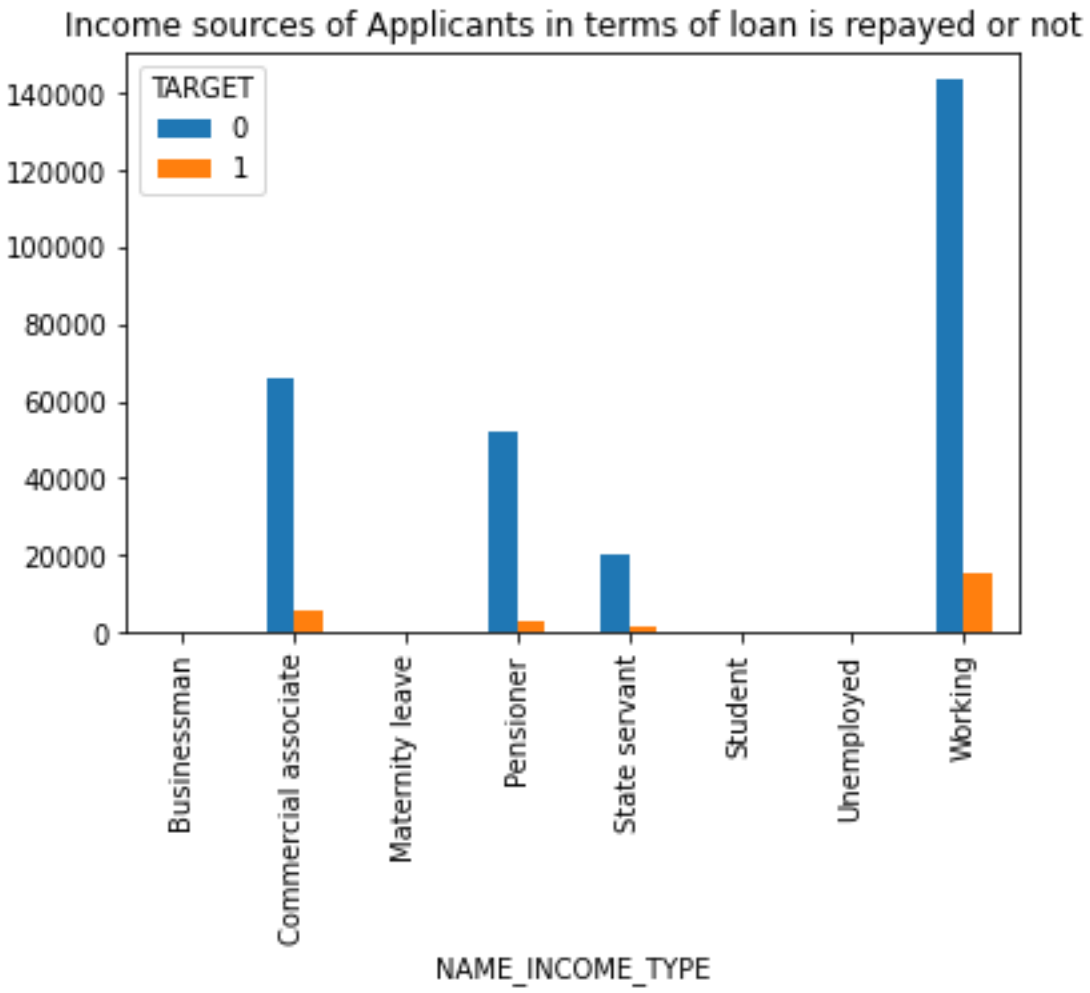
12. Client's working Org type :



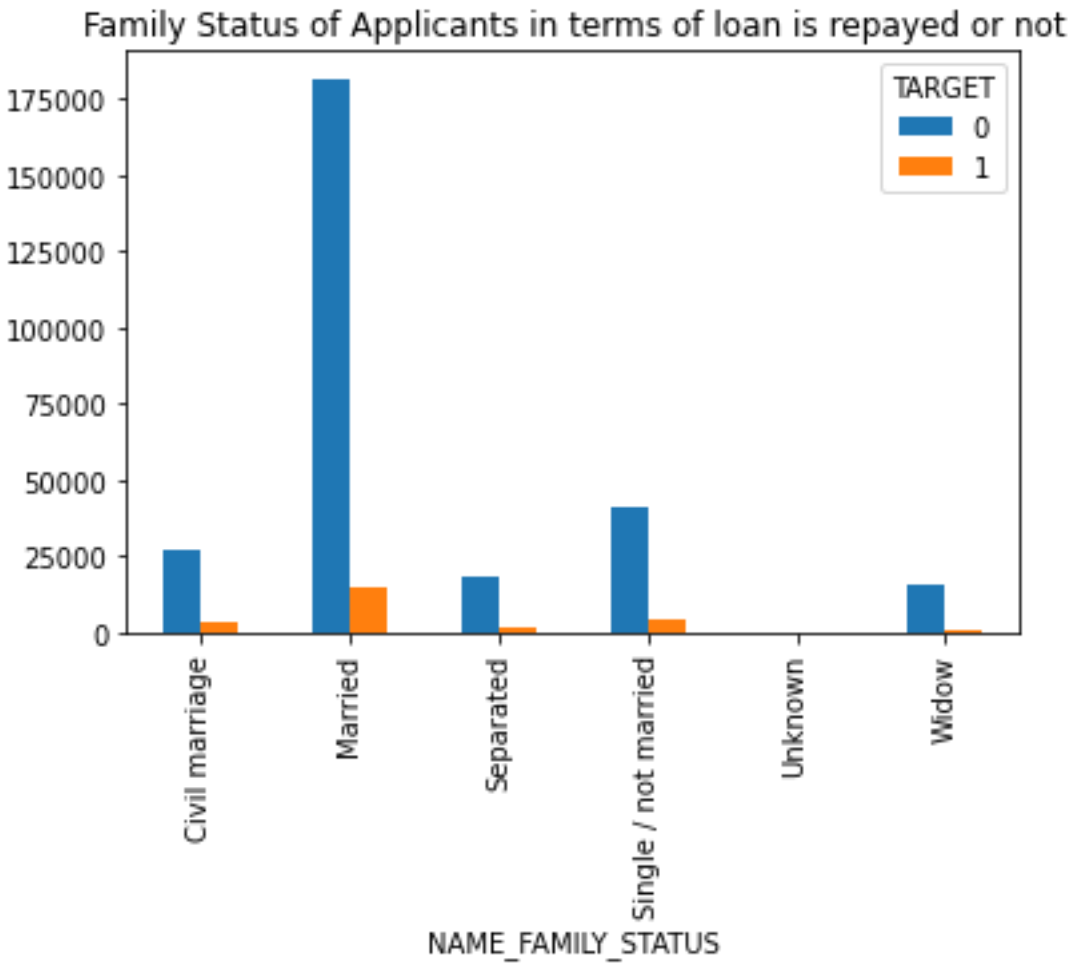
Business Entity Type 3 - Approx. 68 K
XNA - Approx. 55 K
Self employed - Approx. 38 K
Others - Approx. 17 K
Medicine - Approx. 11 K

EDA Target Vs Features: [0- Paid , 1 – Not Paid] Also Assumptions

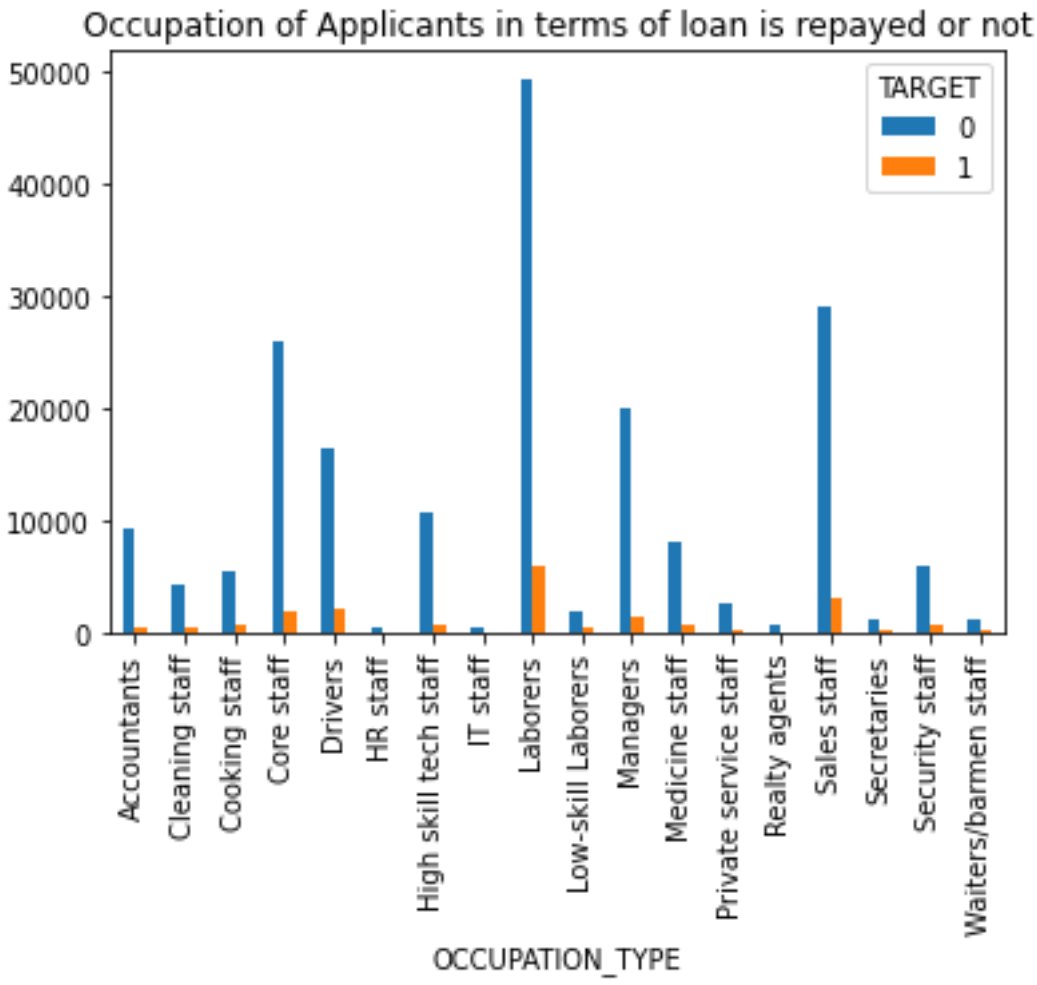
1. Income Source Vs Target – State Servant seems to be very High chance to repaid than others



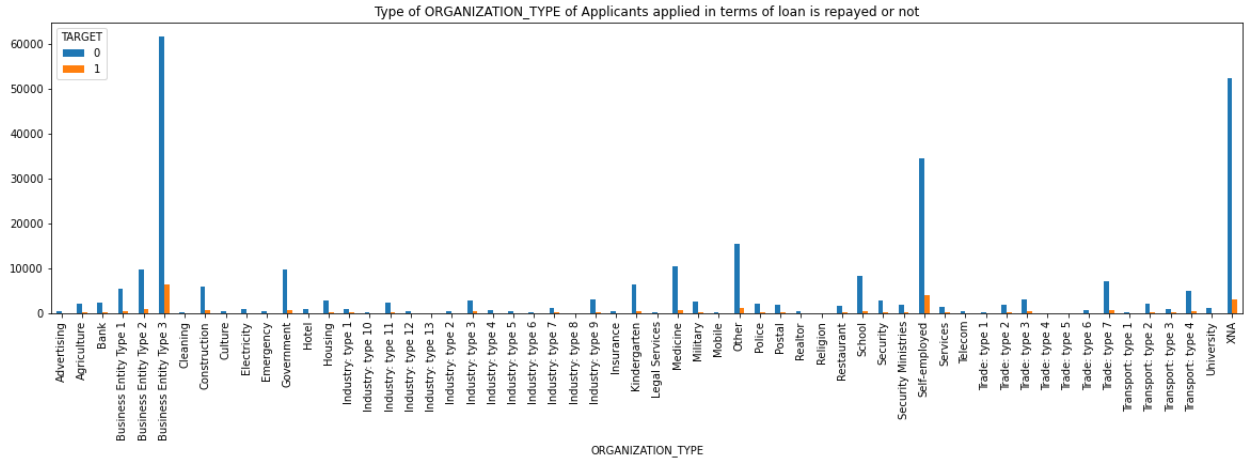
2. Family Status Vs Target**:** Widow & Separated records has very low non repaid reputation than others ,



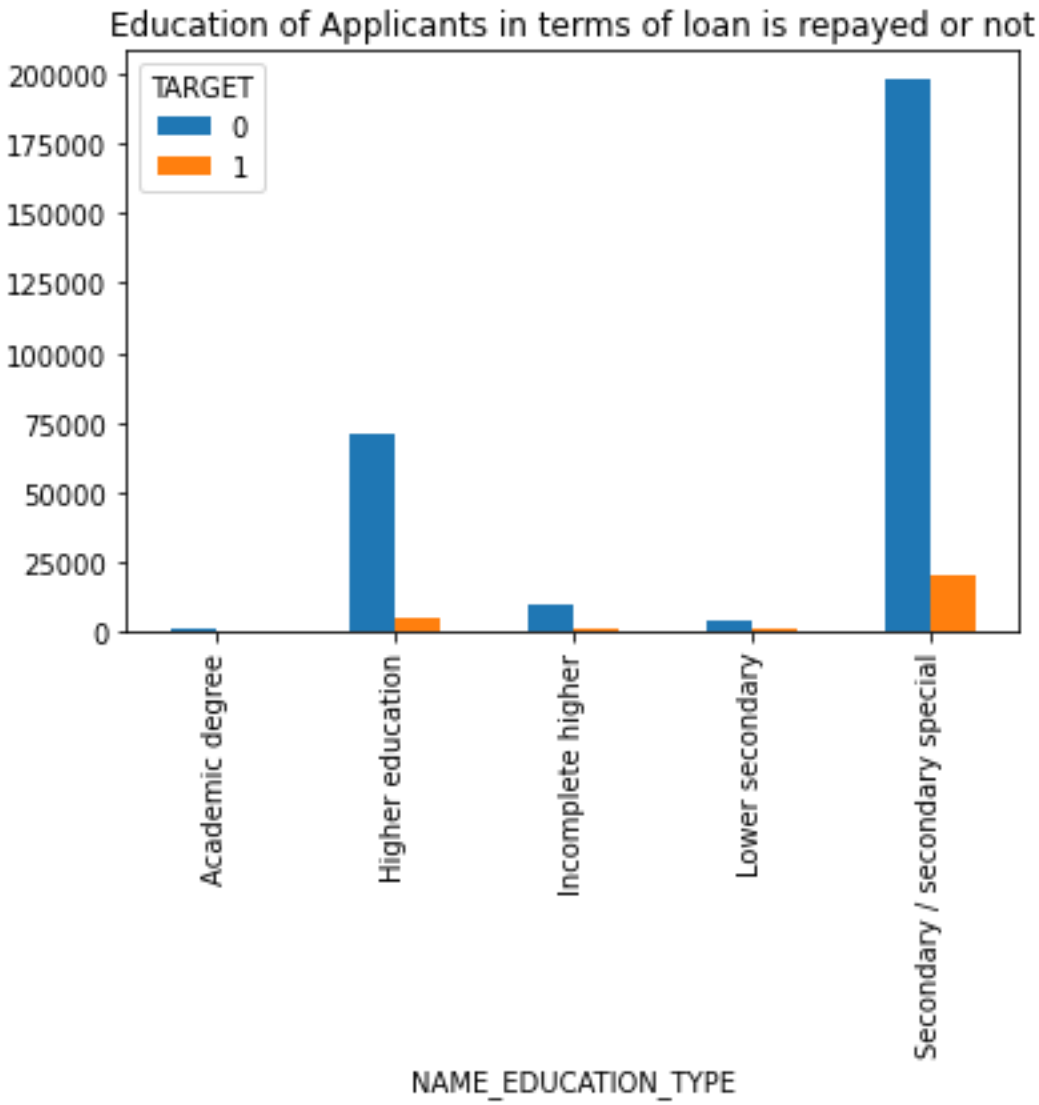
3. Occupation Type Vs Target**:** Realty agents & IT Staffs are taking loan rarely & they tend to repay much better than others



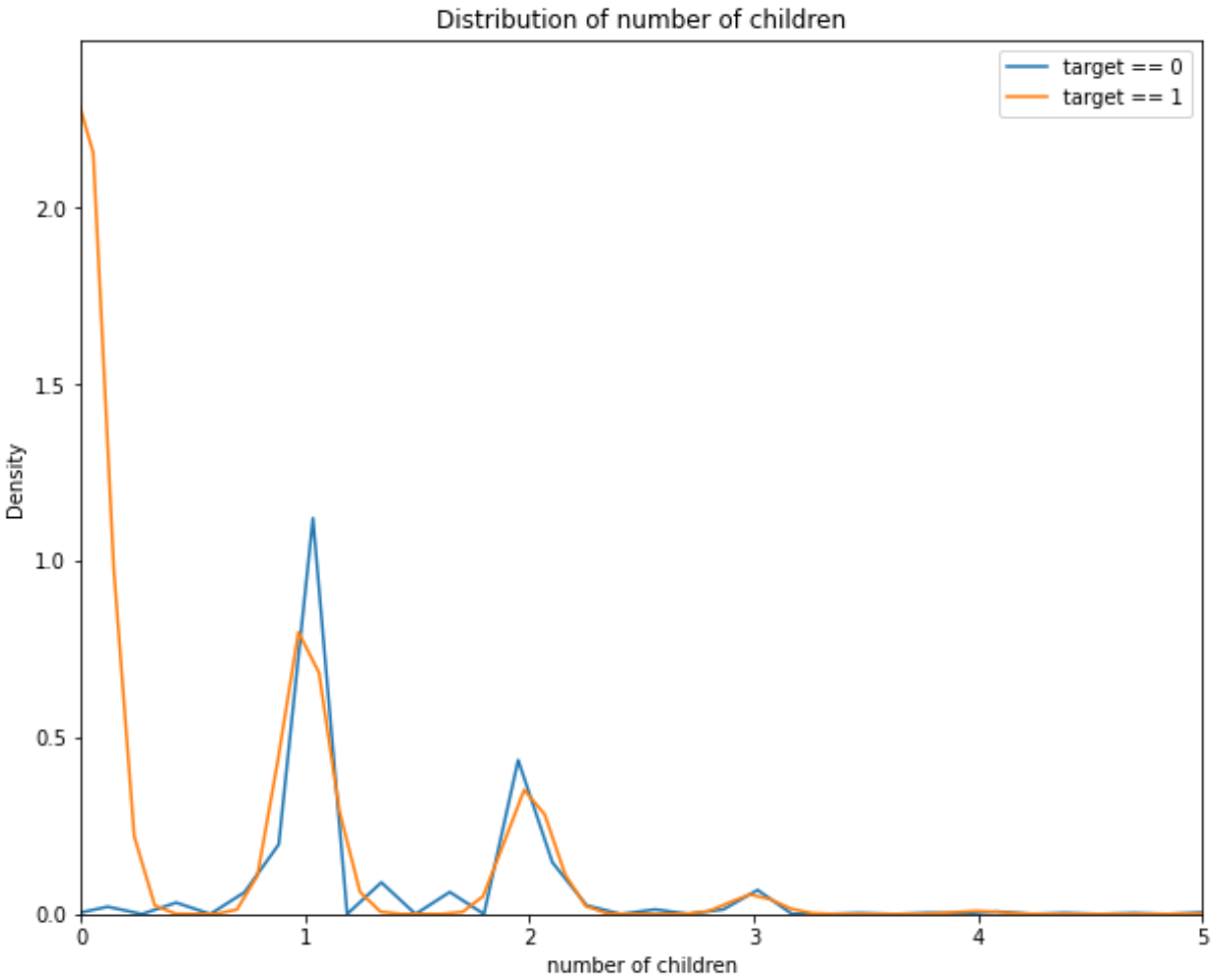
4. Org Type Vs Target: Compared to others Business Entity 3 has High Repaid among others



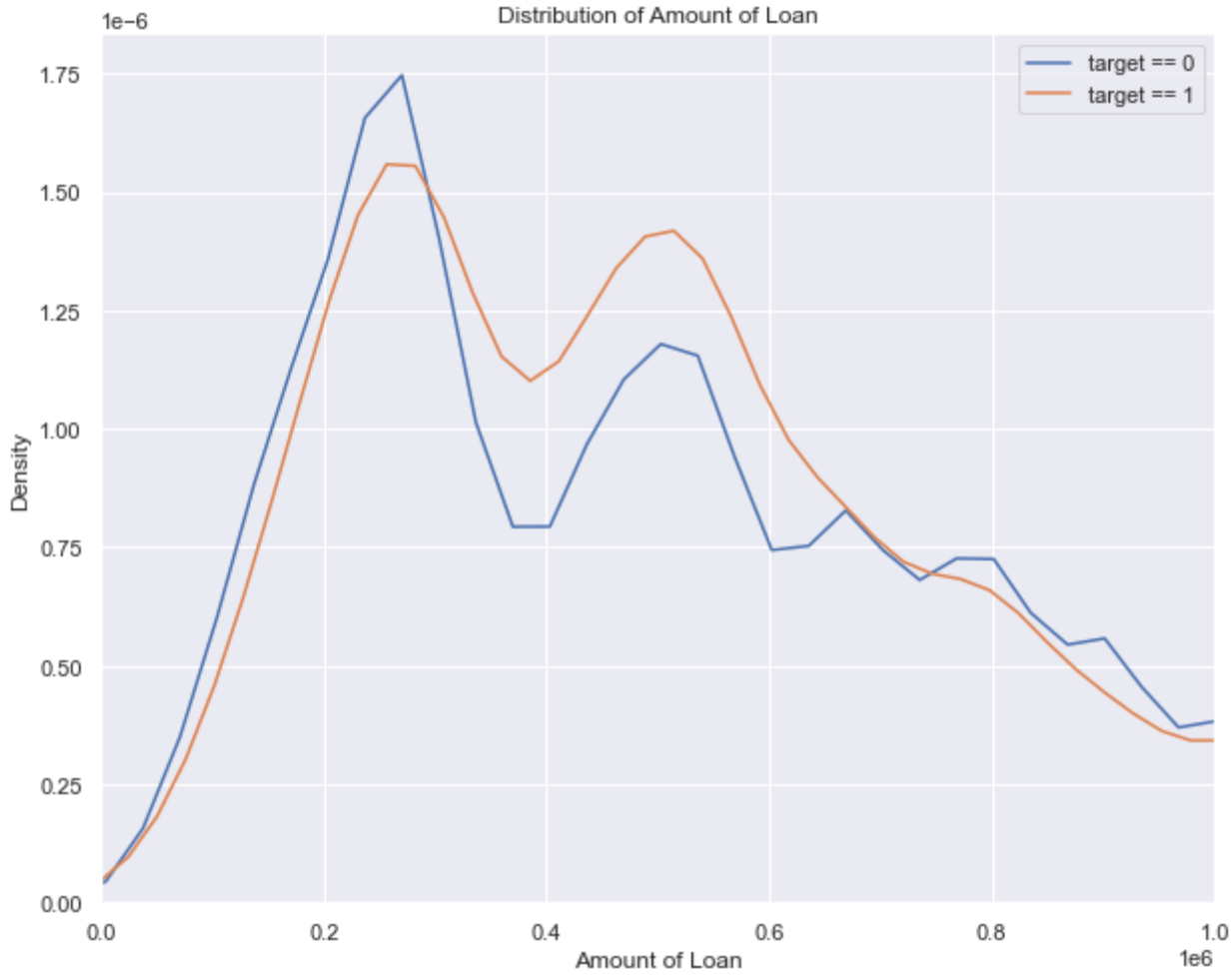
5. Education vs Target: 90 % applicants has done Secondary. Higher education, in that higher education candidates has low repaid ratio .



6. Number of Children Vs Target: Interesting to see that those who have had trouble repaying loans have mainly had zero children, although the distributions look to be similar for other children counts



7. **Amount Credit Vs Target:**



Inferential Statistics:

Assumption1:

Whether Housing type & Family status feature has significant impact for Target prediction or not – OLS Method – F stats, P value

Results:

OLS Regression Results

Dep. Variable:	TARGET	R-squared:	0.003
Model:	OLS	Adj. R-squared:	0.003
Method:	Least Squares	F-statistic:	**82.93**
Date:	Sat, 03 Oct 2020	Prob (F-statistic):	1.77e-171
Time:	15:13:27	Log-Likelihood:	-36033.

```
No. Observations:              307511   AIC:                          7.209e+04
Df Residuals:                  307500   BIC:                          7.220e+04
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================================================
                                                 coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------------------
Intercept                                      0.0960      0.008     11.606      0.000       0.080       0.112
NAME_HOUSING_TYPE[T.House / apartment]         0.0006      0.008      0.075      0.940      -0.015       0.017
NAME_HOUSING_TYPE[T.Municipal apartment]       0.0073      0.009      0.858      0.391      -0.009       0.024
NAME_HOUSING_TYPE[T.Office apartment]         -0.0119      0.010     -1.226      0.220      -0.031       0.007
NAME_HOUSING_TYPE[T.Rented apartment]          0.0422      0.009      4.682      0.000       0.025       0.060
NAME_HOUSING_TYPE[T.With parents]              0.0343      0.008      4.073      0.000       0.018       0.051
NAME_FAMILY_STATUS[T.Married]                 -0.0228      0.002    -13.466      0.000      -0.026      -0.019
NAME_FAMILY_STATUS[T.Separated]               -0.0174      0.002     -6.958      0.000      -0.022      -0.012
NAME_FAMILY_STATUS[T.Single / not married]    -0.0037      0.002     -1.824      0.068      -0.008       0.000
NAME_FAMILY_STATUS[T.Unknown]                 -0.0999      0.192     -0.519      0.603      -0.477       0.277
NAME_FAMILY_STATUS[T.Widow]                   -0.0391      0.003    -14.683      0.000      -0.044      -0.034
==============================================================================
Omnibus:                   185893.783   Durbin-Watson:                   2.002
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1190628.319
Skew:                           3.066   Prob(JB):                         0.00
Kurtosis:                      10.438   Cond. No.                         595.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

As it has high F statistics & P value almost 0 , so we can say above features are statistically significant for prediction.

Assumption 2: Feature: Amount Income Total – using stats model, T stats & P value

Null Hypo: Income does not have significant importance for repaying loan

Alt Hypo: Income has significant impact for Repaying loan

Results:

```
(t_critical_stats= 2.2081011084696076, p_value_stats=0.027237960879676462)
```

As we got 2.2 as T value and p value 0.02 which is <=0.05, so we can reject Null Hypothesis . That mean amount income total has significant impact on repaying loan

Assumption 3: Feature No of children – Replications by random choice, P value from shifted Mean

Null Hypo: There is no impact for repaying loan based on no of child

Alt Hypo: There is significant impact on repaying loan based on child

Results:

As p =0 we can reject null hypo, there should be Highly significant impact on repaying loan based on child

Feature Selection Methodology: Refer – EDA.ipynb

Mutual Info Classifier & K best has been used to identify the Best correlated feature which matches with TARGET variable based on fs score

Results: Top features

	Feature	Score
10	FLAG_MOBIL	0.035830
13	FLAG_CONT_MOBILE	0.034916
393	Cash loans	0.029763
279	House / apartment	0.027121
11	FLAG_EMP_PHONE	0.024349
321	Unaccompanied	0.023997
18	REGION_RATING_CLIENT_W_CITY	0.022159
17	REGION_RATING_CLIENT	0.020673
35	FLAG_DOCUMENT_3	0.018630

Random Forest Feature Selection: Trained model with Random Forest bagging algorithm & fitted model provides feature importance

	Feature	Importance
0	Cleaning	0.029922
1	Sent proposal	0.028931
2	AMT_CREDIT_SUM_LIMIT	0.014392
3	Unknown type of loan	0.013537
4	POS mobile with interest_y	0.013247
5	5	0.012757
6	Cleaning staff	0.012634
7	Hotel	0.012609
8	SK_ID_CURR	0.012303
9	POS industry without interest_y	0.011720
10	Industry: type 7	0.011497
11	POS industry with interest_x	0.011242
12	AMT_PAYMENT_CURRENT	0.011184
13	Cash Street: high_y	0.011157
14	AMT_RECEIVABLE_PRINCIPAL	0.011041
15	AMT_CREDIT_x	0.010865
16	AMT_DRAWINGS_OTHER_CURRENT	0.010835
17	Furniture_x	0.010822
18	THURSDAY	0.010782
19	Cash Street: low_x	0.010740

Preprocessing data set before creating ML modeling:

1. Imputer – Added SimpleImputer & strategy to calculate ‘’Median” for the set , transformed data set.
2. Scaler – Added MinMaxScaler for all feature value range from 0-1 for computation

Created Train & Test data set with 33 % test size. Which is going to be applied in ML Models

Machine Learning Modelling & algorithm Applied: As data set belongs to Supervised Machine learning

Machine Learning Algorithm Used:

- Logistic Regression
- Random Forest
- Naïve Bayes
- Ensemble Modeling (combined above 3)

1. **Logistic Regression Results:**

```
The accuracy score :  0.9223047288018181


The classification report is as follows:
              precision    recall  f1-score   support

           0       0.92      1.00      0.96     75895
           1       0.49      0.02      0.05      6388

    accuracy                           0.92     82283
   macro avg       0.71      0.51      0.50     82283
weighted avg       0.89      0.92      0.89     82283



ROC AUC score is:  0.5107913000044594
```

2. **Random Forest Algorithm Results:**

```
The accuracy in general is :  0.9223776478738014


The classification report is as follows:
              precision    recall  f1-score   support

           0       0.92      1.00      0.96     75895
           1       0.67      0.00      0.00      6388

    accuracy                           0.92     82283
   macro avg       0.79      0.50      0.48     82283
weighted avg       0.90      0.92      0.89     82283

ROC AUC score is:  0.5001499554698198
```

3. Naïve Bayes Algorithm Results:

```
The accuracy in general is :  0.12568817374184219


The classification report is as follows:
              precision    recall  f1-score   support

           0       0.95      0.05      0.10     75895
           1       0.08      0.97      0.15      6388

    accuracy                           0.13     82283
   macro avg       0.52      0.51      0.13     82283
weighted avg       0.89      0.13      0.11     82283

ROC AUC score is:  0.5112823025318859
```

4. Ensemble Method approach & Results :

Ensemble model combines multiple individual models together and delivers superior prediction power.

Basically, an ensemble is a supervised learning technique for combining multiple weak learners/ models to produce a strong learner. Ensemble model works better when we ensemble models with low correlation.

https://blog.statsbot.co/ensemble-learning-d1dcd548e936

Here we used Ensemble method of Voting approach – Bagging

```
The accuracy in general is :  0.9223411883378098


The classification report is as follows:
              precision    recall  f1-score   support

           0       0.92      1.00      0.96     75895
           1       0.50      0.02      0.05      6388

    accuracy                           0.92     82283
   macro avg       0.71      0.51      0.50     82283
weighted avg       0.89      0.92      0.89     82283

ROC AUC score is:  0.5108827478625658
```