Home Loan Credit Risk – Milestone Report

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.

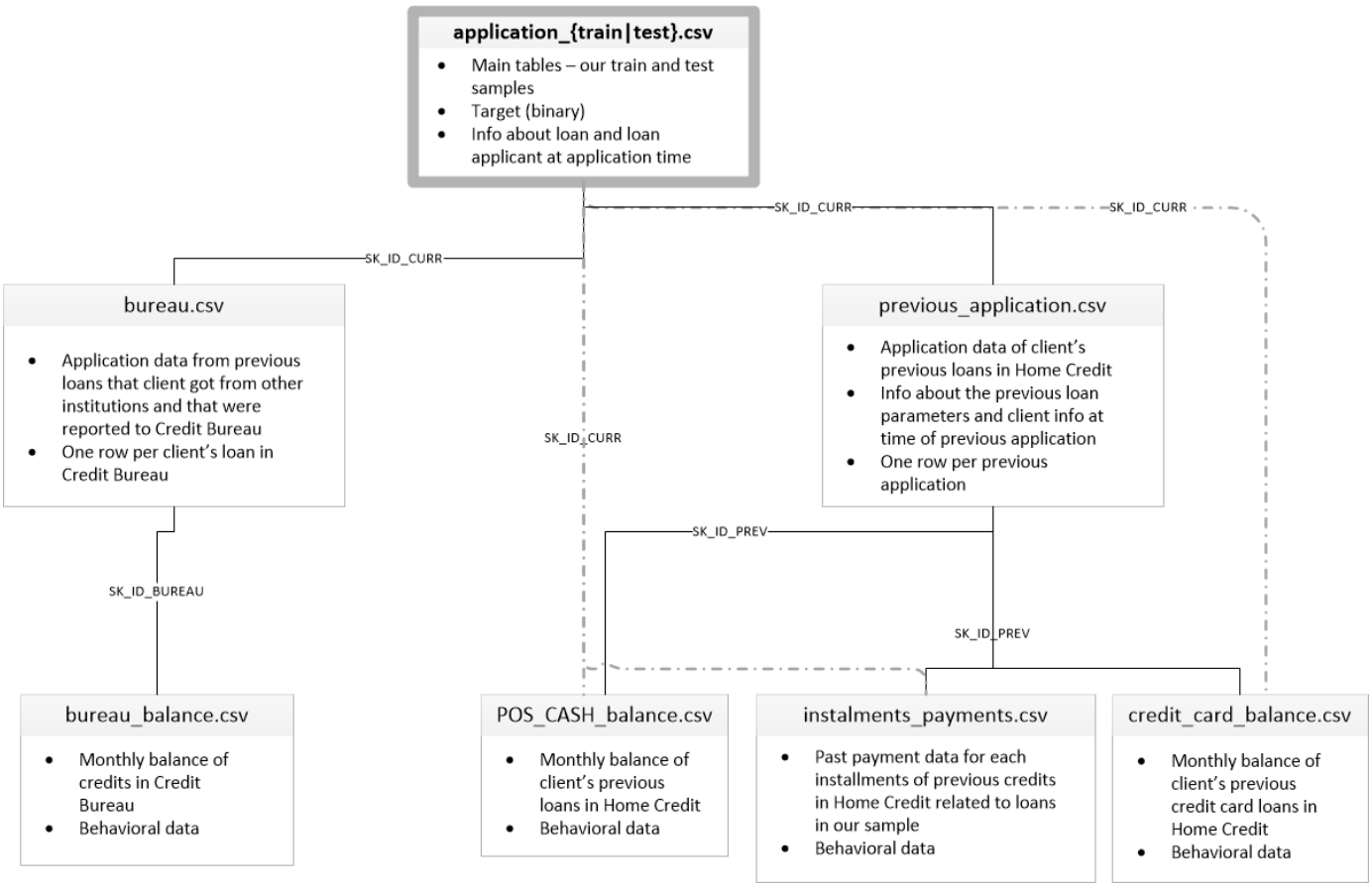
bureau_balance.csv - monthly information per credit, per loan for users in the sample..

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:

Home Loan Credit Risk – Milestone Report



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 . used ordinal encoder for numeric 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
- 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
- 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

E) POS_CASH_BALANCE.csv

- Transforming Categorical variables to Numerical variables

F) installments_payments.csv

- Transforming Categorical variables to Numerical variables
- 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 Installment 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.

Categorical variables encoding Technique:

- OrdinalEncoder from sklearn being used for conversion

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.

Home Loan Credit Risk – Milestone Report

- 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 either 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.
- Negative values being identified &removed

Application_train_merged.csv & Application_test_merged.csv – Preparation:

- Identified categorial variables
- Transforming Categorical variables to Numerical variables
- Replaced all negative values to 0

Conclusion:

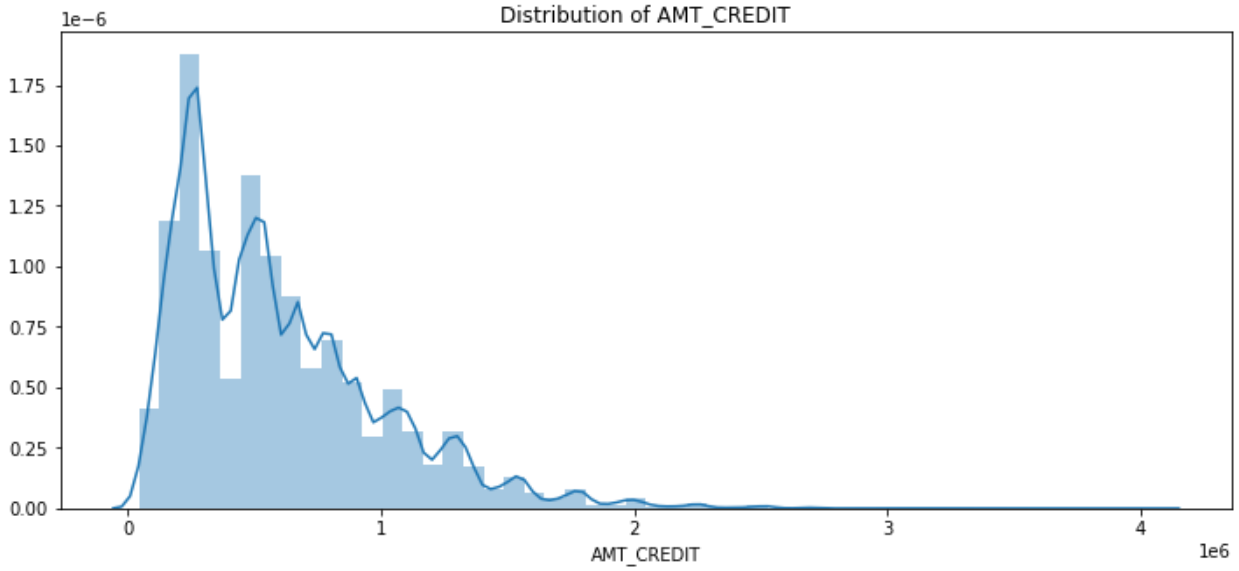
- After all, above steps dataset stored to new file for EDA process

Home Loan Credit Risk – Milestone Report

Data Story with Application train data set:

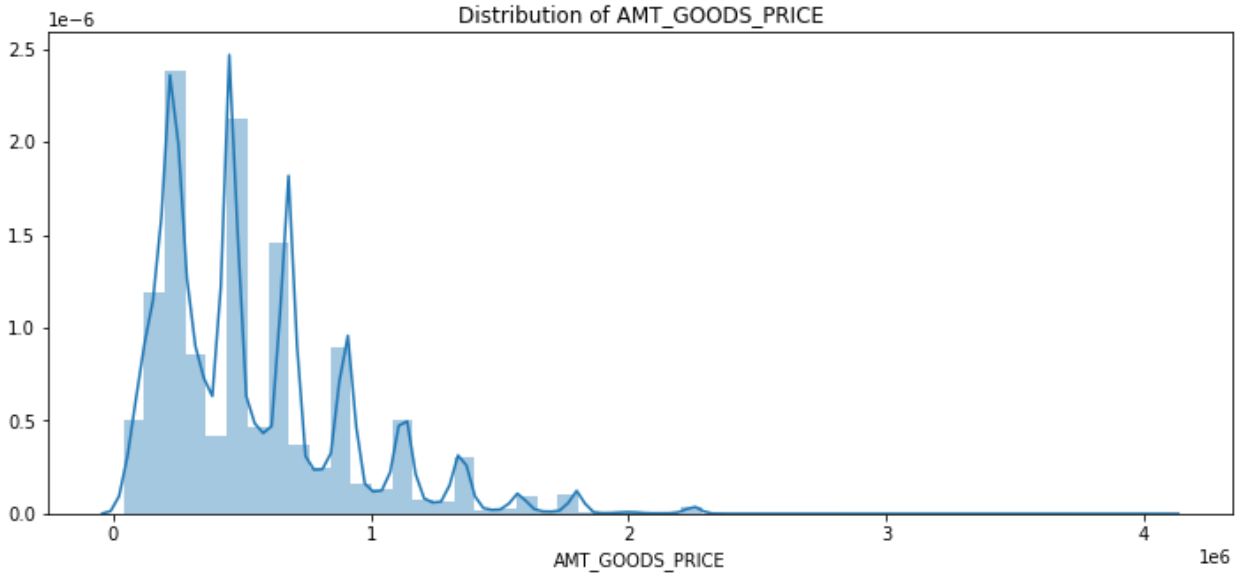
Data Exploration:

1. Distribution of Amount Credit



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

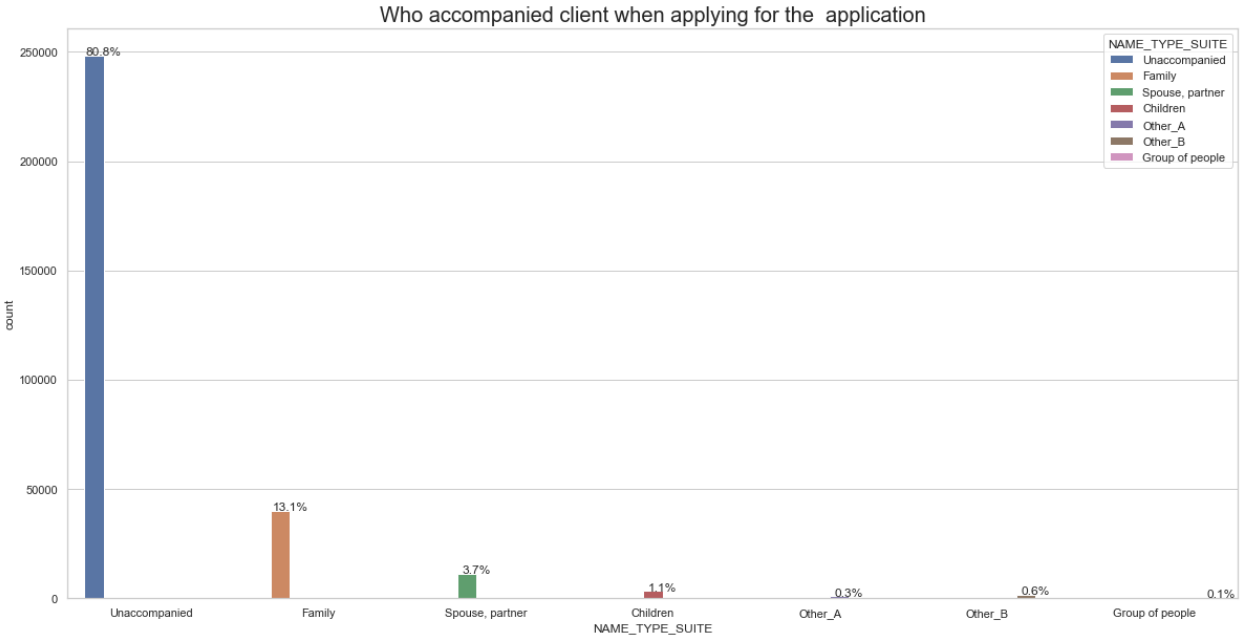
2. Distribution of Amount Goods Price



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

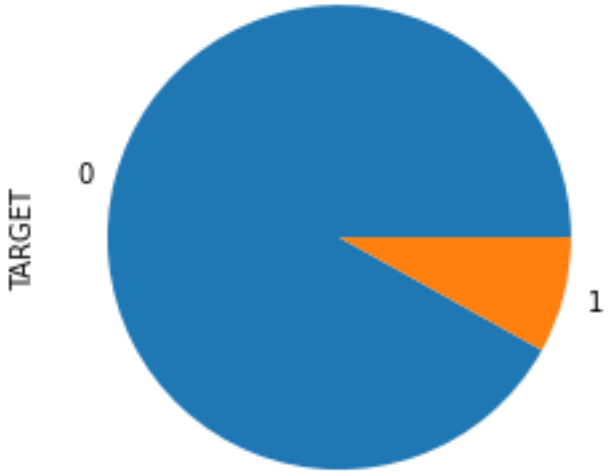
Home Loan Credit Risk – Milestone Report

3. Who accompanied client when applying loan ?



Majority of the applicants are unaccompanied

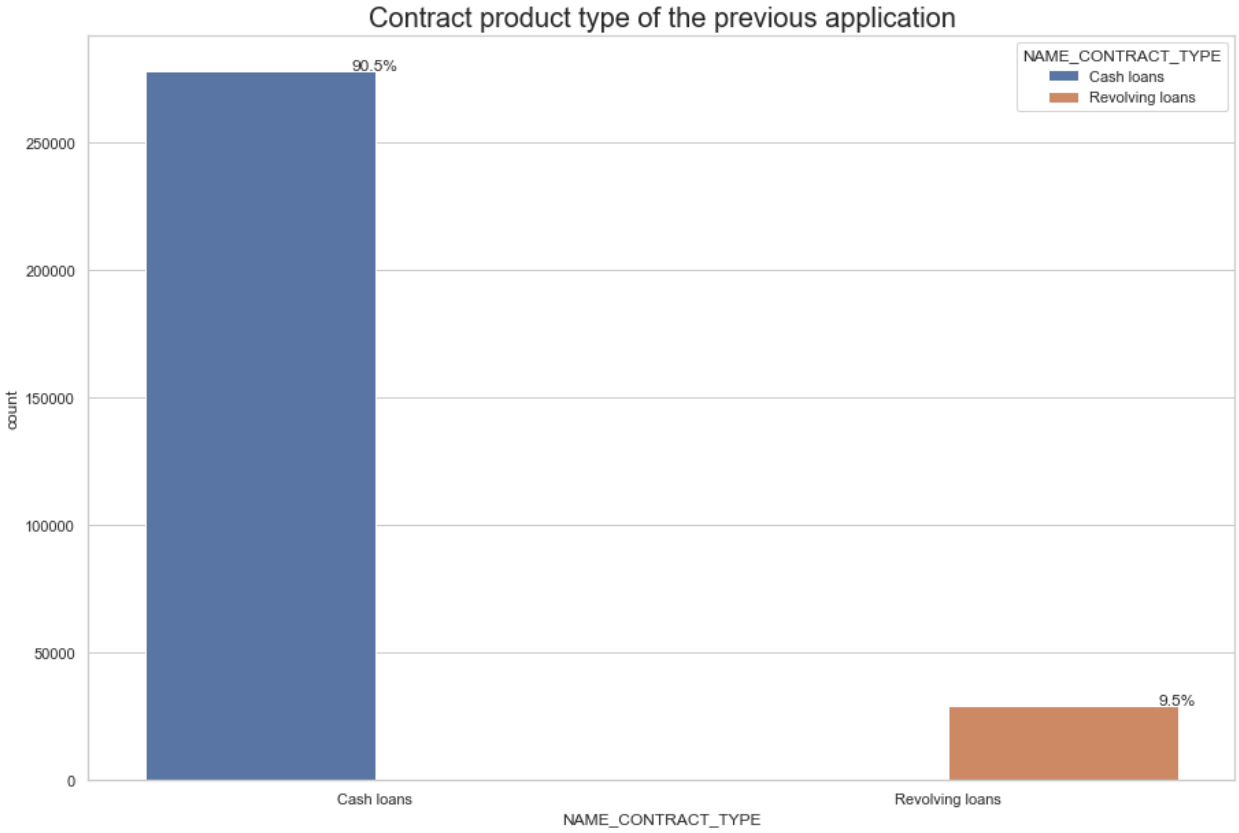
4. Highly imbalanced data!



As we can see data is highly imbalanced.

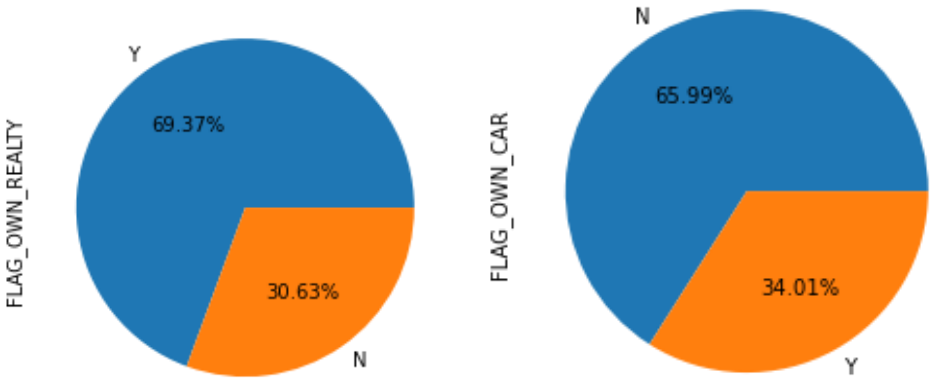
Home Loan Credit Risk – Milestone Report

5. Contract Type of Previous Loan app :



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

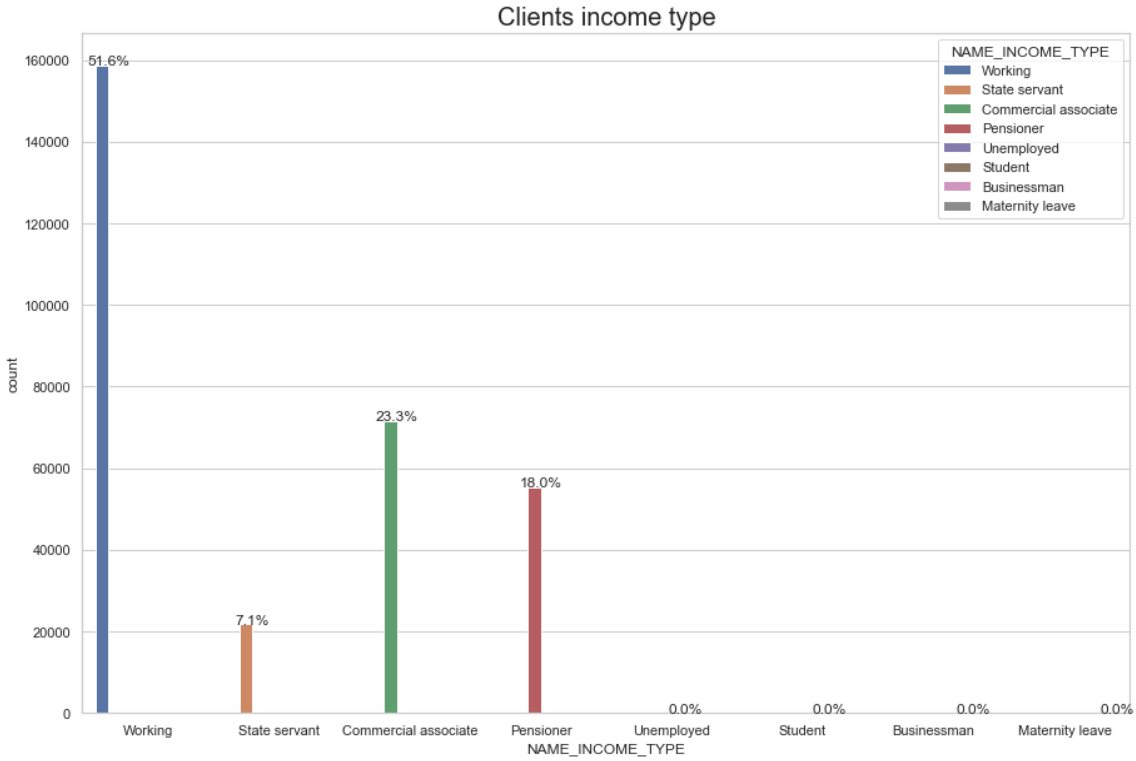
6. Own Relaty & Own Car



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

Home Loan Credit Risk – Milestone Report

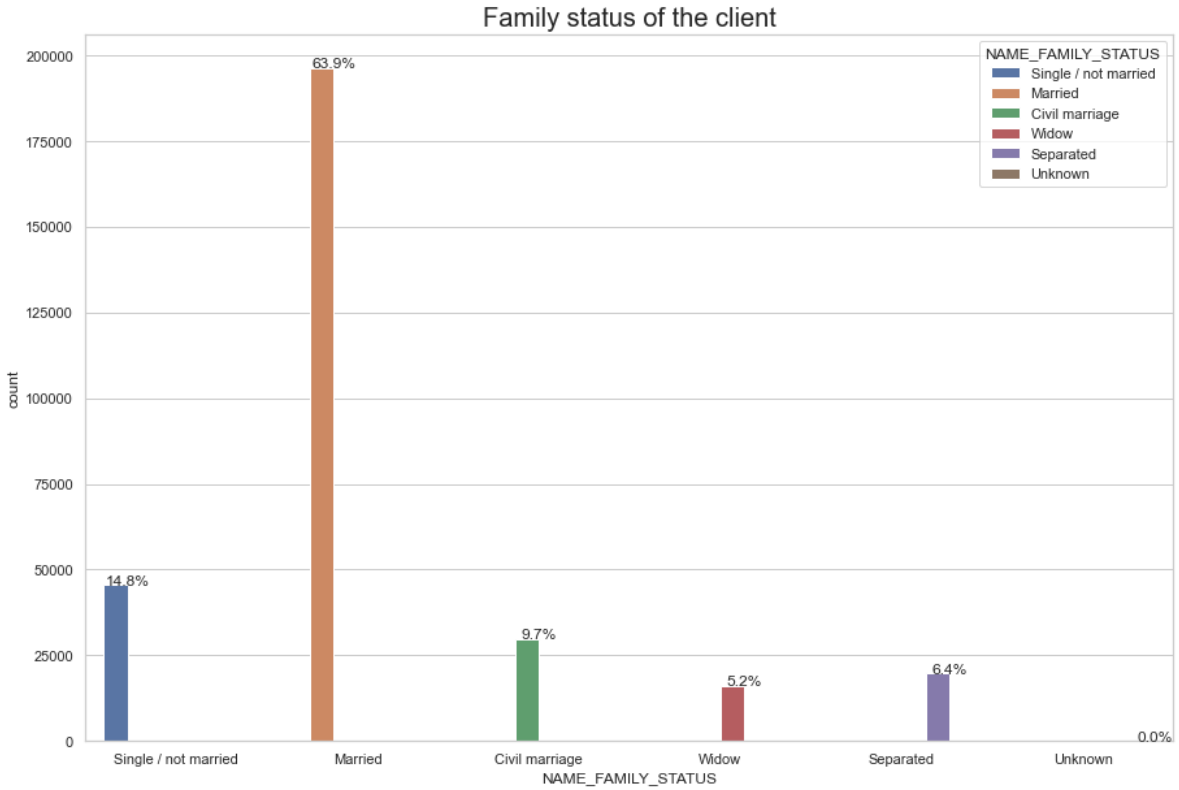
7. Client's Income Type:



Top 3 ratios go like Working , Commercial & Pesnioner

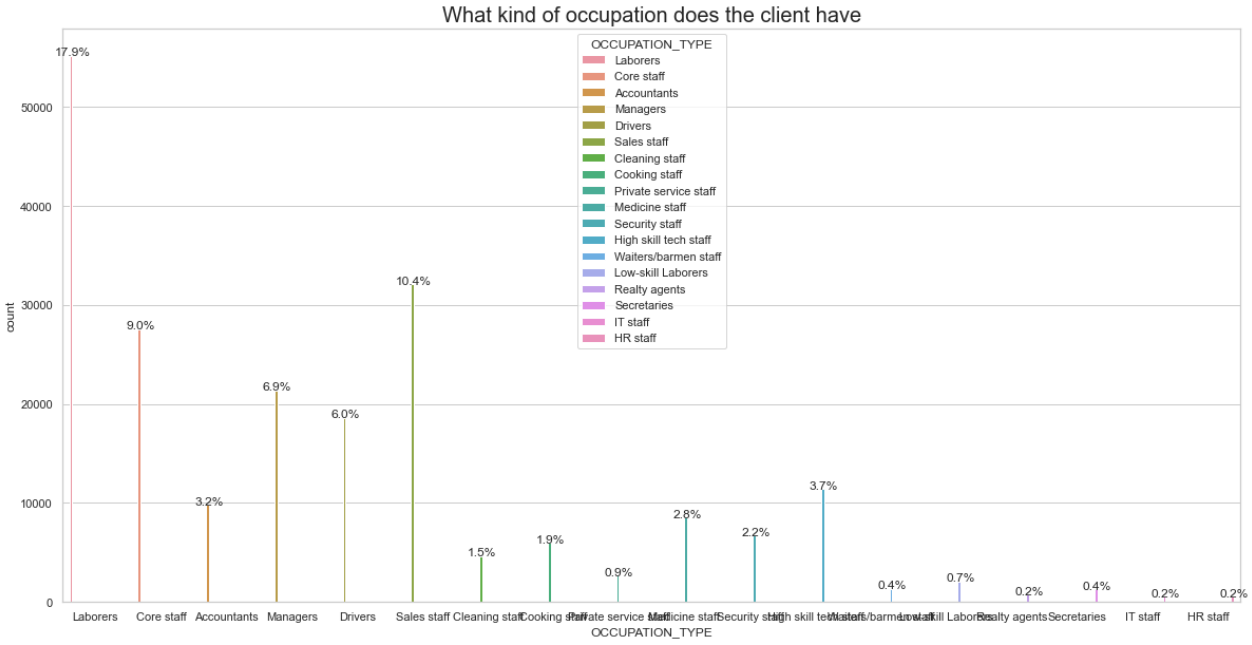
8. Family Status of the Client :

Home Loan Credit Risk – Milestone Report



#Majority of the applicants are Married

9. Client's Occupational type



Top Applicant's who applied for loan :

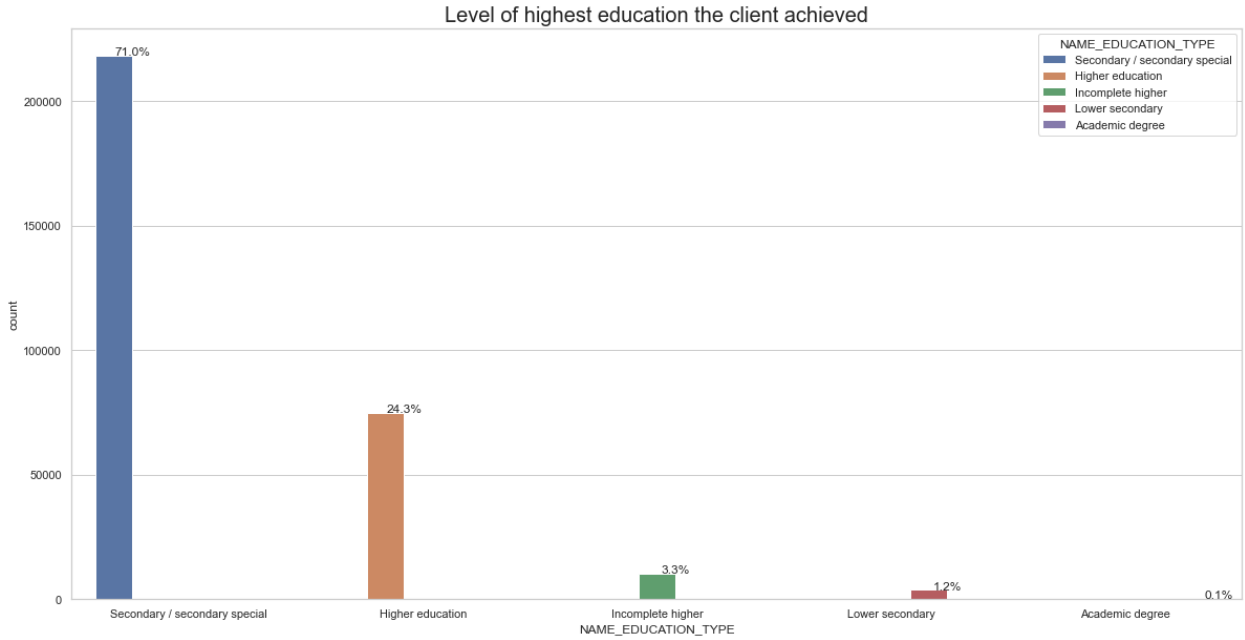
Laborers - Apprx. 55 K

Sales Staff - Approx. 32 K

Home Loan Credit Risk – Milestone Report

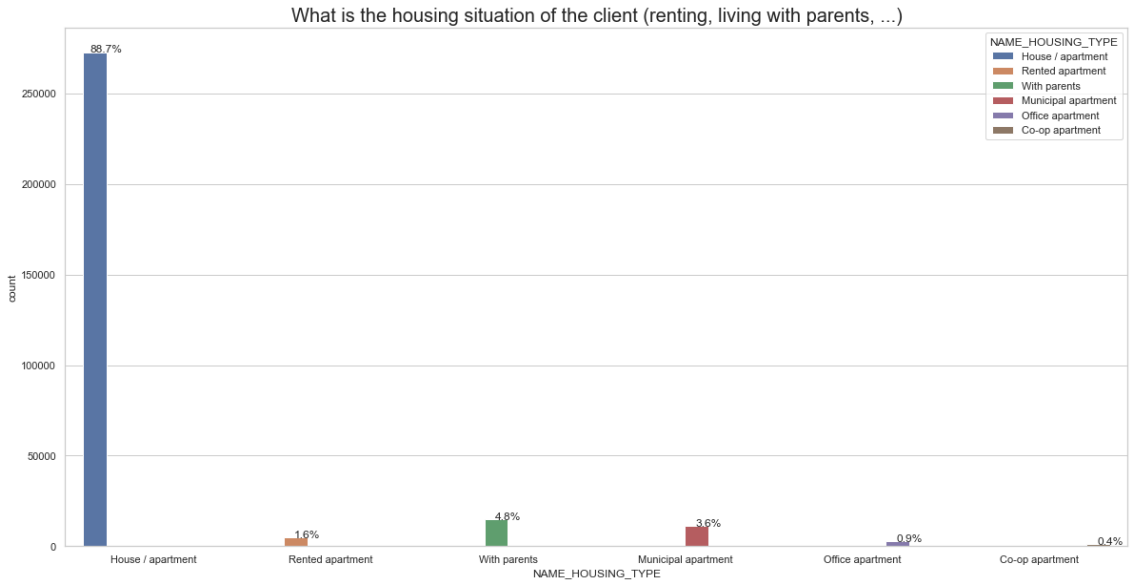
Core staff - Approx. 28 K
Managers - Approx. 21 K
Drivers - Approx. 19 K

10. Client Educational Type



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

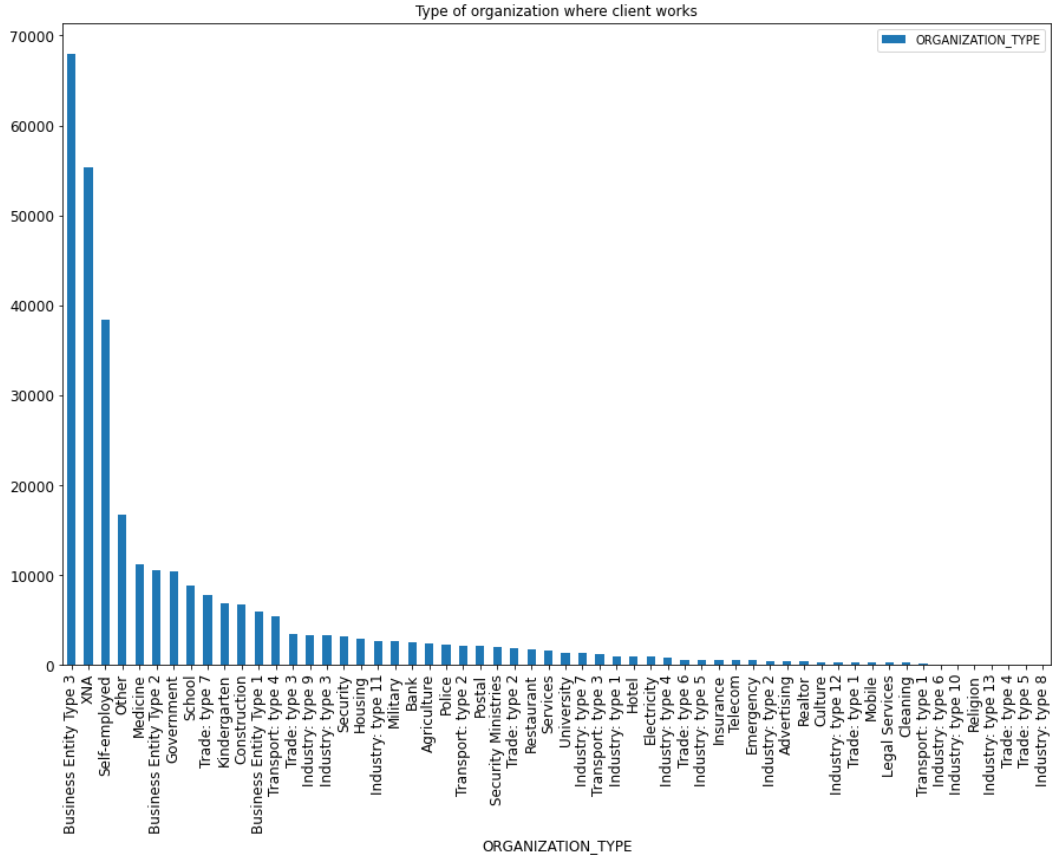
11. Client Housing Type:



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

12. Client's working Org type :

Home Loan Credit Risk – Milestone Report

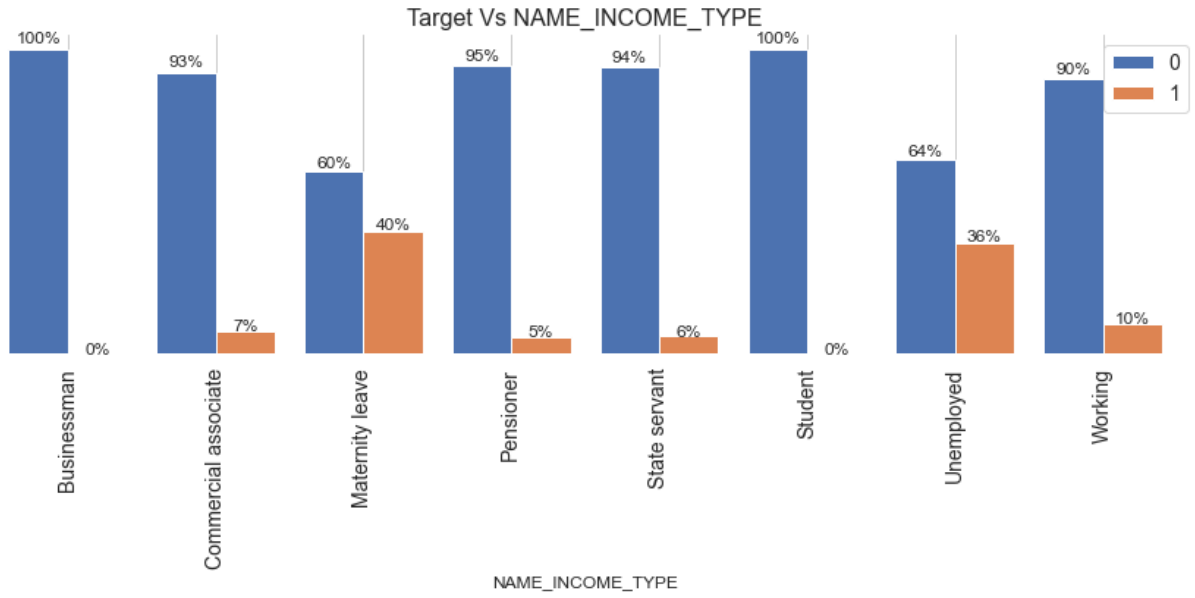


Business Entity Type 3 - Approx. 68 K
XNA - Approx. 55 K
Self employed - Approx. 38 K
Others - Approx. 17 K
Medicine - Approx. 11 K

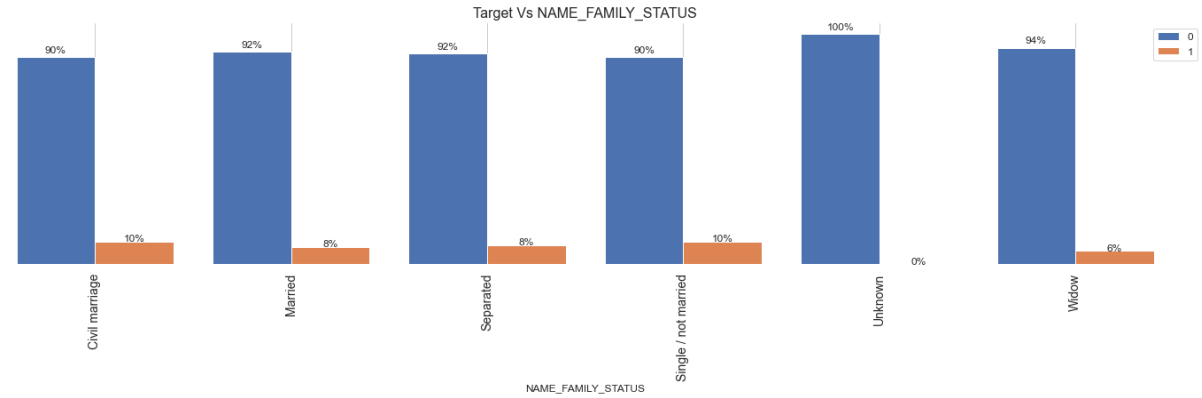
EDA Target Vs Features: [0- Paid , 1 – Not Paid] Also Assumptions

1. NAME_INCOME_TYPE Vs Target – Businessman has highest repaid rate than others

Home Loan Credit Risk – Milestone Report

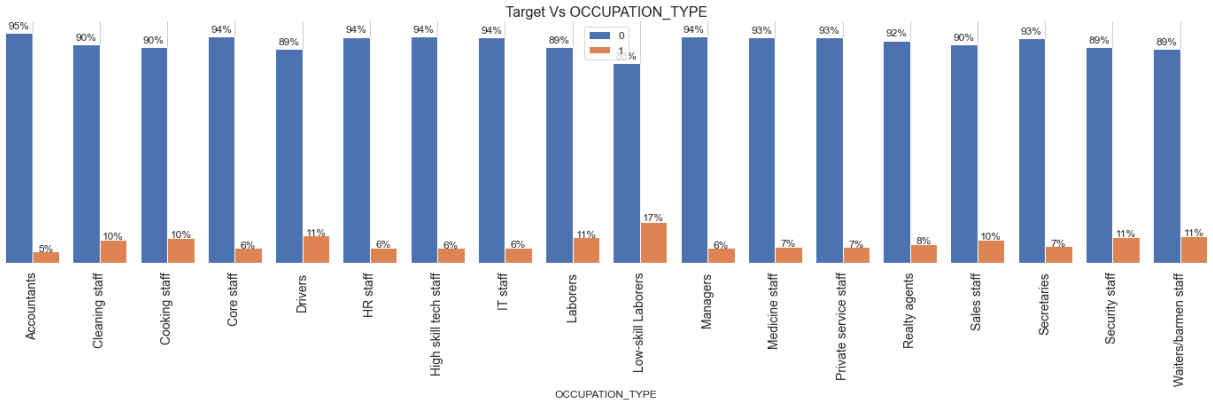


2. Family Status Vs Target**:** Married , Widow , Separated has more repaid ratio than others

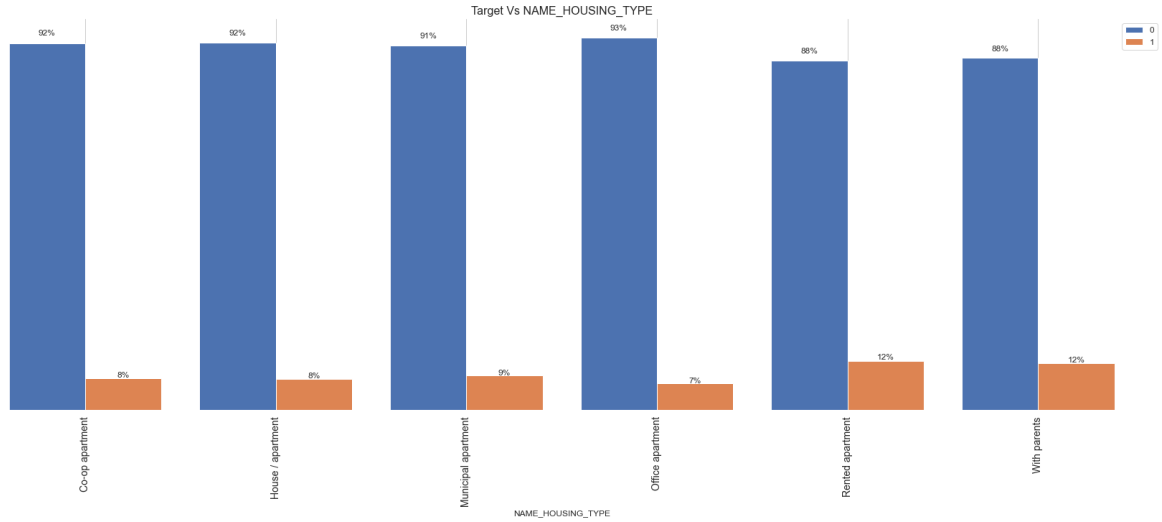


3. Occupation Type Vs Target**:** Accountants tends to repaid more ,low skill worked struggled to repay the loan

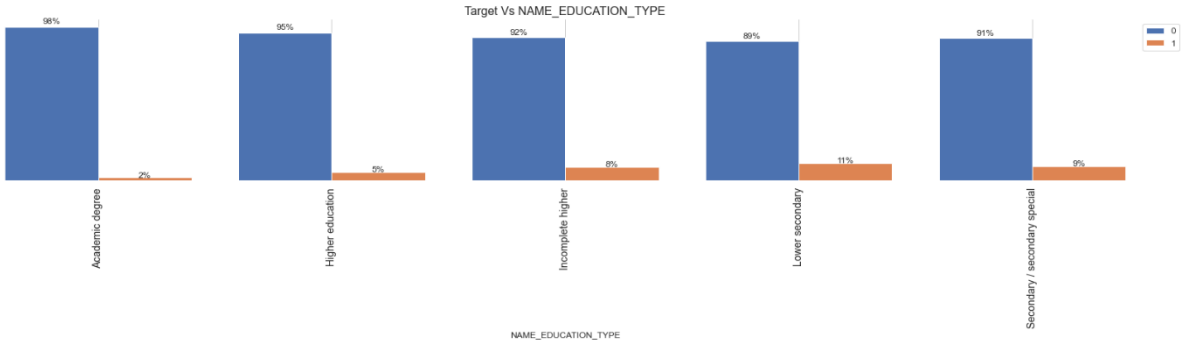
Home Loan Credit Risk – Milestone Report



4. NAME_HOUSING_TYPE Vs Target: Those who are in Rented apartments & accompanied with parents struggled to repay the loan

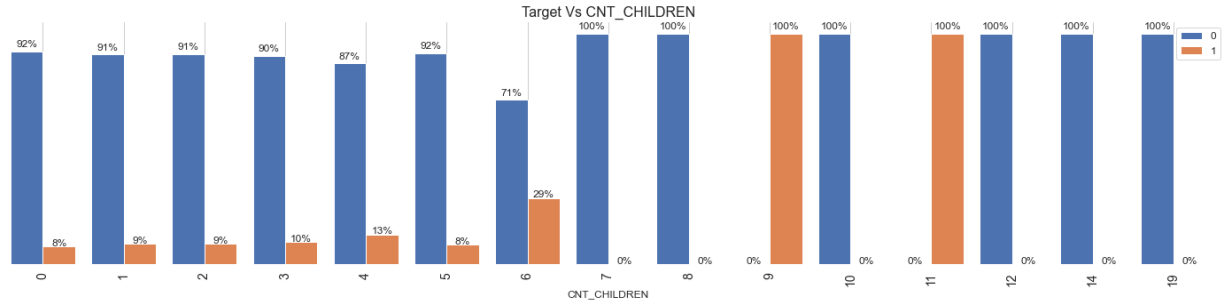


5. Education type vs Target: Those who have academic degree repays loan , other side secondary education type holders struggles to pay loan.

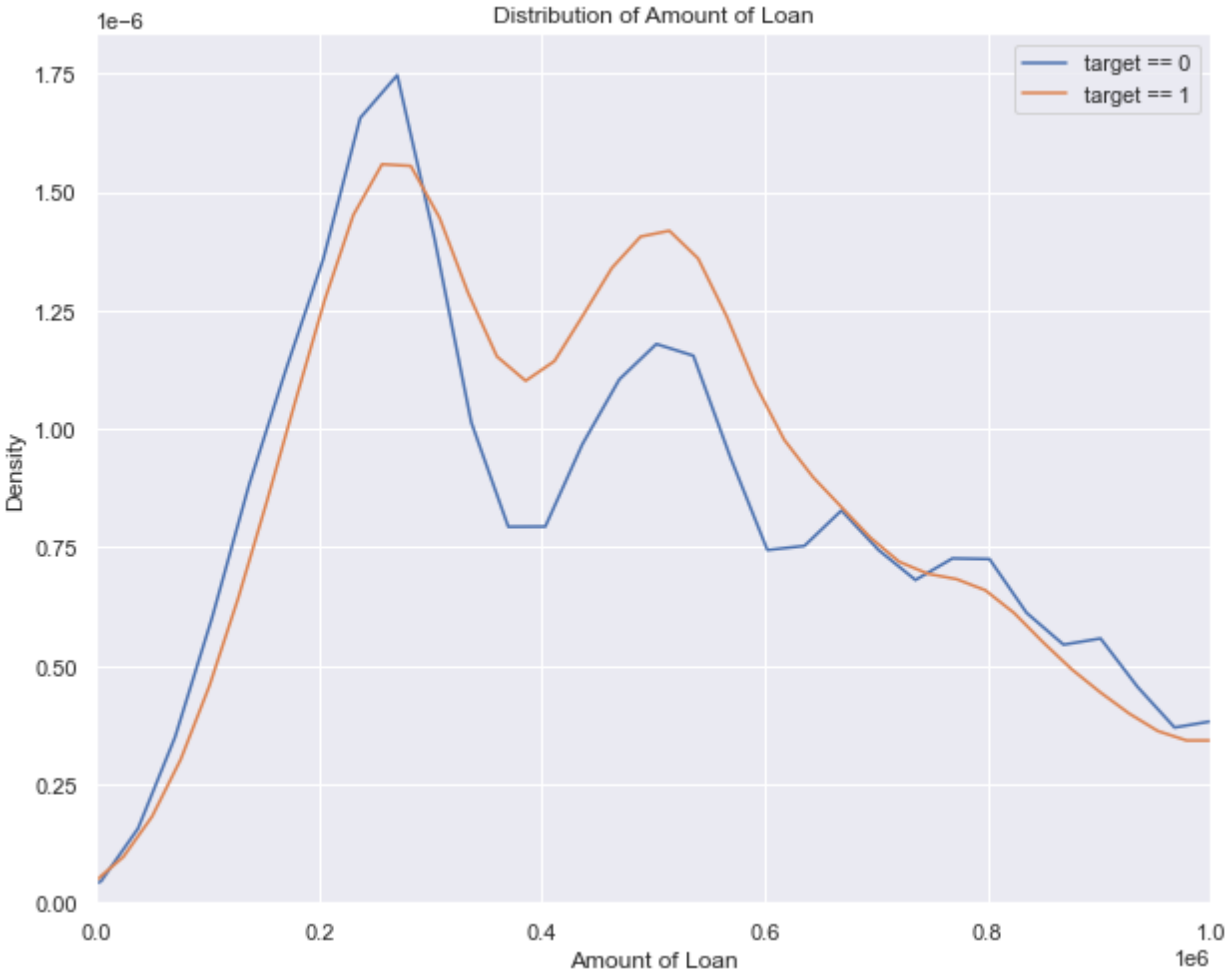


6. Number of Children Vs Target: It's unclear information about repaying history vs CNT_CHILDREN

Home Loan Credit Risk – Milestone Report

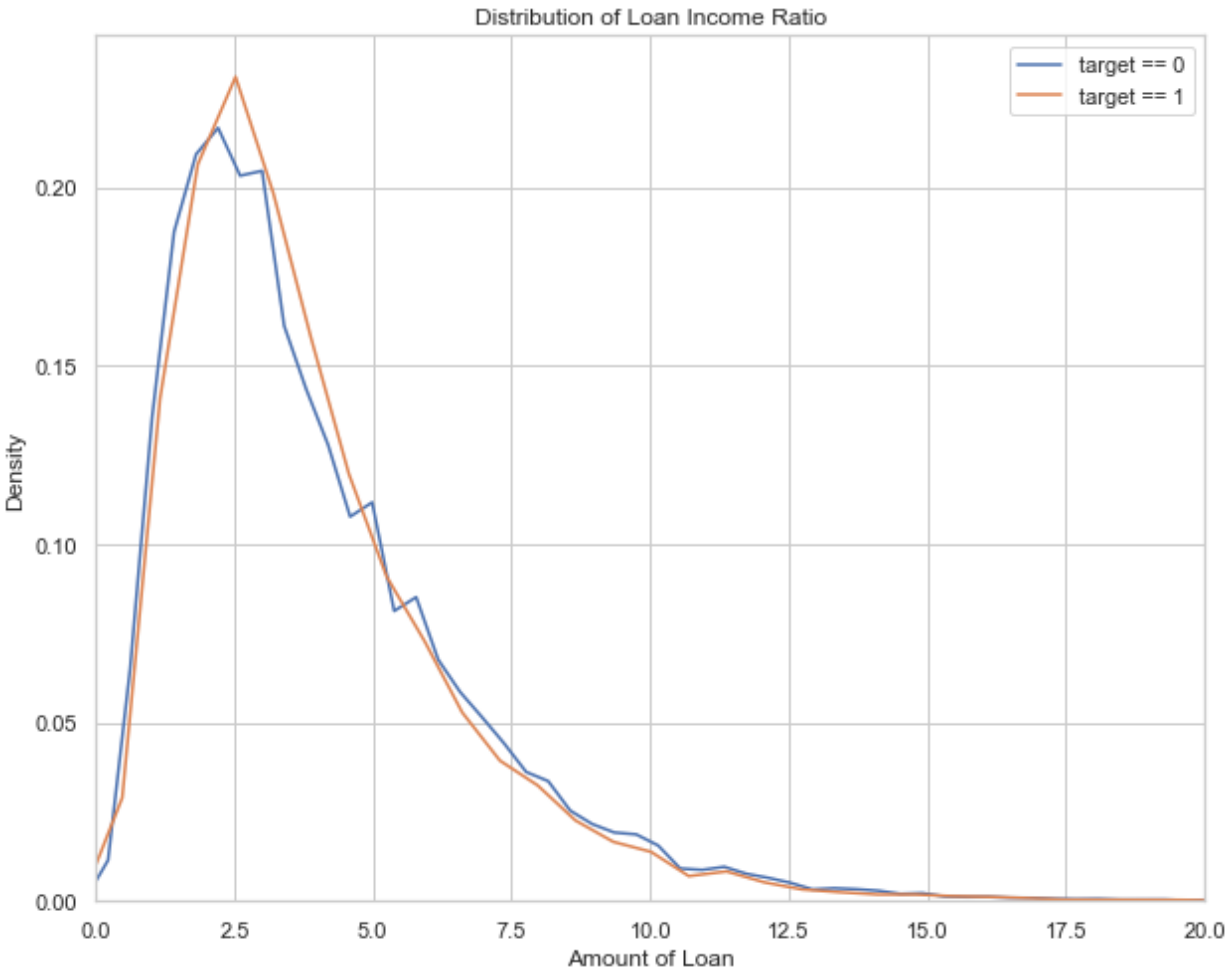


7. Amount Credit Vs Target:



8. **LOAN_INCOME_RATIO VS Target:**

Home Loan Credit Risk – Milestone Report



Home Loan Credit Risk – Milestone Report

Exploratory Data Analysis:

Data Preparation:

- Taken Data wrangling Merged data for EDA

Resampling technique:

- Application train has imbalanced data
- Has more repaid history & lower non repaid record
- We need to down sample our datasets prior modeling
- We have applied sklearn.resample for down sampling our dataset

Feature Selection:

- We have more than 150 features in dataset, so we going to pick relevant features for predicting the loan repayment
- **Mutual Info Classifier & K best** has been used to identify the Best correlated feature which matches with TARGET variable based on fs score
- Top 20 features identified from above technique

Top 20 Features:

- FLAG_OWN_REALTY
- AMT_CREDIT_x
- AMT_ANNUITY_x_1
- AMT_GOODS_PRICE_x
- NAME_TYPE_SUITE
- NAME_INCOME_TYPE
- NAME_EDUCATION_TYPE
- NAME_FAMILY_STATUS
- NAME_HOUSING_TYPE
- FLAG_MOBIL
- FLAG_EMP_PHONE
- FLAG_CONT_MOBILE
- OCCUPATION_TYPE
- CNT_FAM_MEMBERS
- REGION_RATING_CLIENT
- REGION_RATING_CLIENT_W_CITY
- EXT_SOURCE_1
- EXT_SOURCE_2
- EXT_SOURCE_3
- FLAG_DOCUMENT_3

Home Loan Credit Risk – Milestone Report

Assumption1: Above top features are statistically significant or not for Prediction – Using Stats Model

Results:

#Test 1

OLS Regression Results

Dep. Variable:	TARGET	R-squared:	0.003
Model:	OLS	Adj. R-squared:	0.003
Method:	Least Squares	F-statistic:	945.3
Date:	Thu, 22 Oct 2020	Prob (F-statistic):	3.02e-207
Time:	15:19:04	Log-Likelihood:	-35977.
No. Observations:	291057	AIC:	7.196e+04
Df Residuals:	291055	BIC:	7.198e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0708	0.001	113.502	0.000	0.070	0.072
CODE_GENDER	0.0330	0.001	30.746	0.000	0.031	0.035

Omnibus:	173929.323	Durbin-Watson:	2.000
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1084056.111
Skew:	3.034	Prob(JB):	0.00
Kurtosis:	10.250	Cond. No.	2.41

#Test 2

OLS Regression Results

Dep. Variable:	TARGET	R-squared (uncentered):	0.069
Model:	OLS	Adj. R-squared (uncentered):	0.069
Method:	Least Squares	F-statistic:	7150.
Date:	Thu, 22 Oct 2020	Prob (F-statistic):	0.00
Time:	15:19:05	Log-Likelihood:	-38539.
No. Observations:	291057	AIC:	7.708e+04
Df Residuals:	291054	BIC:	7.712e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
FLAG_OWN_REALTY	0.0381	0.001	41.765	0.000	0.036	0.040
AMT_CREDIT_x	-3.054e-08	2.04e-09	-14.985	0.000	-3.45e-08	-2.65e-08
AMT_ANNUITY_x_1	2.229e-06	5.11e-08	43.589	0.000	2.13e-06	2.33e-06

Omnibus:	171824.280	Durbin-Watson:	1.997
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1050853.481
Skew:	2.995	Prob(JB):	0.00
Kurtosis:	10.125	Cond. No.	1.26e+06

#Test 3

OLS Regression Results

Dep. Variable:	TARGET	R-squared:	0.002
Model:	OLS	Adj. R-squared:	0.002
Method:	Least Squares	F-statistic:	611.0
Date:	Thu, 22 Oct 2020	Prob (F-statistic):	9.47e-135
Time:	15:19:05	Log-Likelihood:	-36143.
No. Observations:	291057	AIC:	7.229e+04
Df Residuals:	291055	BIC:	7.231e+04
Df Model:	1		

Home Loan Credit Risk – Milestone Report

```
Covariance Type:            nonrobust
================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept            0.0635      0.001     70.442      0.000       0.062       0.065
NAME_INCOME_TYPE     0.0050      0.000     24.718      0.000       0.005       0.005
==============================================================================
Omnibus:                   174197.029   Durbin-Watson:                   2.001
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1088170.548
Skew:                           3.039   Prob(JB):                         0.00
Kurtosis:                      10.265   Cond. No.                         8.25
==============================================================================
```

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

Assumption 2: Feature: OCCUPATION_TYPE significance for prediction by Hypothesis – Through scipy.stats , T stats & P value

H_0: OCCUPATION_TYPE does not have significant importance for repaying loan

H_1: OCCUPATION_TYPE has significant impact for Repaying loan

Results:

```
(9.994138148092723, 1.631066018444426e-23)
```

As we got 9.9 as T value and p value almost 0 which is <=0.05, so we can reject H_0. That means occupation type has significant impact on repaying loan

Assumption 3: Feature NAME_INCOME_TYPE – Replications by random choice, P value from shifted Mean

H_0: There is no impact for repaying loan based on no of Income Type

H_1: There is significant impact on repaying loan based on Income Type

Results:

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

Conclusion:

- We can conclude above picked features are statistically significantly for loan repayment prediction modeling
- Top 20 featured in down sampled (column) dataset saved for modeling purpose

Home Loan Credit Risk – Milestone Report

Modeling & In-Depth Analysis:

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

Early Models: As data set belongs to Supervised Machine learning, applied below Algorithms

Metrics to be used: ROC-AUC Score

Machine Learning Algorithm Used:

- Logistic Regression
- Random Forest
- Naïve Bayes
- XGB Boost
- Ensemble Modeling (combined above 4)
- Deep Learning – Keras

1. Logistic Regression Results:

```
The accuracy score :  0.6712972562029166
The classification report is as follows:
              precision    recall  f1-score   support

           0       0.65      0.61      0.63      6578
           1       0.69      0.73      0.71      7891

    accuracy                           0.67     14469
   macro avg       0.67      0.67      0.67     14469
weighted avg       0.67      0.67      0.67     14469

Taregt Values:
1    8327
0    6142
Name: target, dtype: int64
ROC AUC score is:  0.6658108026204868
```

2. Random Forest Algorithm Results:

```
The accuracy score :  0.6746838067592784
The classification report is as follows:
              precision    recall  f1-score   support

           0       0.65      0.61      0.63      6578
           1       0.69      0.73      0.71      7891
```

```
    accuracy                           0.67     14469
   macro avg       0.67      0.67      0.67     14469
weighted avg       0.67      0.67      0.67     14469

Taregt Values:
1    8326
0    6143
Name: target, dtype: int64
ROC AUC score is:  0.6692317960672662
```

3. Naïve Bayes Algorithm Results:

```
The accuracy score :  0.6523602183979542
The classification report is as follows:
              precision    recall  f1-score   support

           0       0.62      0.62      0.62      6578
           1       0.68      0.68      0.68      7891

    accuracy                           0.65     14469
   macro avg       0.65      0.65      0.65     14469
weighted avg       0.65      0.65      0.65     14469

Taregt Values:
1    7803
0    6666
Name: target, dtype: int64
ROC AUC score is:  0.6500302040198895
```

4. XG Boost Results:

```
The accuracy score :  0.6703987836063308
The classification report is as follows:
              precision    recall  f1-score   support

           0       0.65      0.60      0.62      6578
           1       0.69      0.73      0.71      7891

    accuracy                           0.67     14469
   macro avg       0.67      0.66      0.67     14469
weighted avg       0.67      0.67      0.67     14469

Taregt Values:
1    8370
0    6099
Name: target, dtype: int64
ROC AUC score is:  0.6646329460239638
```

5. Ensemble Method approach & Results:

Home Loan Credit Risk – Milestone Report

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 score :  0.6767572050590919
The classification report is as follows:
              precision    recall  f1-score   support

           0       0.64      0.66      0.65      6578
           1       0.71      0.69      0.70      7891

    accuracy                           0.68     14469
   macro avg       0.67      0.68      0.68     14469
weighted avg       0.68      0.68      0.68     14469

Taregt Values:
1.0    7634
0.0    6835
Name: target, dtype: int64
ROC AUC score is:  0.6756984867435408
```

Observation: Ensemble provides slight better result than others, though not significant improvement

6. **Deep Learning -Keras Results – 500 epochs:**

```
Epoch 250/250
29376/29376 [==============================] - 3s 105us/sample - loss: 0.4
820 - acc: 0.7574 - auroc: 0.8446
```

Validation set score

```
ROC AUC score is:  0.6491528271390306
```

In [28]:

Observation: With NN we can achieve 0.64 ROC-AUC score after 500 echoes

Hyper Parameter Tuning:

1. **Logistic Regression – Grid Search CV**
   - Used Grid Search CV technique for tuning C value

- Tuned 'C' value by passing param as np.logspace(0.01,0.01,1)
- We generated best C score as 0.6668709123299972
- Results:

```
The accuracy score :  0.6710208030962748
The classification report is as follows:
              precision    recall  f1-score   support

           0       0.65      0.61      0.63      6578
           1       0.69      0.73      0.71      7891

    accuracy                           0.67     14469
   macro avg       0.67      0.67      0.67     14469
weighted avg       0.67      0.67      0.67     14469

Taregt Values:
1    8323
0    6146
Name: target, dtype: int64
ROC AUC score is:  0.6655573493192576
```

Observation: ROC-AUC Score not much improved after tuning the C value - let us try with XG Boost Tuning

2. **XG Boost – Bayes Search CV**
   - Used Bayes Search CV for Hyper tuning XG Boost
   - ***Best ROC-AUC: 0.744***
   - Best params:

```
OrderedDict([
('colsample_bylevel', 0.8015579071911014),
('colsample_bytree', 0.44364889457651413),
('gamma', 3.811128976537413e-05),
('learning_rate', 0.2700390206185342),
('max_delta_step', 18),
('max_depth', 36),
('min_child_weight', 2),
('n_estimators', 83),
('reg_alpha', 1.5057560255472018e-06),
('reg_lambda', 659),
('scale_pos_weight', 256),
('subsample', 0.8835665823899177)])
```

Observation: After Hyper Tuning XGB ROC-AUC Score improved from 0.666 to 0.740

Conclusion*:*

Bayesian Optimization was used to tune the XG Boost Classification models. Subsequently, a kfold cross validation with 8 splits was conducted to evaluate the validity of the models.

Home Loan Credit Risk – Milestone Report

The averaged CV result was a roc_auc of 0.740 with a standard deviation of 0.005.

Also, we can refer Neural Network Keras model for the prediction which has0.844 roc_auc score

This is a significant improvement on the early models that did not include as many variables and only relied on the mean groupings of past financial transactions.

It can therefore be concluded that the new variables significantly help in the identification of features that might make someone more likely to have challenges in repaying a loan.

Future Scope of Work:

In this Capstone Project, we could use below methodological to further improve the models. Here are a few ideas of what I would try if time allowed:

- We made prediction with 10-15 % of features, if we make prediction with all features together, we may get better results.
- We used ordinal encoder for the categorical conversion for all algorithms, if we use specific encoder for specific algorithms, we may get better accuracy.
- In Neural Network keras model we may try with Embedded layer for better learning & accuracy improvement