CAPSTONE - Predicting Home Credit Loan Default

MILESTONE REPORT

Introduction

Credit defaults pose a major threat to the profitability of banks and other lending institutions. The problem of bad loans has taken gigantic proportions in India and all over the world. In the past few years,we have seen a few high-profile cases of defaults worth thousands of crores, and innumerable instances of medium to minor level loan defaults. Non-performing loans have the potential to impact the GDP of a nation. Hence, it's imperative to find a robust method of credit appraisal, to improve the overall health of the credit portfolio.

Problem Statement

The current credit appraisal process includes analysing past credit history and taking decisions based on the Credit Rating provided by external rating agencies to borrowers (for example, CIBIL score in India). Though this method is somewhat effective, it has many loopholes. We cannot be cent per cent sure about the future, based on just the past records of a borrower's repayment habits. A borrower with a good past record may default, and another with a shady record, may turn out to be a diligent loanee. Hence, we need to build a more detailed and robust appraisal system that takes the general credit health scenario into account in addition to the specific attributes of the customer, and predicts a probability of default, based on mathematical modelling. This helps extend credit only to deserving customers, and weed out shady clients.

Value to client

Financial institutions could utilise ML models in adjudicating borrowers and improving their credit portfolios. This will ensure proper and timely credit delivery to deserving customers, while weeding out untrustworthy and erratic customers. The total NPA amount in India is estimated to be about 6.93 lakh crore (as on March 2020). If a model helps reduce this by just 0.1%, NPA amount will decrease by 693000 * 10000000 *

0.0001 = 693 crore rupees. It is a relatively small amount when considered on a national level, but continuous improvement may help increase predictive accuracy.

The stakeholders

The top management of financial institutions are the first deciding and reviewing point. On approval, new processes trickle down to branch level.

Source of the dataset

The dataset has been obtained from a past Kaggle competition organised by a HOME CREDIT, a non-banking financial company that lends primarily to people with very less or non-existent credit history. Hence, judging the 'credit-worthiness' of the borrower and quantifying credit risk is very important to minimize losses.

Broad Methodology for EDA and modelling

1. Cleaning and preprocessing - data provided to us maybe in incorrect format, containing garbage values, missing values, or in some other unusable form. Machine Learning models cannot work with data in this form. Hence, we need to preprocess the data and "clean" it to bring it into the desired form for modelling
2. Carrying out visual EDA for getting a general sense of data and understanding the relationships between various attributes- the dataset has a large number of columns. A careful and in-depth study of the distribution of the features can provide hidden trends in customer behaviour. In addition to aiding us in developing a model, uncovering such trends can help credit appraisal teams look for specific traits in customers, or discard other traits.
3. Ascertaining critical attributes and their effect on target - being given a large number of attributes may be good or a bad thing. Proper utilization of all features to improve modelling accuracy, can result in better results. On the other hand, incorporating unimportant and insignificant features into our model can make our model more complex than is optimally required, and decrease predictive accuracy
4. Applying ML models to predict default probability - the final aim of this capstone is to develop a model to predict probability of default. Though we are not classifying a customer as "good" or "bad" directly, this is a classification problem because we are given the training data target as binary values of 0 and 1 and we have to train a classifier using this data. At the first look, it seems that LogisticRegression with optimal hyperparameters may suffice for the task, as it

can output probabilities and is good for binary classification. As the modelling progresses, we will try out other models, taking into account various combinations of features (dropping or keeping features) based on significance tests.

Brief description of the dataset

- The main data is present in the file “application_train.csv”. There are many auxiliary files that contain details about the customers credit card balances, previous inquiries for credit, credit rating bureau data, etc. The primary dataset is the one with the core features of the customers, and containing the TARGET values, representing whether it was a “good” or “bad” loan. This dataset will be used for our modelling, as the attributes present in this dataset contain rich and varied information about a customer’s demographics and personal history.

- Each loan has been given a unique ID, called the SK_ID_CURR, that represents the application ID of the current loan of the customer in the Home Credit database. This field is insignificant for modelling, and hence will be used only to analyse trends in exploratory data analysis.

- WIth common sense and domain knowledge, it is evident that some attributes may turn out to be more important than others in modelling. For example, the default probability may be heavily influenced by the customer’s income. The loan amount of the customer may be dependent on the income. We may observe more defaults by unemployed people. People working in a particular type of industry may be better overall borrowers. The social circle of the borrower may affect his/her repayment habits. A particular type of loan may be getting paid better than others. More educated people may be less prone to inconsistent repayment habits. As we carry out EDA, we will be able to find out the truth about these assumptions.

- There are 123 columns in the main dataset, comprising categorical and numeric features. The pandas types allocated to the features are “int64”, “float64” and “object”. But we need to apply appropriate conversions, depending on the feature’s meaning.

- Features like client’s income, credit amount, etc are numeric, and no preprocessing is required for them as they have been identified correctly by pandas. Null values in these columns are replaced with either a central

statistic(mean/median/mode), or 0, depending on the significance of the feature on the TARGET. The detailed process for handling null values has been described below.

- Some columns are stored as type ‘object’, but in reality, they are categorical. The values in these fields are in the correct format, and no string preprocessing is required as these will be converted to categorical type for modelling. Null values will be handled based on the importance of the feature on the TARGET, as described below.

- Some fields contain negative as well as positive values, for example, DAYS_ID_PUBLISH. This field represents the days passed since the client got their ID changed. Negative and positive values will be analysed coherence and consistency with the real world. They will be handled accordingly.

Preprocessing and data cleaning techniques used

(i) Removal of null values

- Columns having a small number of null values - some features like customer income have a very few number of null values. The rows having null values contain valuable information about other features. Hence, the null values were replaced by the median values of those columns. The data contains some outliers, hence the median was used instead of the median.
- Some columns have a huge number of null values. For example, the feature ‘OWN_CAR_AGE’ has more than two-thirds of the values as null. This is expected, since many customers of Home-Credit do not own cars. To handle such features, we evaluated the statistical significance of the feature on the target variable. If the feature was statistically significant, we will proceed with replacing the null values, with either 0, or a central number representative of the feature (mean or median). To test the statistical significance of the feature, we will use either F-test in regression or ANOVA analysis. Here, since the output is categorical variable and inputs are numeric, ANOVA would be a better choice. If the feature was not significant, the it was dropped.
- In many cases, numerical data was present in fields that could actually be treated as categorical, for example, the rating of the region where the client lives. The rating can be 1,2, or 3. Clearly, this can be handled better by separating these into categories.

- In many categorical attributes, imputation using the default sklearn class “SimpleImputer” was unsuccessful in completing (the code went into an infinite loop). In these cases, manual imputation was carried out. Most of the imputation in categorical variables was done using the mode of a column.

(ii) Handling of data in incorrect format

- The numerical columns contain numbers in the correct format. There are no unnecessary commas, or other special characters. Hence, no preprocessing is required initially. For modelling, these features may be normalized depending on type of model used and feature variability.

- The categorical columns contain data in correct format. The values are informative and meaningful. As these columns will be converted to ‘category’ type from ‘object’ type, no preprocessing is required on the values. As feature selection progresses, the correct method for handling the categorical columns (one-hot encoding, label encoding, etc.) will be decided.

(iii) Handling of outliers

- Since, the dataset has 123 columns, hence, complex relationships may exist between the input columns and the TARGET. As such, handling of outliers needs to be done with caution.
- As the dataset has been obtained from a credible source, very few outliers were observed. Visual and descriptive EDA was carried out to statistically analyse the features and remove outliers. For example, boxplot of the feature ‘AMT_INCOME_TOTAL’ was plotted and data points beyond a limit of 2 times the IQR were removed.

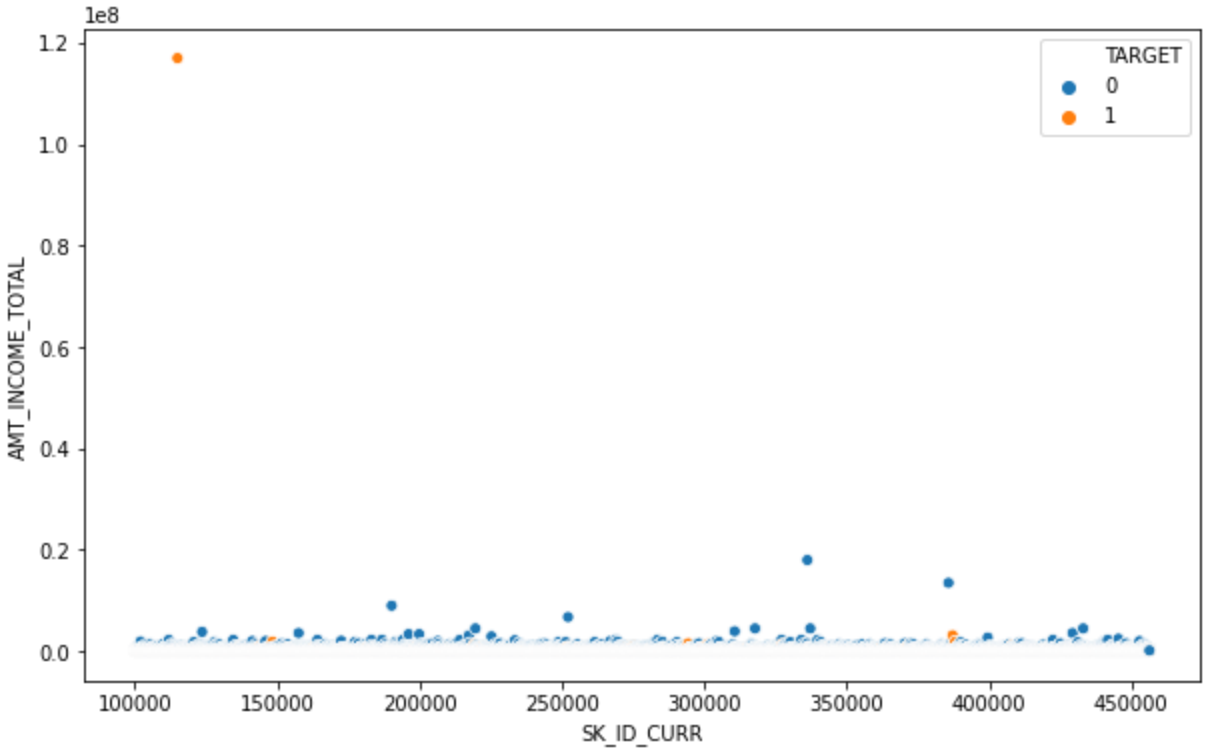
(iv) Removal of insignificant features (feature selection)

- Since the dataset contains a large number of features, there may be multi-collinearity in the features. This is readily evident in some cases, for example, a higher amount of loan will have a higher annuity (EMI). In other cases, it may not be visible. Such multicollinearity will be handled using statistical analysis. Appropriate methods like univariate regression analysis and ANOVA will be used.

Basic Data Story

The dataset can be broadly divided into numerical and categorical features. Features like income, credit amount, external rating, etc are numeric. Features like organization type of client, education level of the client, industry where the client works, type of loan taken, etc are categorical features. There are more categorical features in the data. The following the are the broad findings from basic trend analysis:

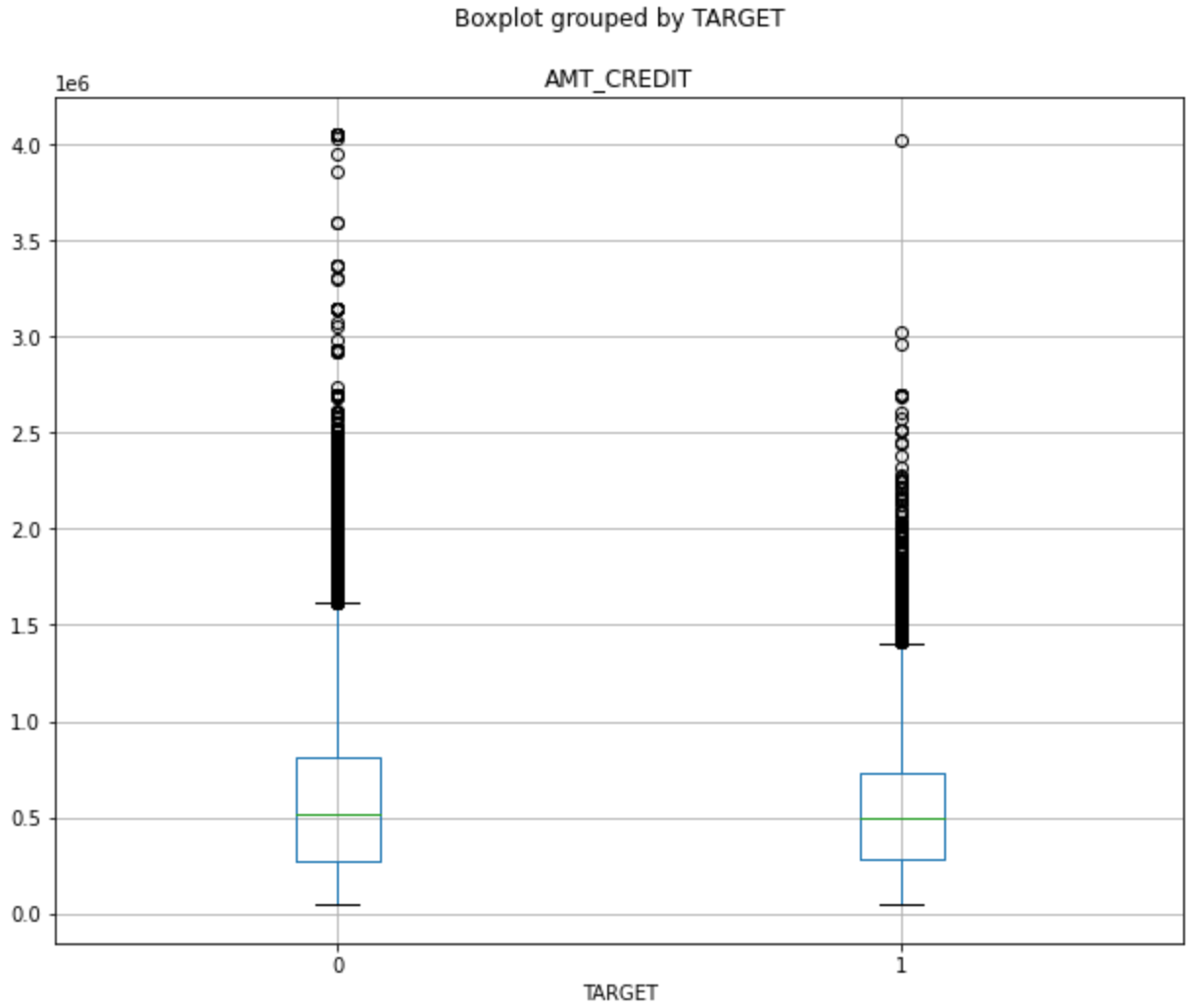
- The median income is 147150 units. The 1st and 3rd quartile of the income are 112500 and 202500 respectively. This seems to indicate that there isn't much spread in the data. But there are many outliers in the data. The highest income is in the range of 1.1 million, and the lowest is 25000 units. Hence, there is a great variability in the income. Upon closer analysis, its revealed that majority of the income values are spread within a standard range of median +- (1.5*IQR). The higher income values may need to be dropped for modelling, Including outliers in the data may result in poor modelling.
- Incidentally, the loan by the borrower with the highest income has turned bad !!!



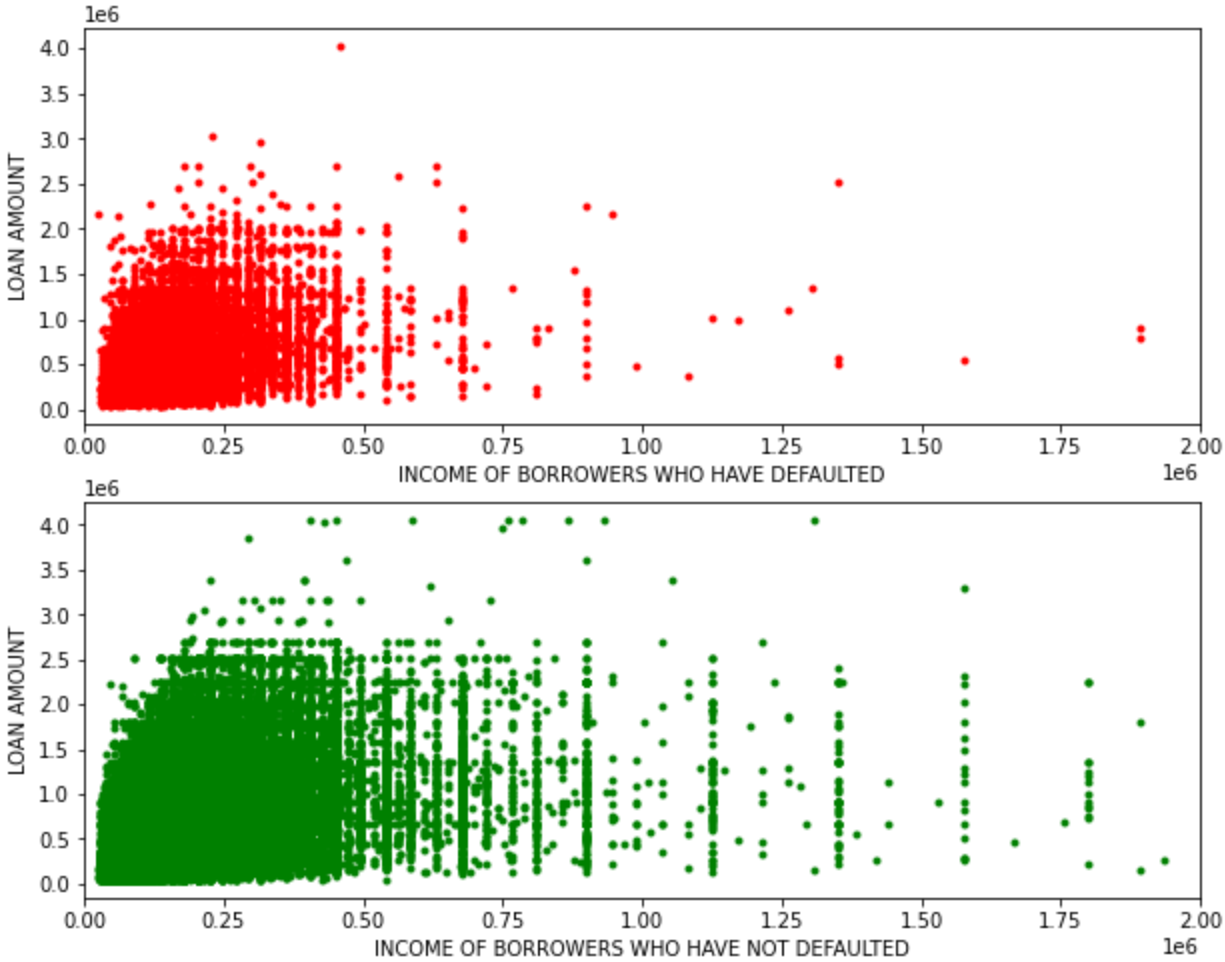
Here, “1” represents bad loans. As we can see, the highest value of income is for a loan defaulter.

- After removing the outliers, the income is distributed somewhat normally.

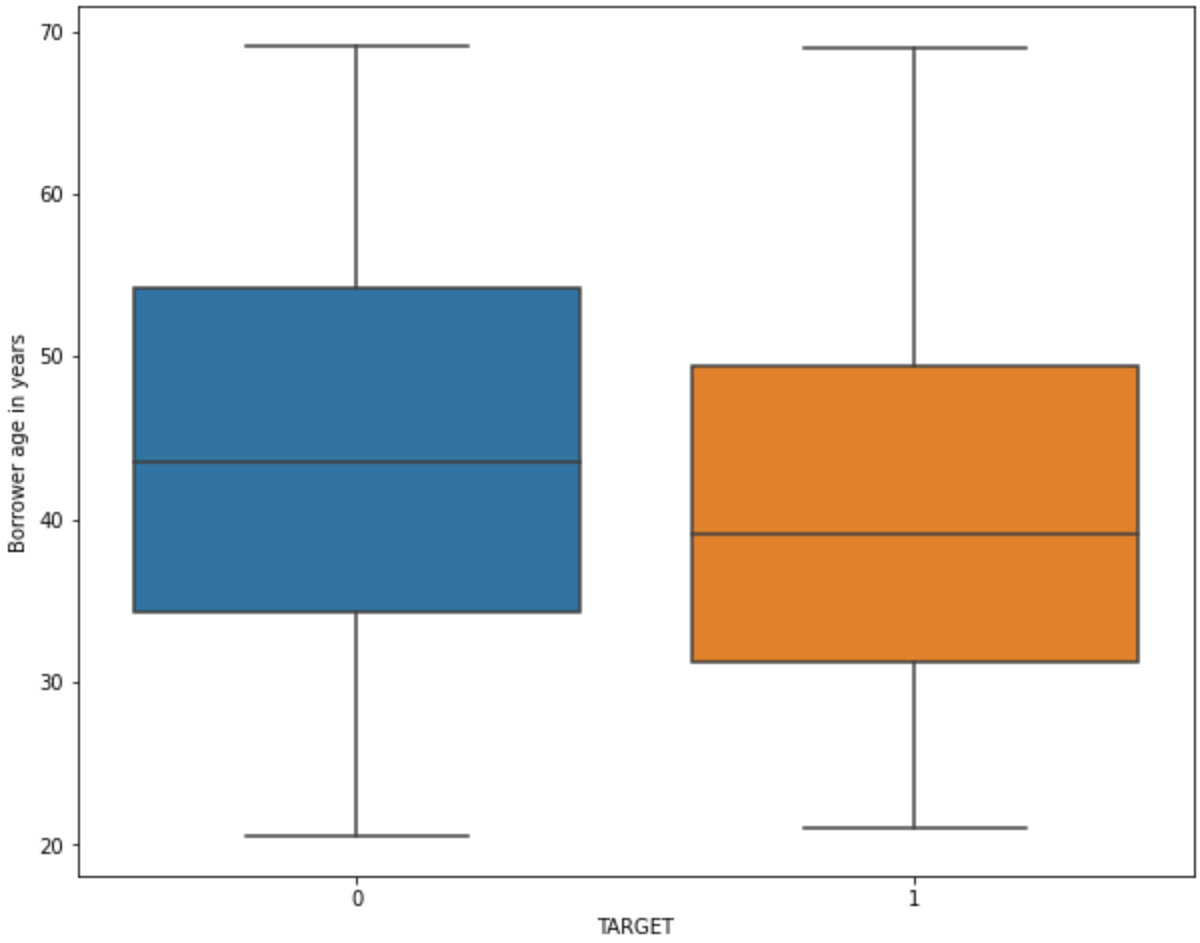
- People with higher amounts of loans are repaying better !! This is an interesting observation. It means that repayment habits are not overly dependent on the amount people have to pay. Some borrowers are just more credit-worthy.



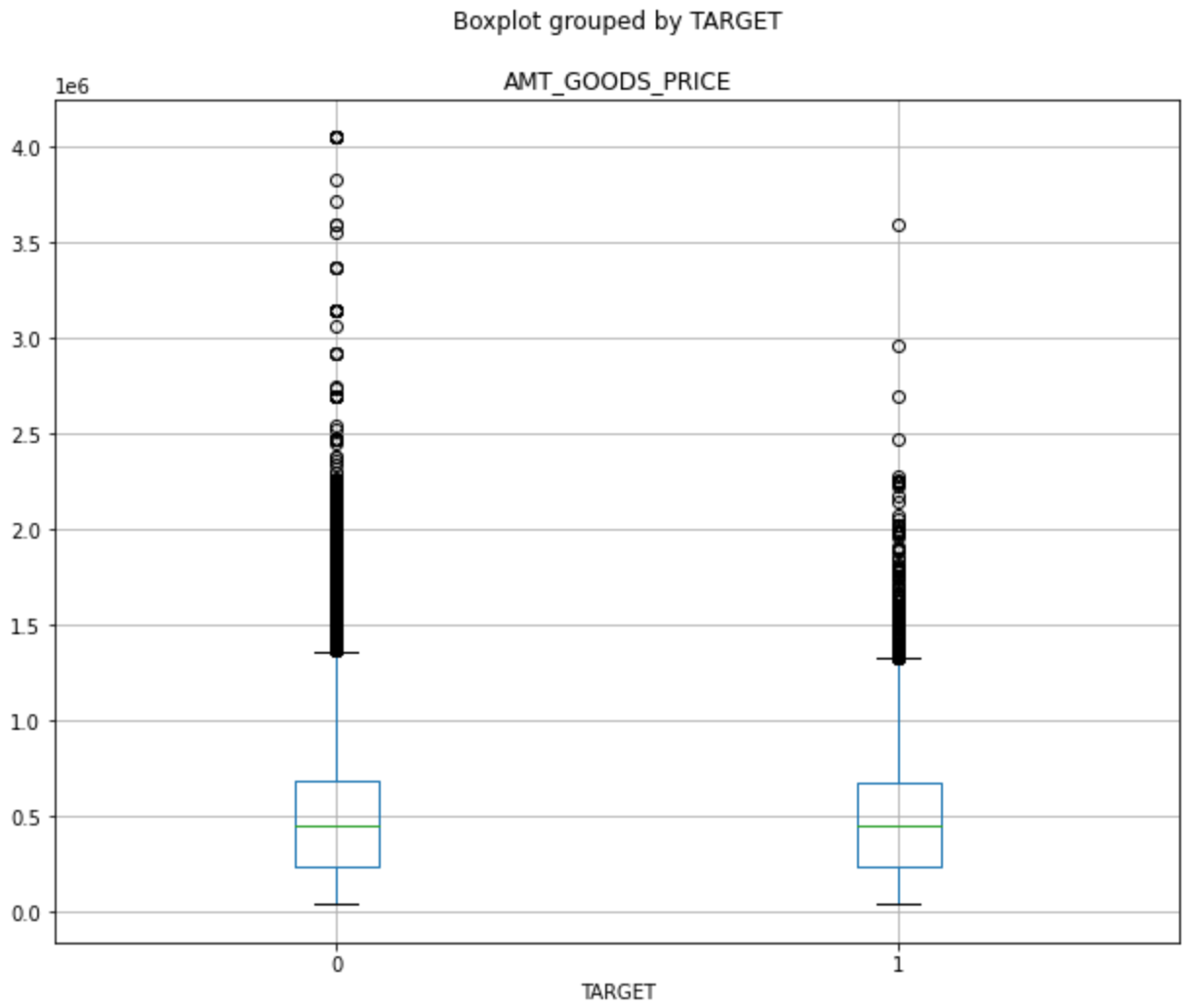
- The majority of the borrowers are in the low-income, low-credit category. Borrowers with high income are more widely present in the good-loans category. This is expected. But we would also expect the borrowers with high incomes to be provided high loans. That is not the case. The loans amounts are fairly similar for low-income as well as high-income borrowers.



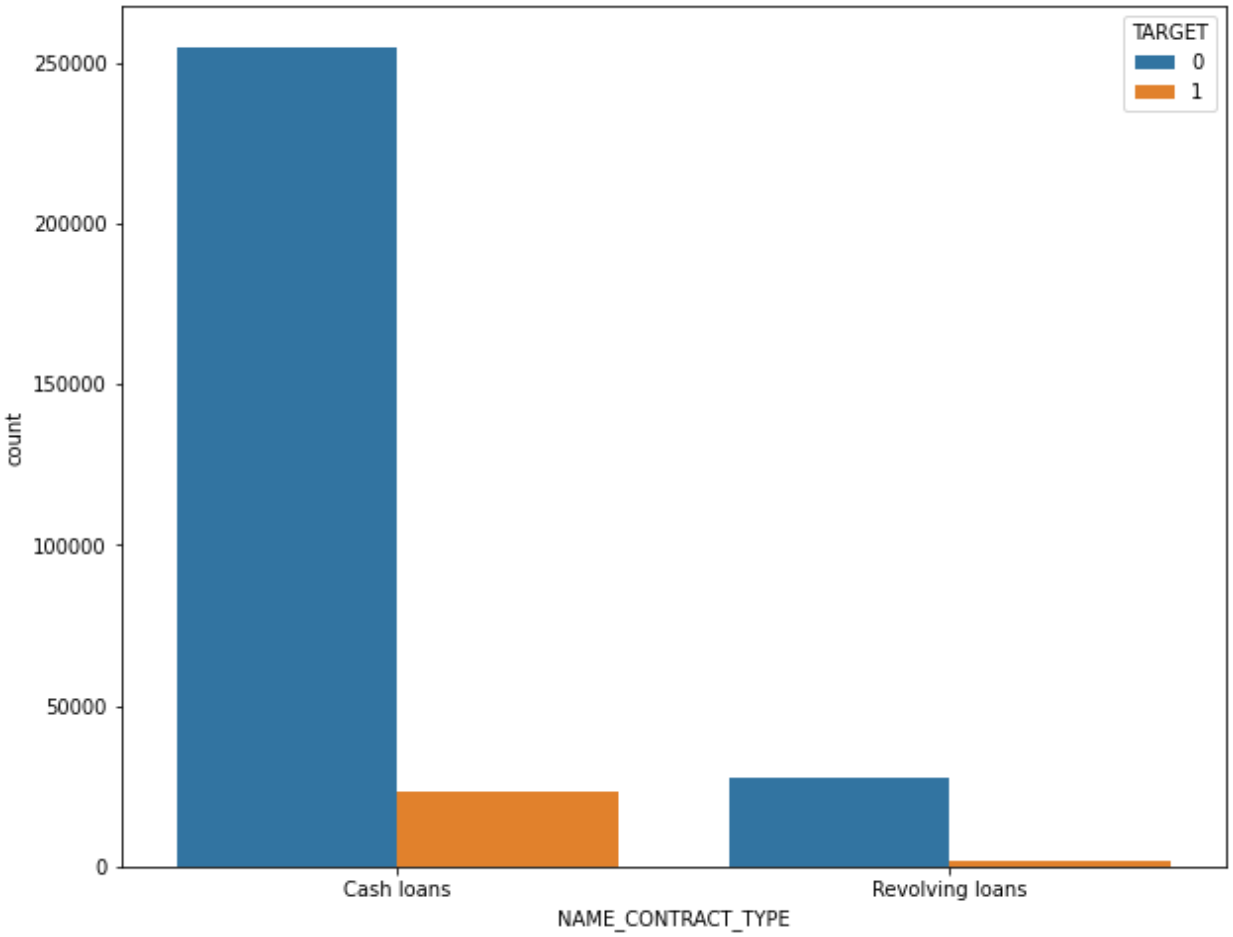
- Younger borrowers seem to default more. This may be due to low-paying jobs (but we have seen that income is not a deciding factor !) or a lack of financial discipline. Its seen that as people grow and have more responsibilities, they develop financial discipline.

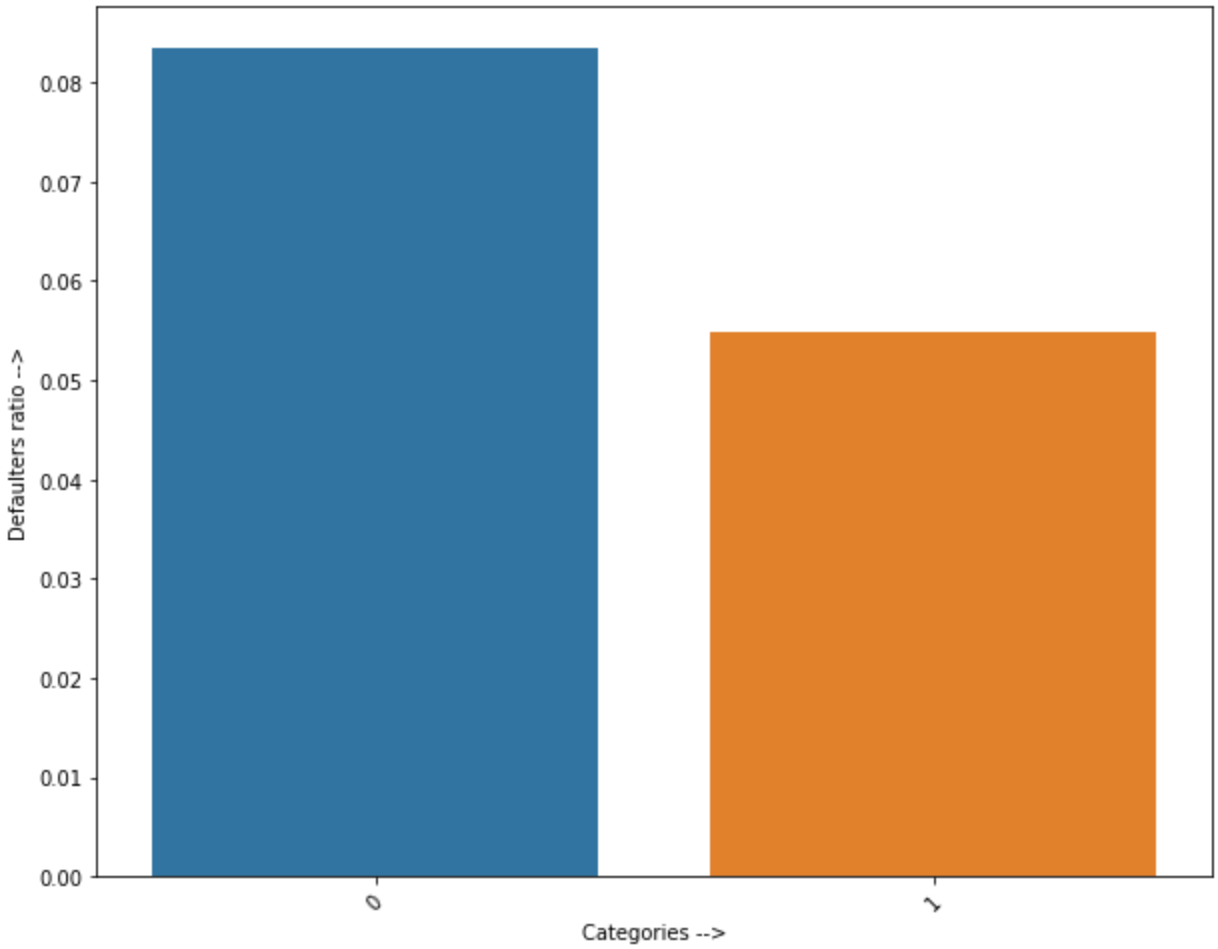


- The median value price of goods purchased with the loan money is fairly similar. This can be seen in the boxplots. But good loans have higher density of high priced items. This too is an interesting observation.



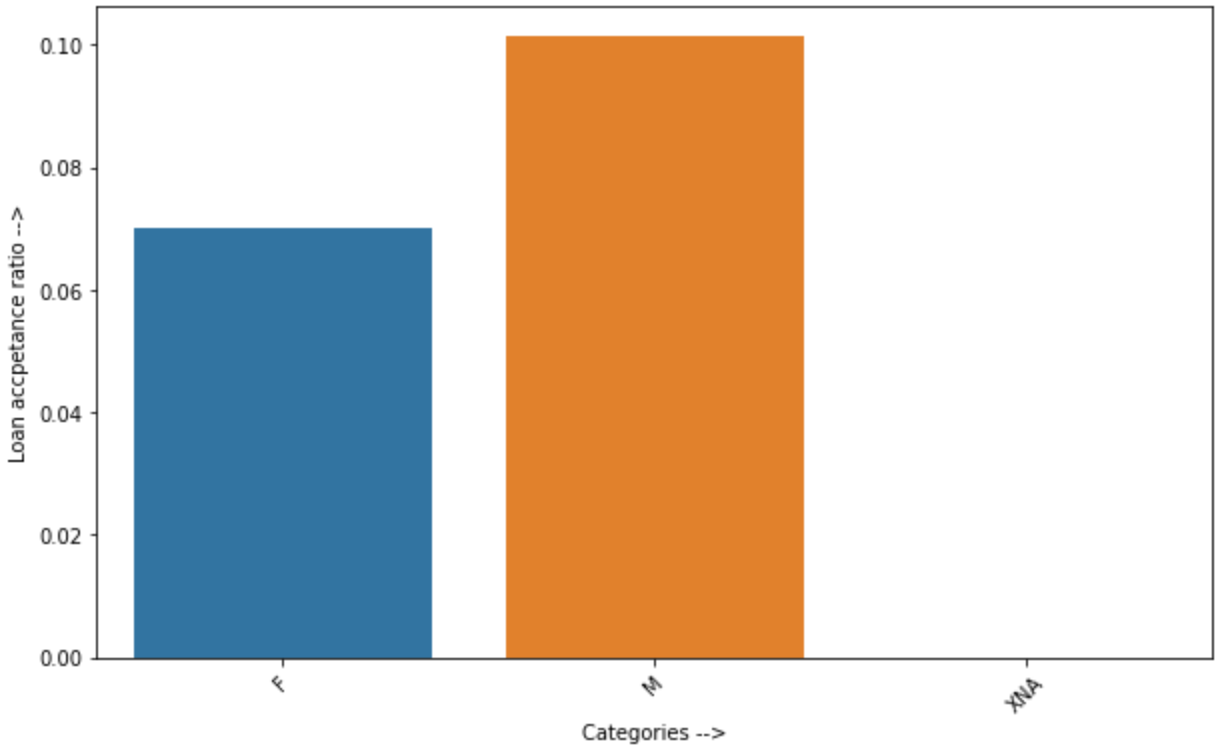
- Cash loans are generally provided for one-time purchase of goods, for example, purchasing a car, mobile phone, etc. They have to repaid in regular intervals, generally as EMIs. When the full amount of interest + principal is repaid, the loan is closed.
- Revolving loans are those where the client is provided a limit upto which he/she can borrow. Repayment renews the limit, it does not generally close the loan. Example is credit cards.
- The dataset has a relatively large number of cash loans. The defaults ratios are as shown


TARGET
0
1
250000
200000
150000
100000
50000
0
count
Cash loans
Revolving loans
NAME_CONTRACT_TYPE



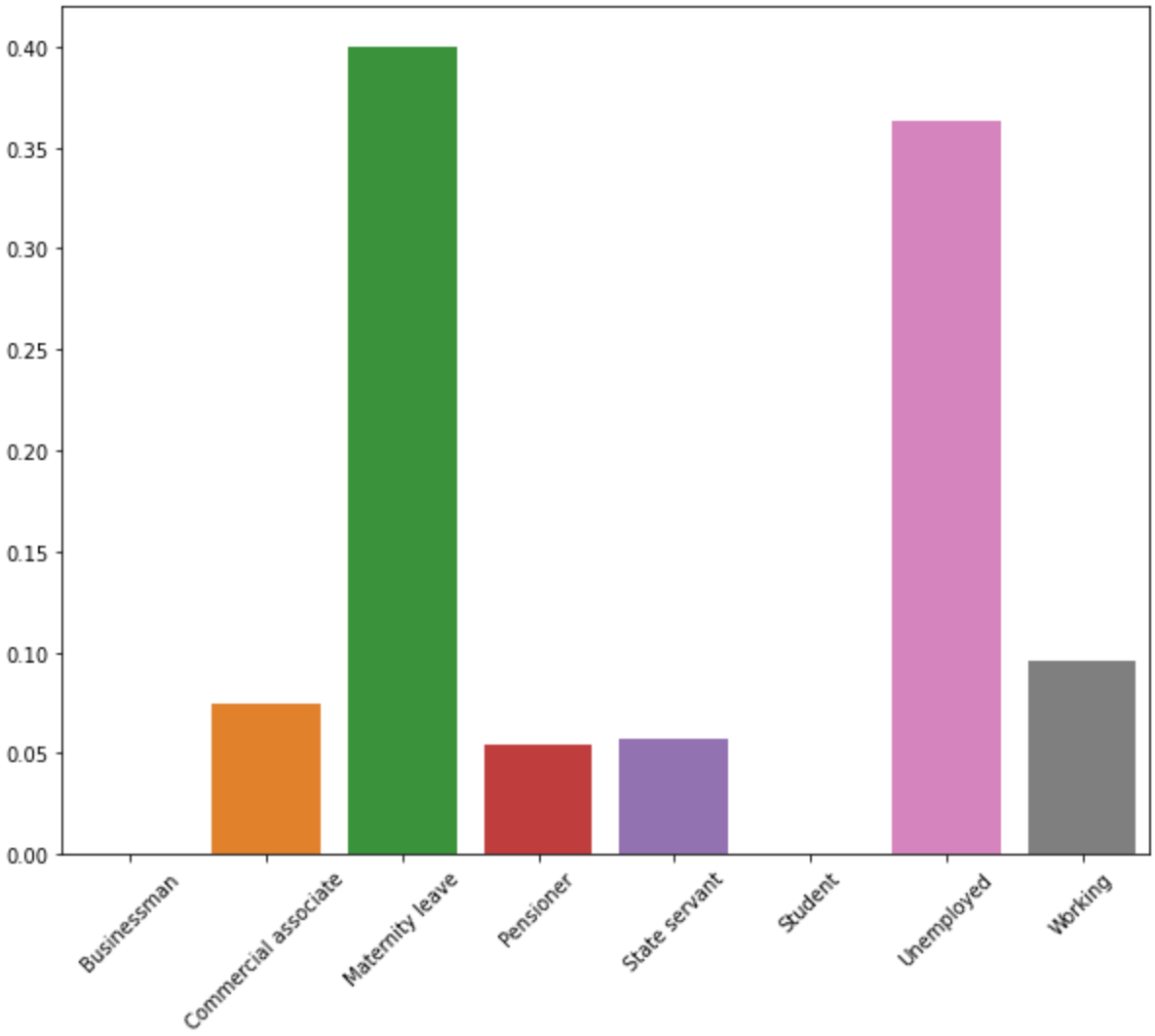
- In the above figure, blue color represents Cash Loans. Thus cash loans seem to have a higher default ratio.

- Under normal circumstances, one assumes that gender does not play much of a role in the default rate, i.e. there is no particular reason why a person of a particular gender will tend to default more. Interestingly, the data presents a different story. Males tend to default more than other genders.



XNA here represents genders other than M or F, or absence of data.

- Unemployed persons have the highest default rate. This is expected. It's interesting that unemployed people were even provided a loan in the first place.



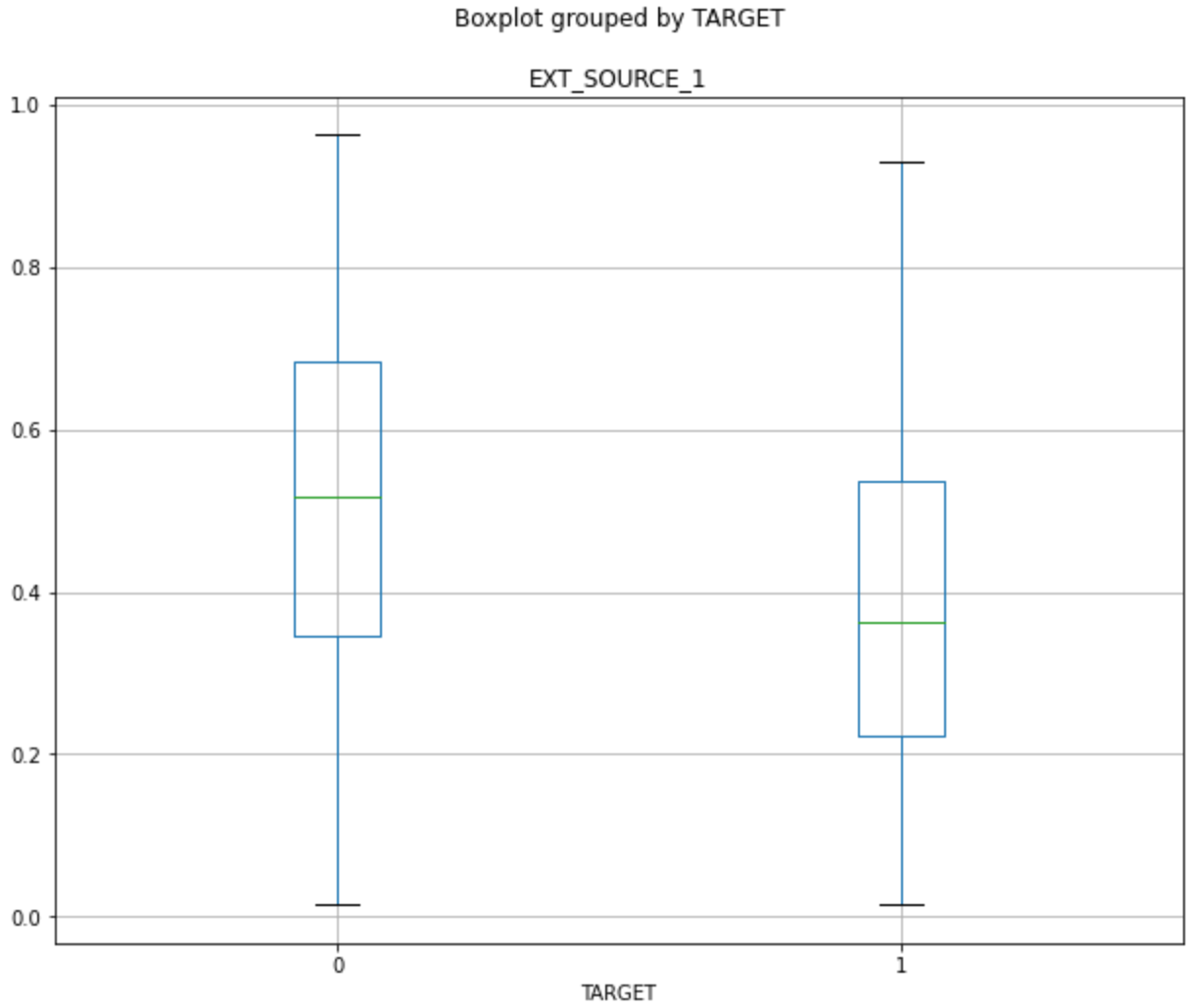
- Businessmen and students have extremely few accepted loan applications .
- Default ratios of pensioners, state servants and salaried class are fairly low.

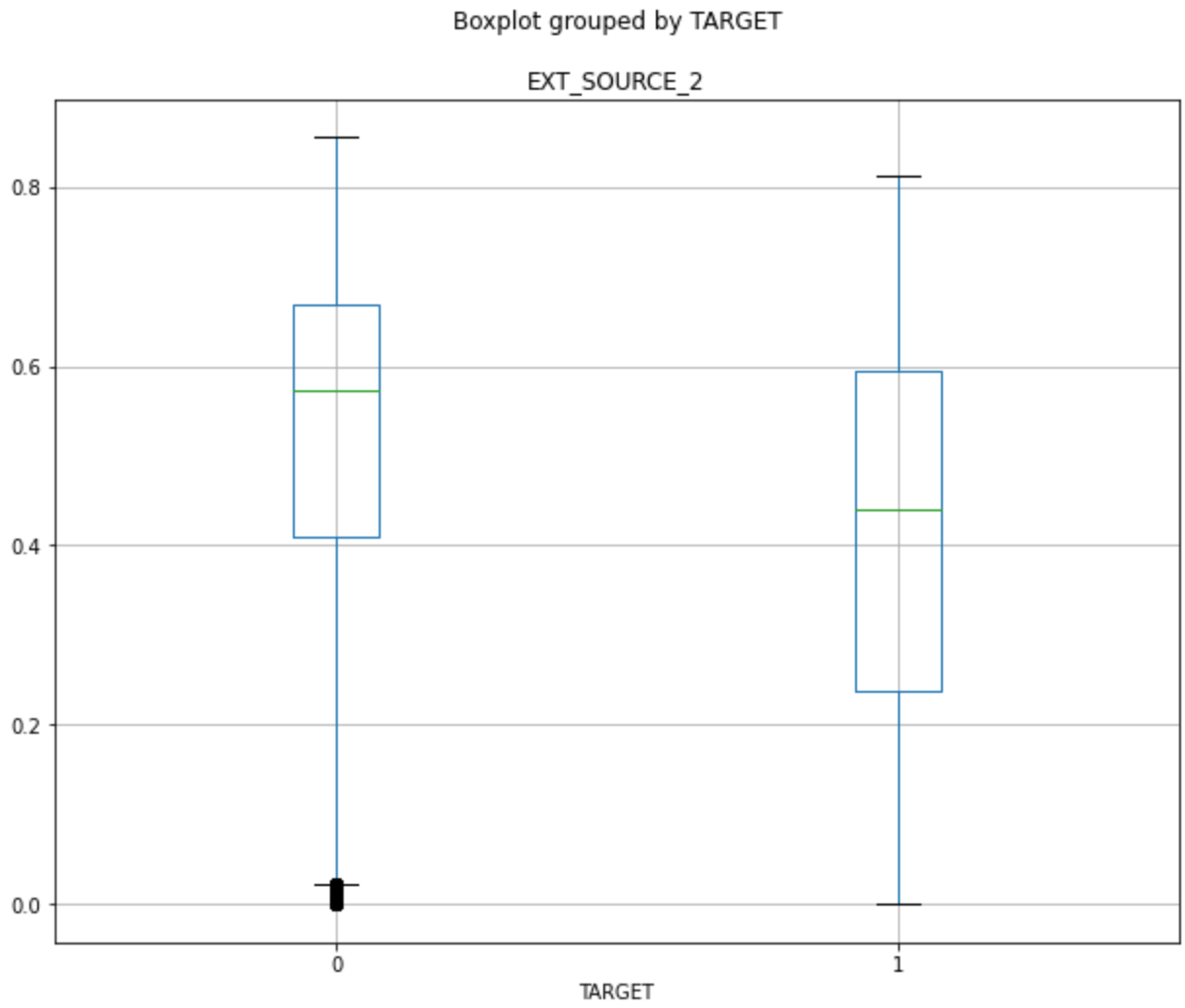
- Apart from the above observations, there are some other interesting observations, as outlined below.

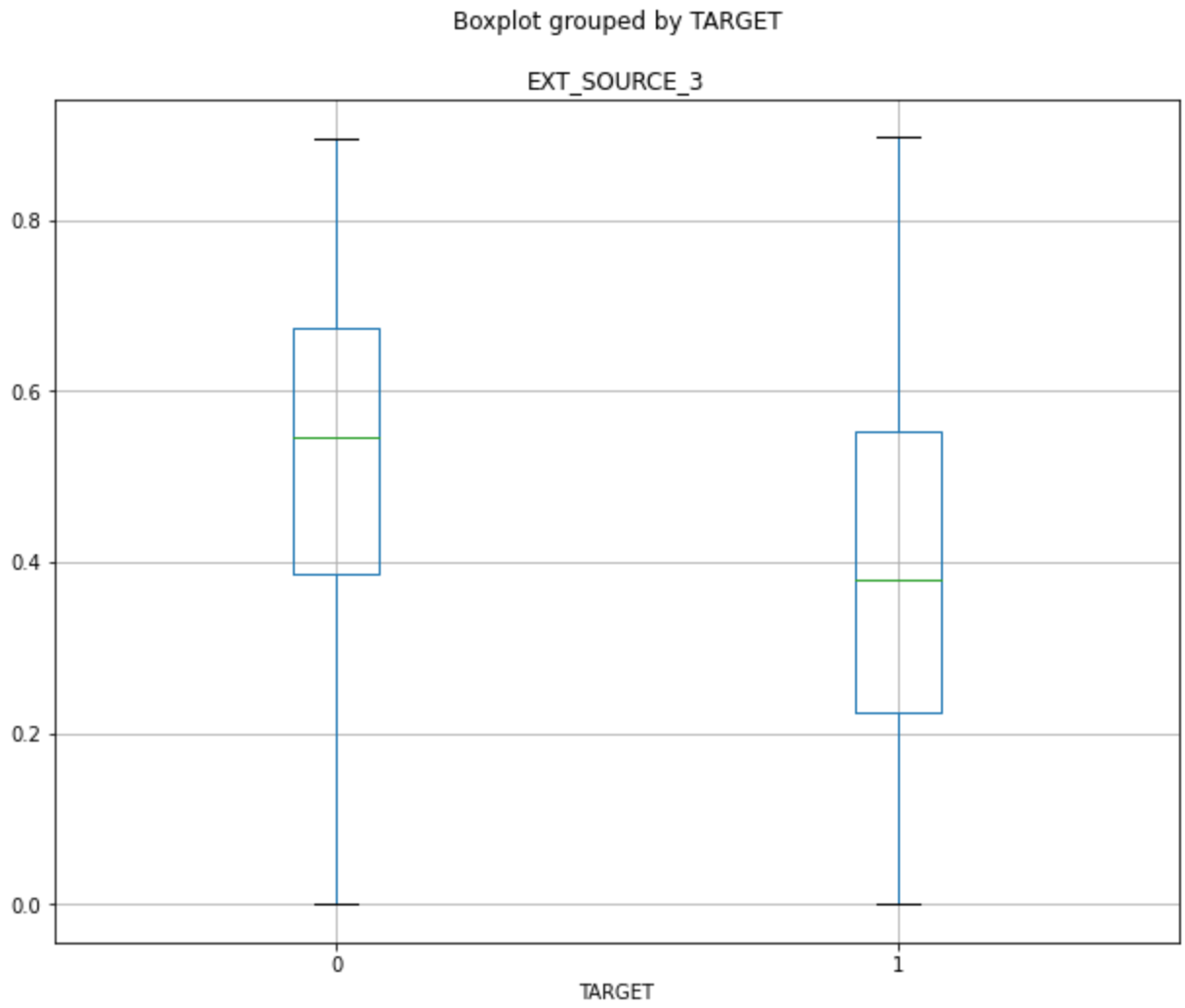
- Less educated people have a lower default ratio.
- Owning a house does not seem to impact the probability of default.
- It does not matter whether the client has provided all the documents or not. Generally, we would expect the people providing all the documents to be diligent about repayment. But there is hardly any significant difference between people who did and did not provide all relevant documents.

- People living in rented apartments or with parents seem to have slightly more tendency to default. This may indirectly be related to income.

- Finally, external ratings seem to provide a very good, although not complete, picture of the credit-worthiness of a borrower. In general, defaulters have been given lower ratings by all 3 external agencies. But some ratings are more reliable than others.




Boxplot grouped by TARGET
EXT_SOURCE_2
0.8
0.6
0.4
0.2
0.0
0
1
TARGET



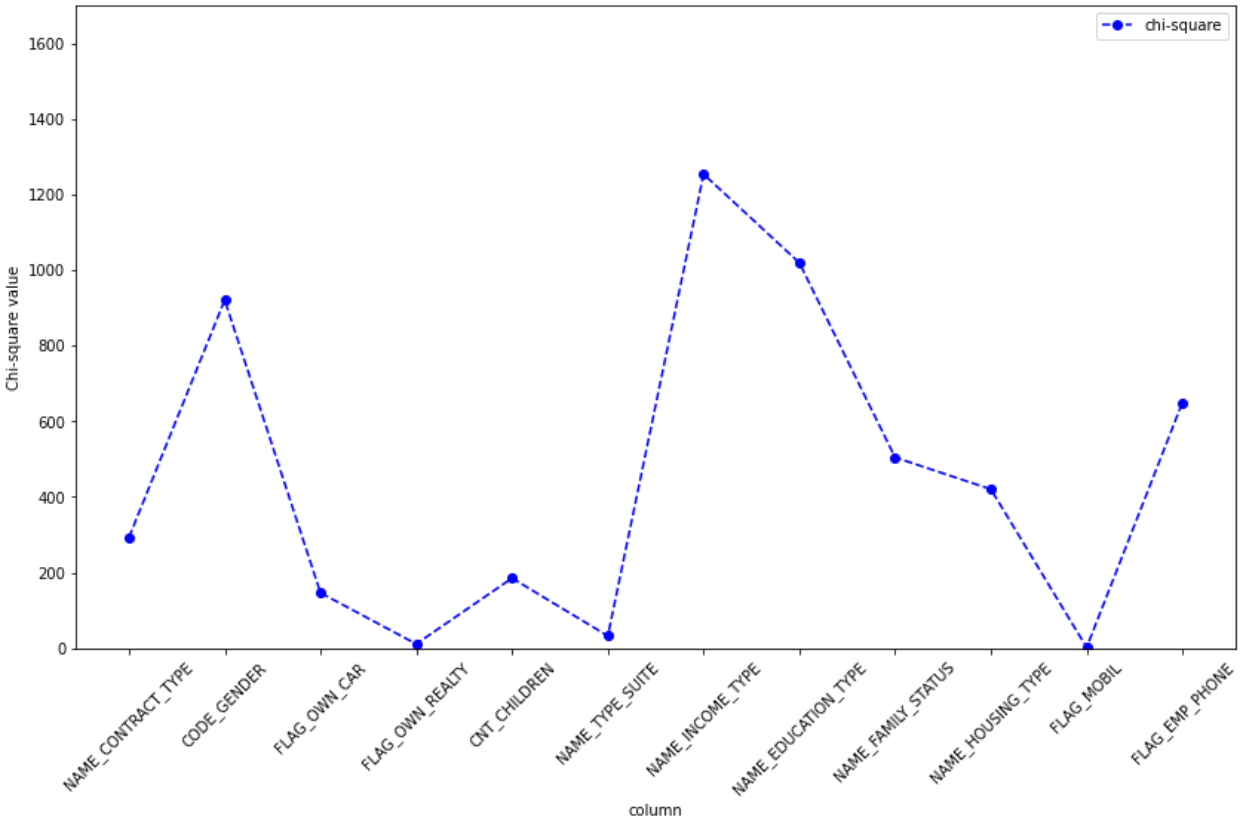
- All the above points present a general behaviour of the customers, when it comes to healthy or unhealthy credit repayment practices.

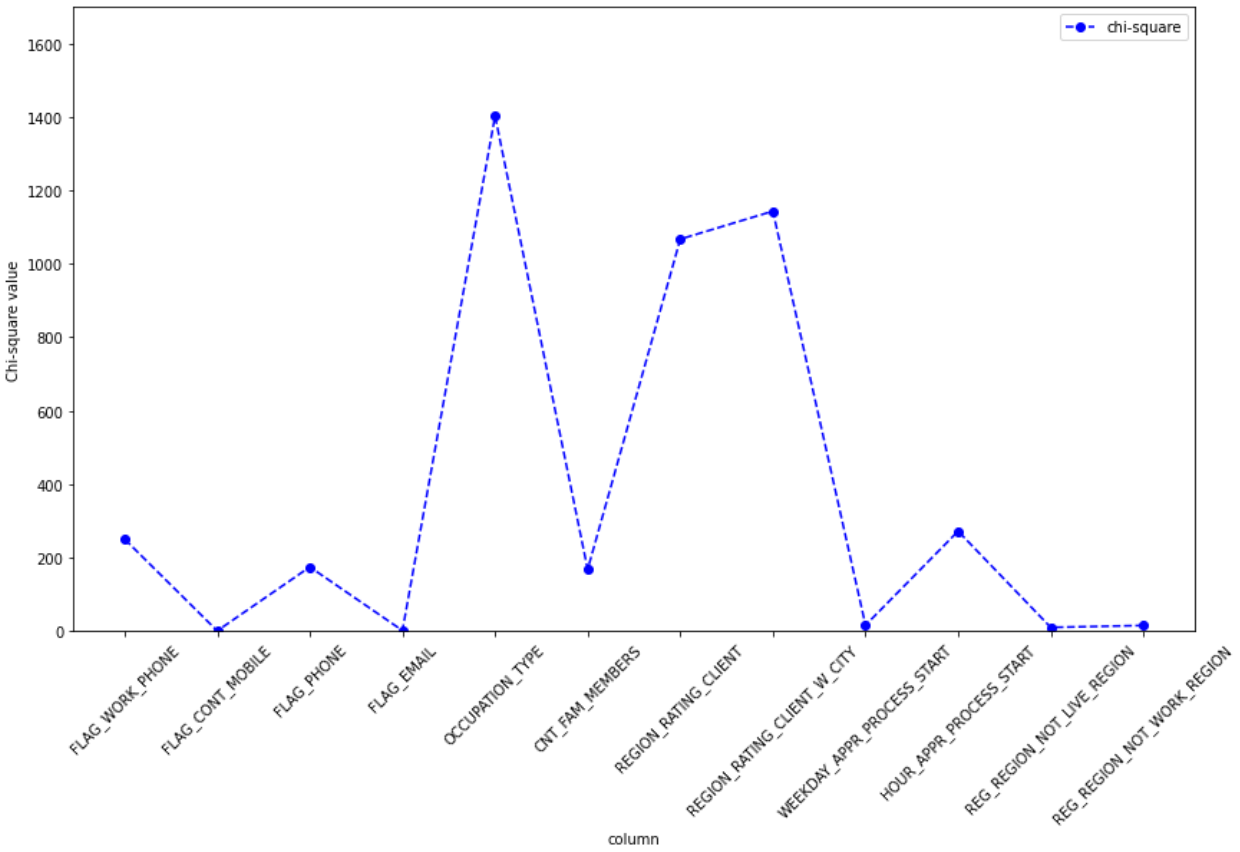
Use of Inferential Statistics to analyse attributes

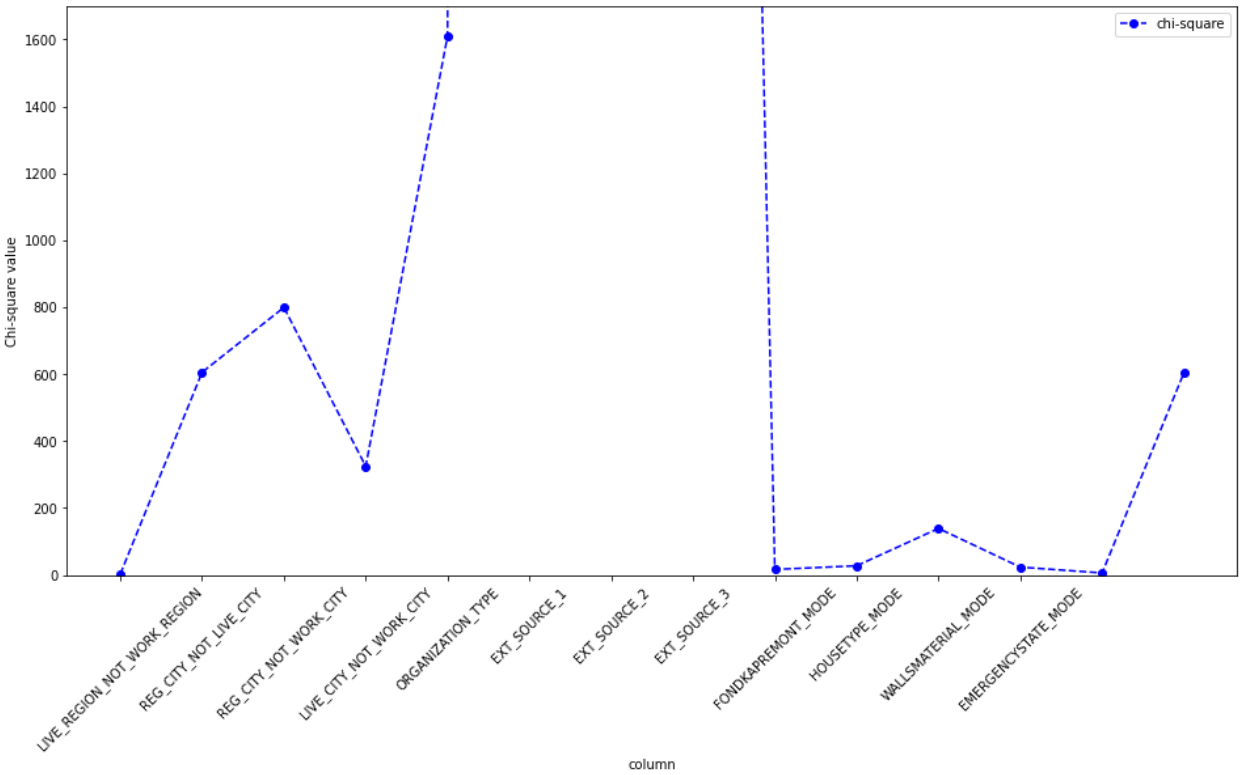
The presence of a large number of features necessitates significance analysis of columns. This will help decrease complexity and overfitting, and more accurate modelling.

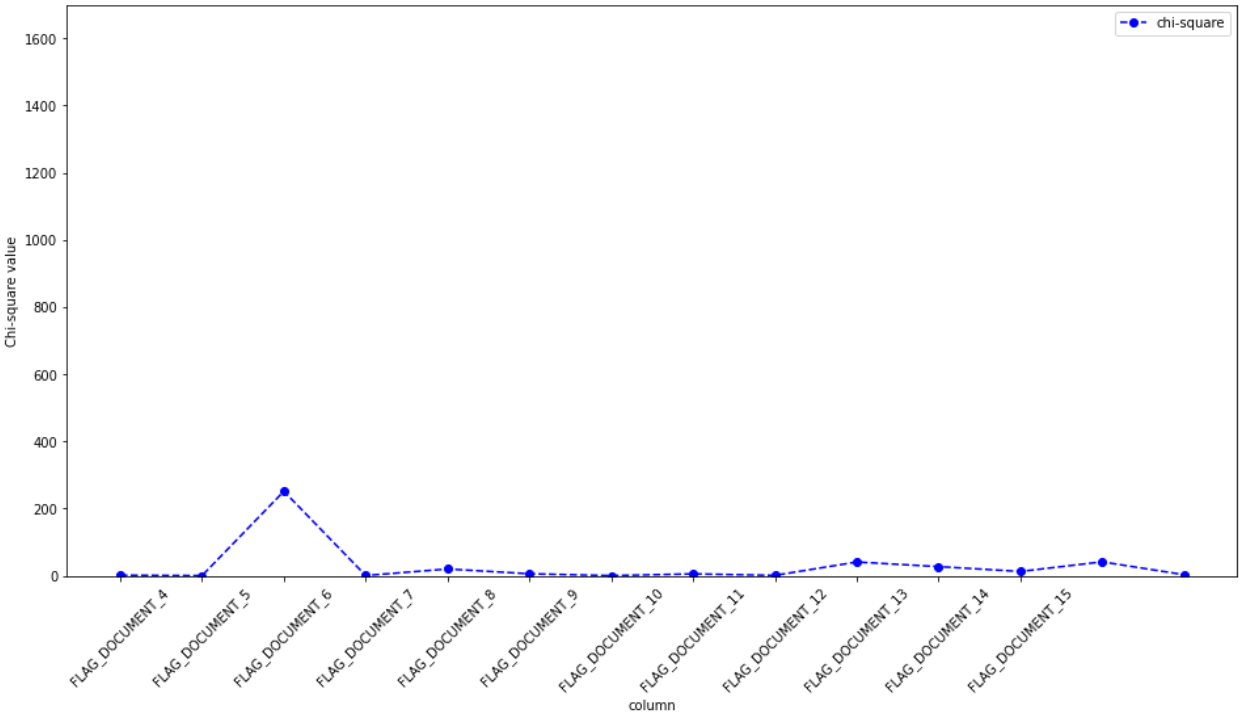
- CATEGORICAL FEATURES - as the TARGET is either 0 or 1, this is a classification problem. Hence, the counts of various categories, distributed by the TARGET were calculated using pandas contingency_table. We then calculated

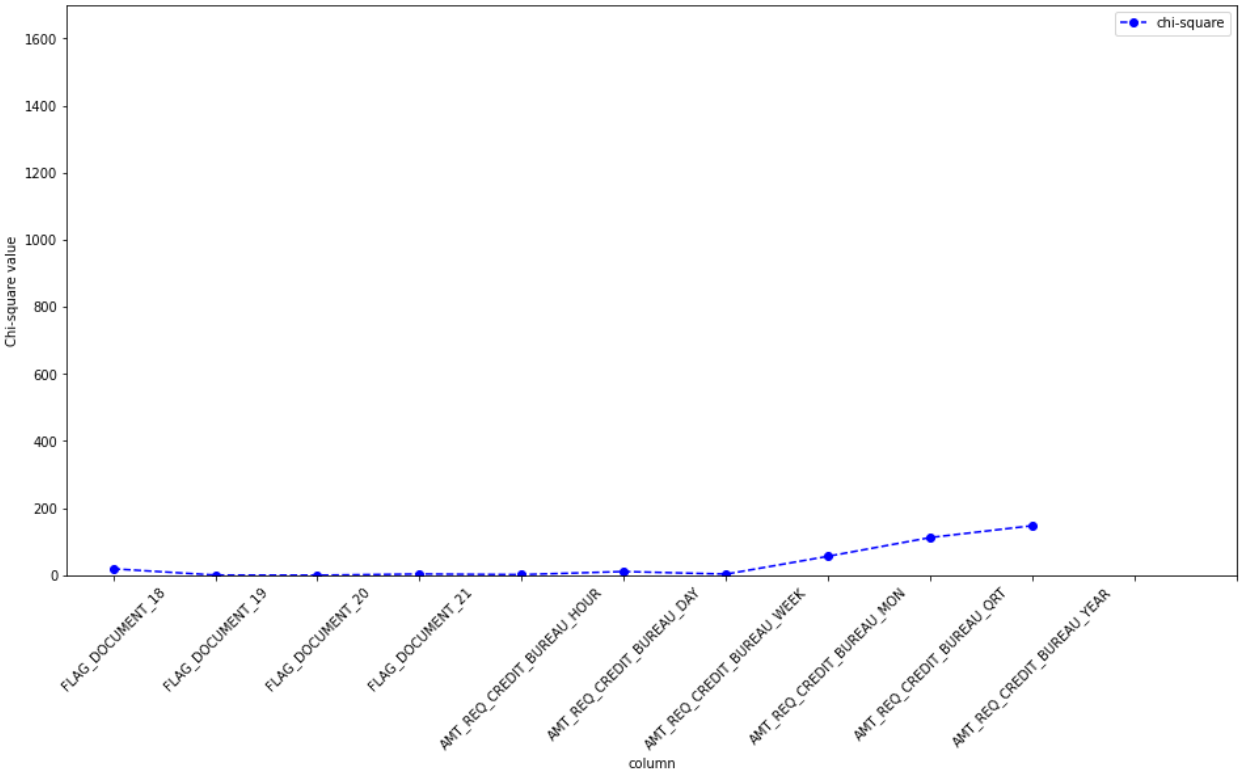
the chi-square values for all the columns to test the significance. The following plots show this.




chi-square
Chi-square value
column
FLAG_WORK_PHONE
FLAG_CONT_MOBILE
FLAG_PHONE
FLAG_EMAIL
OCCUPATION_TYPE
CNT_FAM_MEMBERS
REGION_RATING_CLIENT
REGION_RATING_CLIENT_W_CITY
WEEKDAY_APPR_PROCESS_START
HOUR_APPR_PROCESS_START
REG_REGION_NOT_LIVE_REGION
REG_REGION_NOT_WORK_REGION


chi-square
Chi-square value
column
LIVE_REGION_NOT_WORK_REGION
REG_CITY_NOT_LIVE_CITY
REG_CITY_NOT_WORK_CITY
LIVE_CITY_NOT_WORK_CITY
ORGANIZATION_TYPE
EXT_SOURCE_1
EXT_SOURCE_2
EXT_SOURCE_3
FONDKAPREMONT_MODE
HOUSETYPE_MODE
WALLSMATERIAL_MODE
EMERGENCYSTATE_MODE


chi-square
Chi-square value
0
200
400
600
800
1000
1200
1400
1600
FLAG_DOCUMENT_4
FLAG_DOCUMENT_5
FLAG_DOCUMENT_6
FLAG_DOCUMENT_7
FLAG_DOCUMENT_8
FLAG_DOCUMENT_9
FLAG_DOCUMENT_10
FLAG_DOCUMENT_11
FLAG_DOCUMENT_12
FLAG_DOCUMENT_13
FLAG_DOCUMENT_14
FLAG_DOCUMENT_15
column


chi-square
Chi-square value
0
200
400
600
800
1000
1200
1400
1600
FLAG_DOCUMENT_18
FLAG_DOCUMENT_19
FLAG_DOCUMENT_20
FLAG_DOCUMENT_21
AMT_REQ_CREDIT_BUREAU_HOUR
AMT_REQ_CREDIT_BUREAU_DAY
AMT_REQ_CREDIT_BUREAU_WEEK
AMT_REQ_CREDIT_BUREAU_MON
AMT_REQ_CREDIT_BUREAU_QRT
AMT_REQ_CREDIT_BUREAU_YEAR
column

- As we had observed during basic data story analysis, provision of documents by clients seems to be insignificant in deciding the target.
- Ratings by external agencies are very important predictors. Their chi-square values exceed the plot limits.
- Again, as we had seen in EDA, income type and occupation type of applicants are very important features
  .
- NUMERICAL FEATURES - these will be analysed using ANOVA analysis of Student's t-test for significance. The results will be utilised in developing a robust model for default prediction.

Conclusion

Basic EDA and significance testing of features confirm some of our assumptions. At the same time, they present some unexpected and interesting results. We will keep in mind all these observations while carrying out in-depth analysis and developing the final model for prediction.