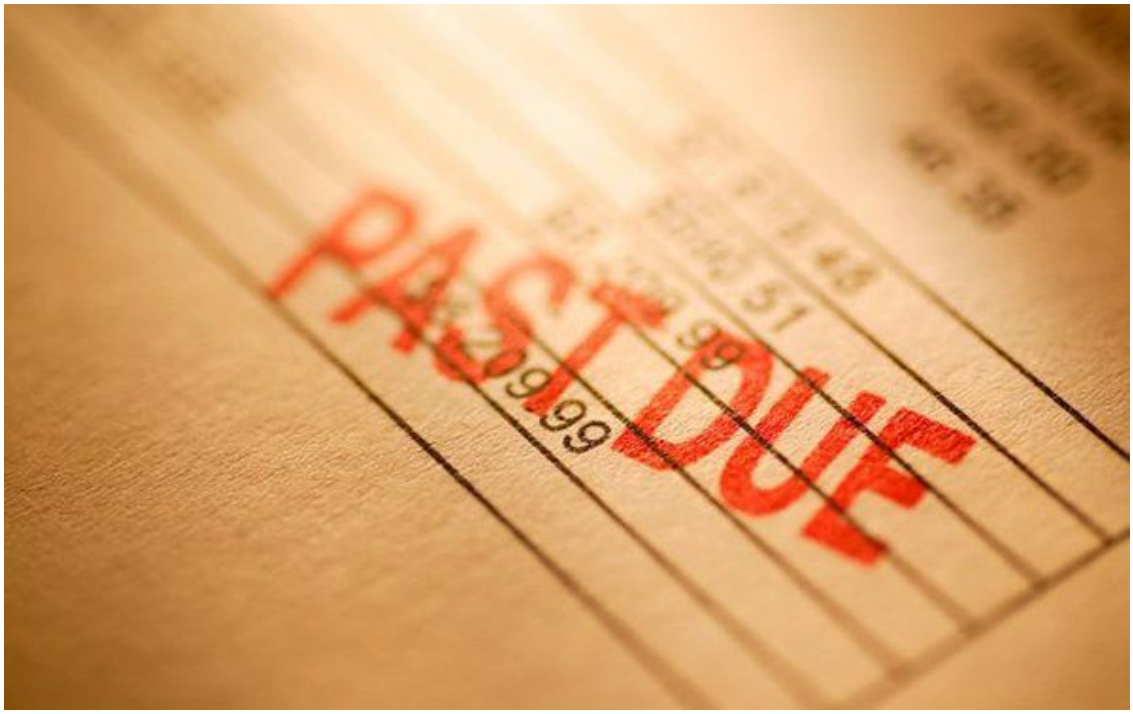
PREDICTING LOAN DEFAULT - HOME CREDIT

DETAILED ANALYSIS



1. 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.

2. 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.

3. **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.

4. **The stakeholders**

The top management of financial institutions are the first deciding and reviewing point. On approval, new processes trickle down to branch level. Ultimately, as our citizens are depositing their funds in banks, everyone who uses the services of a bank is indirectly a stakeholder.

5. **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.

6. **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.

7. 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.

- Categorical features are stored as type ‘object’. We will convert the categorical features into type ‘category’ for modelling

- 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.

- Let us take a look at the different types of columns in our dataset.

```
df.dtypes.value_counts()
```

```
float64    65
int64      41
object     16
dtype: int64
```

Thus we have 65 float columns, 41 integer columns and 16 columns of type 'object', which are actually of categorical type. The details of these columns can be found in the notebook.

8. 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. Numerical features will be tested with t-test, and categforical features with chi-square tests.
- 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

- Some numerical columns contain numbers in the incorrect format, or physically impossible values. For example, the "DAYS_EMPLOYED" column has some values that convert to 1000 years! Such values are appropriately handled and replaced with either the mean or NULL values, depending on the scenario. 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, box-plot of the feature 'AMT_INCOME_TOTAL' was plotted and data points beyond a limit of 2 times the IQR were removed.

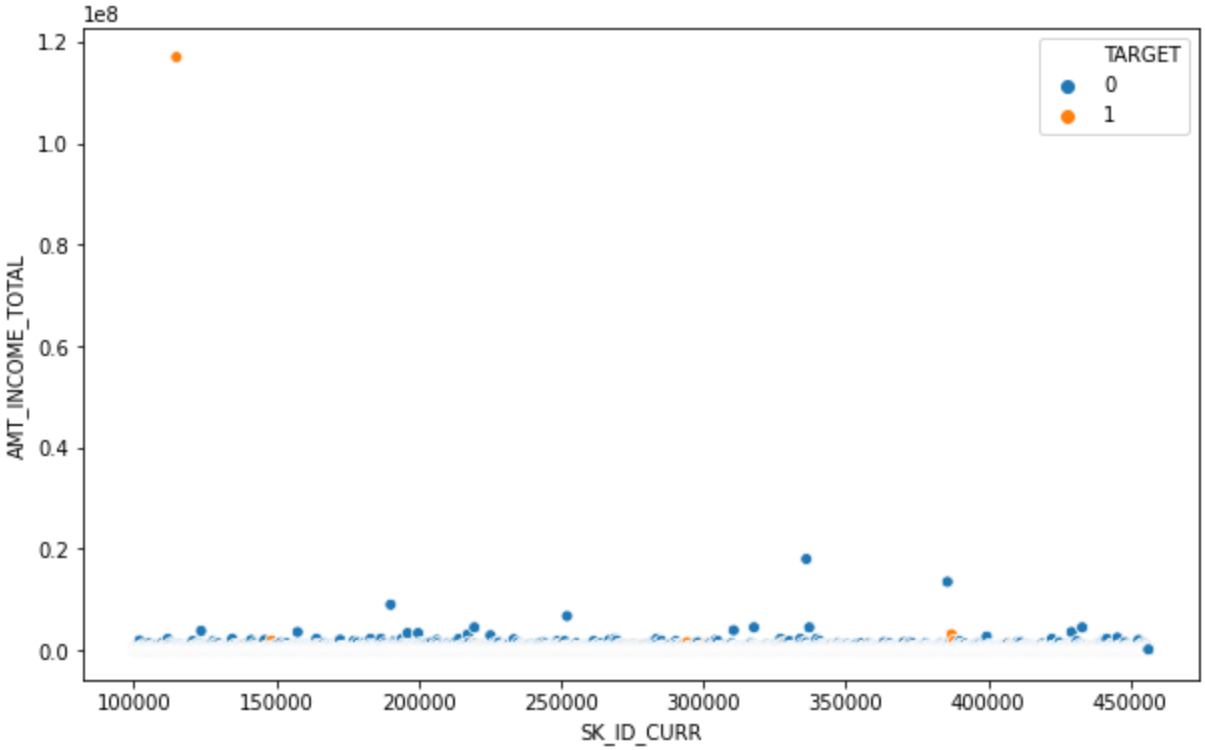
(iv) Evaluating feature importance and feature selection

- 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.
- Many ML models have inherent feature selection capability by the way they fit the data. For example, Logistic Regression gives us the coefficients of the features when it is fit on data. We will use such properties of models to discuss the relative importance of features.
- Statistical tools like t-test and chi-square tests are used to compare the relative importance of features. T-test will be used to compare the means of defaulters and non-defaulters, and chi-square will be used to compare categorical features.

9. Basic Data Story and Visualization

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:

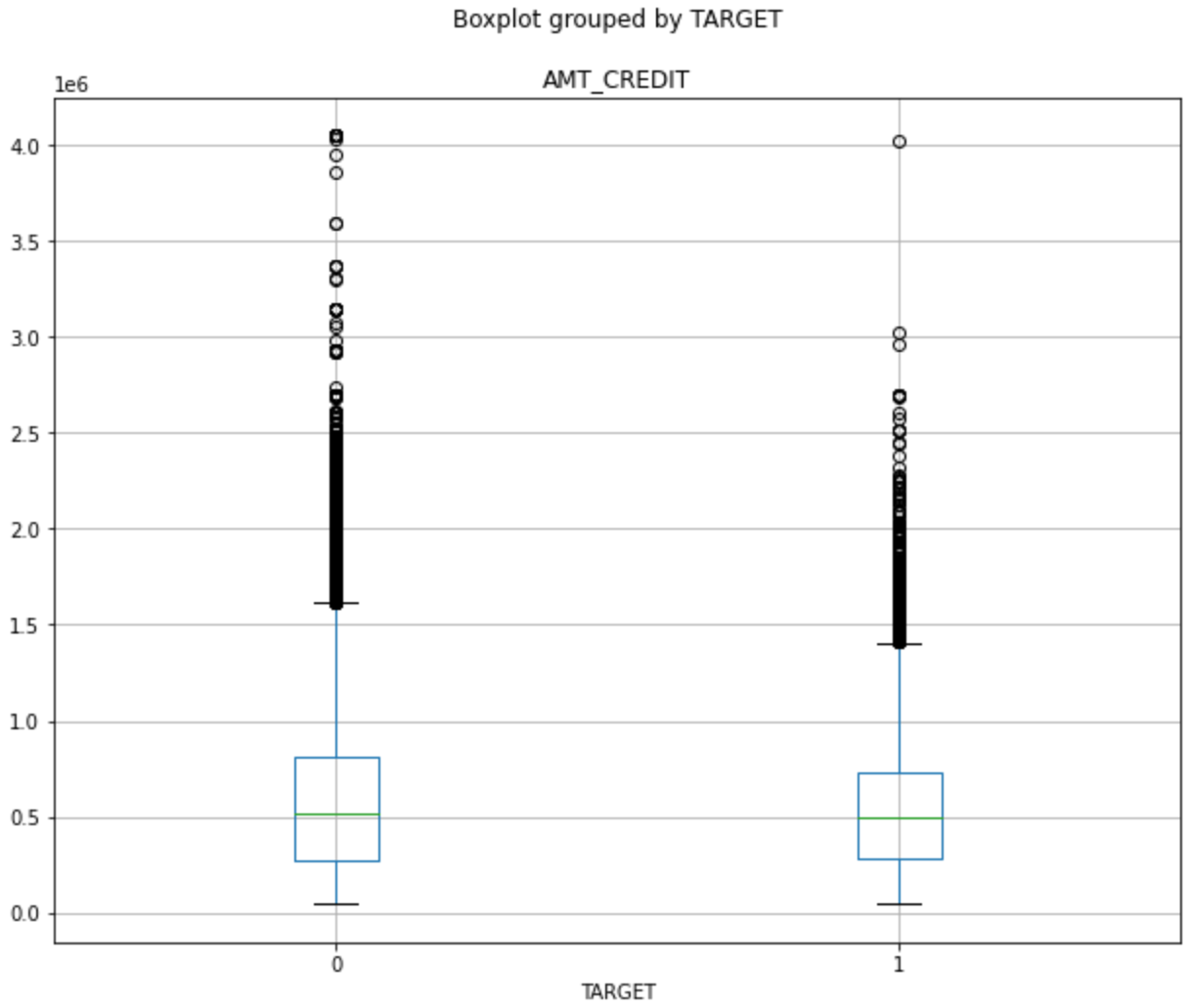
- The median income is 147150 units. The 1st and 3rd quartile of the income are 112500 and 202500 respectively. This seems to indicate that there isn't much spread in the data. But there are many outliers in the data. The highest income is in the range of 1.1 million, and the lowest is 25000 units. Hence, 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 !!!



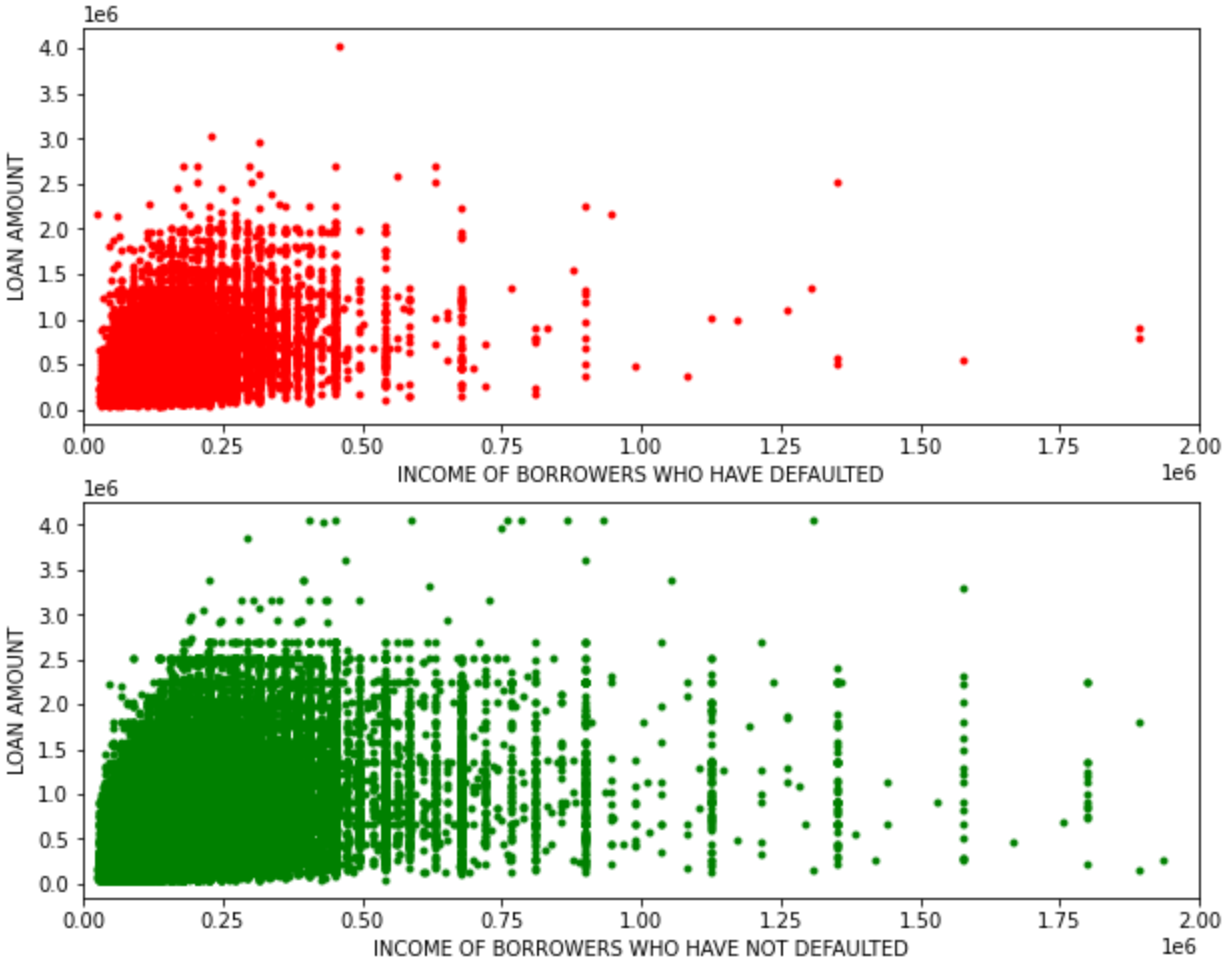
Here, “1” represents bad loans. As we can see, the highest value of income is for a loan defaulter.

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

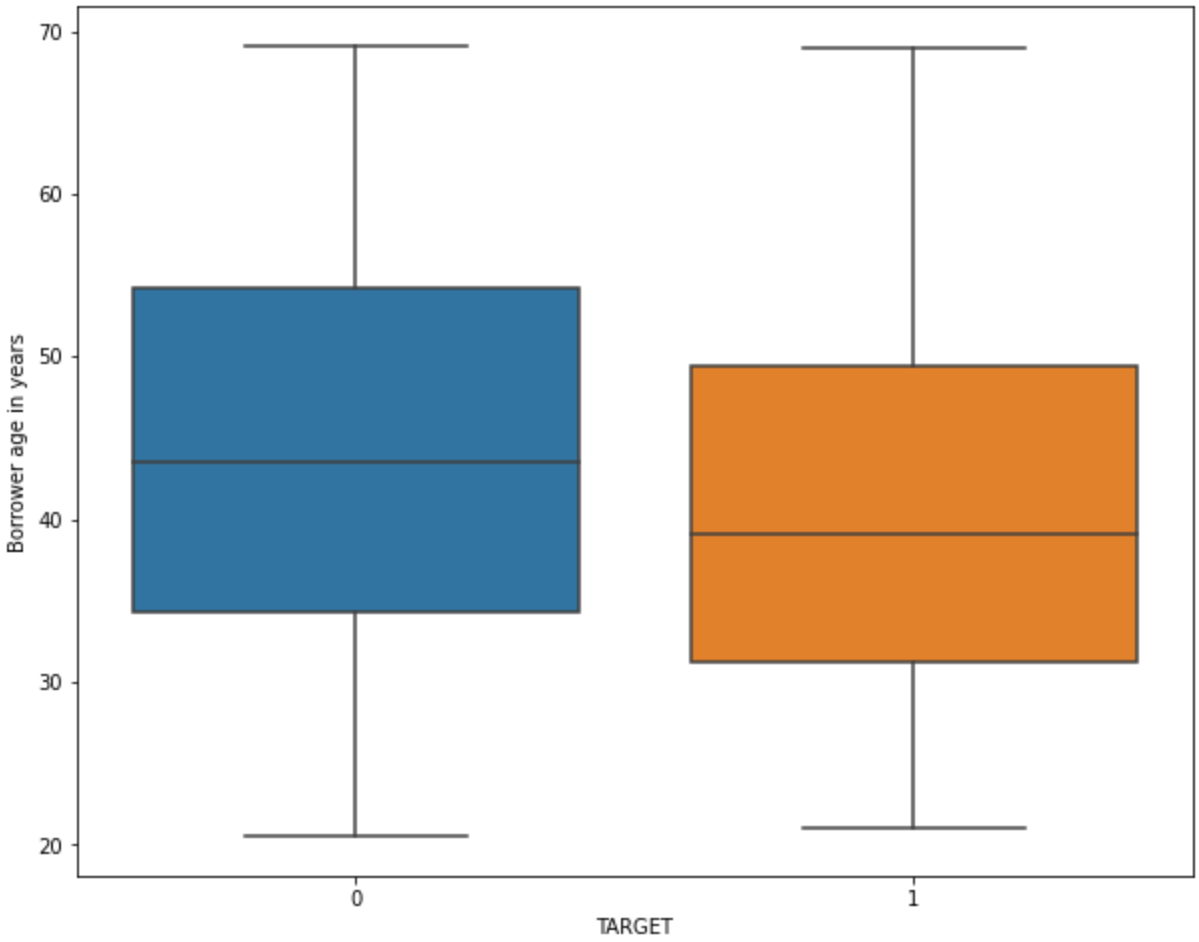
- People with higher amounts of loans are repaying better !! This is an interesting observation. It means that repayment habits are not overly dependent on the amount people have to pay. Some borrowers are just more credit-worthy.



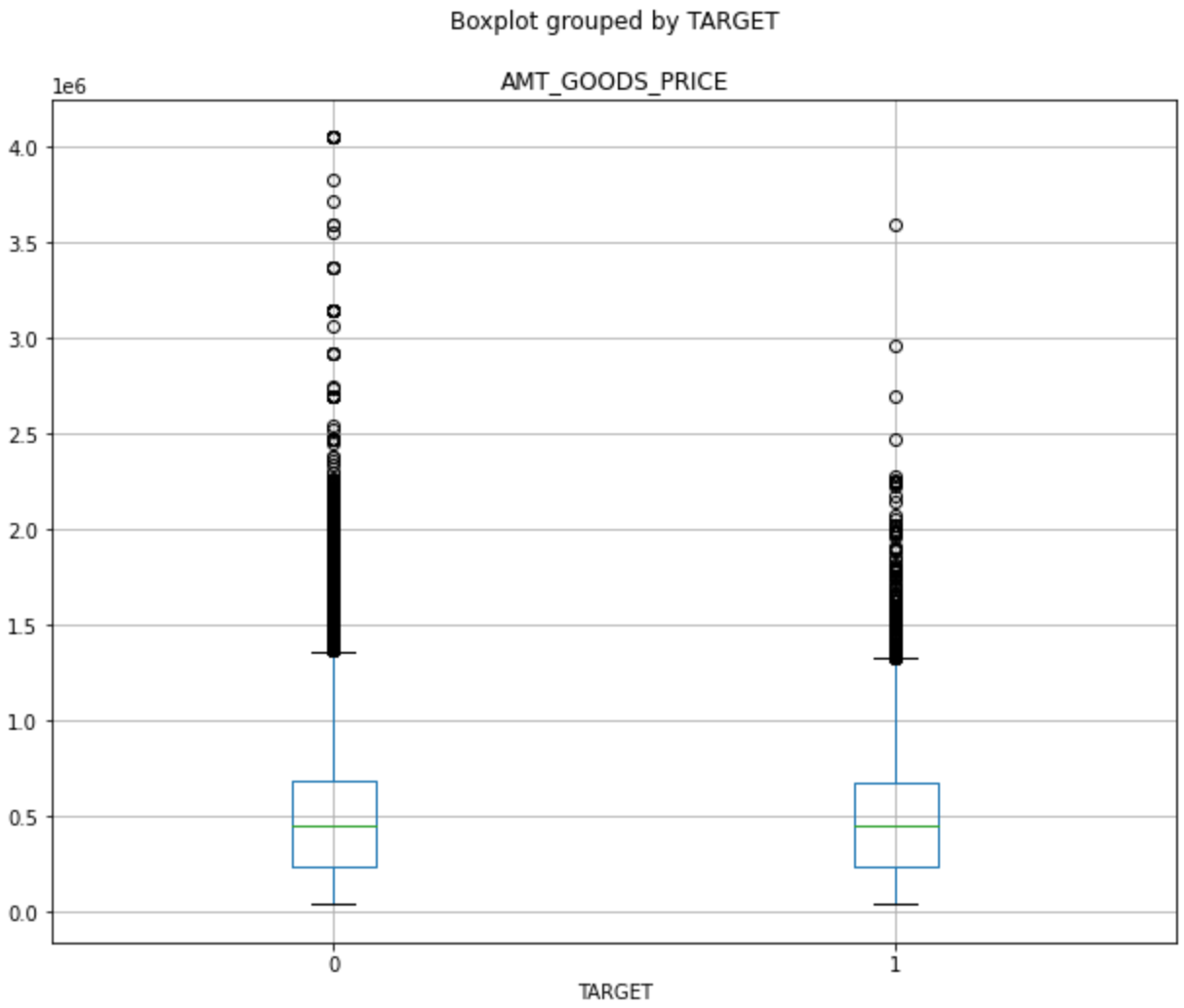
- The majority of the borrowers are in the low-income, low-credit category. Borrowers with high income are more widely present in the good-loans category. This is expected. But we would also expect the borrowers with high incomes to be provided high loans. That is not the case. The loans amounts are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



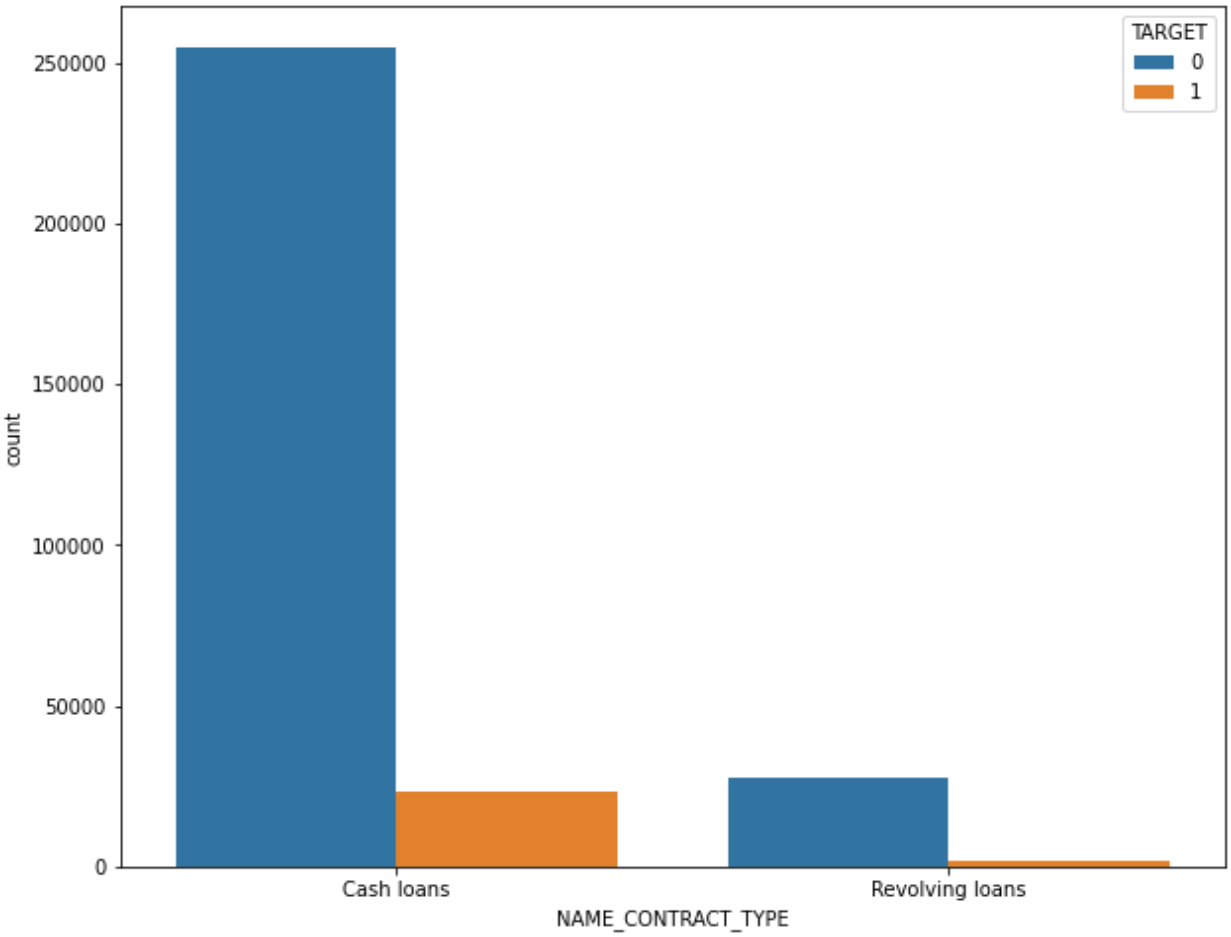
- Younger borrowers seem to default more. This may be due to low-paying jobs (but we have seen that income is not a deciding factor !) or a lack of financial discipline. It is seen that as people grow and have more responsibilities, they develop financial discipline.

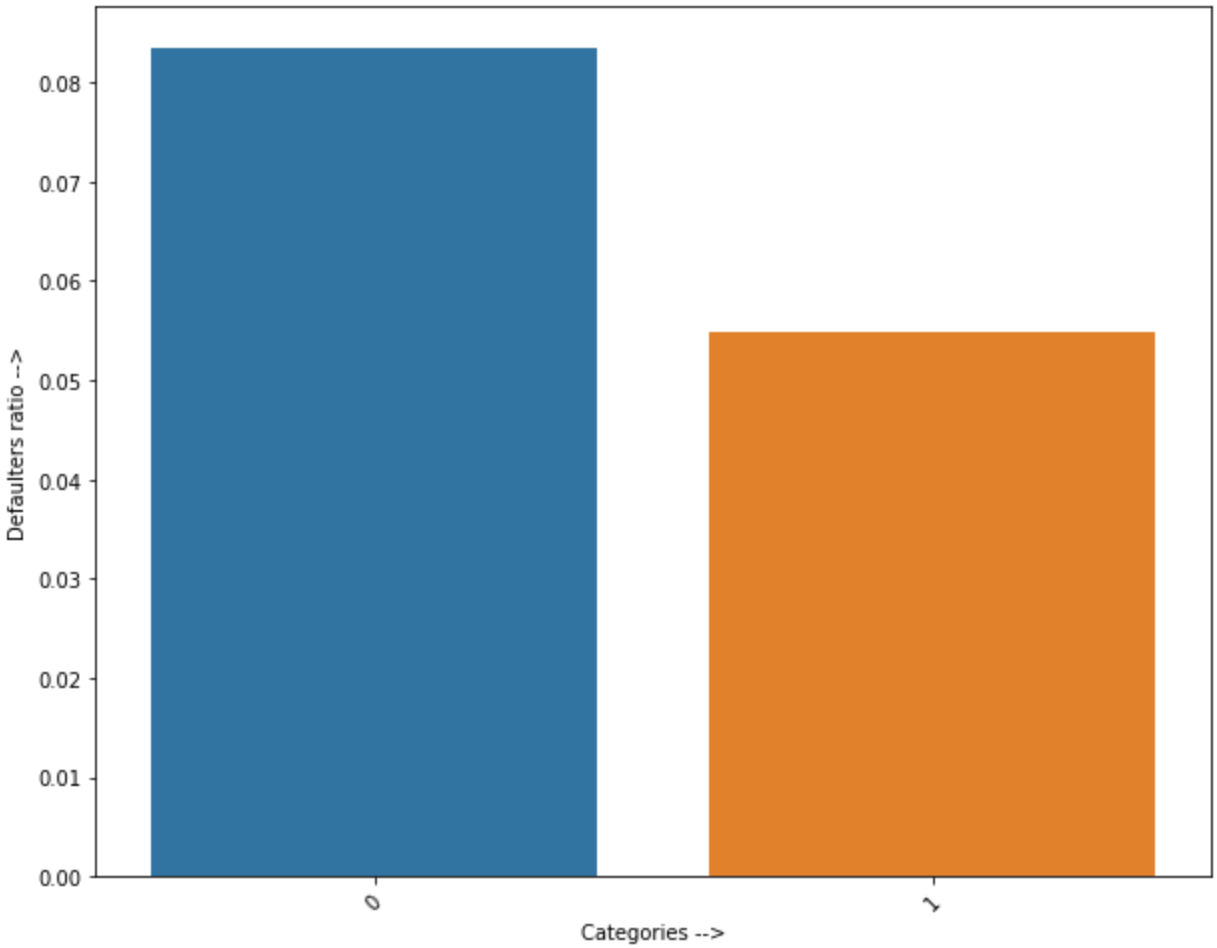


- The median value price of goods purchased with the loan money is fairly similar. This can be seen in the boxplots. But good loans have higher density of high priced items. This too is an interesting observation.



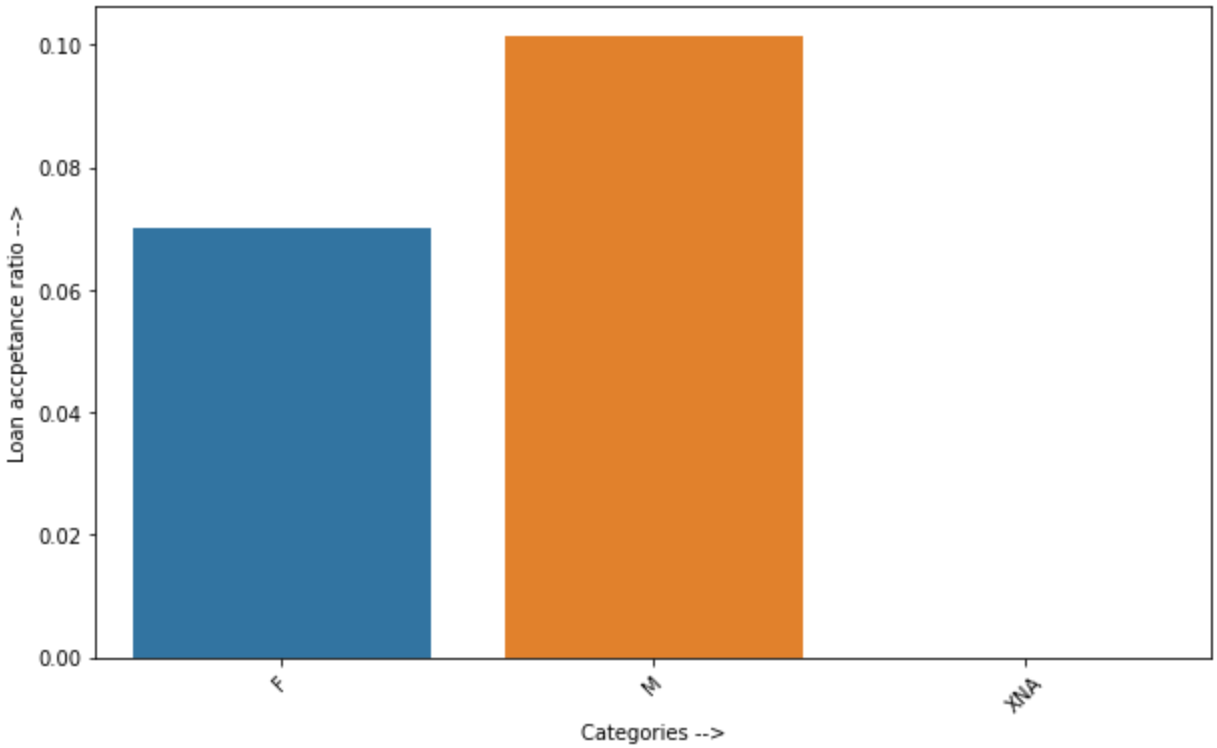
- Cash loans are generally provided for one-time purchase of goods, for example, purchasing a car, mobile phone, etc. They have to be repaid in regular intervals, generally as EMIs. When the full amount of interest + principal is repaid, the loan is closed.
- Revolving loans are those where the client 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



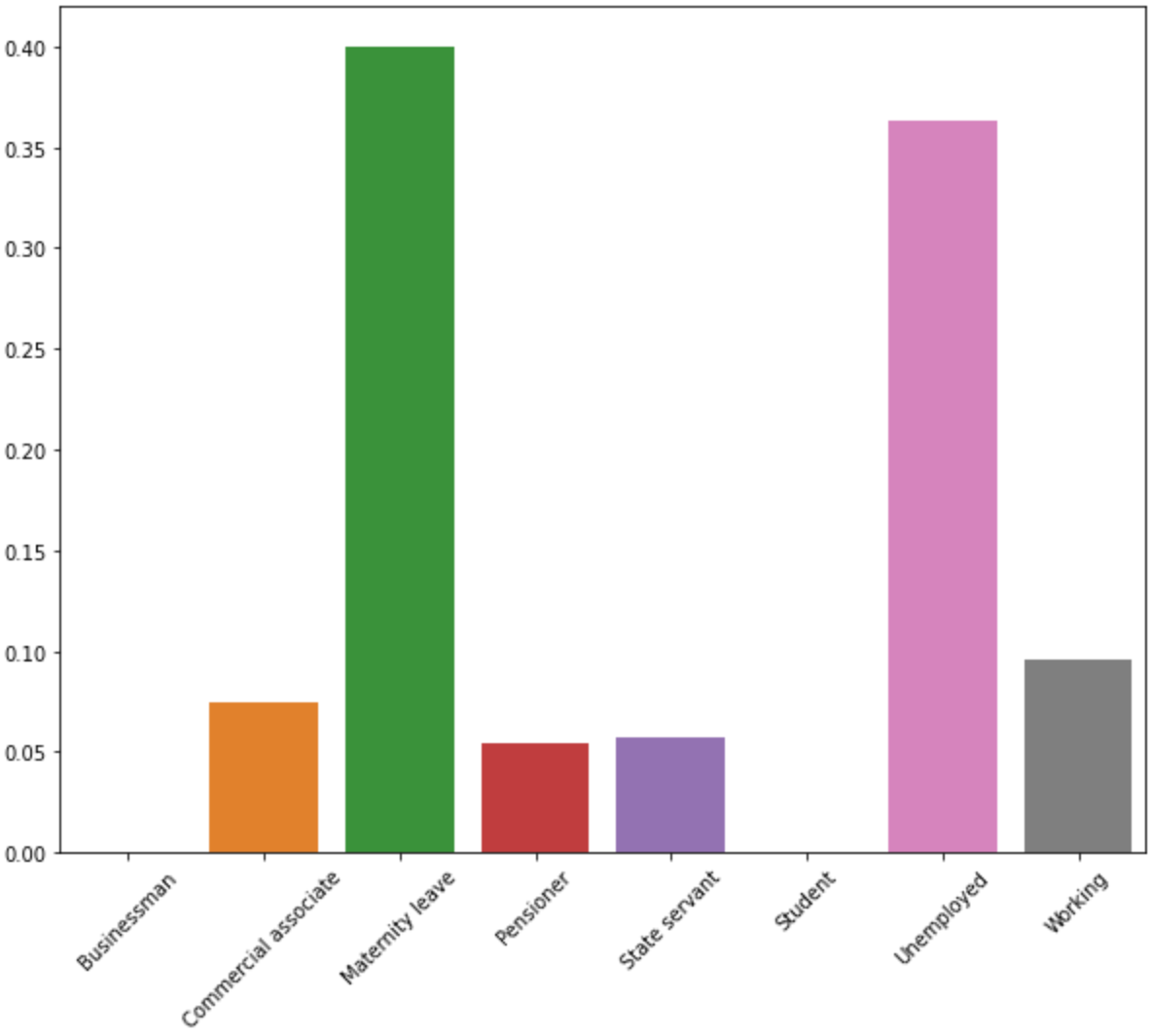
- In the above figure, blue color represents Cash Loans. Thus cash loans seem to have a higher default ratio.

- 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.

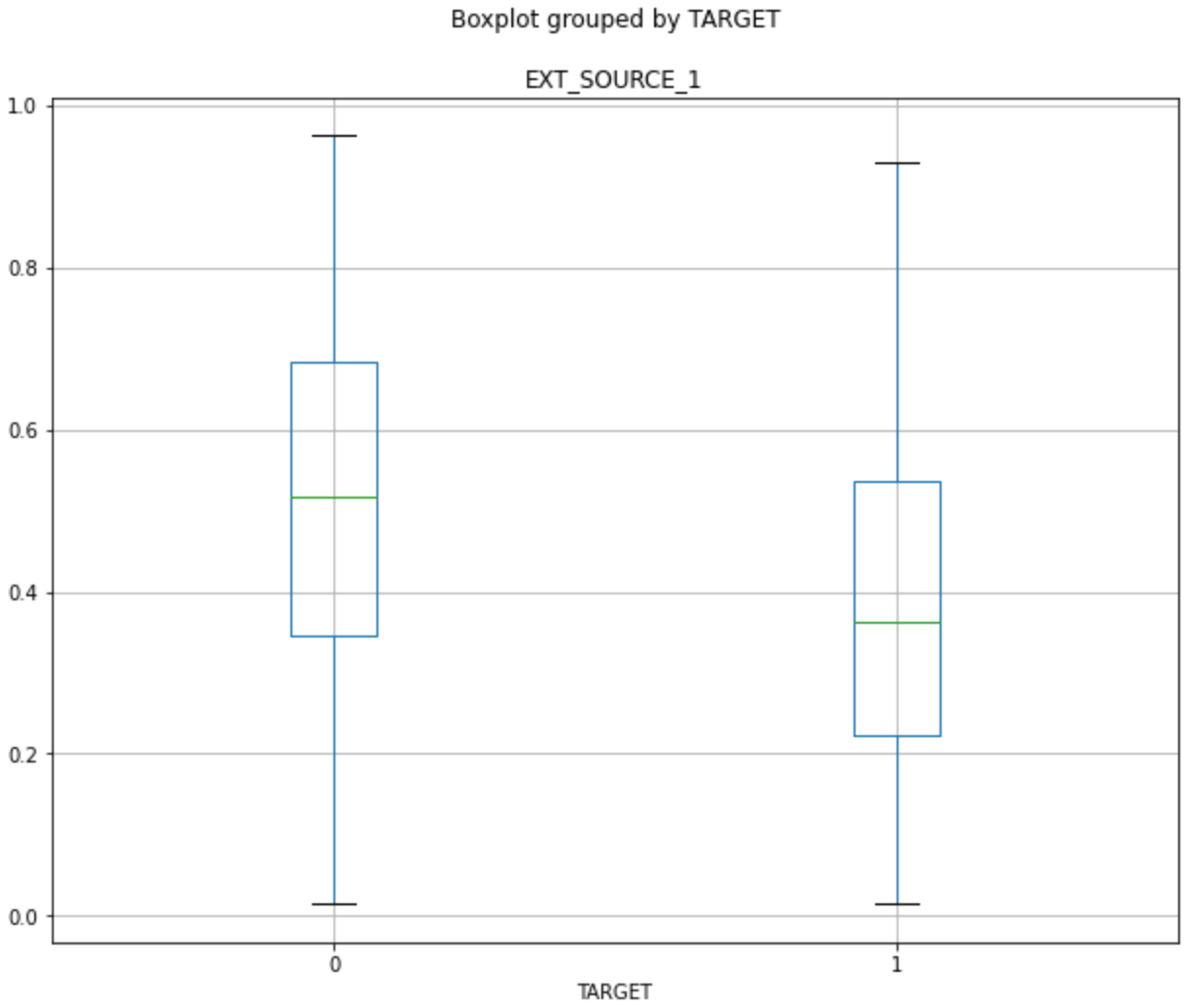


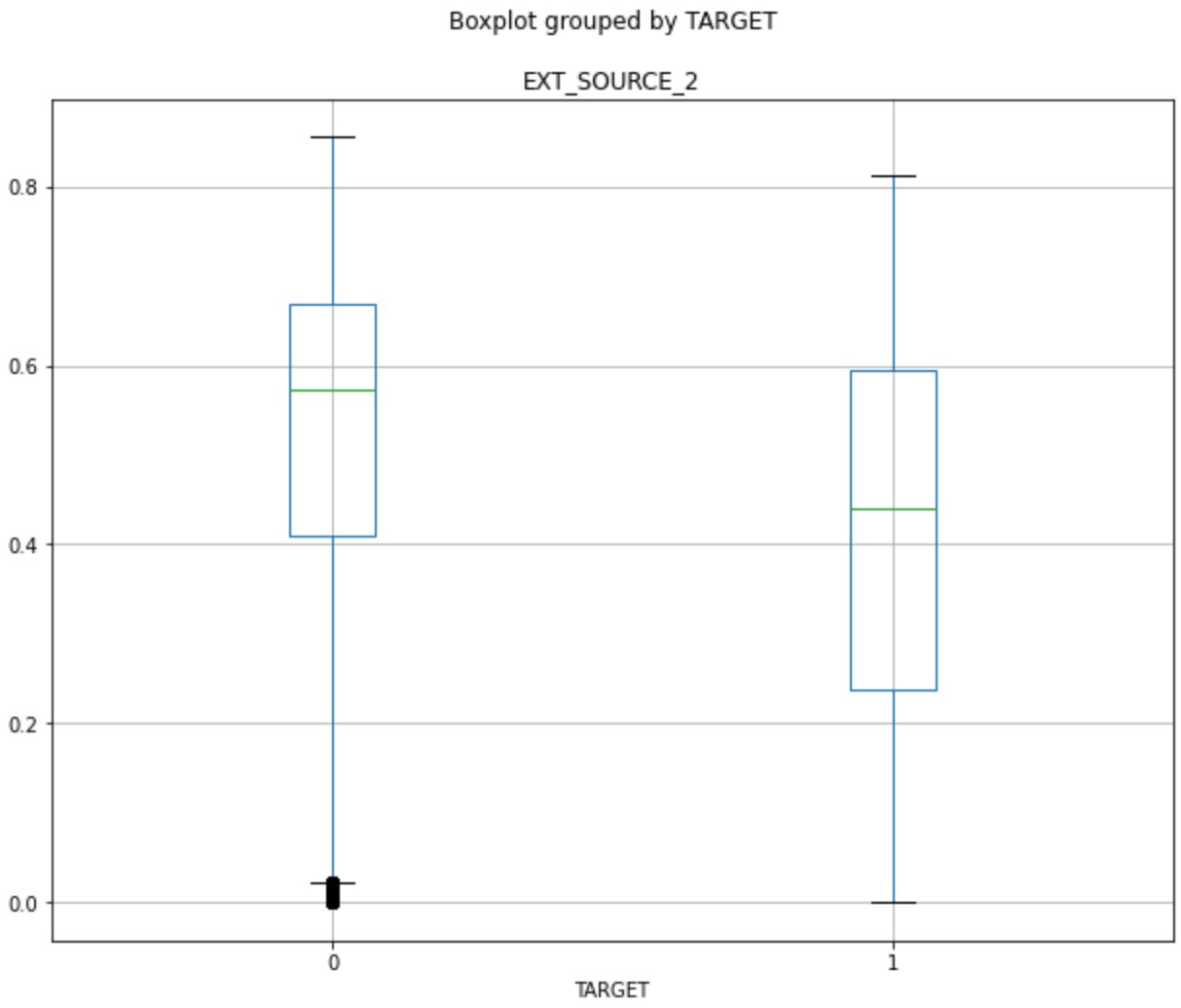
- Businessmen and students have extremely few accepted loan applications
  .
- Default ratios of pensioners, state servants and salaried class are fairly low.

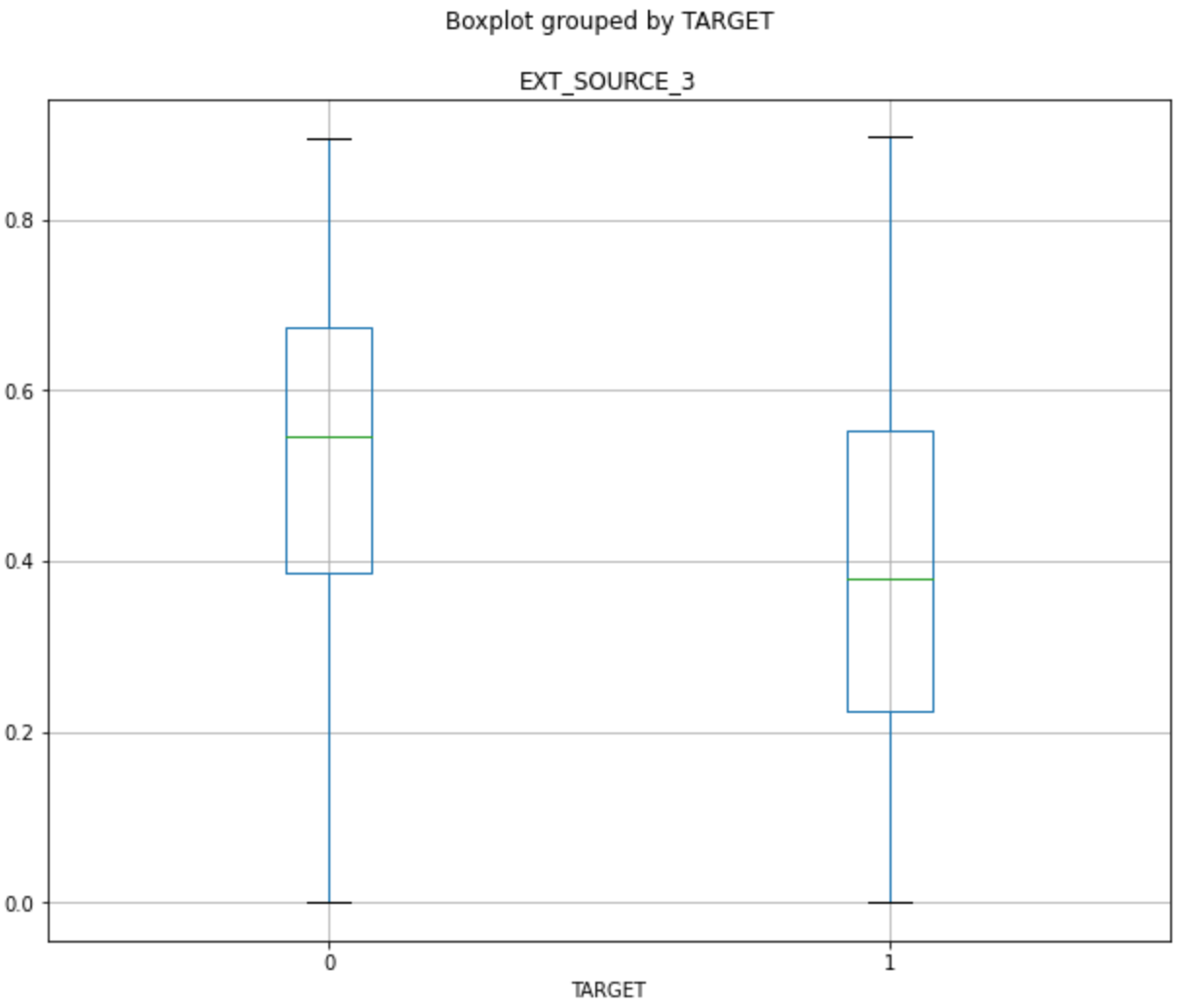
- Apart from the above observations, there are some other interesting observations, as outlined below.

- 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_1
1.0
0.8
0.6
0.4
0.2
0.0
0
1
TARGET


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.

10. Using statistical tests to evaluate feature importance

To analyse the distribution of numerical features between ba and good loans, we will perform the t-test for independent samples, with unequal variances. We will take each feature, separate the values for bad and good loans, and perform t-test for these values. We will then plot the values.

Let us take a look the highest and lowest t-statistics that we have got:

	t-statistic	p-value
DAYS_BIRTH	45.006188	0.000000e+00
DAYS_LAST_PHONE_CHANGE	33.126917	2.020162e-236
DAYS_ID_PUBLISH	28.408916	3.736621e-175
DAYS_REGISTRATION	24.702226	2.169540e-133
DEF_30_CNT_SOCIAL_CIRCLE	15.614027	9.938508e-55
DEF_60_CNT_SOCIAL_CIRCLE	14.855095	9.992832e-50
AMT_CREDIT	-19.273175	2.721911e-82
REGION_POPULATION_RELATIVE	-23.626701	2.446249e-122
AMT_GOODS_PRICE	-25.601850	4.280023e-143
DAYS_EMPLOYED	-28.962056	4.919699e-182
EXT_SOURCE_1	-58.554994	0.000000e+00
EXT_SOURCE_2	-80.465534	0.000000e+00
EXT_SOURCE_3	-84.578411	0.000000e+00

Let us take a look at the plots of t-statistics for all features.

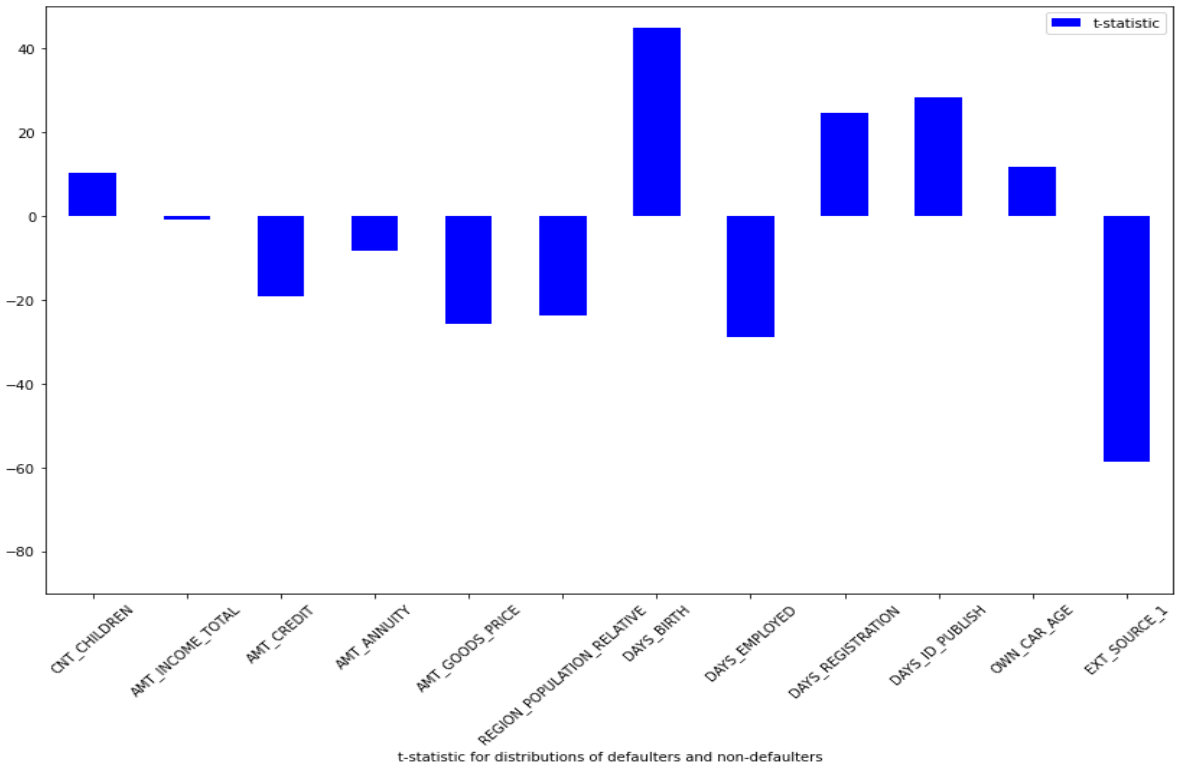


Figure: X-axis → Columns in dataset | Y-axis → T-statistic for independent t-test

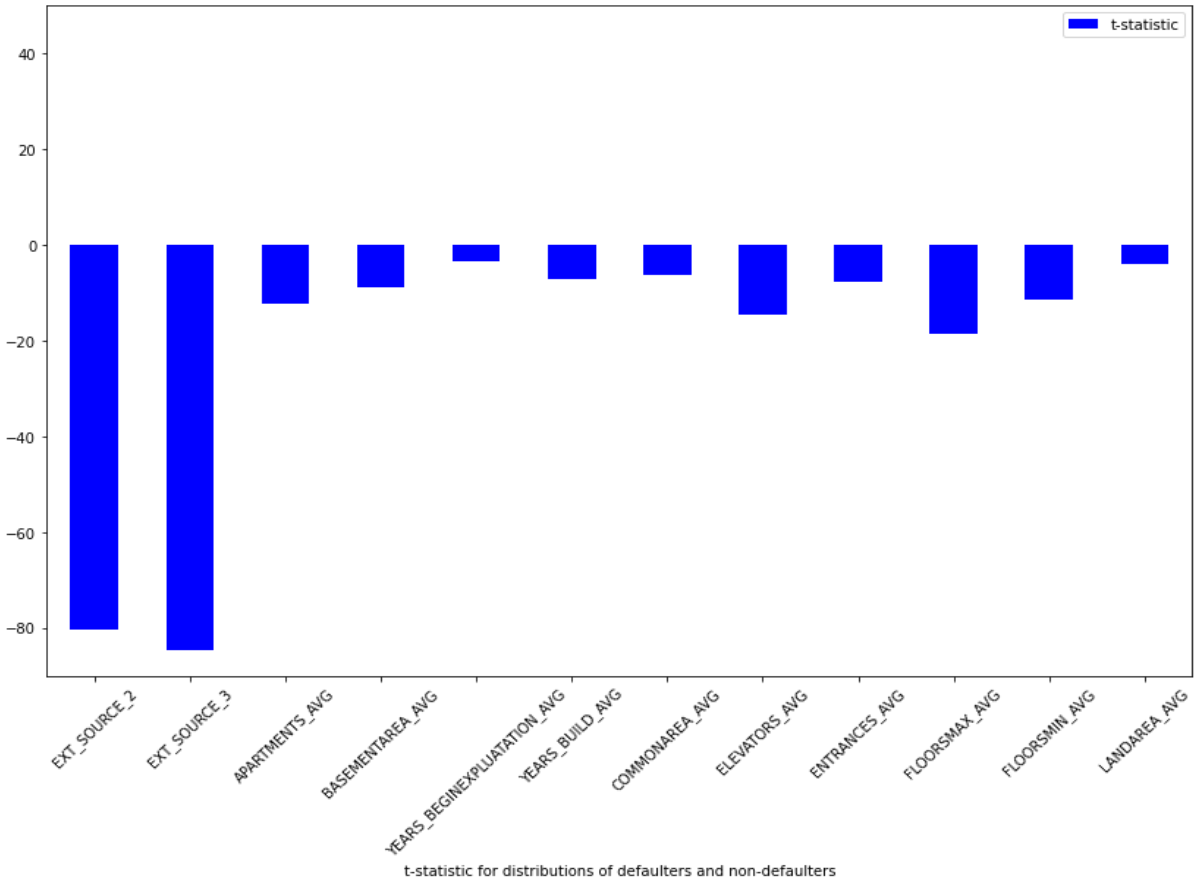
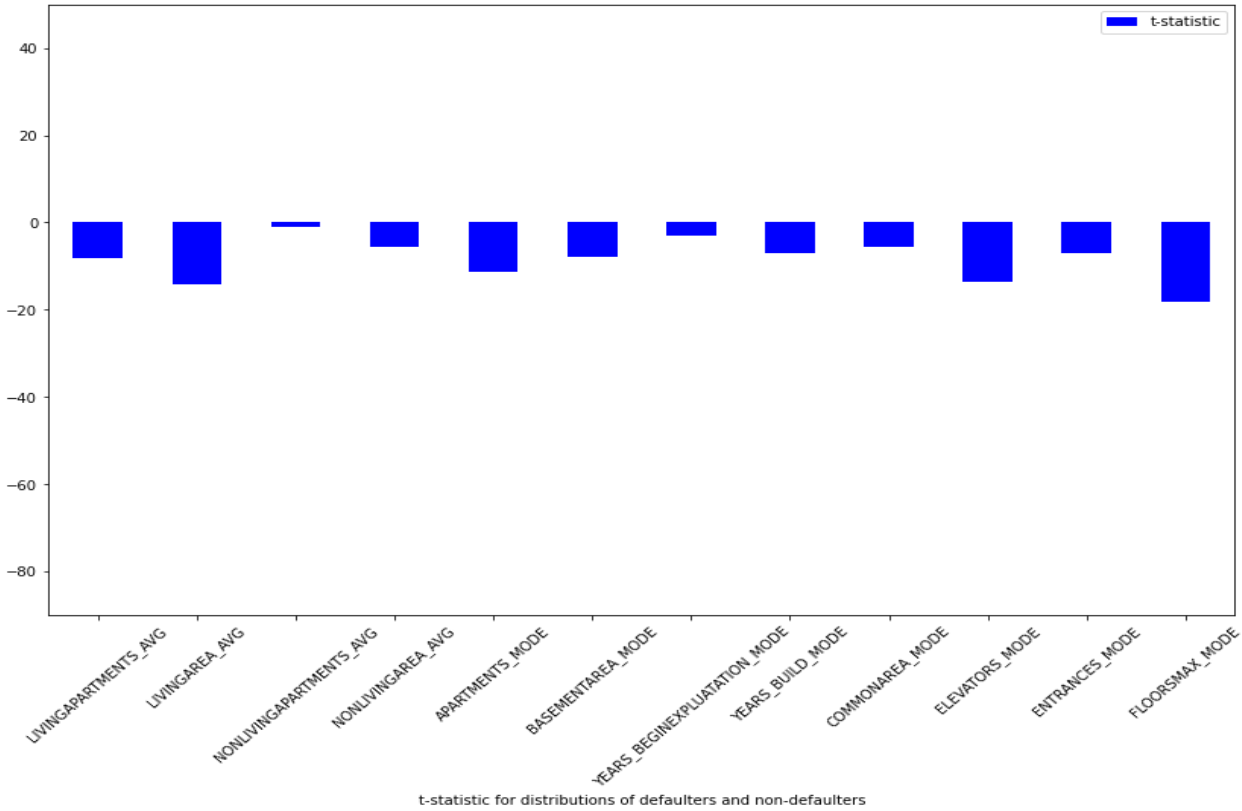


Figure: X-axis → Columns in dataset | Y-axis → T-statistic for independent t-test

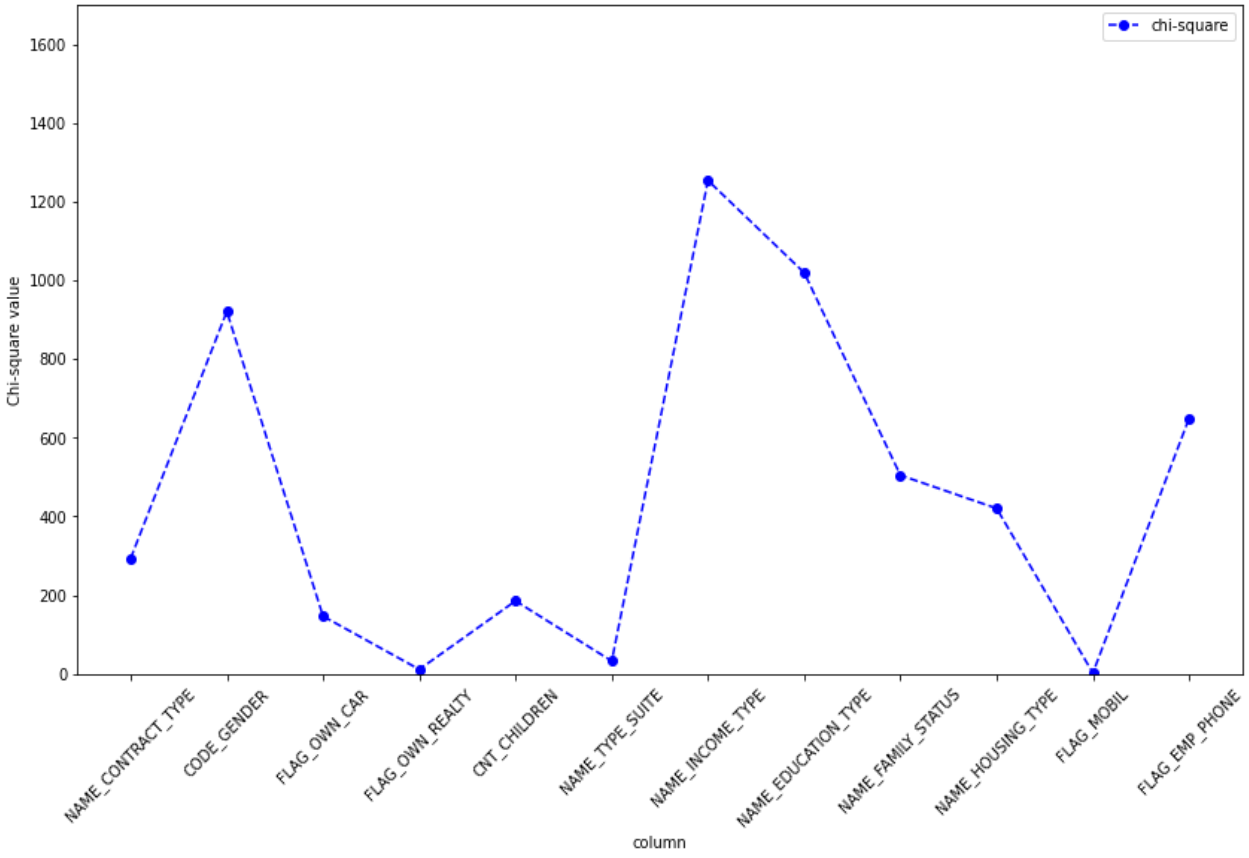
Figure: X-axis → Columns in dataset | Y-axis → T-statistic for independent t-test

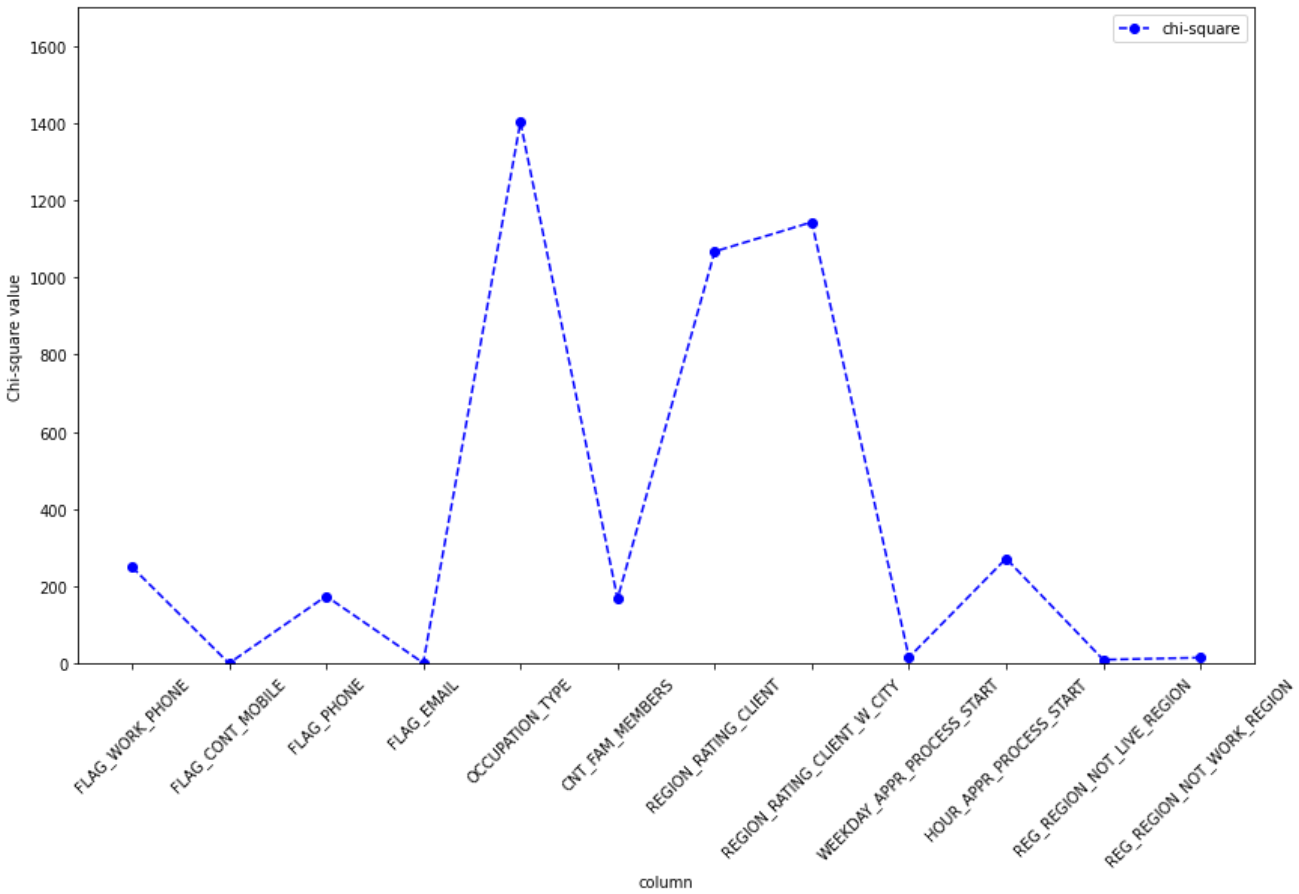


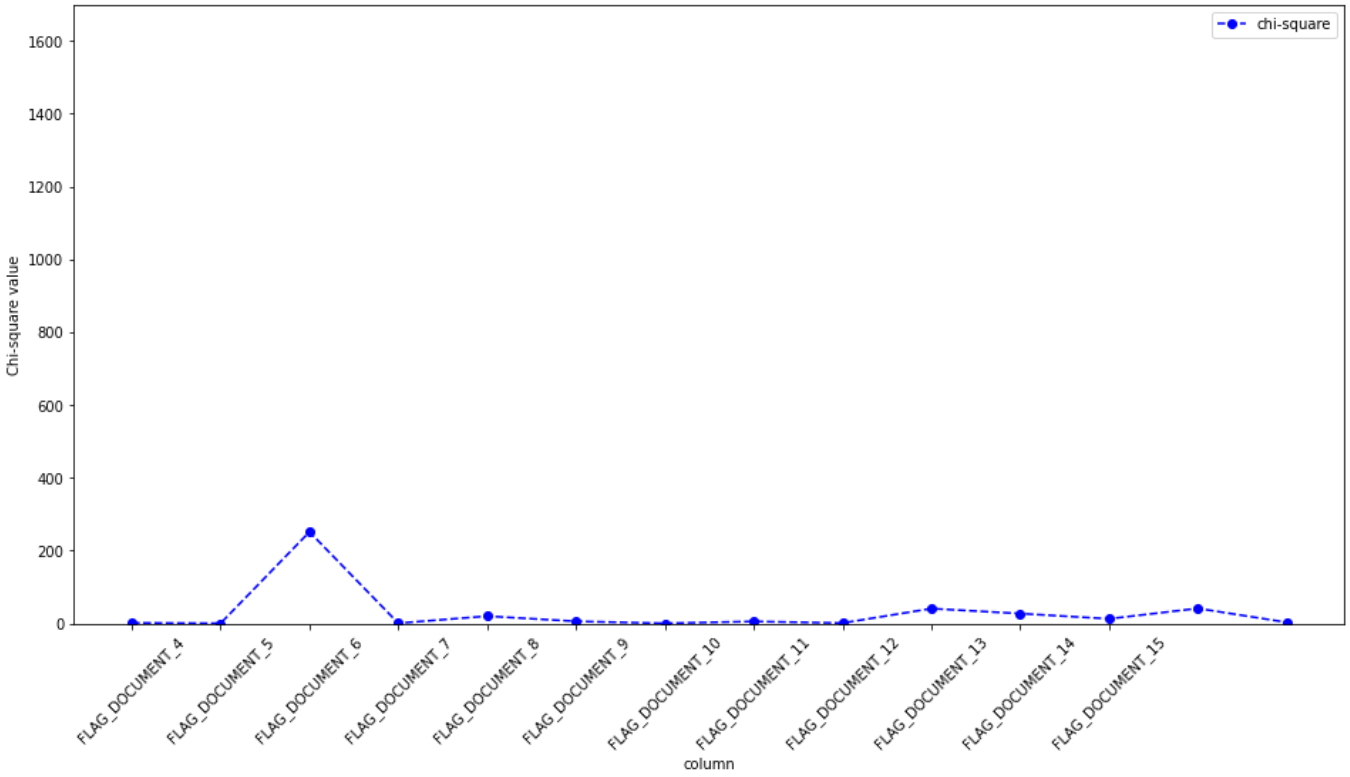
- The age of the applicant seems to be a major factor in deciding default.
- Also, whether the applicant has recently changed their ID or phone number, is also significant in deciding default
- The number of defaults in the applicant's social circle are also important
- As we had assumed earlier, the external ratings are also very significant in predicting the outcome of a loan.
- The period of employment of the borrower, and the amount of loan taken are very important predictors.
- Also, the relative population of the region where the client lives seems to be a major deciding factor.

For categorical features, we will perform the chi-square test to test significance.

Let us take a look at the plots.


chi-square
Chi-square value
column
NAME_CONTRACT_TYPE
CODE_GENDER
FLAG_OWN_CAR
FLAG_OWN_REALTY
CNT_CHILDREN
NAME_TYPE_SUITE
NAME_INCOME_TYPE
NAME_EDUCATION_TYPE
NAME_FAMILY_STATUS
NAME_HOUSING_TYPE
FLAG_MOBIL
FLAG_EMP_PHONE


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



We are seeing some interesting trends here:

- The gender of the applicants is a major factor in deciding default. As depicted in the plots in previous sections on Exploratory Data Analysis, we remember that Male borrowers tend to default more.
- The family status is important in predicting default
- The INCOME_TYPE, EDUCATION_TYPE and OCCUPATION_TYPE of the borrower are significant
- Whether the borrower has provided a document or not, does not seem to be playing a major role in deciding default.

11. Evaluating the relative importance of features for various models

a. Relative importance of Features using Logistic Regression

As our dataset has a large number of columns, we will process step by step. Let us first fit a basic logistic regression model on only the numerical features, and see the feature importances. We first run GridSearchCV on the various possible logistic regression parameters, and then select the best parameters.

```
lr_model = LogisticRegression(solver='lbfgs', C = 0.001, penalty = 'l2', max_iter=3000)
lr_model.fit(X_train, y_train)
```

Below are the features with the most negative values:

	LogReg coefficient
DAYS_LAST_PHONE_CHANGE	-2.326008e-04
DAYS_ID_PUBLISH	-9.744821e-05
DAYS_EMPLOYED	-9.597625e-05
DAYS_BIRTH	-8.256772e-05
DAYS_REGISTRATION	-2.593505e-05
AMT_GOODS_PRICE	-3.741725e-06
OWN_CAR_AGE	-1.444482e-06
AMT_INCOME_TOTAL	-9.396423e-07
EXT_SOURCE_2	-3.345025e-07
EXT_SOURCE_3	-3.094166e-07
YEARS_BEGINEXPLUATATION_MEDI	-1.835258e-07
YEARS_BEGINEXPLUATATION_AVG	-1.833951e-07
YEARS_BEGINEXPLUATATION_MODE	-1.833651e-07
YEARS_BUILD_MODE	-1.482221e-07

From the above, we can make some interesting observations:

- People in the dataset who have changed their phone recently, seem to have defaulted more on their loans. Thinking about it, we can imagine that if someone has a stable job, steady source of income, they would generally not want to change their phone number much. Hence, it makes sense. (Here, more negative value means the feature is working against the output being “1” i.e. more negative value means less default probability).
- Older people tend to default less.
- People earning more are defaulting less.
- A higher value in the “EXT_SOURCE_X” columns, that represent the ratings provided by external credit rating agencies (like CRISIL and ICRA in India) works against default. This is expected, as credit-worthy borrowers get higher ratings from external agencies.

Let us look at the most important features in default, as per the above model:

COMMONAREA_MEDI	-1.169324e-08
COMMONAREA_AVG	-1.161136e-08
COMMONAREA_MODE	-1.079076e-08
NONLIVINGAREA_AVG	-8.263508e-09
NONLIVINGAREA_MEDI	-8.240192e-09
NONLIVINGAREA_MODE	-7.915762e-09
REGION_POPULATION_RELATIVE	-7.035393e-09
NONLIVINGAPARTMENTS_AVG	-2.050797e-09
NONLIVINGAPARTMENTS_MEDI	-1.961924e-09
NONLIVINGAPARTMENTS_MODE	-1.691081e-09
AMT_ANNUITY	3.740627e-08
DEF_60_CNT_SOCIAL_CIRCLE	9.451098e-08
DEF_30_CNT_SOCIAL_CIRCLE	1.128363e-07
AMT_CREDIT	3.021641e-06

- Unsurprisingly, “AMT_CREDIT” (the amount of loan taken by a borrower) works in favour of prediction of “1” or YES to default. High amounts of loans put a financial burden on the borrower. Hence, this seems to be the most important feature as per the coefficients of the logistic regression model.

Although we can see some differences in the relative values of the coefficients of features, the actual values of the coefficients are very small. Let us try to visualise this:

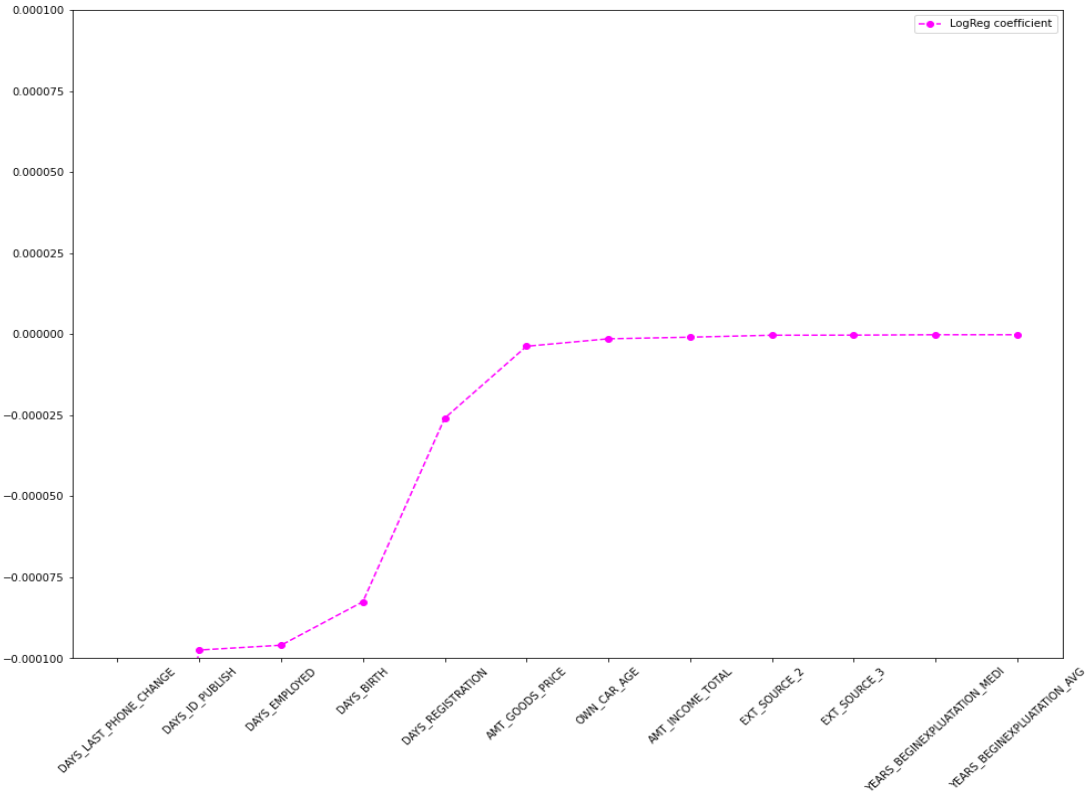


Figure: X-axis → Columns in dataset | Y-axis → Coefficient in Logistic Regression

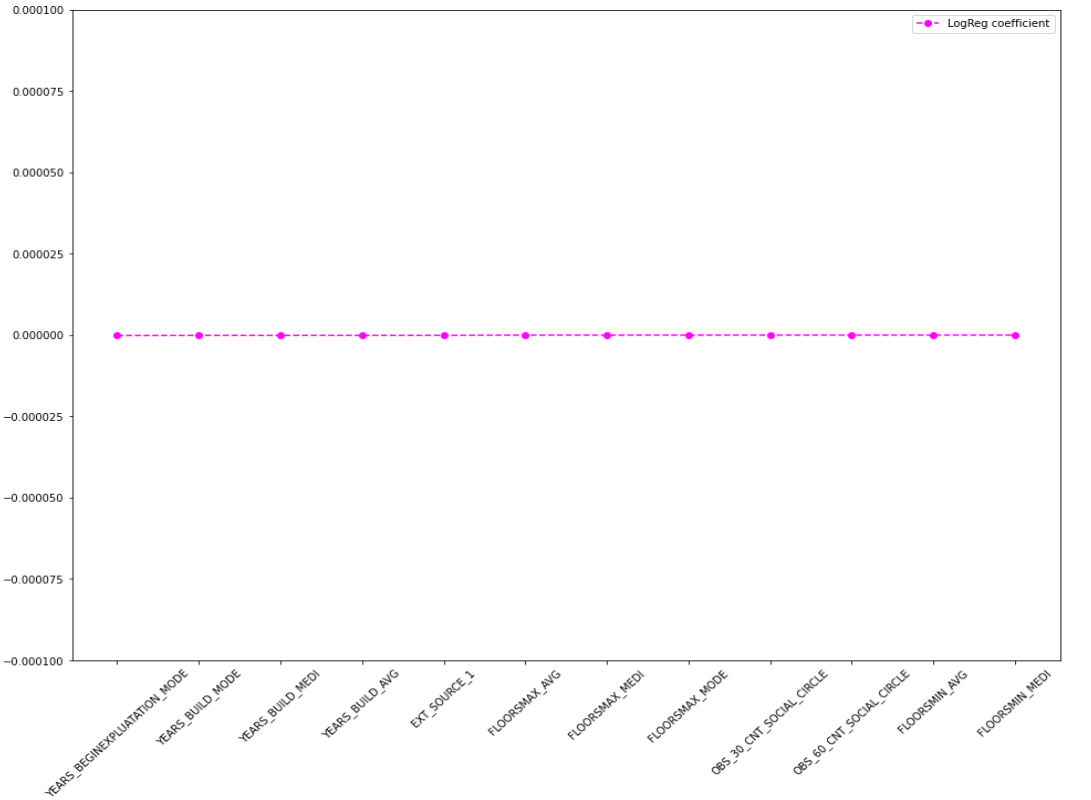


Figure: X-axis → Columns in dataset | Y-axis → Coefficient in Logistic Regression

As visible above, the coefficients of most of the features are extremely close to 0. This trend follows for other features. (Please refer to the notebook).

Next, we will fit a logistic regression model with the full dataset, and see the relative importance.

Here are the top 10 most important features working in favour of loan default.

	Coefficient
CODE_GENDER_M	0.250753
FLAG_DOCUMENT_3_1	0.170059
NAME_EDUCATION_TYPE_Secondary / secondary special	0.137770
AMT_CREDIT	0.131846
ORGANIZATION_TYPE_Self-employed	0.111949
AMT_ANNUITY	0.107856
OCCUPATION_TYPE_Drivers	0.097735
NAME_INCOME_TYPE_Working	0.095087
REGION_RATING_CLIENT_W_CITY_3	0.094016
FLAG_WORK_PHONE_1	0.090524

- The GENDER of the borrower appears to be the most important deciding factor. Male applicants in the dataset are defaulting more.
- People who have provided DOCUMENT_3 are less likely to default. As Home Credit has not disclosed the nature of this document, we cannot comment more.
- Higher loan amount lead to more defaults
- Borrowers with education only upto secondary/secondary special are defaulting more.
- Higher EMIs lead to more default. This is expected.
- Working class people tend to default more

Let us look at features which are working against borrower default:

DAYS_ID_PUBLISH	-0.062162
OCCUPATION_TYPE_Core staff	-0.076647
NAME_FAMILY_STATUS_Married	-0.101029
AMT_REQ_CREDIT_BUREAU_QRT_1.0	-0.138153
DAYS_EMPLOYED	-0.142767
NAME_CONTRACT_TYPE_Revolving loans	-0.158643
EXT_SOURCE_1	-0.162649
NAME_EDUCATION_TYPE_Higher education	-0.163749
FLAG_OWN_CAR_Y	-0.187506
AMT_GOODS_PRICE	-0.202572
EXT_SOURCE_2	-0.388777
EXT_SOURCE_3	-0.447027

- As we had seen earlier as well, higher ratings from external agencies mean the borrower is credit worthy. Hence, this factor is working against borrower default.
- Interestingly, people who are purchasing costlier products with the loan amount are defaulting less! This is not what we would normally expect.
- Borrowers with higher education tend to default less
- People who are employed since long back tend to default less.
- People who have not changed their ID recently are defaulting less. In general, if someone is settled and financially stable with a well paying job, they would not change their ID regularly. Thus, we can understand the real world significance.

As in the case with fitting only for numerical features, only some of the features have high coefficient values, rest all are very close to 0.

Let us plot the feature coefficient on the y-axis and see the results.

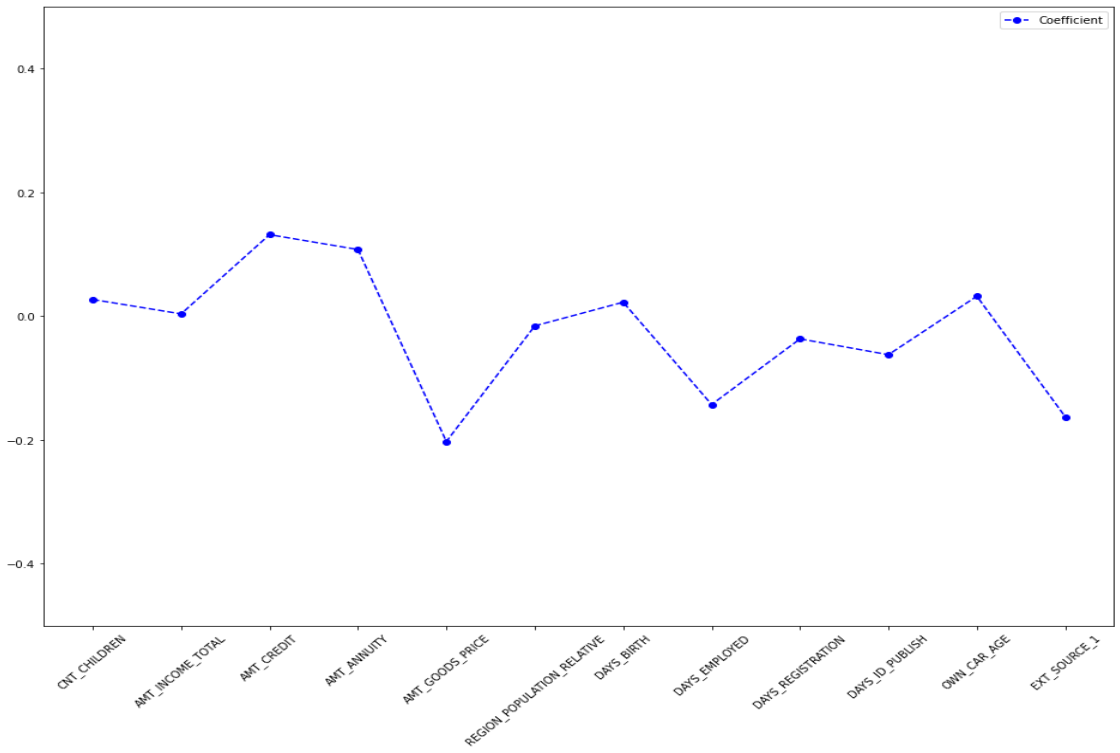


Figure: X-axis → Columns in dataset | Y-axis → Coefficient in Logistic Regression

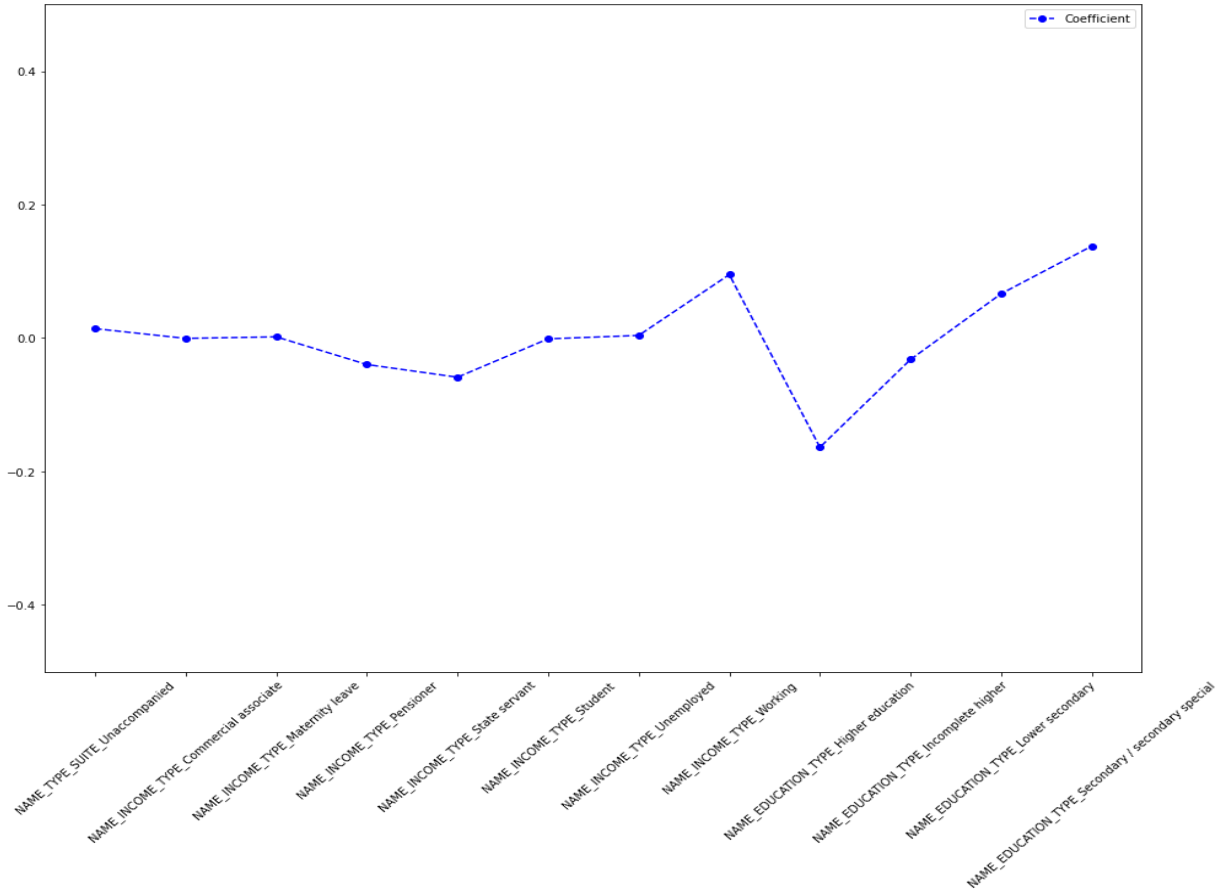


Figure: X-axis → Columns in dataset | Y-axis → Coefficient in Logistic Regression

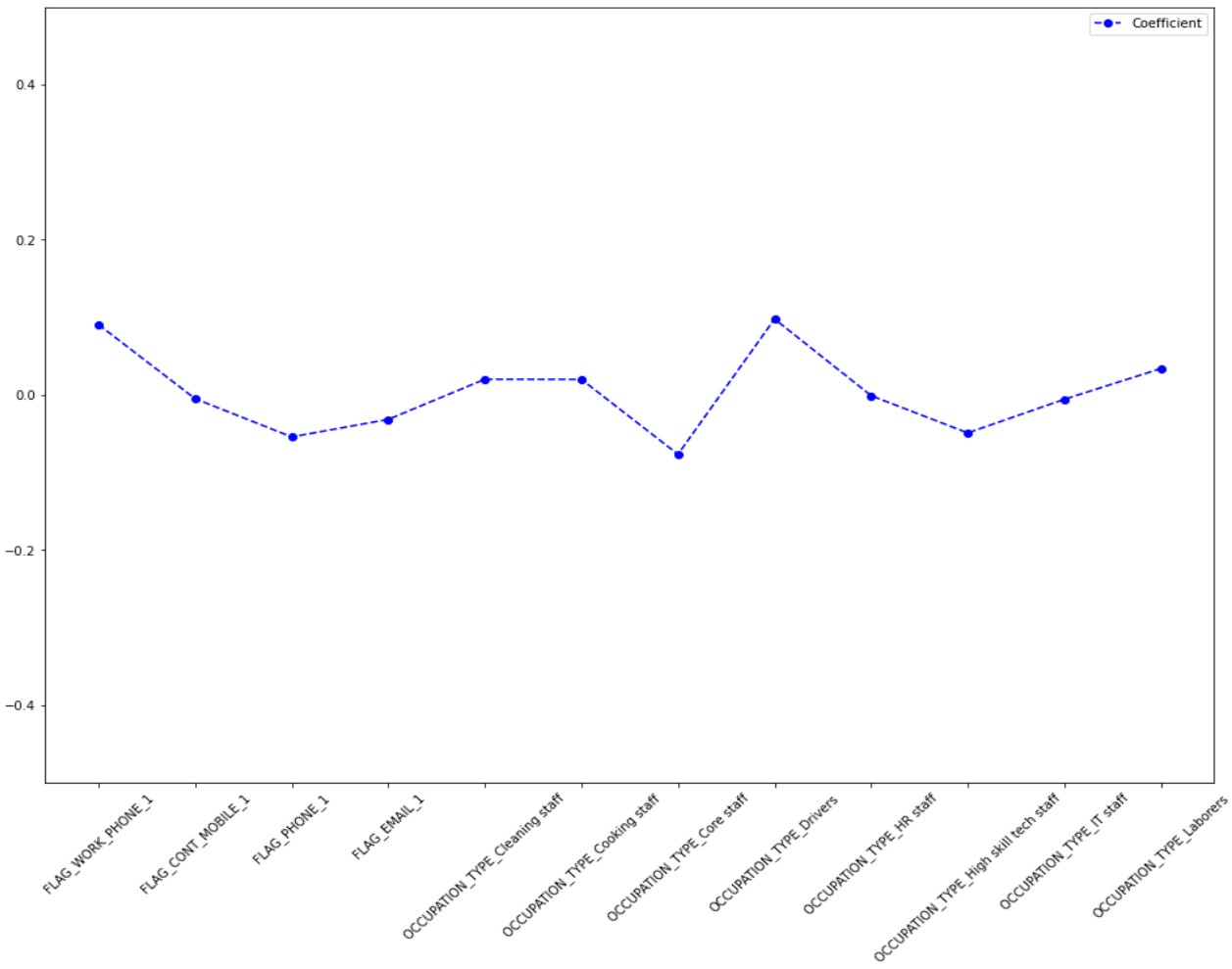


Figure: X-axis → Columns in dataset | Y-axis → Coefficient in Logistic Regression

Here we are only plotting showing some of the features. Please refer to the notebook to see all the plots.

b. Feature importance using Decision Trees

For a decision tree classifier, the split at each node is performed by deciding the value of a feature that will lead to the “purest” possible separation of the TARGET into two groups, i.e. a separation having lowest impurity. Hence, a feature having a high value of feature_importances_ is very important in separating the defaulters from the non-defaulters.

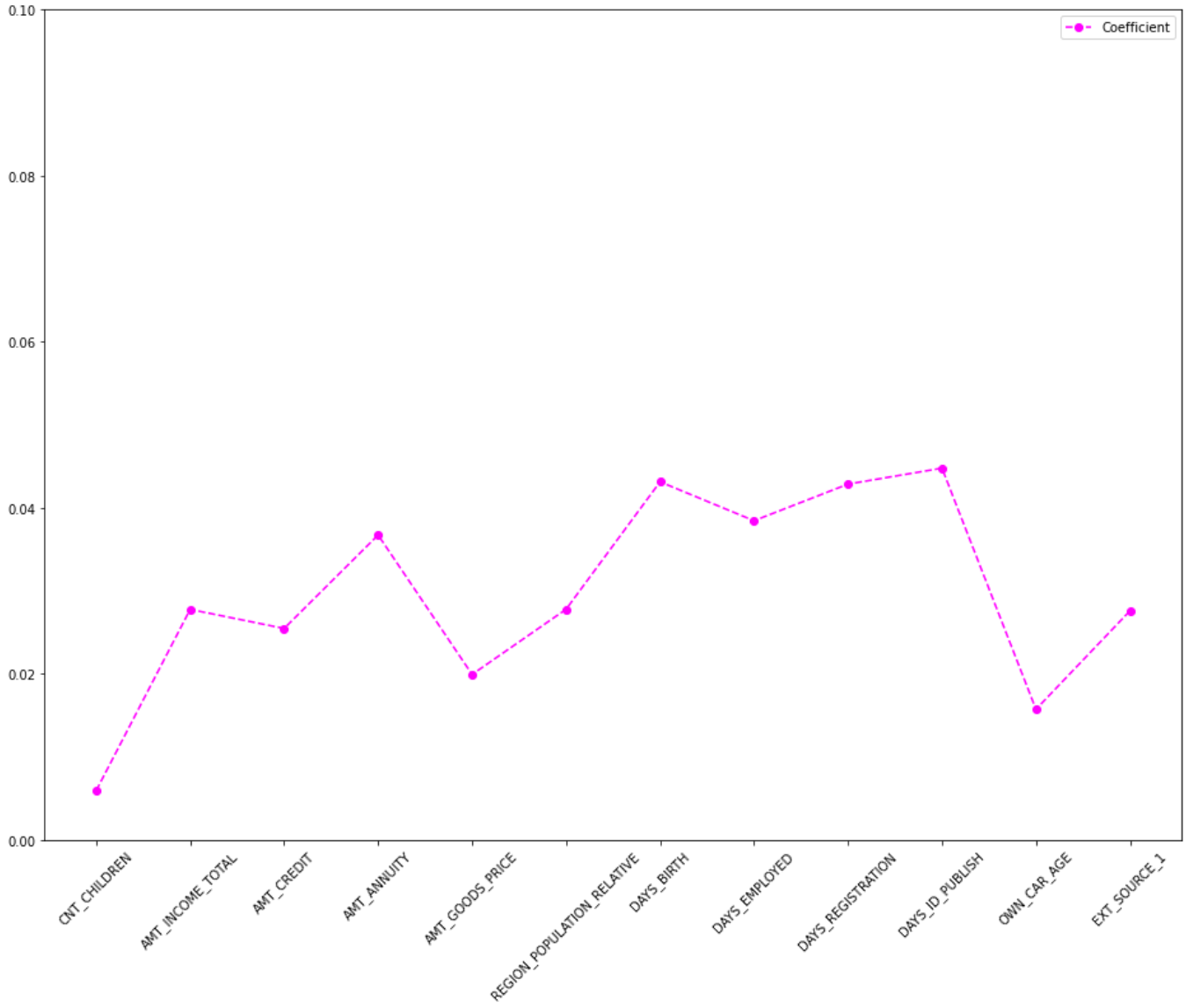


Figure: X-axis → Columns in dataset | Y-axis → Coefficient in Decision Tree model

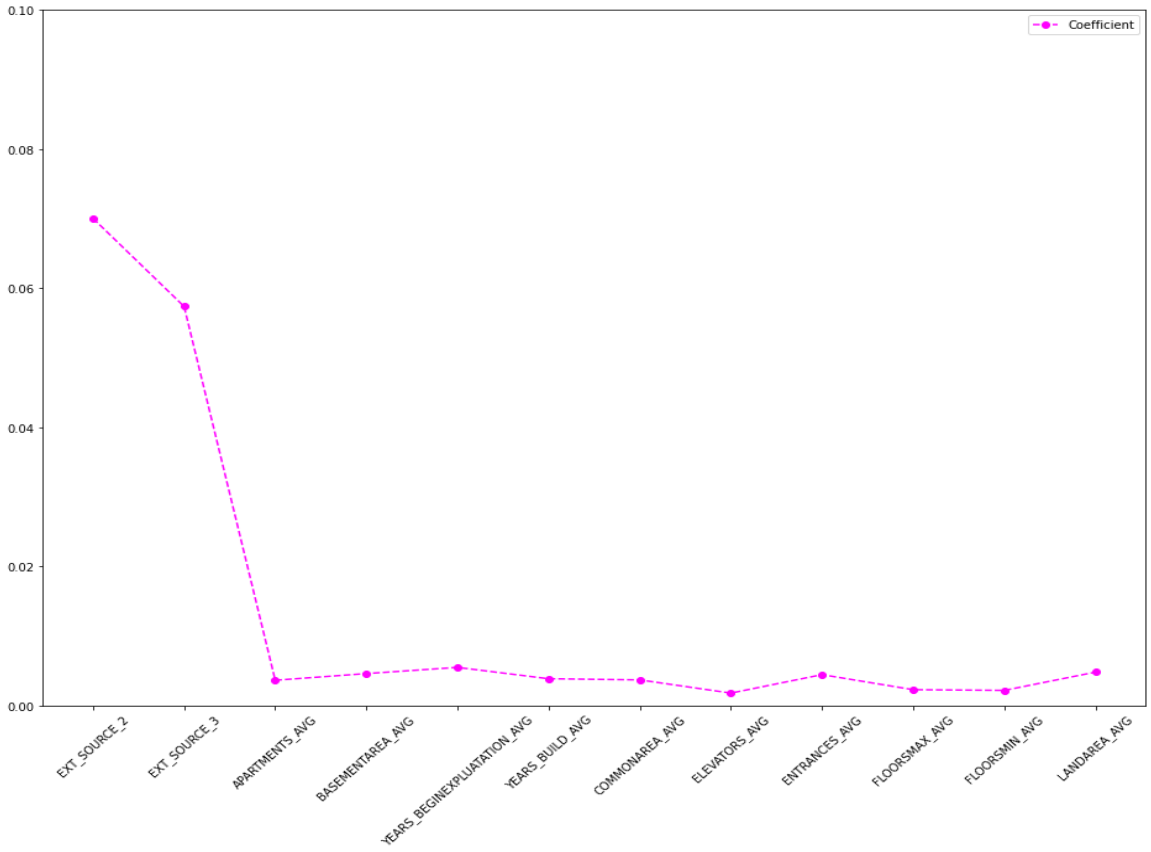


Figure: X-axis → Columns in dataset | Y-axis → Coefficient in Decision Tree model

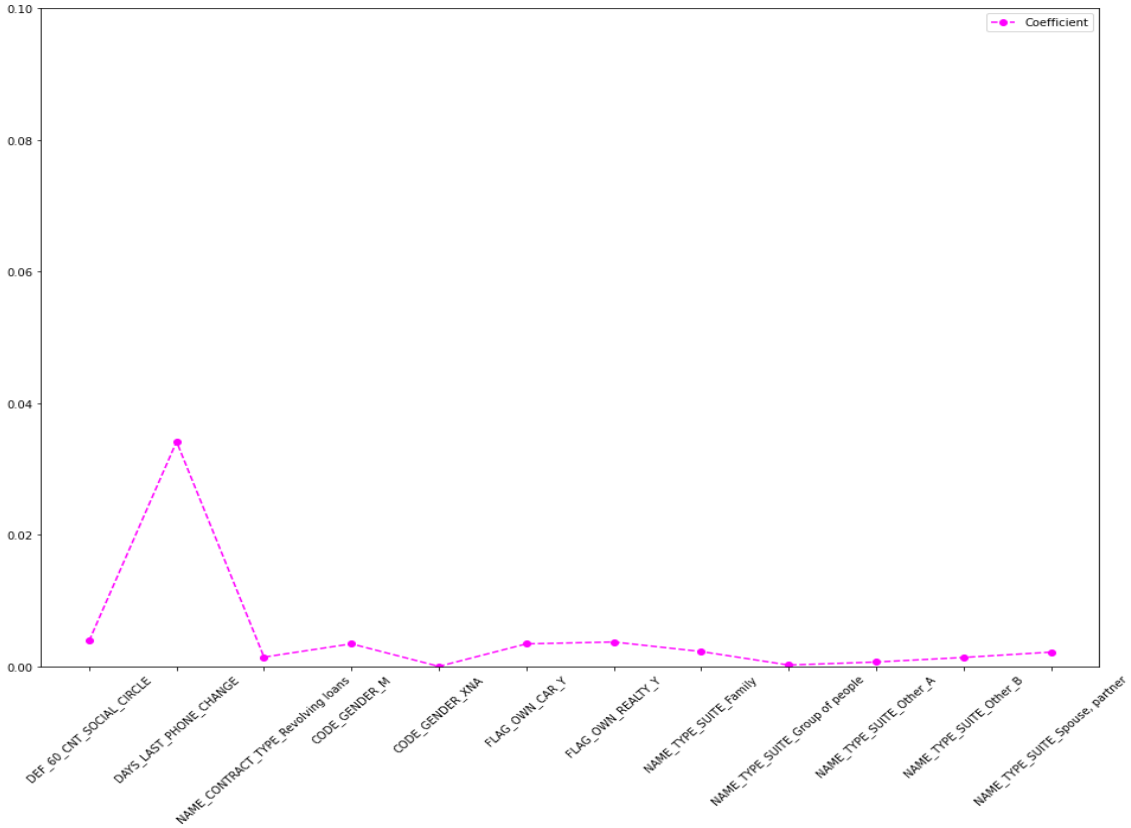


Figure: X-axis → Columns in dataset | Y-axis → Coefficient in Decision Tree model

12. SUMMARY OF OBSERVATIONS

- In all the significance tests, we find a repeating pattern that the amount of loan of the borrower is a deciding factor in default. Higher amounts of loan are being more defaulted on.
- External ratings are highly reliable. Good loans have been rated considerably higher than bad loans.
- The Gender, Age and Family status of the applicant are significant predictors. In general, male applicants are defaulting more.
- People who have changed either their ID or their phone are defaulting more.
- Higher amount of annuity or installment increases the chances of default.
- More educated borrowers are defaulting less.