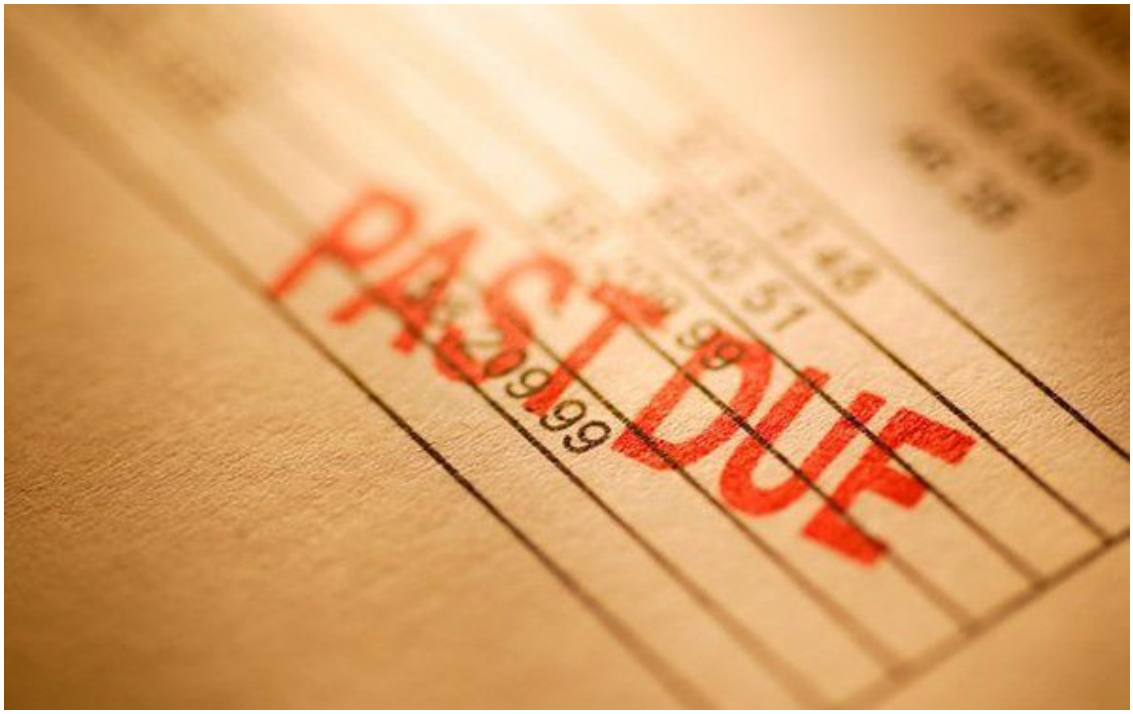
PREDICTING LOAN DEFAULT - HOME CREDIT

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. Carrying out visual EDA for getting a general sense of data - 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.
2. Ascertaining critical attributes and their effect on target - We will use statistical techniques to find the relative importance of attributes.
3. Cleaning and preprocessing - the data has a large number of null values. Depending upon the count of null values an the feature, we will use various techniques to handle null values
4. Applying ML models to predict default probability - the final aim of this capstone is to develop a model to predict probability of default. 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. Dataset Description

The dataset has been collected from a Kaggle competition:

https://www.kaggle.com/c/home-credit-default-risk/data

- The main customer data is present in the file “application_train.csv”. This is the primary dataset containing core customer attributes. We will use this for training and evaluation.

- There are 122 independent features and 1 dependent feature. We need to predict the probability of “TARGET”, to maximise ROC_AUC score.

- 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.

- With some domain knowledge, we make some initial assumptions about the data. For example, the default probability may be heavily influenced by the customer’s income; we may observe more defaults by unemployed people; the social circle of the borrower may affect his/her repayment habits; more educated people may be less prone to inconsistent repayment habits. As we carry out analysis, we will establish the truth about our assumptions.

8. Cleaning and Preprocessing

(i) Removal of null values

- For columns having very few null values, we will use proper imputation techniques.
- 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. We will first build a model without dropping any values. Later, high null features will be dropped to see if we achieve a gain in accuracy.
- Categorical features contain very few nulls. We will impute with the mode.

(ii) Handling of incorrect data

- 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!
- 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.We will convert all such columns to a single sign

9. Exploratory Analysis and Visualization

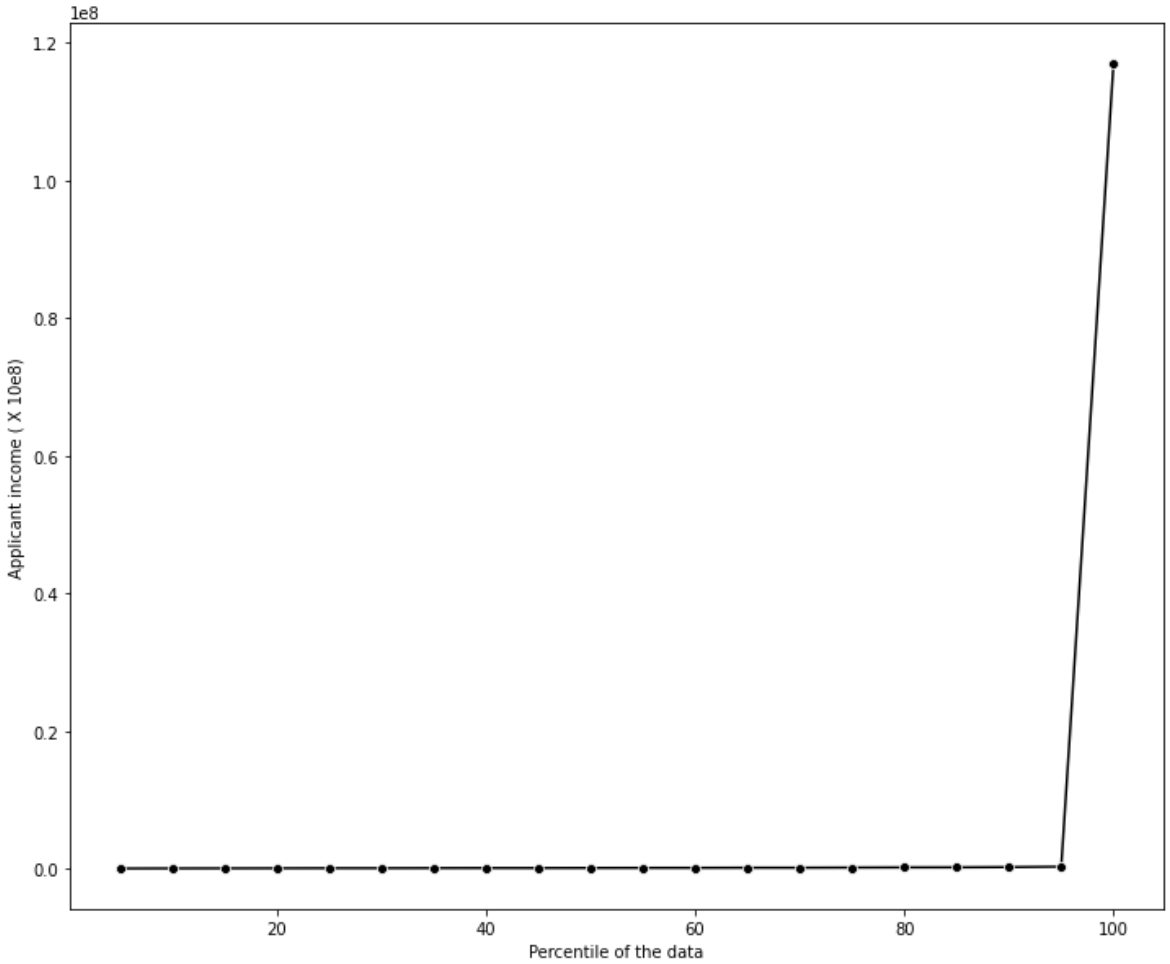
The dataset can be broadly divided into numerical and categorical features. Features like income, credit amount, external rating, etc are numeric. Features like organization type of client, education level of the client, industry where the client works, type of loan taken, etc are categorical features.

We will separate the numerical and categorical columns, and use appropriate exploratory tools for these.

First, let us look at the numerical columns.

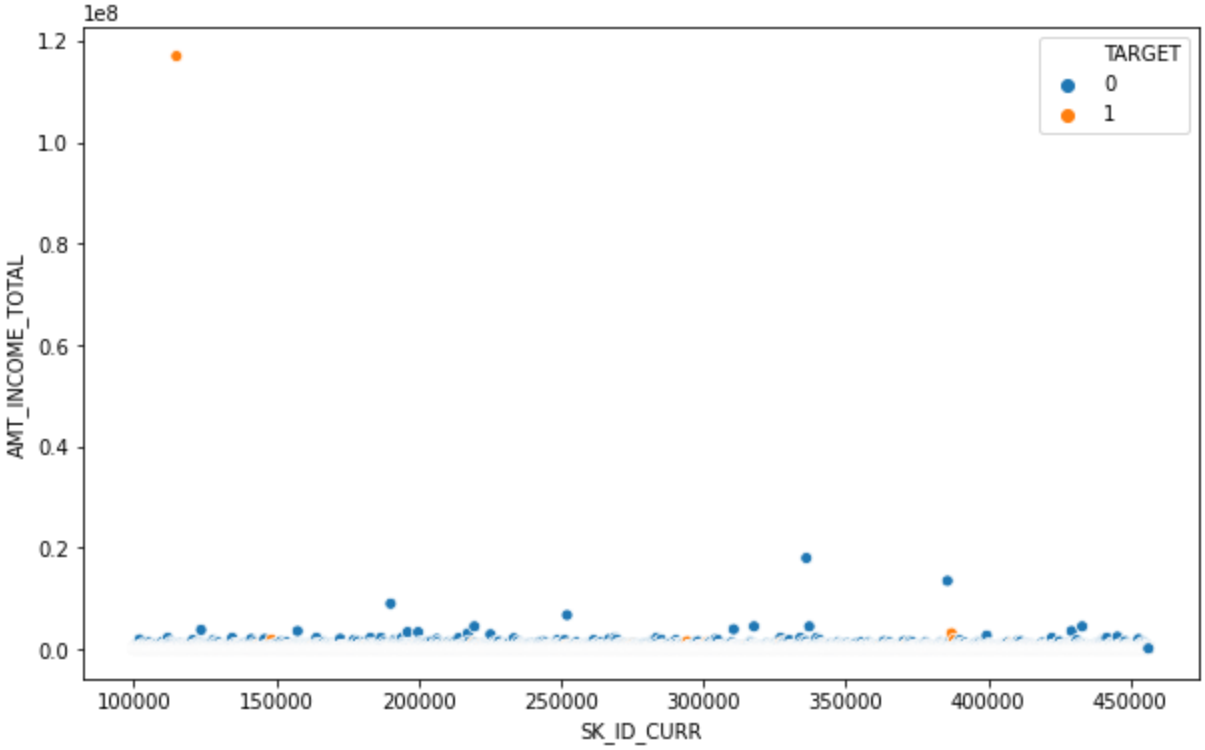
The AMT_INCOME_TOTAL column represents the income of the borrower. Domain knowledge tells me that this is an important feature.

But the incomes are very unevenly distributed.



As we can see, 95% of the income values are relatively small. We have a few very large values. Upon analysis, we fin the following:

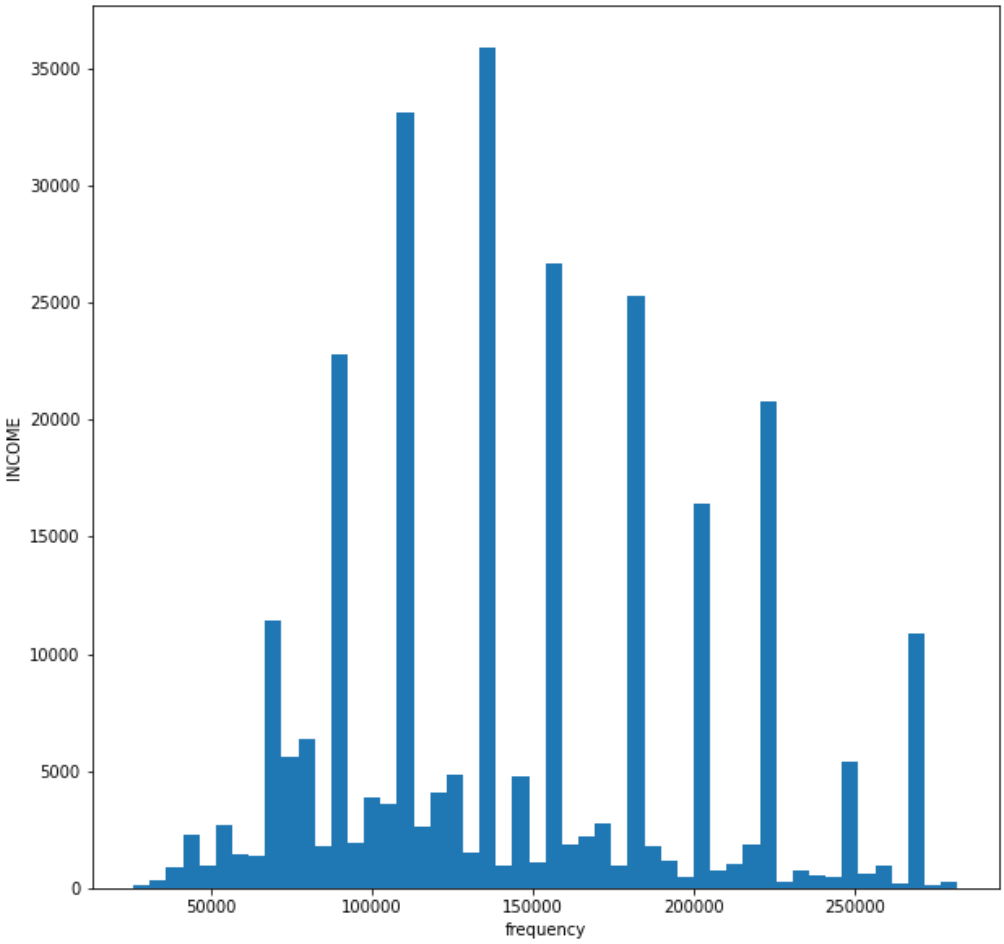
- The median income is 147150 units. The 1st and 3rd quartile of the income are 112500 and 202500 respectively. This seems to indicate that there isn't much spread. But there are many outliers.
- The highest income is in the range of 1.1 million, and the lowest is 25000 units. Hence, there is a great variability in the income. Majority of the income values are spread within a standard range of median +- (1.5*IQR).



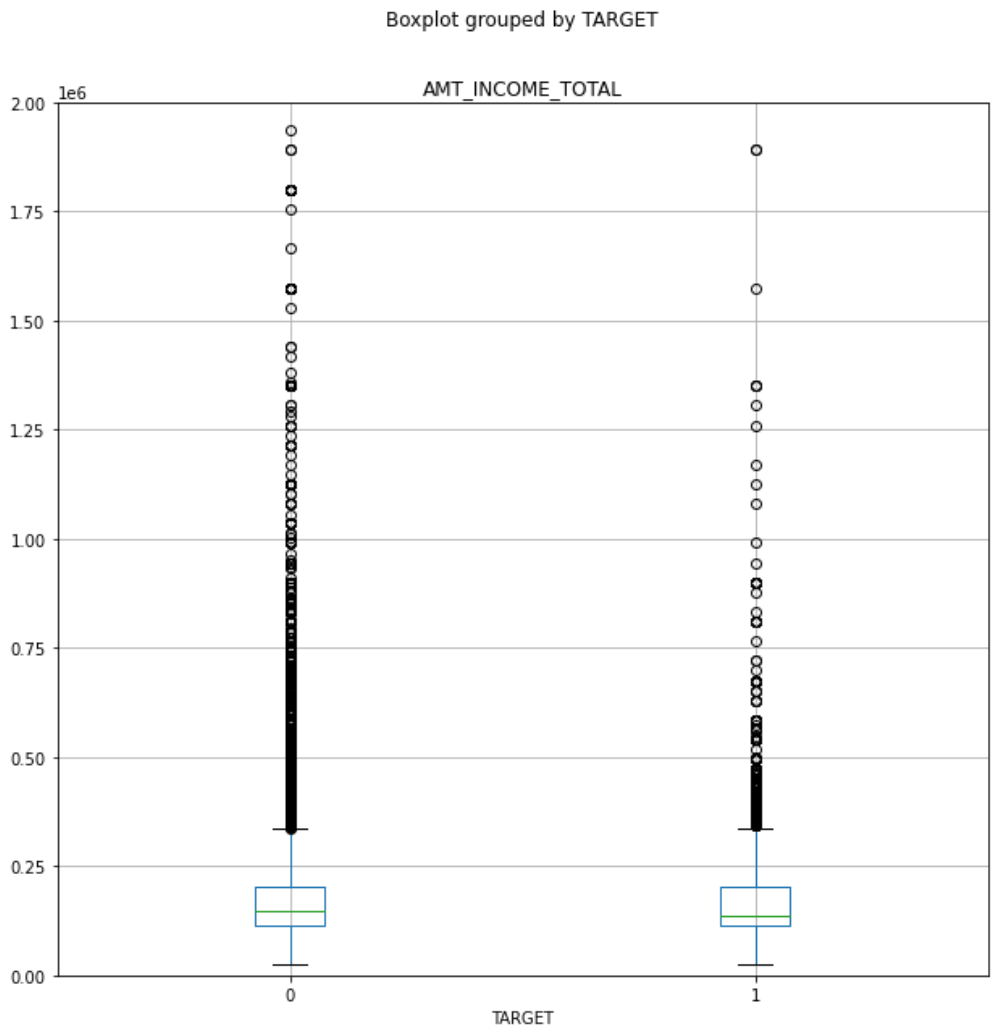
Here, "1" represents bad loans. As we can see, the highest value of income is for a loan defaulter.

- Interestingly, the loan by the borrower with the highest income has turned bad !!!

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

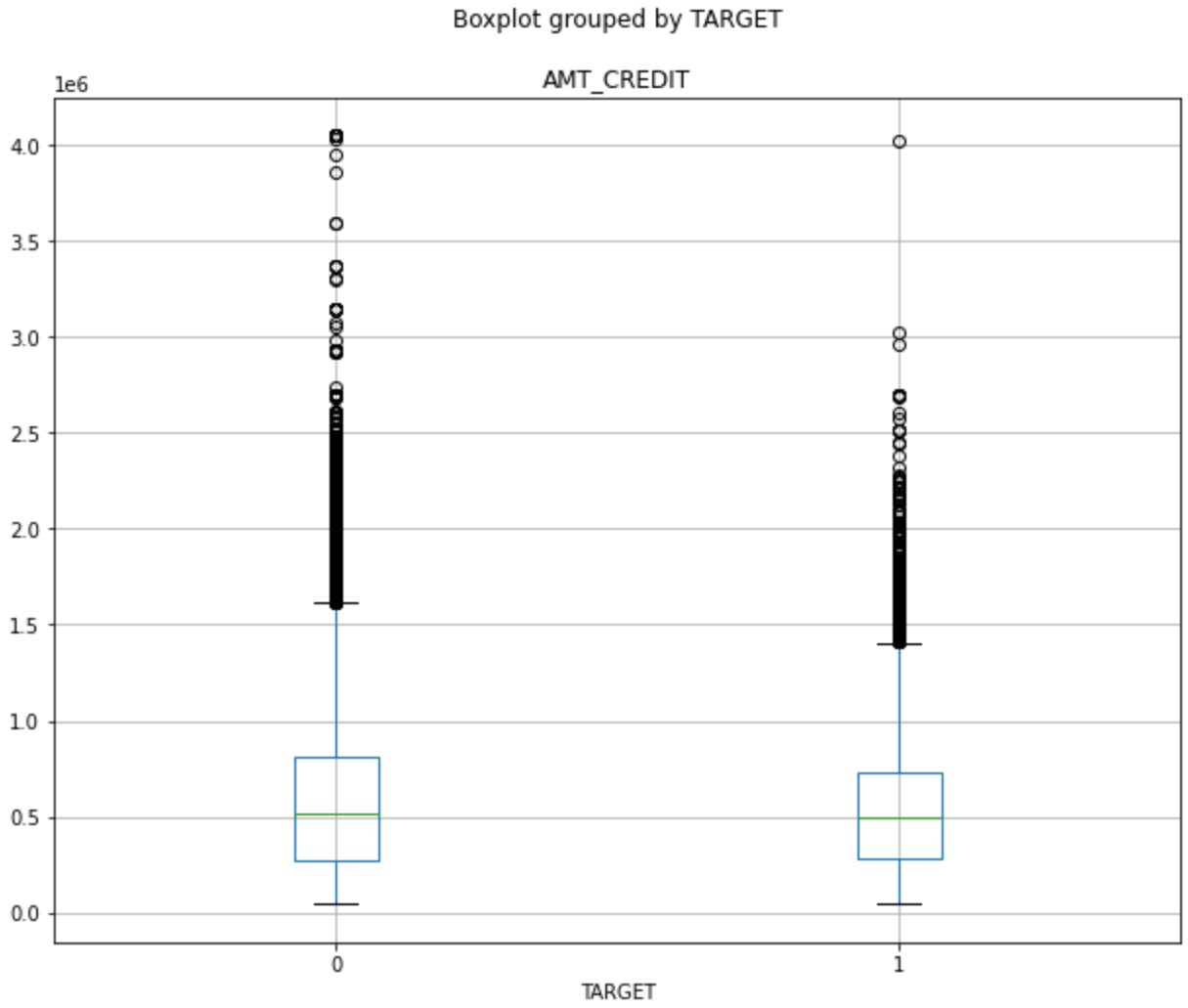


- We see a somewhat “normal-ish” distribution of incomes



- Richer people are clearly more densely present in the good-loans category. This is expected. All else remaining the same, having more money means better repayment.

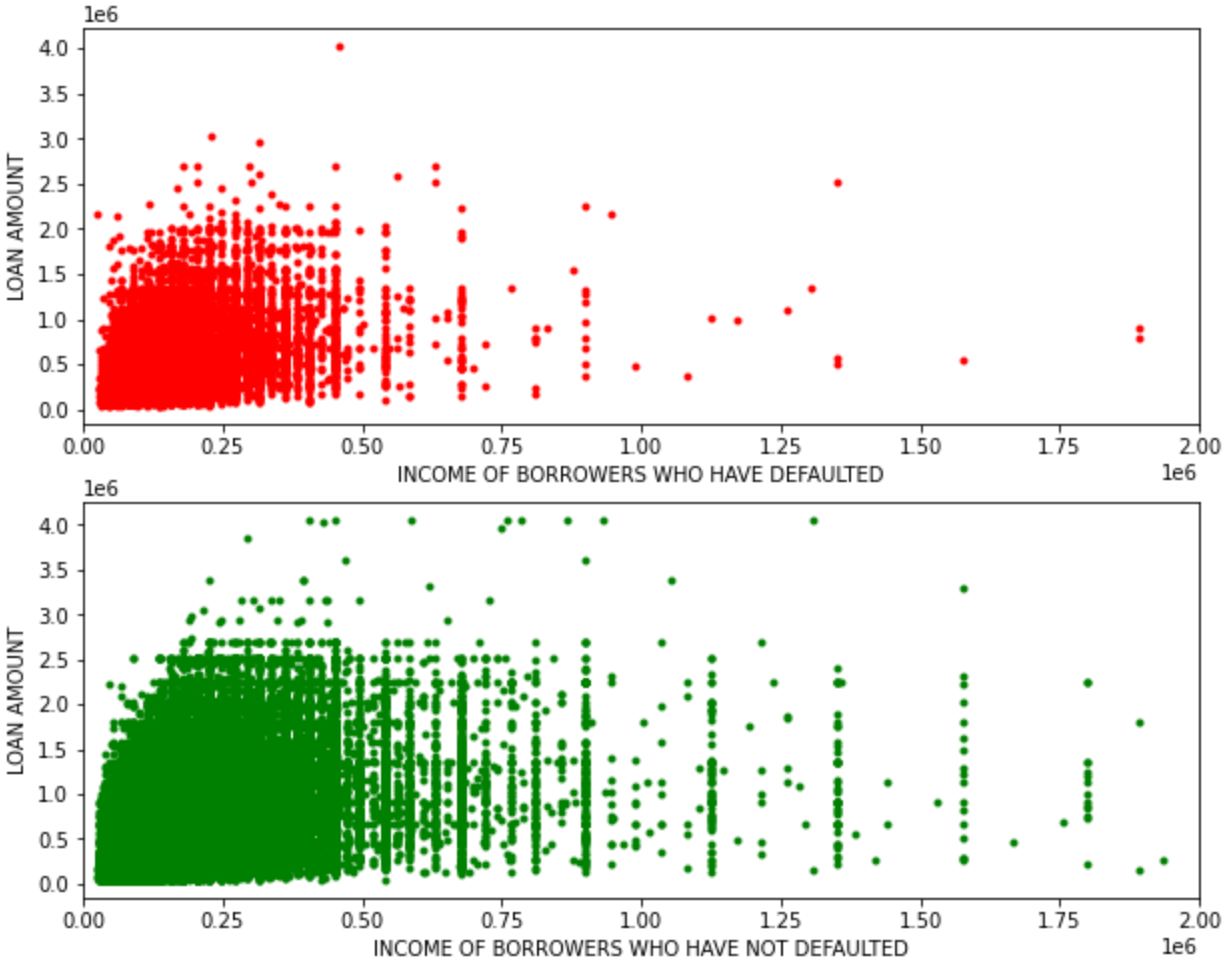
Now, let us take a look at how the amount of loan varies by TARGET.



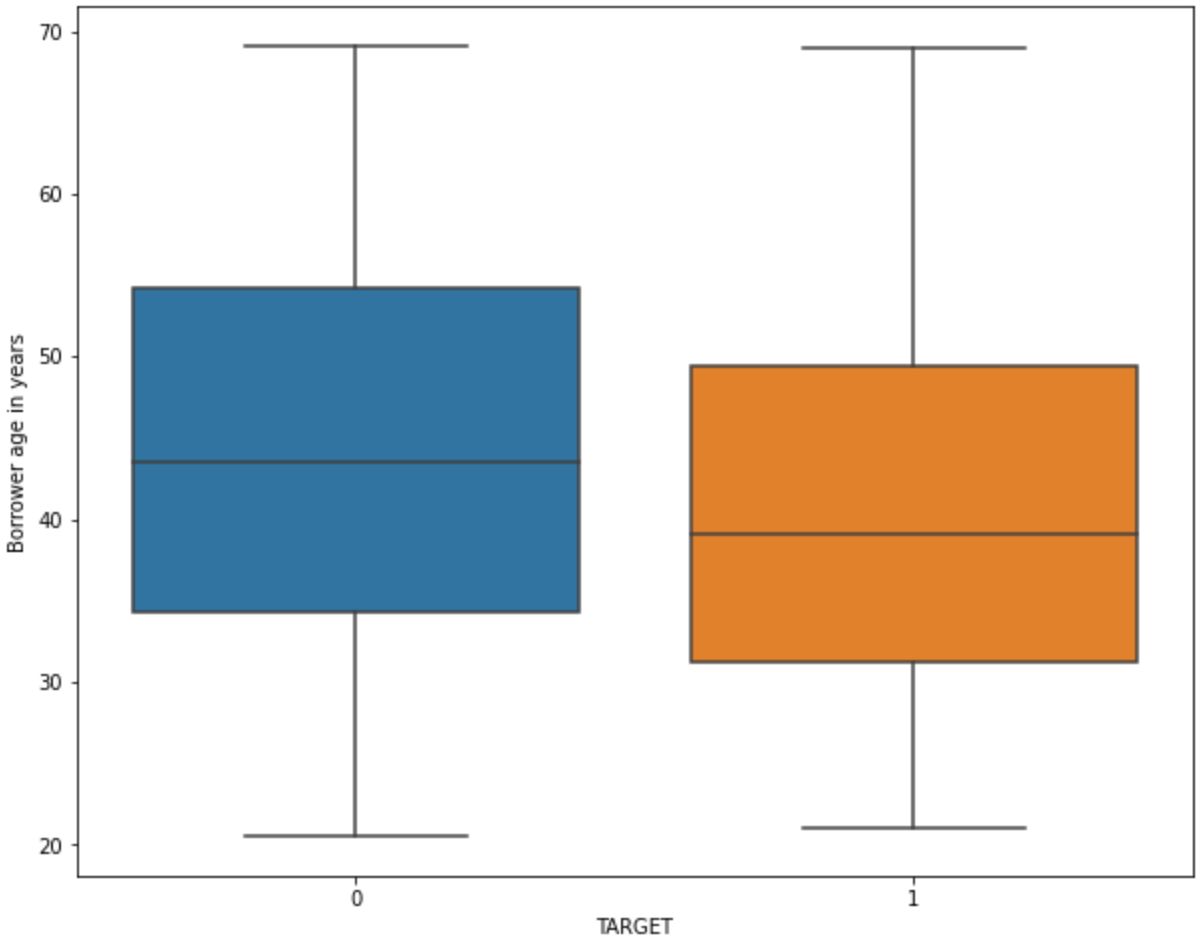
- People with higher amounts of loans are repaying better !! Thus having more debt does not imply bad loan repayment habits. Some borrowers are just more credit-worthy.

We will now analyse some more important columns that are important predictors of credit-health in the banking world, like AMT_GOODS_PRICE (price of the assets acquired using loan money), borrower age (which determines loan limit eligibility) and the loan type.

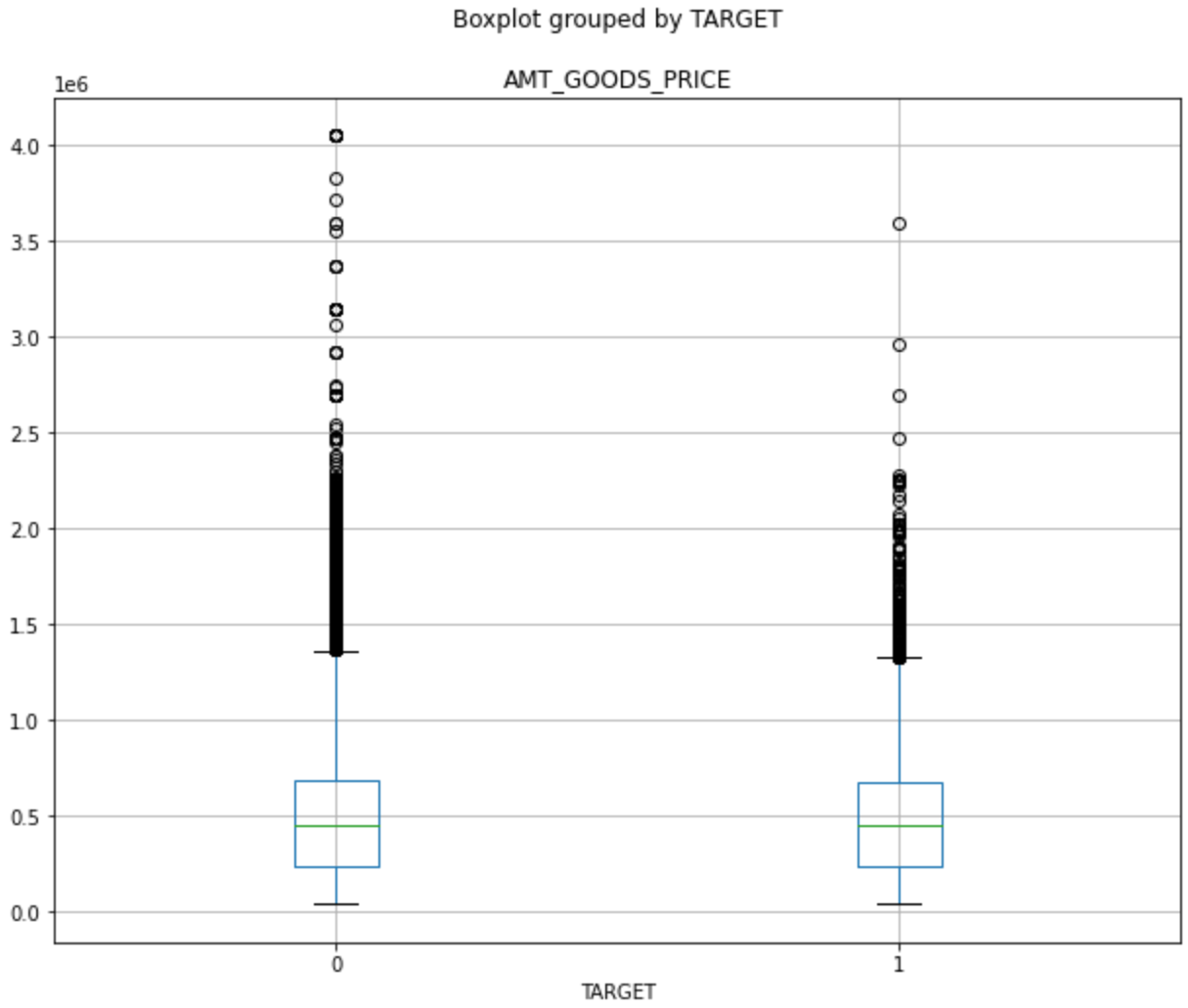
Let us see how income varies by loan amount, for bad and good loans.



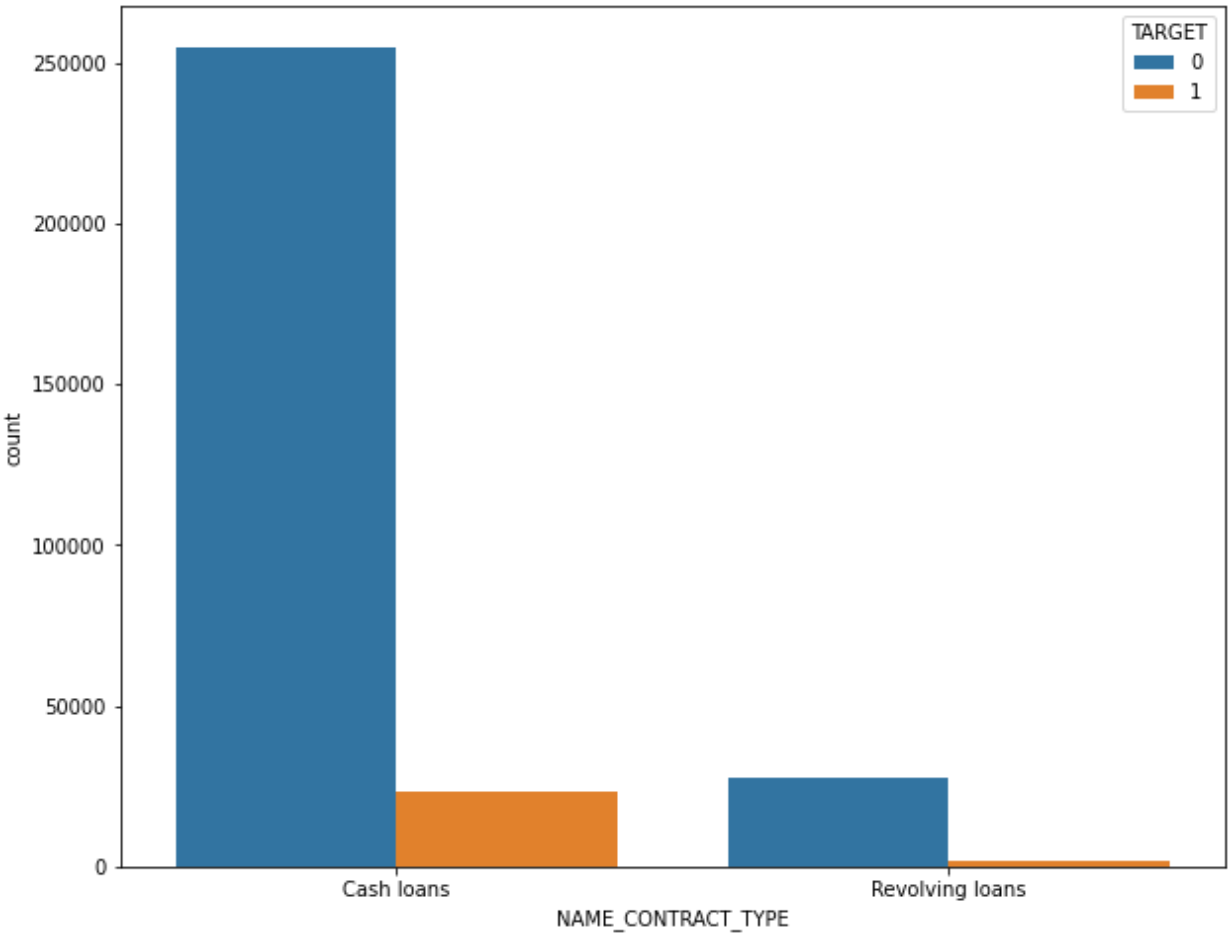
- The majority of the borrowers are in the low-income, low-credit category. Borrowers with high income are more widely present in the good-loans category. But we would also expect the borrowers with high incomes to get greater loans. That is not the case. The loans amounts are fairly similar for low-income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



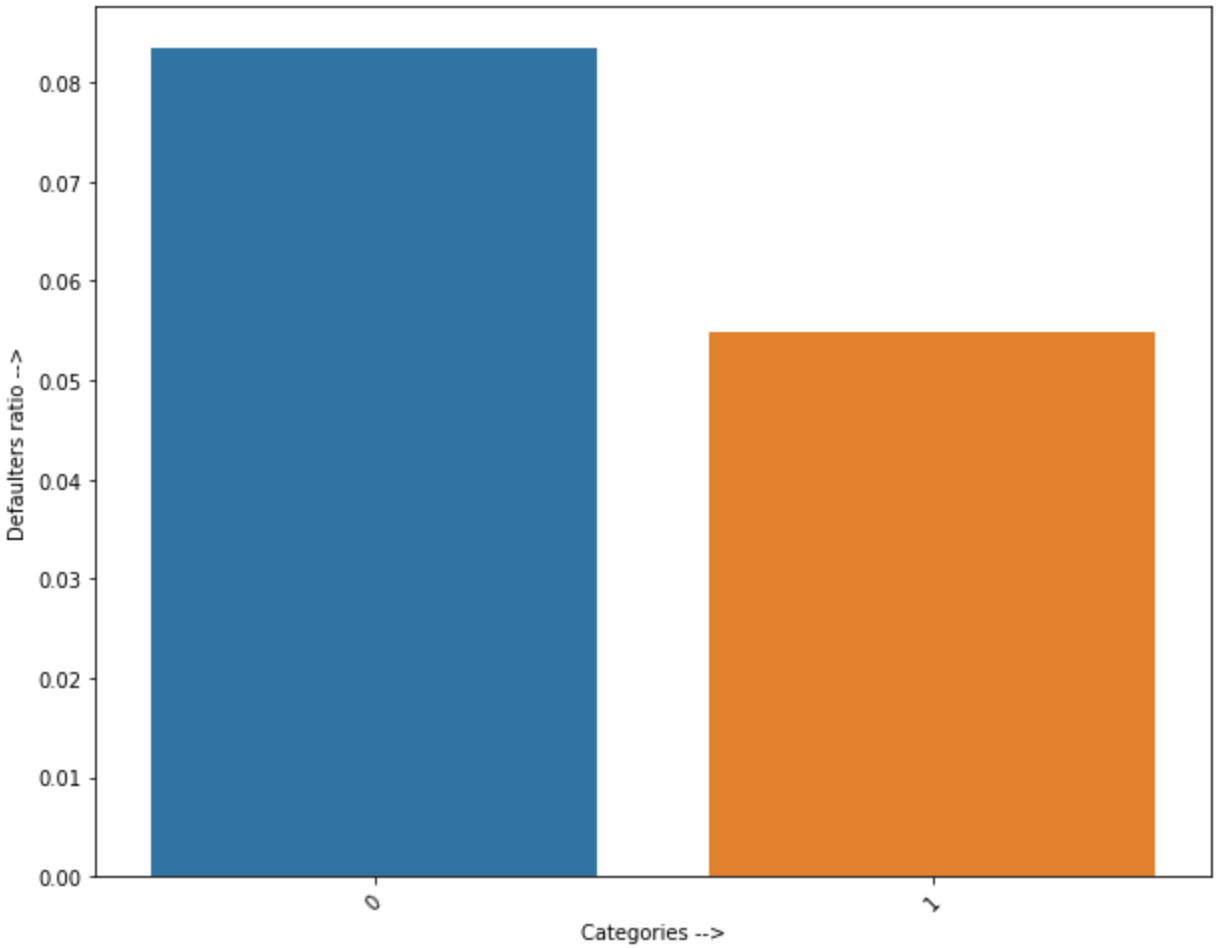
- Younger borrowers seem to default more. This may be due to low-paying jobs (but we have seen that income is not a deciding factor !) or a lack of financial discipline. It is seen that as people grow and have more responsibilities, they develop financial discipline.



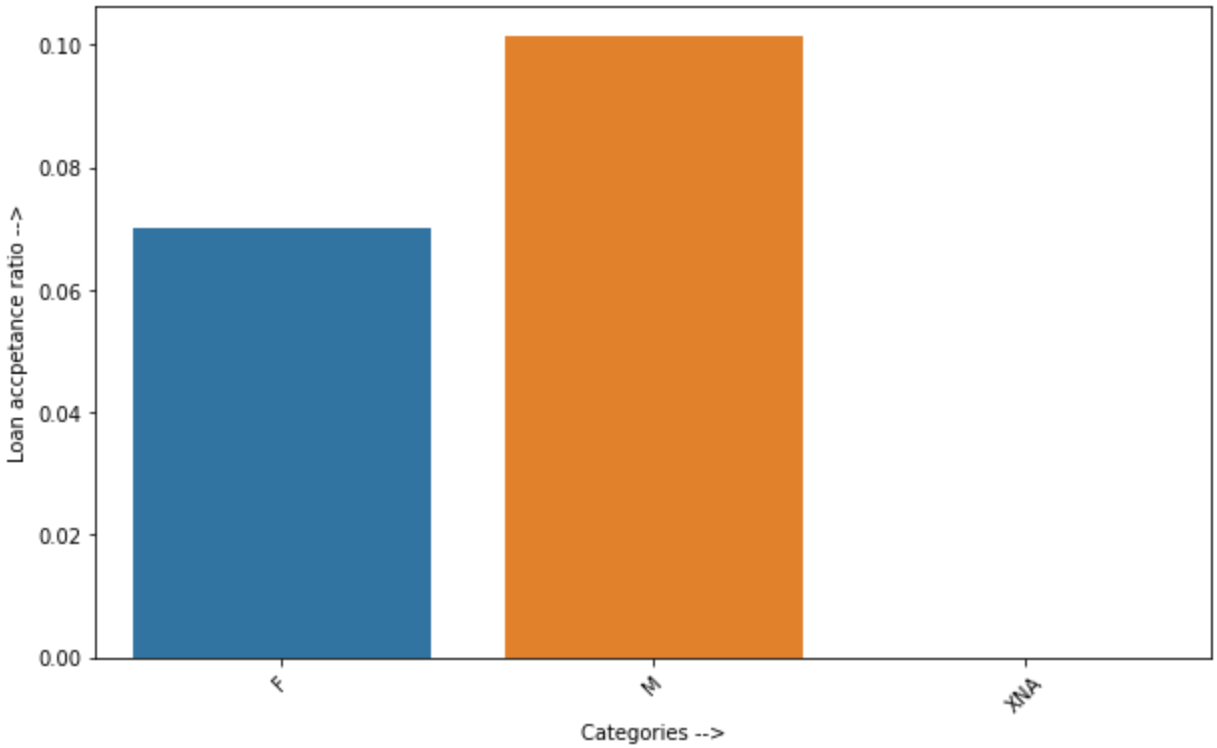
- The median value price of goods purchased with the loan money is fairly similar. This can be seen in the boxplots. But good loans have higher density of high priced items. This too is an interesting observation.



- 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 and 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

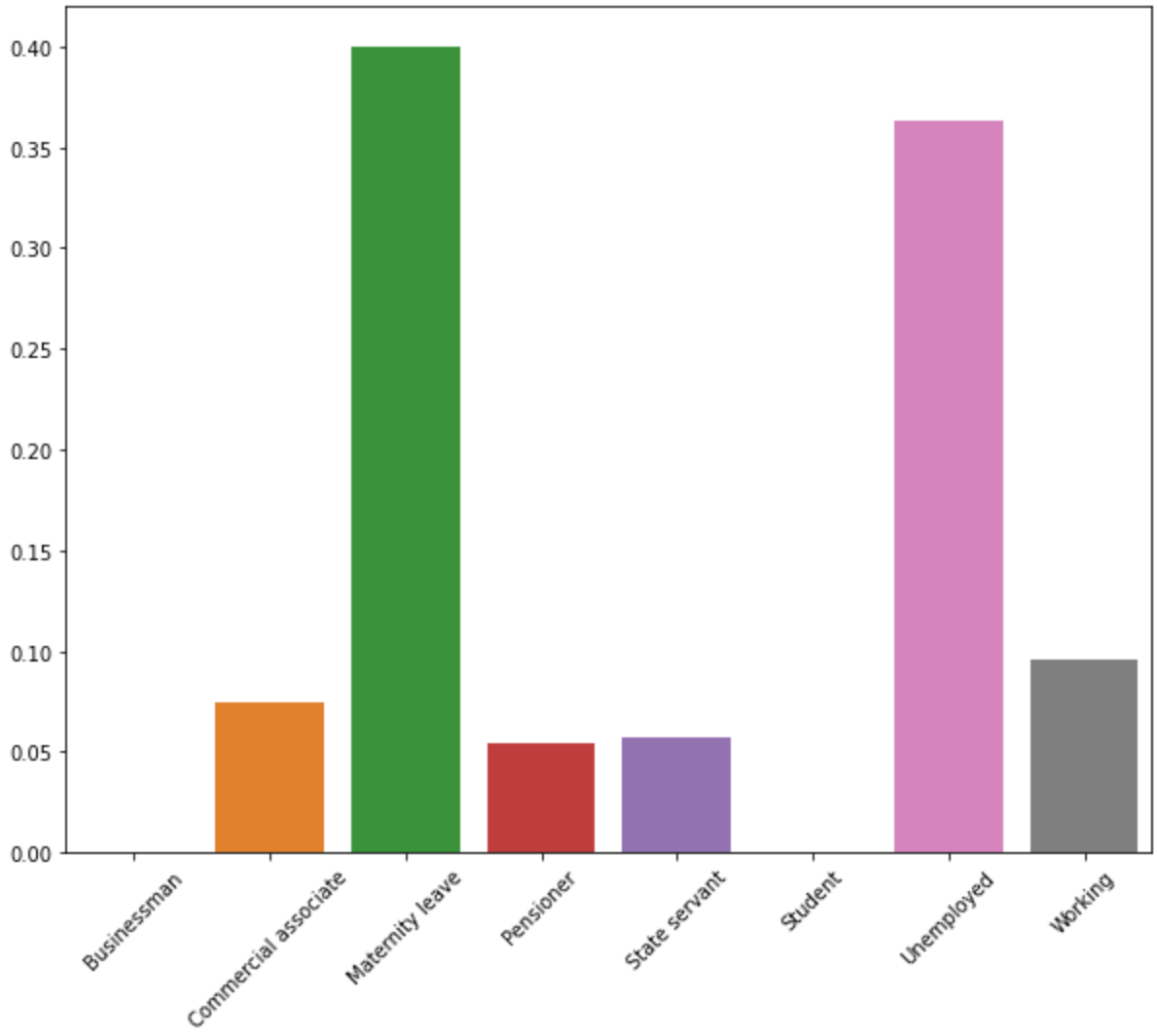


- Cash Loans are in BLUE. Thus cash loans seem to have a higher default ratio.



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

- Under normal circumstances 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.

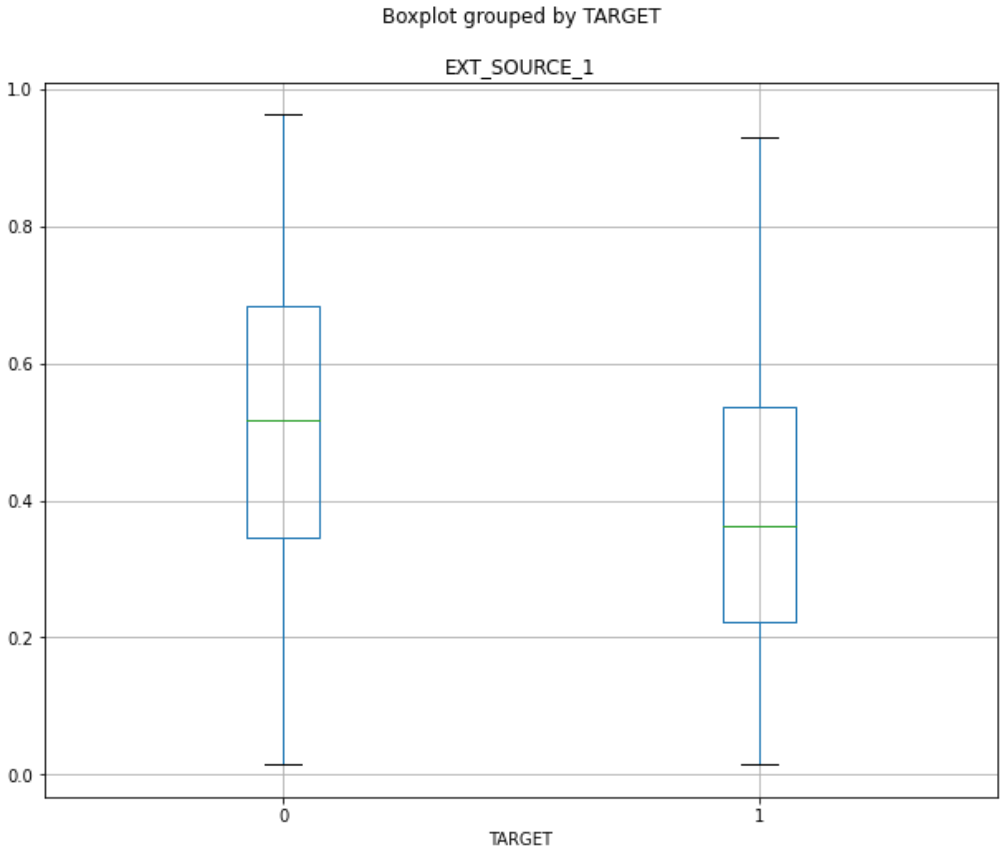


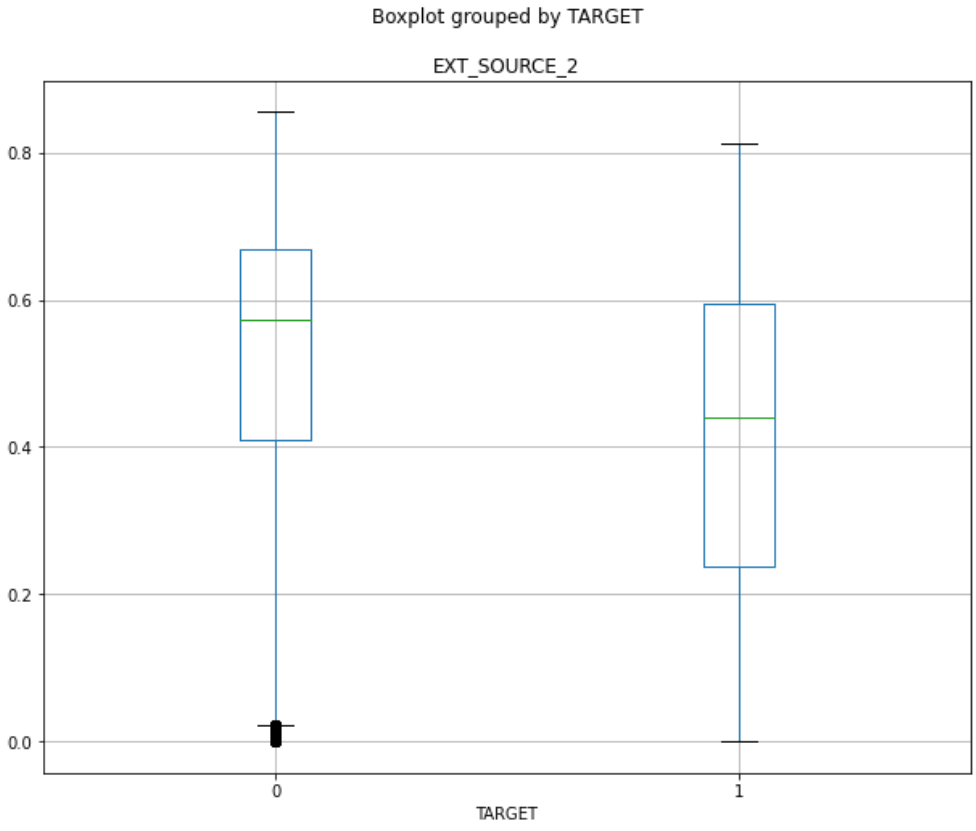
- Unemployed persons have the highest default rate. This is expected. It's interesting that unemployed people were even provided a loan in the first place.

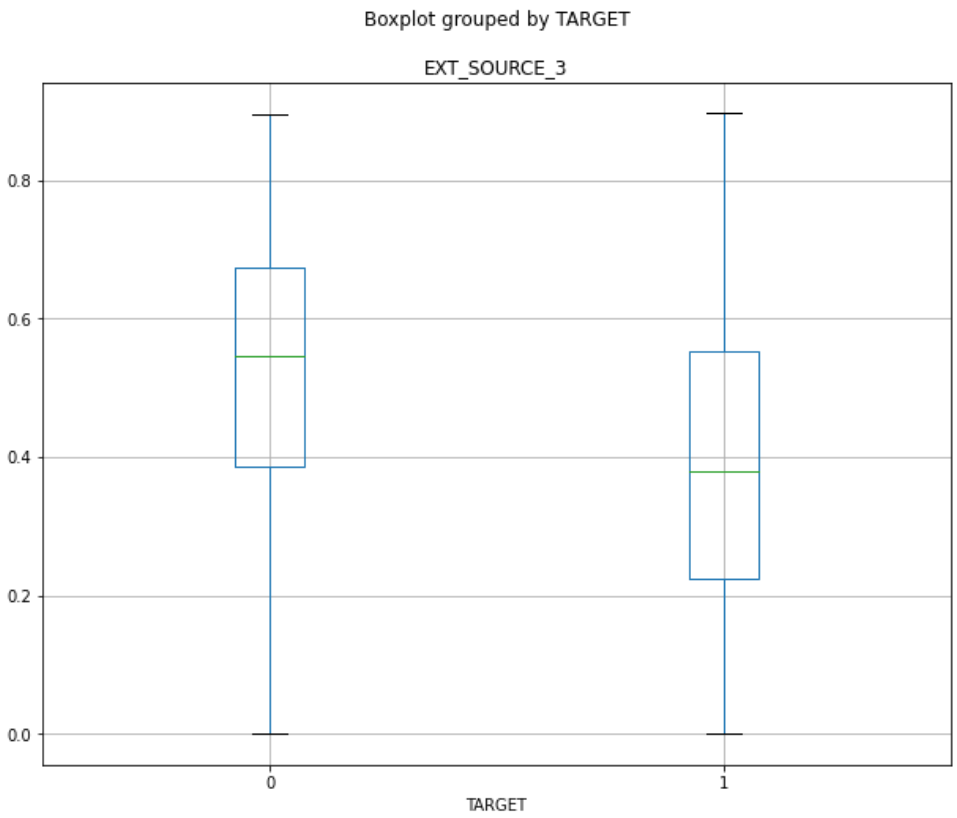
- Businessmen and students have very few defaults.
  .
- Default ratios of pensioners, state servants and salaried class are fairly low.

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

External credit rating agencies provide credibility ratings to borrowers. These are fairly good indices to judge loan health.


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



- Borrowers of good loans have better median ratings on all 3 values.

Apart from the above plots, extensive visual analysis provided some more interesting insights, as below:

- Less educated people have a lower default ratio.
- Owning a house does not seem to impact the probability of default.
- Generally, bankers expect the people providing all the documents to be diligent about repayment. But there it hardly seems to matter, whether the borrower has provided all the 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.

10. Feature Importance using Statistical Analysis

Let us find out which features are relatively more important than others in predicting default.

To analyse the distribution of numerical features between bad 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.

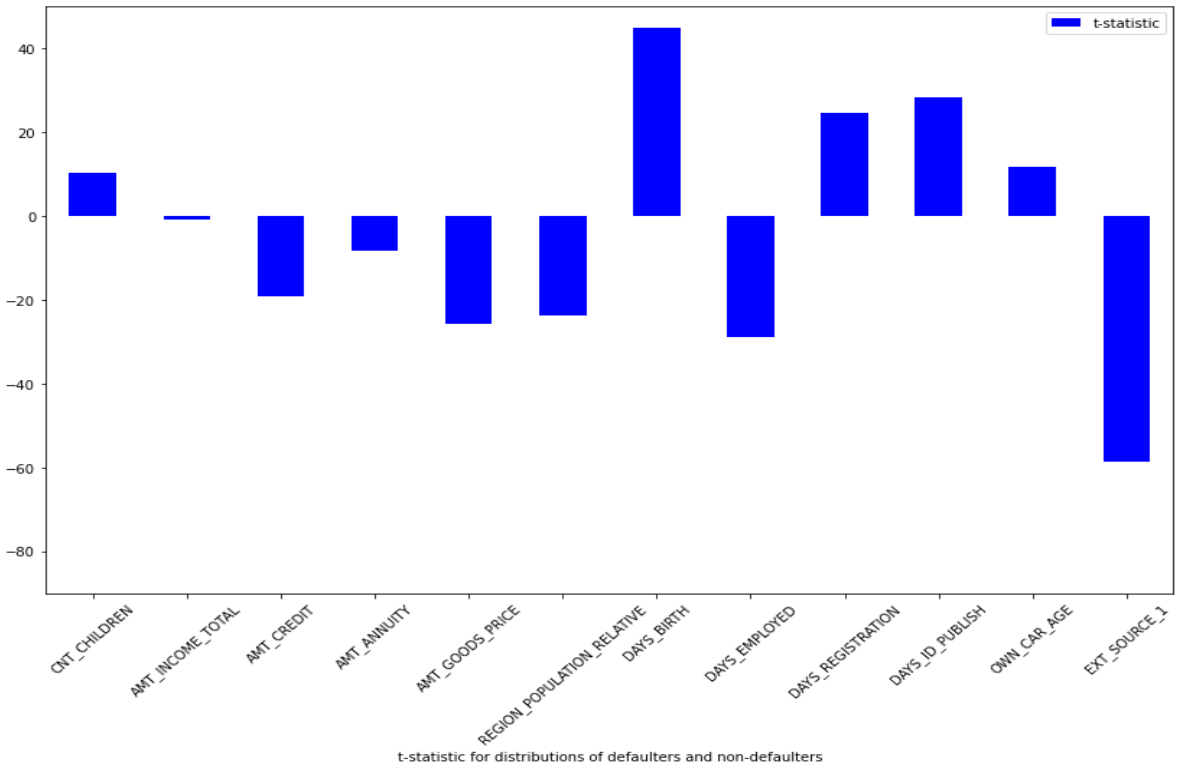


Figure: X-axis → Columns in dataset | Y-axis → T-statistic for independent t-test

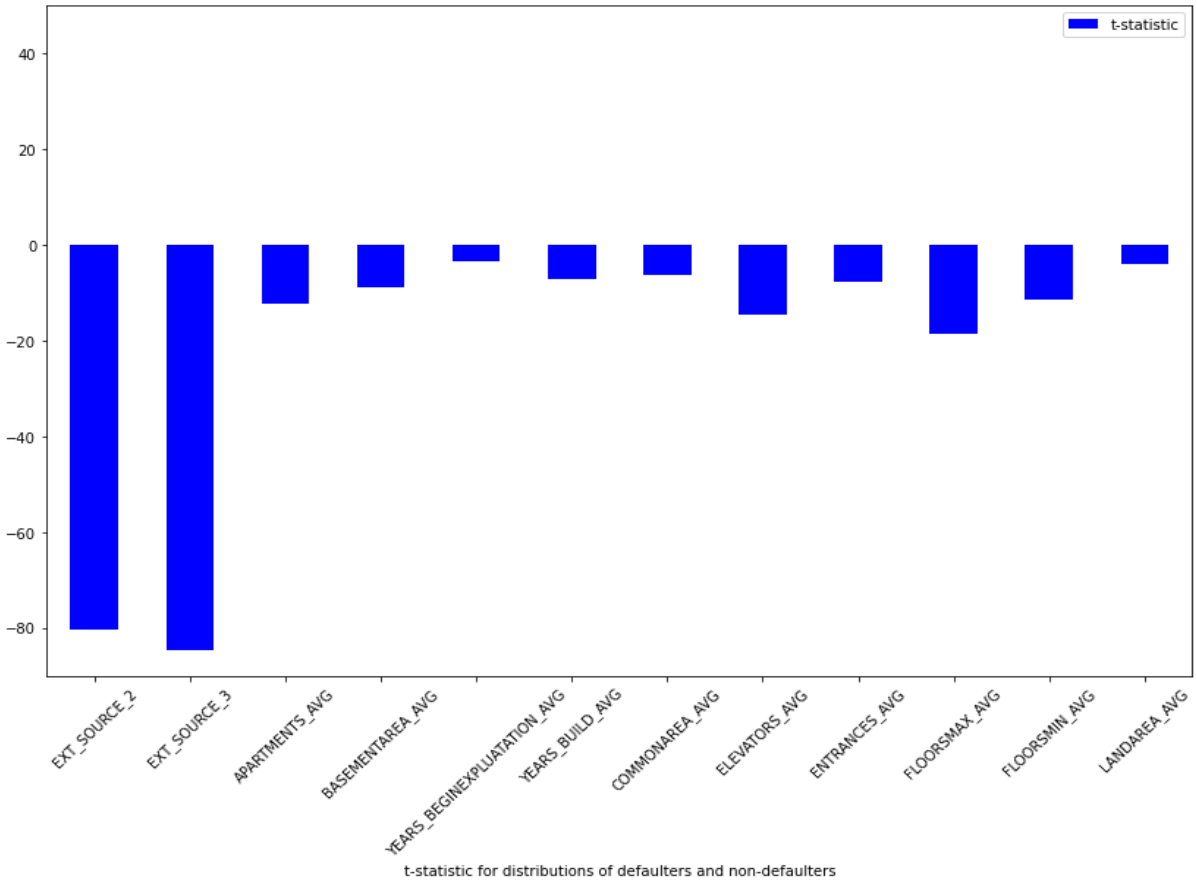
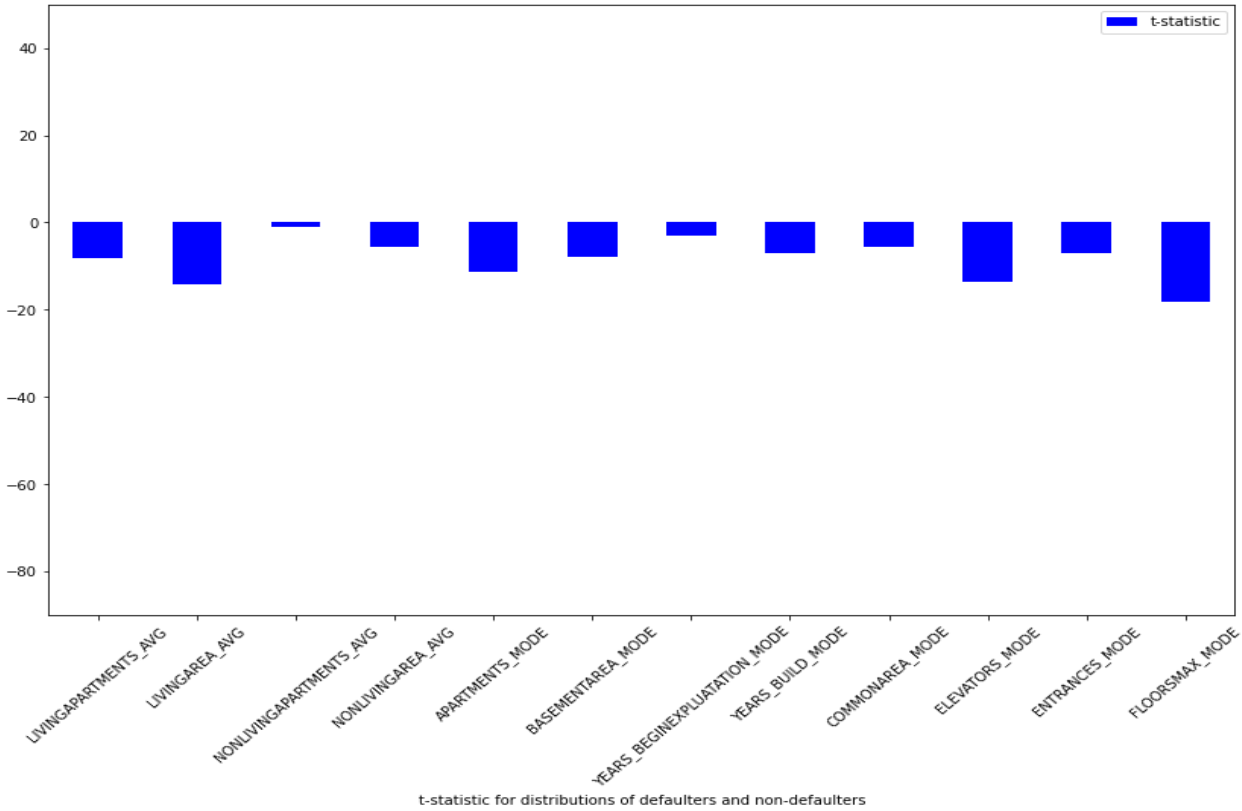


Figure: X-axis → Columns in dataset | Y-axis → T-statistic for independent t-test

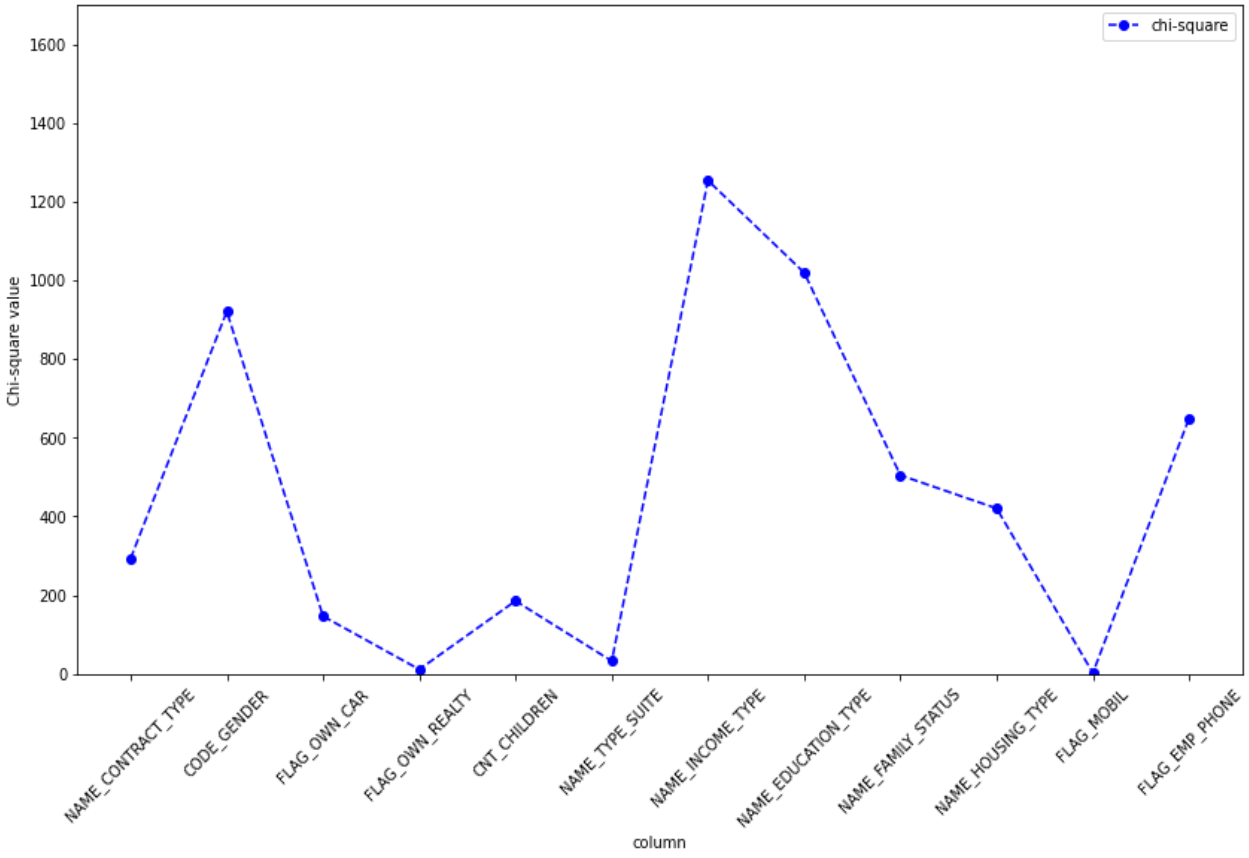
Figure: X-axis → Columns in dataset | Y-axis → T-statistic for independent t-test

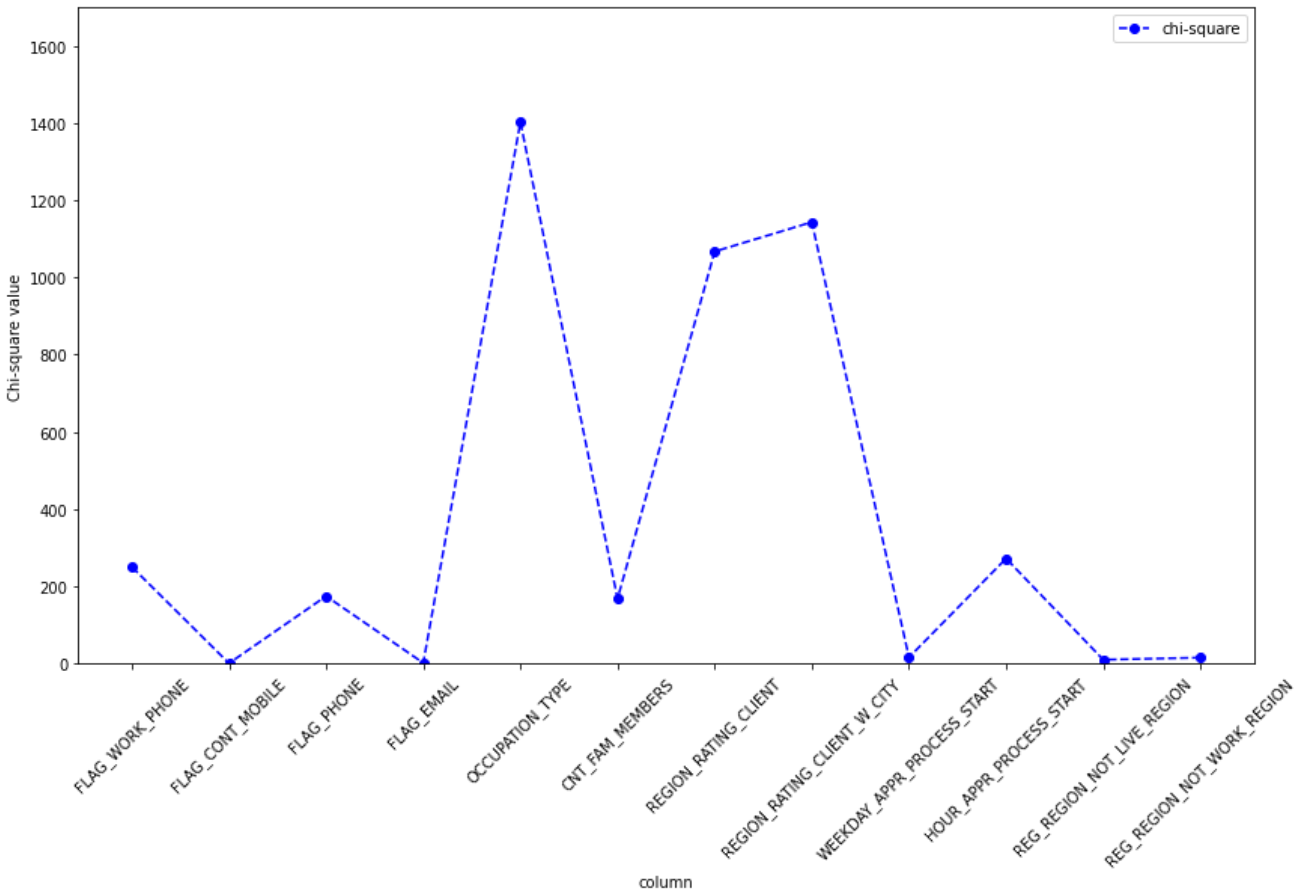


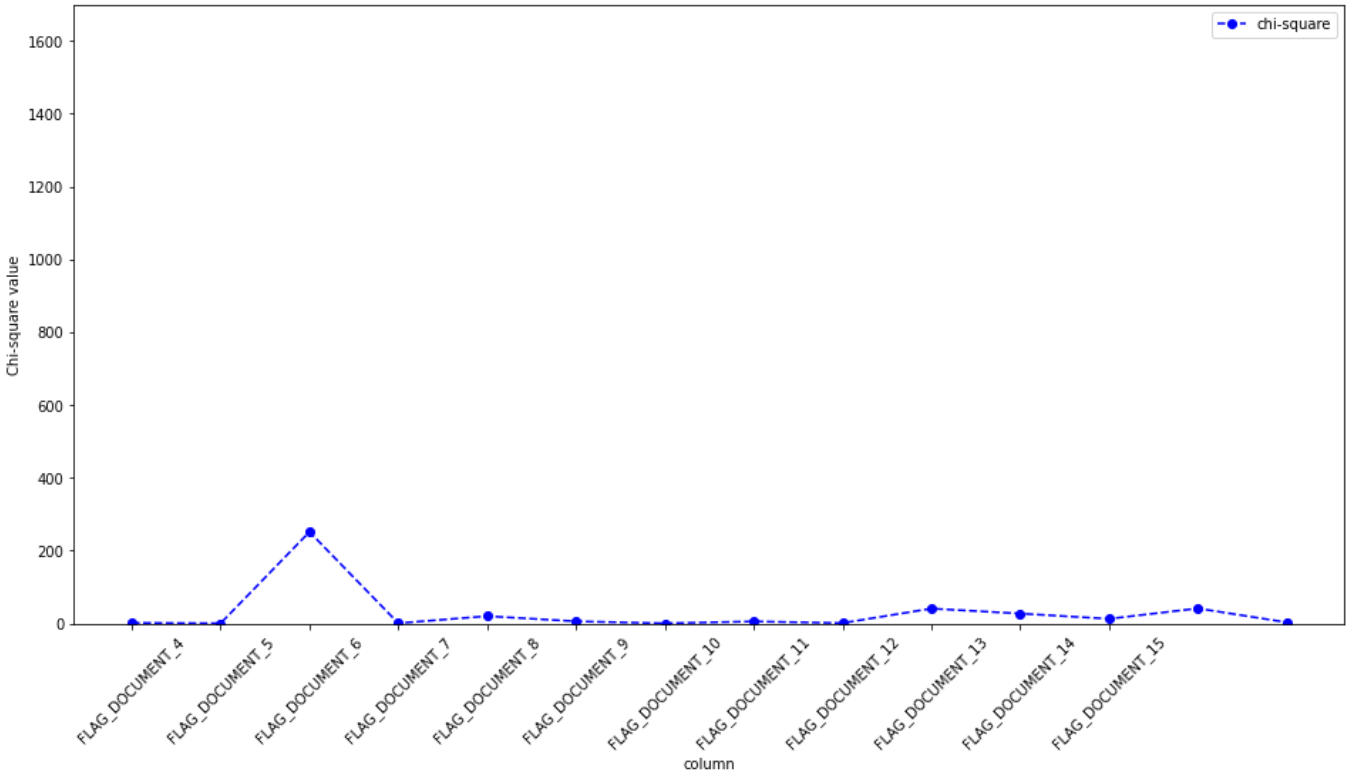
- The age of the applicant seems to be a major factor in deciding default.
- Also, whether the applicant has recently changed their ID or phone number, is also significant in deciding default
- The number of defaults in the applicant's social circle are also important. Bad repayment may be indirectly linked to peer pressure and initiating habits of immediate friends.
- 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. Feature Importance with model fitting

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:

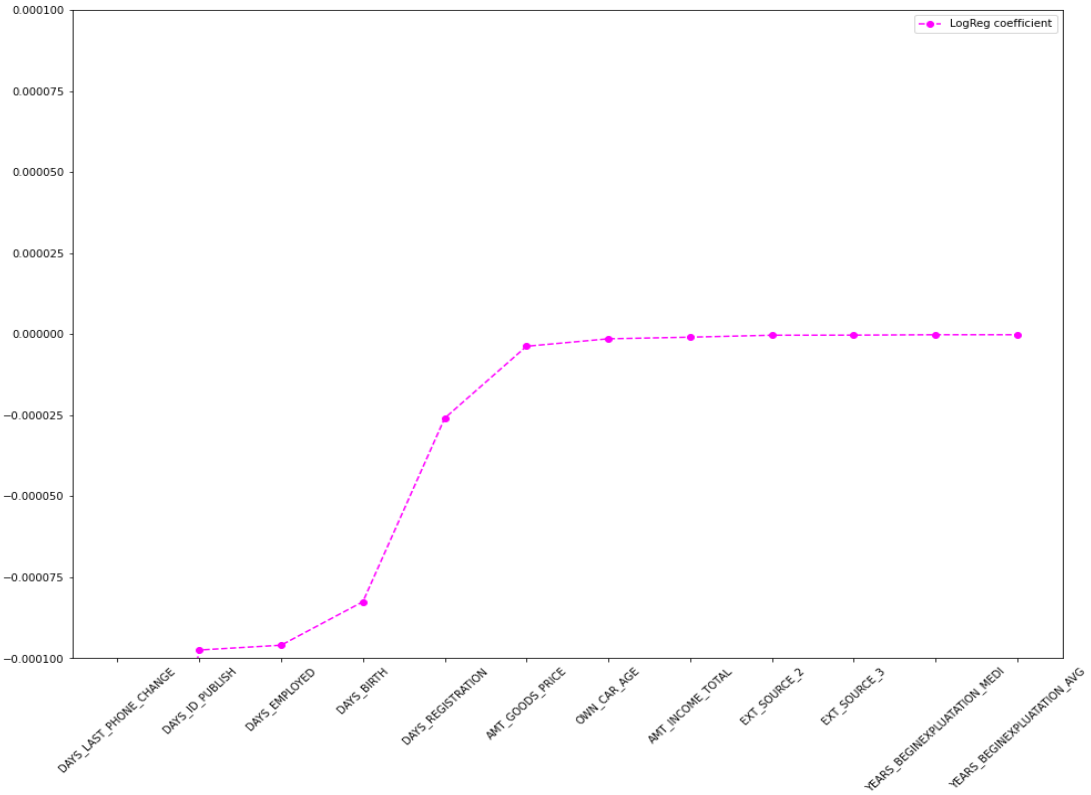


Figure: X-axis → Columns in dataset | Y-axis → Coefficient in Logistic Regression

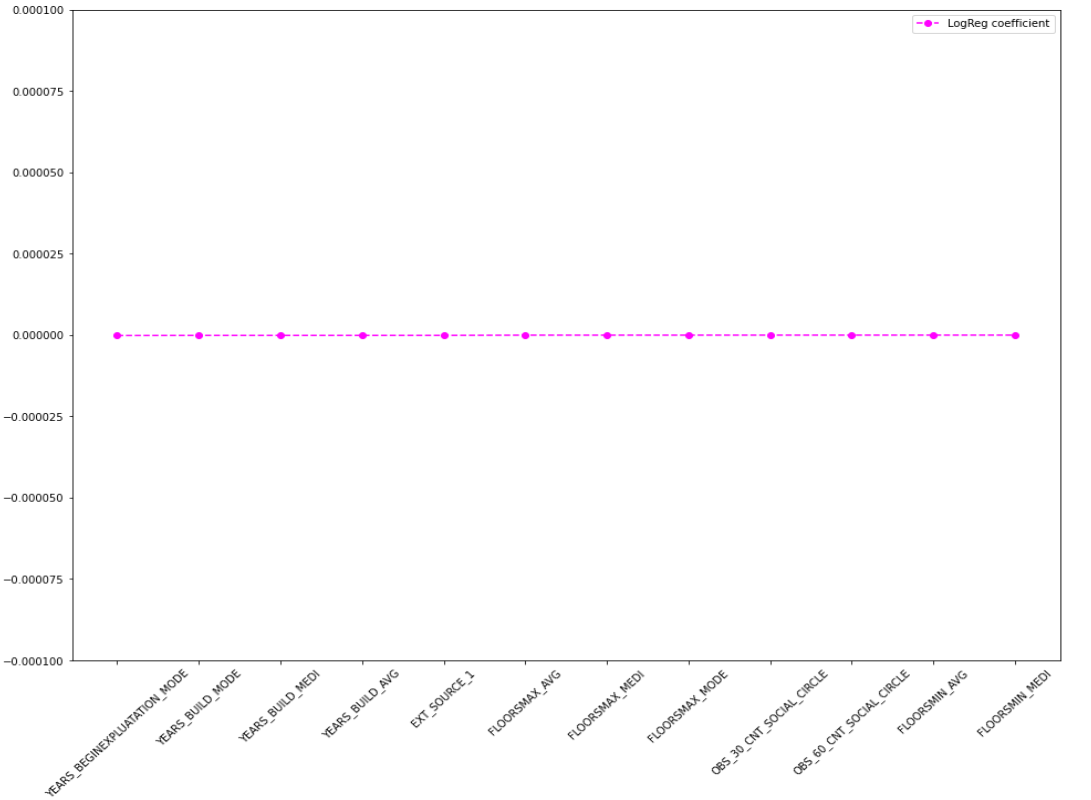


Figure: X-axis → Columns in dataset | Y-axis → Coefficient in Logistic Regression

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.

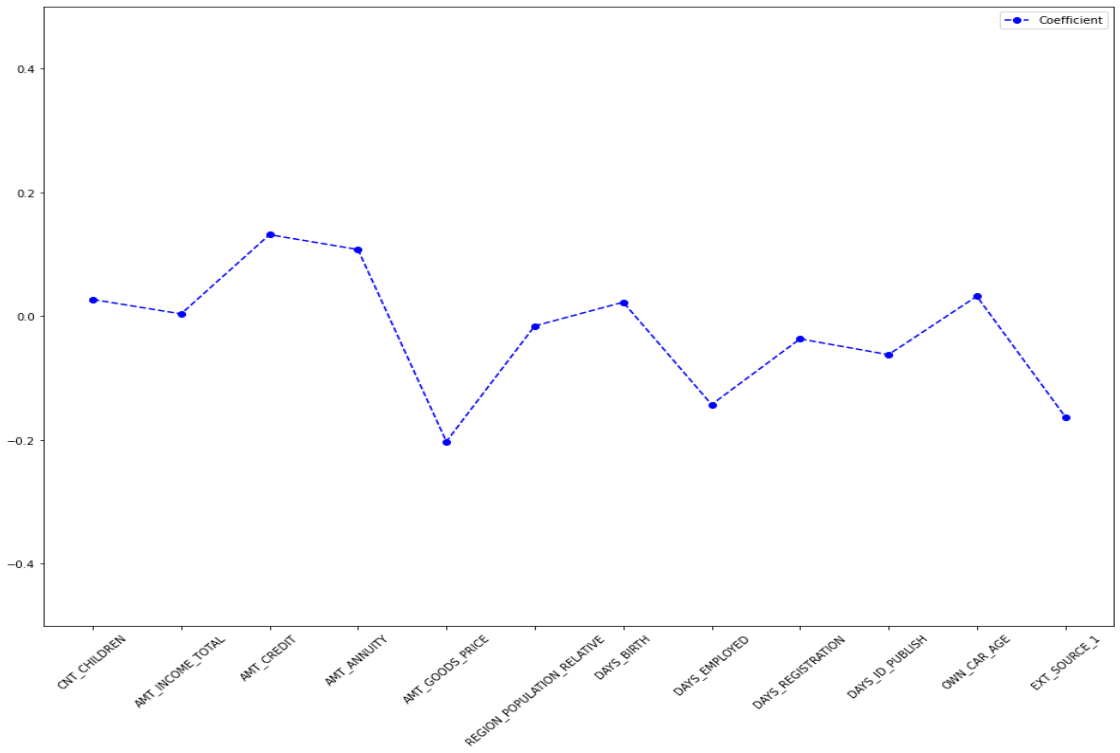


Figure: X-axis → Columns in dataset | Y-axis → Coefficient in Logistic Regression

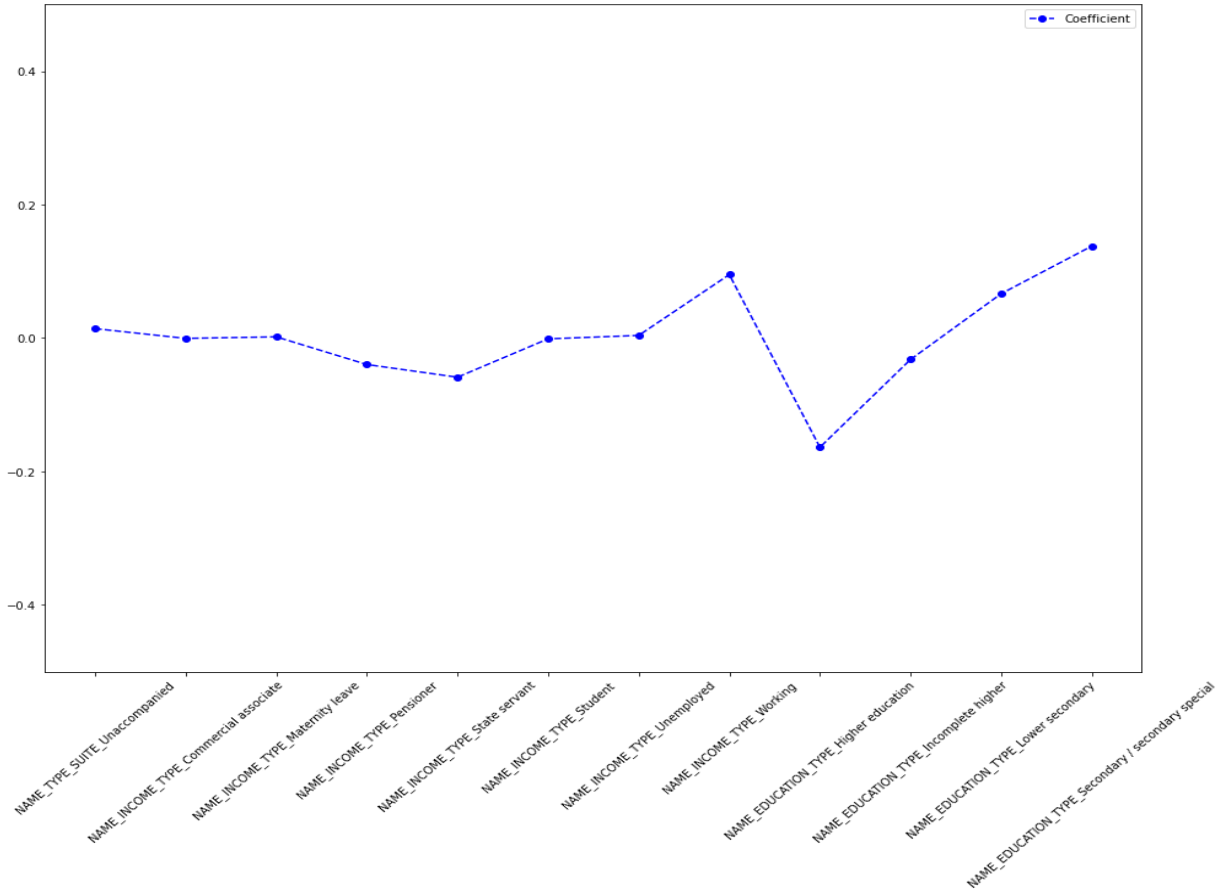


Figure: X-axis → Columns in dataset | Y-axis → Coefficient in Logistic Regression

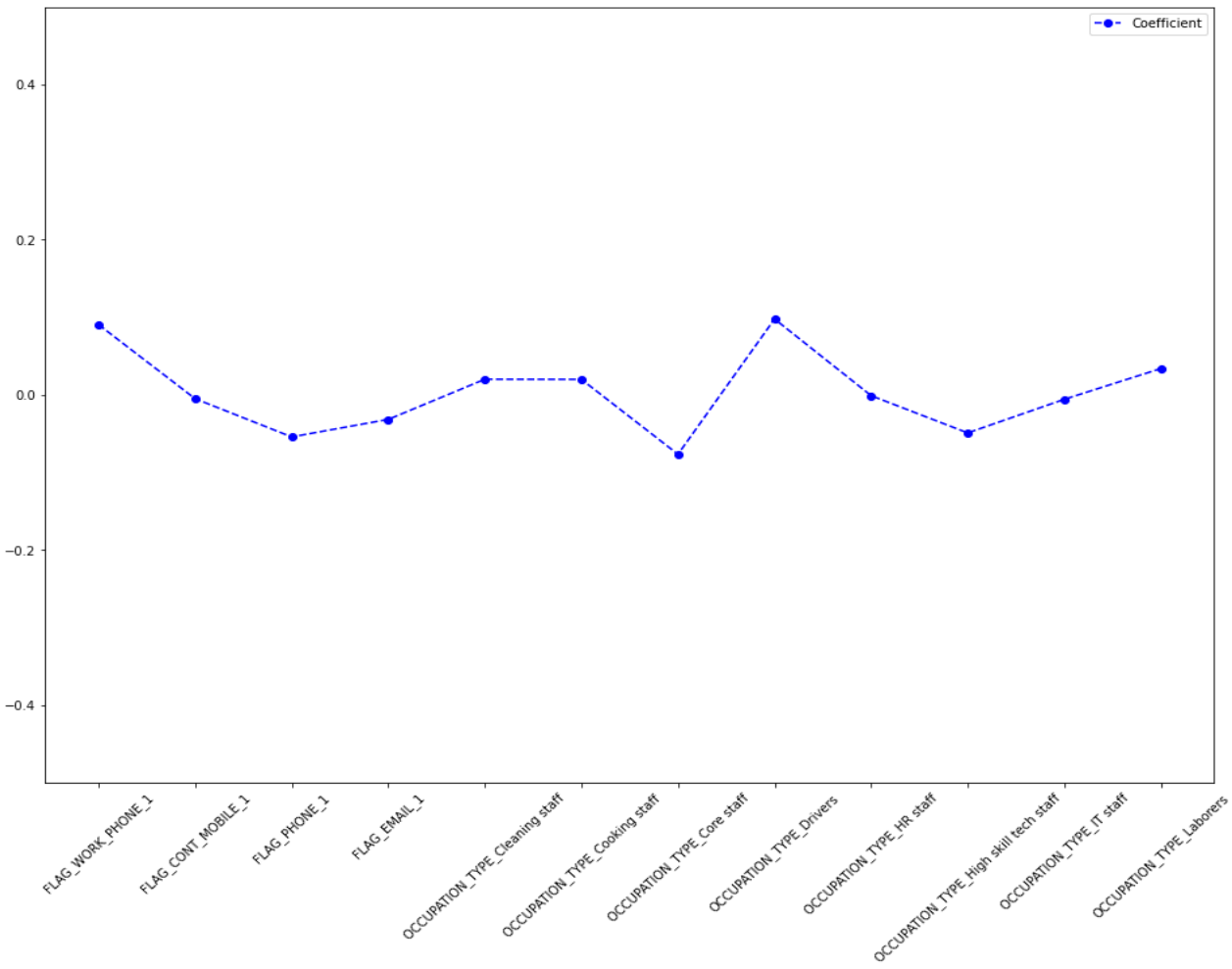


Figure: X-axis → Columns in dataset | Y-axis → Coefficient in Logistic Regression

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.

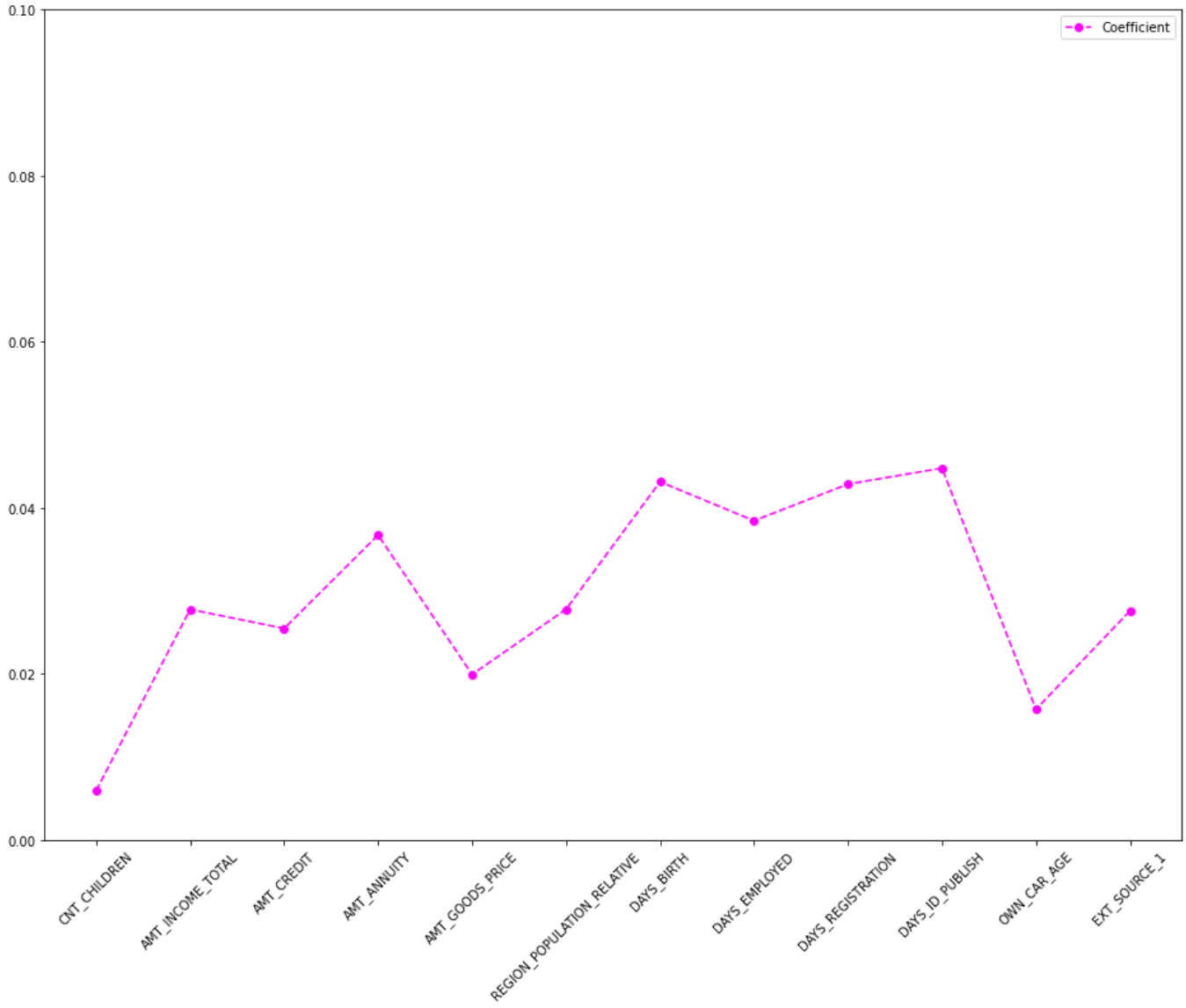


Figure: X-axis → Columns in dataset | Y-axis → Coefficient in Decision Tree model

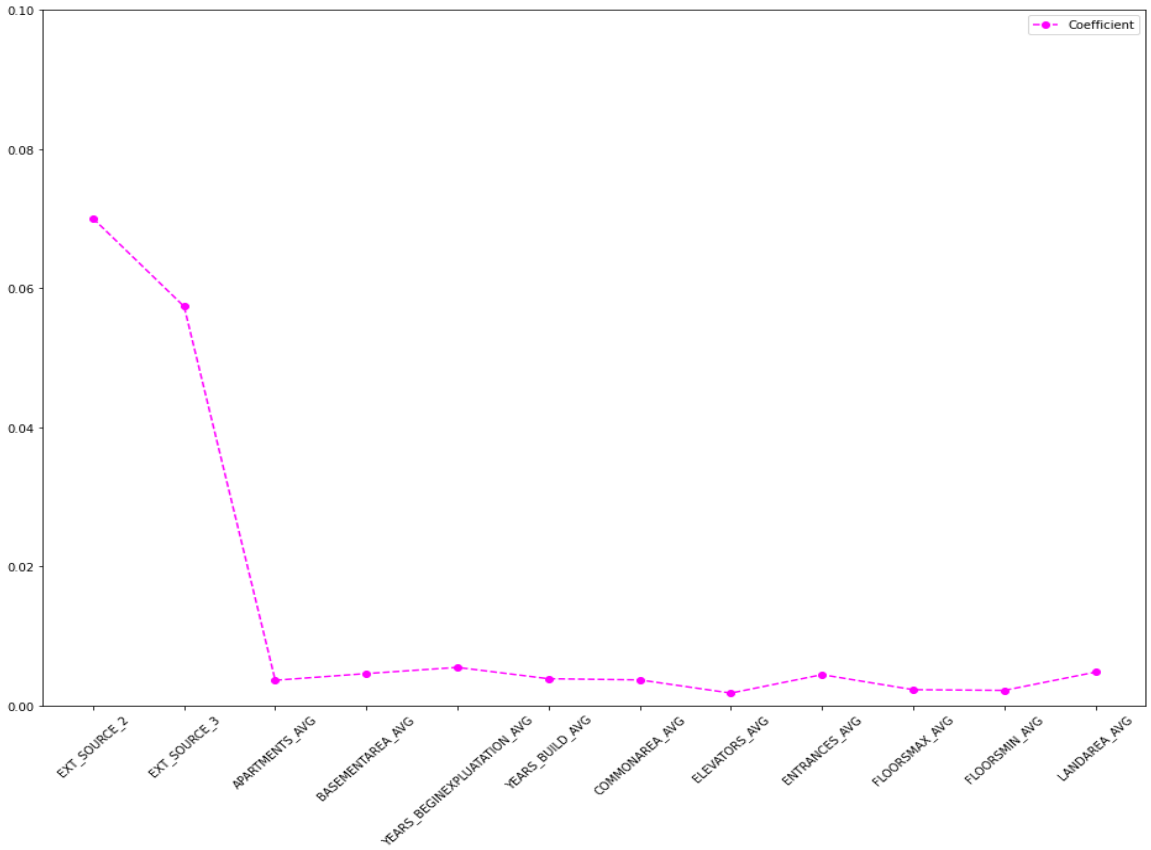


Figure: X-axis → Columns in dataset | Y-axis → Coefficient in Decision Tree model

12. Building classification models.

In this problem, we need to predict the probabilities of the unknown data. This will help us find the roc-auc score.

We will use the logistic regression model.

a. Training a model with all the features:

   We just impute the null values, split the data in the ratio of 0.8 - 0.2 into training and testing sets, and fit a basic logistic regression model on the data.

   The results are as follows:

```
predict_score(lr_model, X_train, y_train)
```

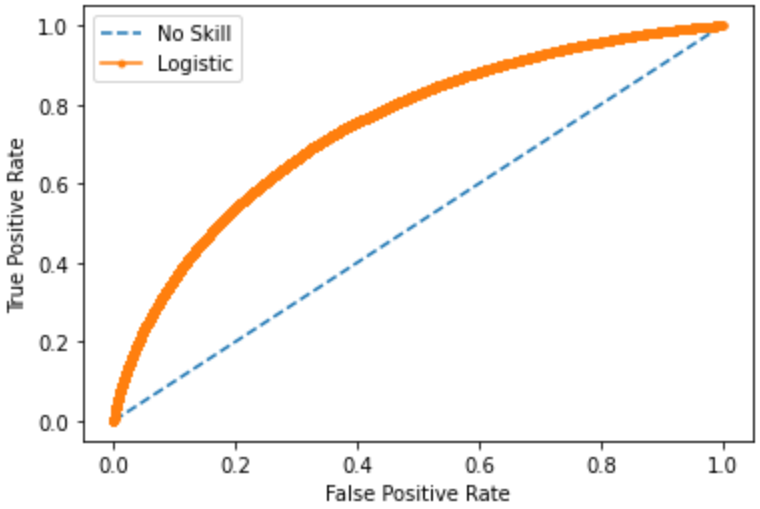
```
Accuracy:  91.92749829273845
recall:  0.5438066465256798
precision:  50.232558139534889
ROC_AUC_SCORE 0.5024824624913444
[[226041    107]
 [ 19752    108]]
```

```
predict_score(lr_model, X_test, y_test)
```

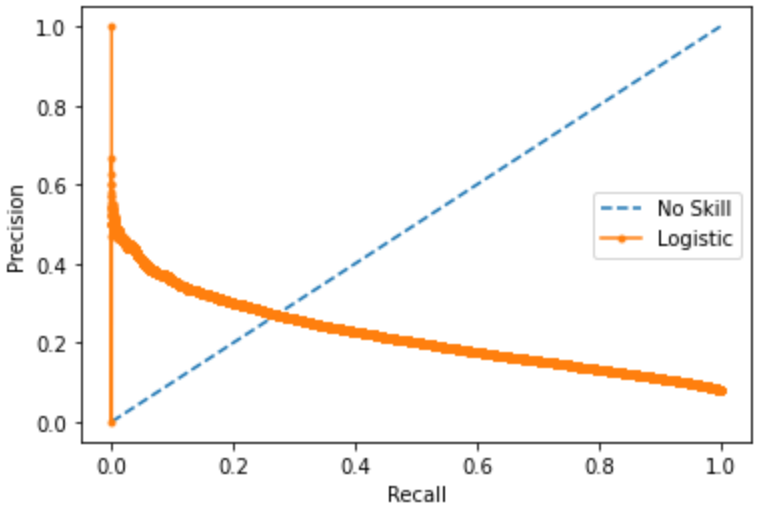
```
Accuracy:  91.9386046209128
recall:  0.6243705941591138
precision:  56.36363636363636
ROC_AUC_SCORE 0.5029096063402109
[[56514    24]
 [ 4934    31]]
```

We are getting a high accuracy of 91%, but very low recall of <1% and precision of 56%.

Let us plot the curves.



The ROC_AUC performance is barely better than random guessing.



As we try to increase precision, recall decreases drastically.

b. Now, let us drop the features that contain more than half null values.

Some numerical features that contain high number of nulls are:

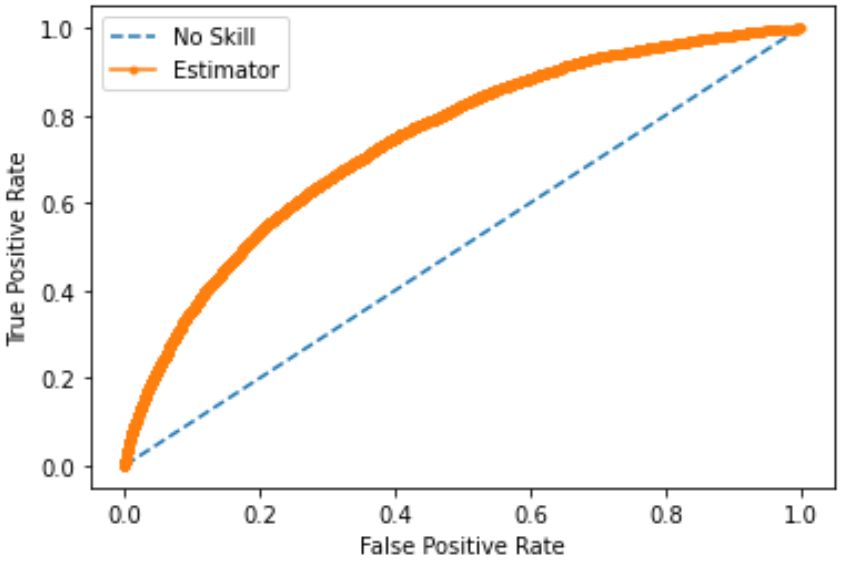
COMMONAREA_AVG	214865
COMMONAREA_MEDI	214865
COMMONAREA_MODE	214865
NONLIVINGAPARTMENTS_MEDI	213514
NONLIVINGAPARTMENTS_MODE	213514
NONLIVINGAPARTMENTS_AVG	213514
LIVINGAPARTMENTS_MEDI	210199
LIVINGAPARTMENTS_MODE	210199
LIVINGAPARTMENTS_AVG	210199
FLOORSMIN_MEDI	208642
FLOORSMIN_MODE	208642
FLOORSMIN_AVG	208642

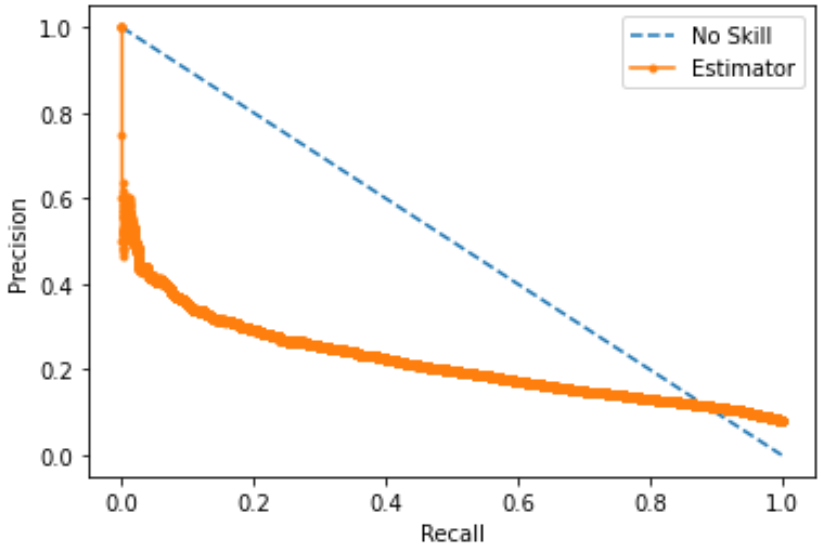
Some categorical features containing a high number of nulls.

FONDKAPREMONT_MODE	210295
WALLSMATERIAL_MODE	156341
HOUSETYPE_MODE	154297

Categorical columns are relatively clean.

On dropping high null features, and fitting again, we get:


No Skill
Estimator
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0


No Skill
Estimator
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0

```
Accuracy:  0.9193196969204254
recall:  0.0050352467270896 27
precision:  0.5319148936170213
ROC_AUC_SCORE 0.5023230605108996
[[226060     88]
 [ 19760    100]]


None
Accuracy:  0.9193047493618197
recall:  0.0040281973816717 02
precision:  0.5263157894736842
ROC_AUC_SCORE 0.5018549137178973
[[56520    18]
 [ 4945    20]]
```

It seems that a large number of bad loans are getting classified as bad. This may be due to the default threshold used by logistics regression to classify into positive and negative classes.

c. Varying the decision threshold

The decision of classifying an instance into positive or negative class is taken by comparing the decision score to a predefined threshold. Upon investigating, we find our model gives the following minimum and maximum scores to instances in the dataset.

```
decision_train.min(), decision_train.max(), decision_test.min(), decision_test.max()
```

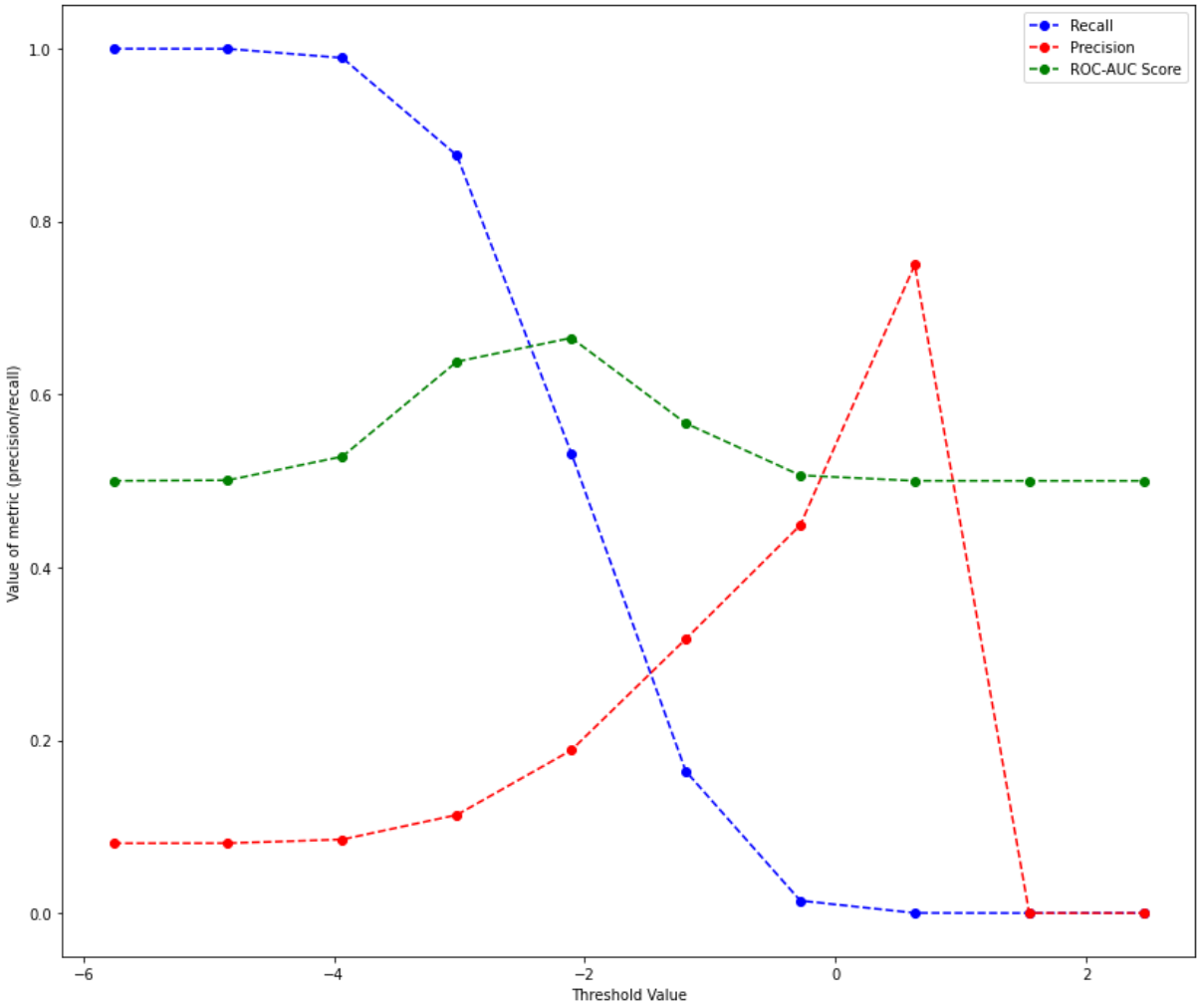
```
(-5.761124556556767,
 2.4609815984859242,
 -14.444117875909015,
 0.6126910881306573)
```

Let us manually set the decision threshold for the probabilities, to be classified into positive and negative classes. We will take 10 decision thresholds for the train data and 15 for the test data, spread evenly (equally spaced) over the range of decision scores, and plot the data.

We will calculate the precision and recall scores and roc-auc scores for all the thresholds, and see the performance.

We get the following results:

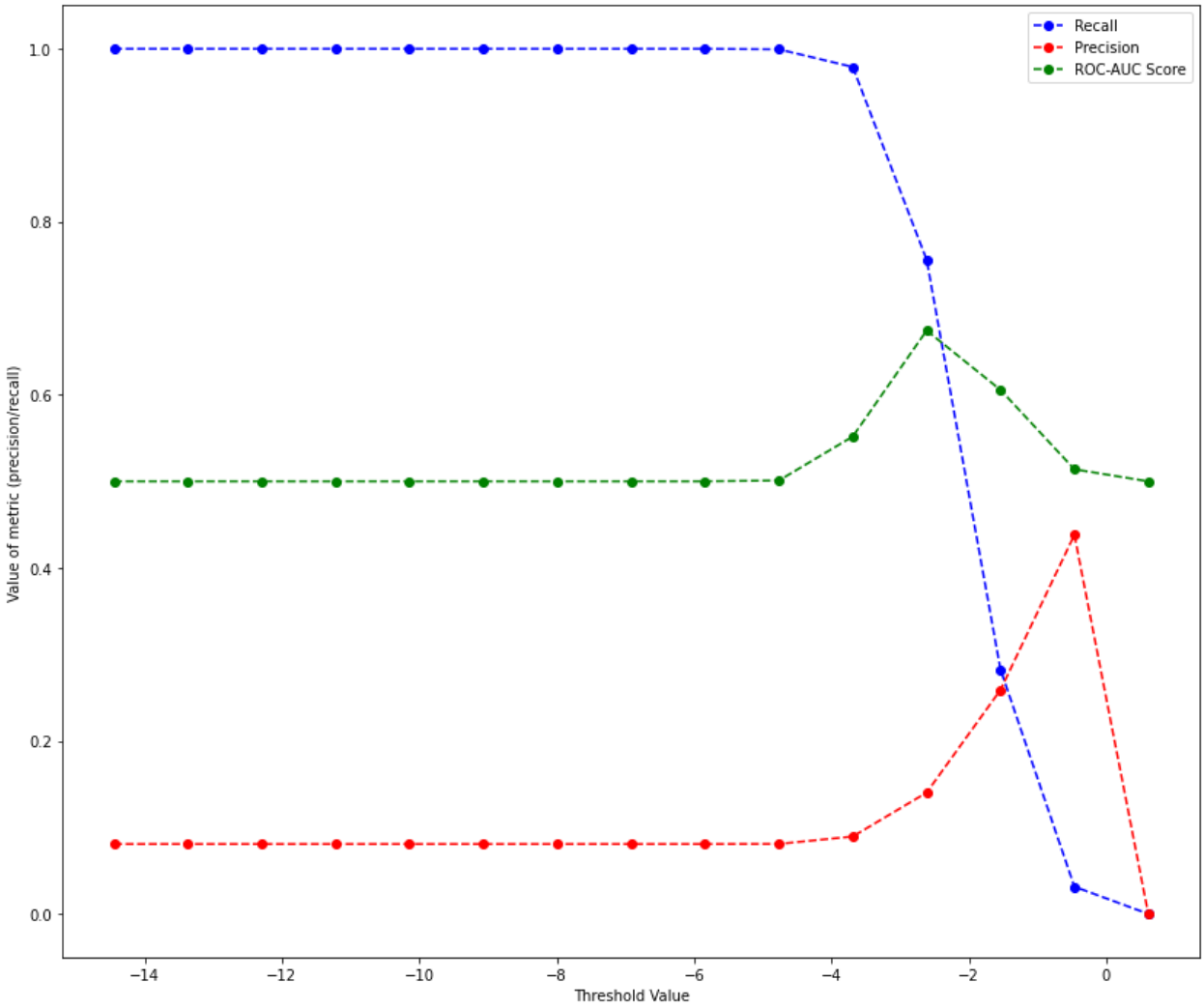
For train data:



The metrics for max roc-auc score are:

```
Threshold: -2.1068551543155705
Accuracy:  0.7772836655718514
recall:  0.5319738167170192
precision:  0.18845879414912595
ROC_AUC_SCORE 0.6654001244824637
Confusion matrix:
 [[180653  45495]
 [  9295  10565]]
```

For TEST data:



Metrics:

```
Threshold: -2.613767975592129
Accuracy:  0.6058728842495488
recall:  0.7560926485397784
precision:  0.14016353657170594
ROC_AUC_SCORE 0.674386838614224
Confusion matrix:
 [[33509 23029]
 [ 1211  3754]]
```

For the TEST data, the maximum ROC_AUC score of 0.67 is obtained at a recall of 74%, and precision of 14%.
In the current problem statement, having a high recall for predicting bad loans is more important.

13. Summary of observations

- In all the significance tests, we find a repeating pattern that the income of the borrower plays an important role in predicting default. Richer people are better repayers.
- 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.
- Our classifier is classifying many bad loans as good loans
- Setting the decision threshold for classification seems to have helped. The maximum ROC-AUC score is obtained at 74% recall.

14. Conclusion:

The menace of non-performing assets is plaguing financial institutions worldwide, but the issue is particularly prevalent in India. Due to the large population and extremely skewed income distribution, weeding out untrustworthy borrowers and lending to credible people is a challenging task. In this project, we have analysed a dataset in detail to find out which factors are critical in a loan turning bad. We have explored customer attributes in detail with statistics and visualisation to make inferences about consumer loan-repayment behaviour. We have also tried to build a basic model to predict the probability of default by a consumer. A model like this, combined with the expertise of bankers and credit appraisers, can help extend credit to deserving people and curb the menace of non-performing assets.