Big Data Project

Submitted To

Eng. Omar Samir

By Team 10

Asmaa Sayed

Aya Samir

Bassant Mohammed

Khalid Ali

Problem Description

Consider you are a Data Analyst with a private bank or a loan distribution firm. Your organization receives many applications in a given day. In order to process the applications, you sometimes miss out on accepting applications from people who are able to pay loans in time and end up sanctioning loans to those who later turn out to be defaulters.
We worked on Current_app data set to analyze loan applications whether or not clients are defaulters. The data set has 307511 rows and 122 columns.

Project Pipeline

Due to the large number of columns and considering the purpose of the project which is classification with Big Data techniques. followed the following flow:

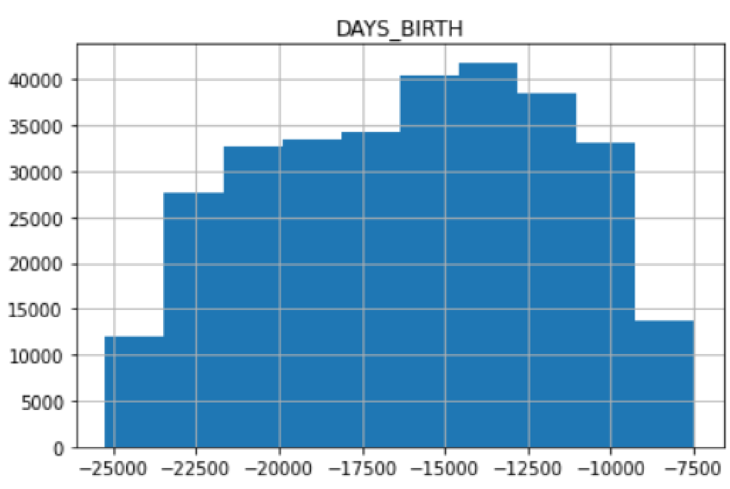
Notice: Due to the large number of features in this dataset we mentioned the most important features for each step so that the report won't be too large. For further details you could visit notebook attached.

1. Data Preprocessing

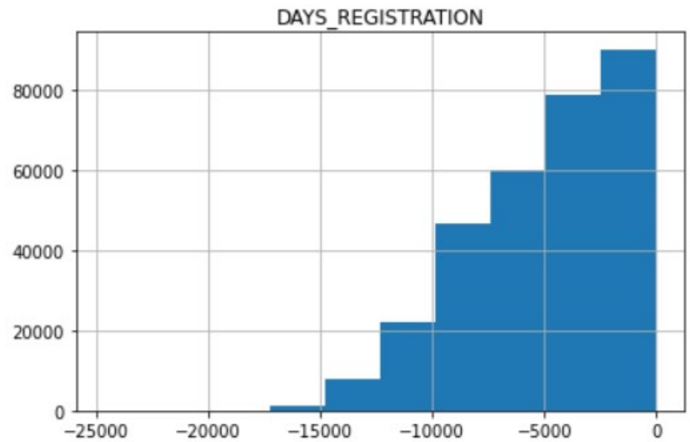
a. Check invalid values

Check if any features have invalid data and replace it with valid data if possible. We Noticed that the following features have invalid data.

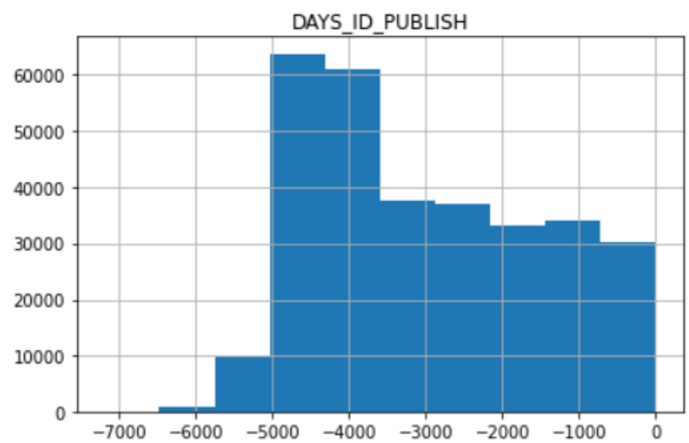
i. DAYS_BIRTH have values from -25000 to -7500



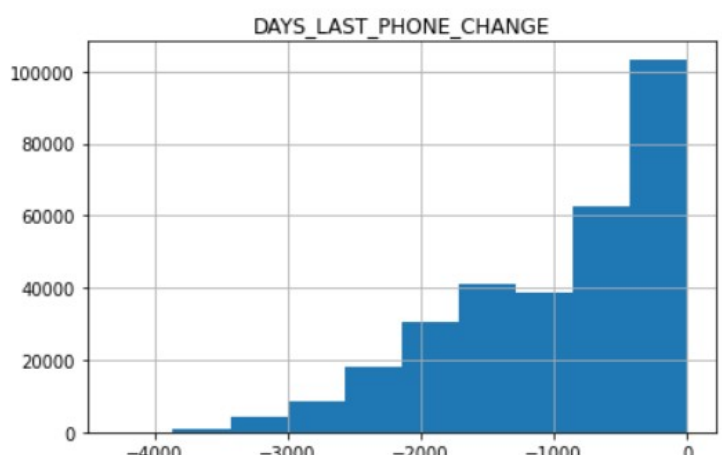
ii. DAYS_REGESTRATION have values from -25000 to 0



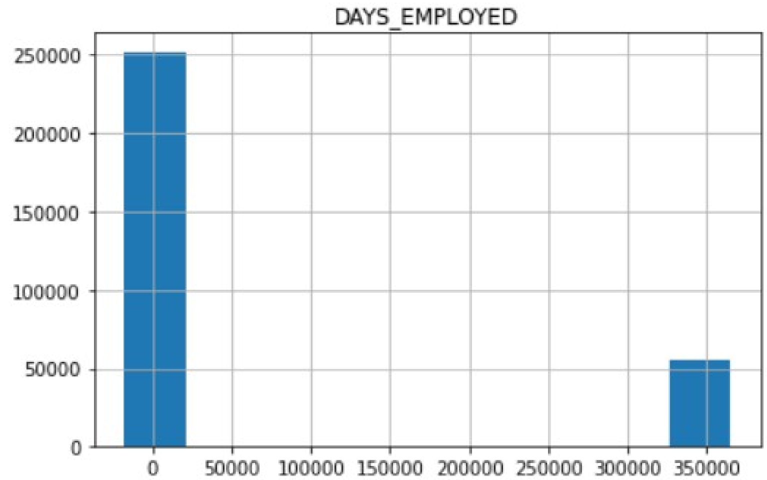
iii. DAYS_ID_PUBLISH have values from -7000 to 0



iv. DAYS_LAST_PHONE_CHANGE has values from -4000 to 0



v. DAYS_EMPLOYED have values less than 0

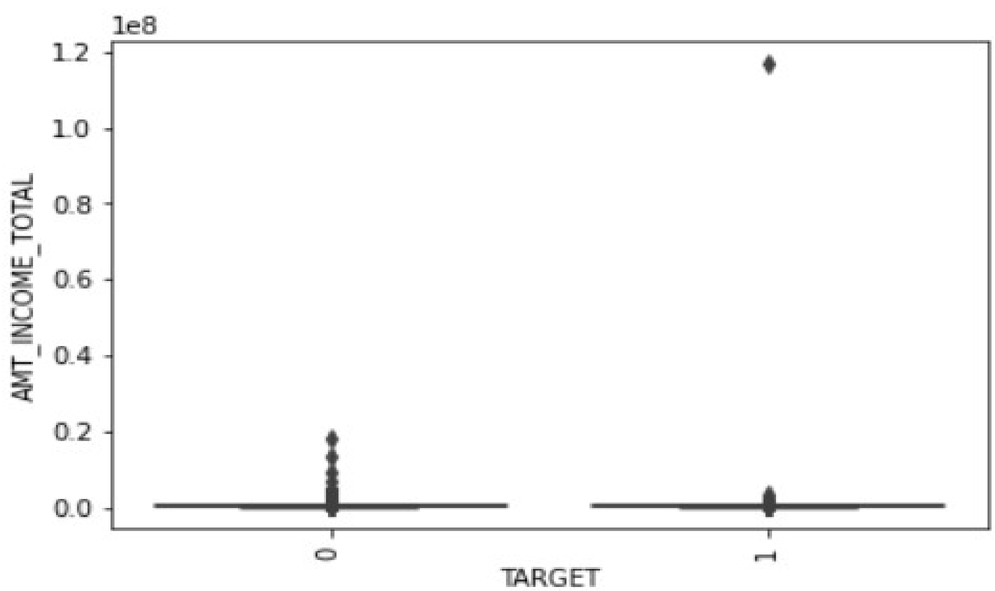


b. *Outliers Detection*

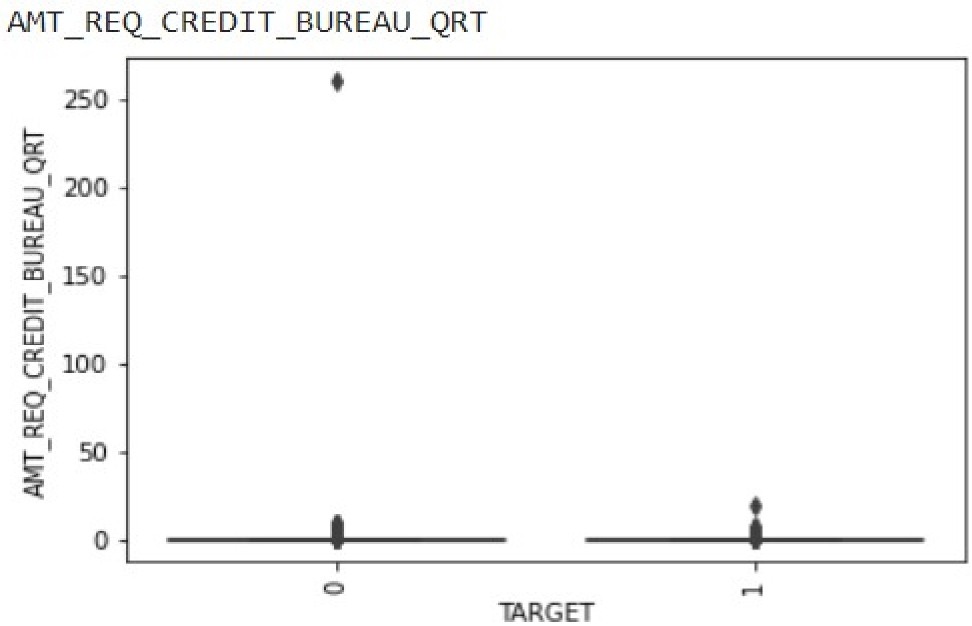
Draw boxplot for continuous features and check/remove outliers.
From the plotted figures we notice the following features have outliers

i. AMT_INCOME_TOTAL has outliers, so we removed rows with values larger than 0.2e8

AMT_INCOME_TOTAL

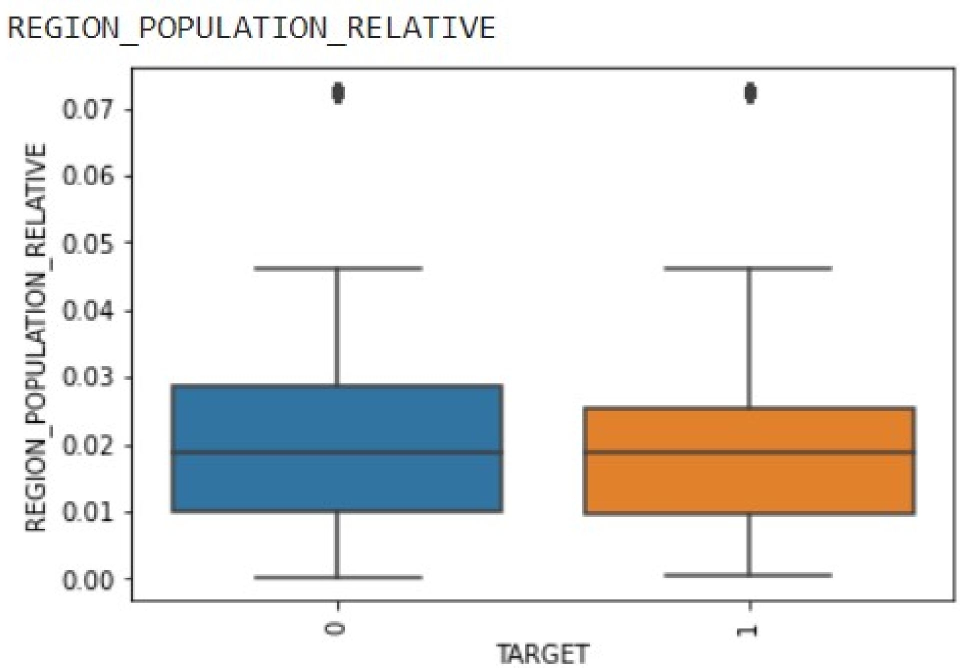


ii. AMT_REQ_CREDIT_BUREAU_QRT
Has outliers, so we removed rows with values larger than 200



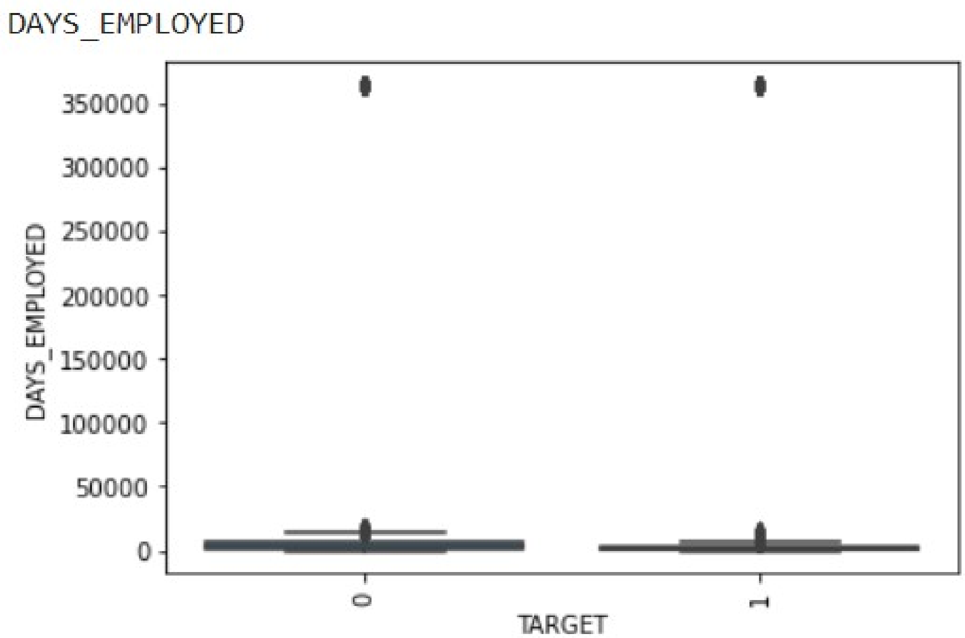
iii. REGION_POPULATION_RELATIVE

Has outliers, so we removed rows with values larger than 0.05



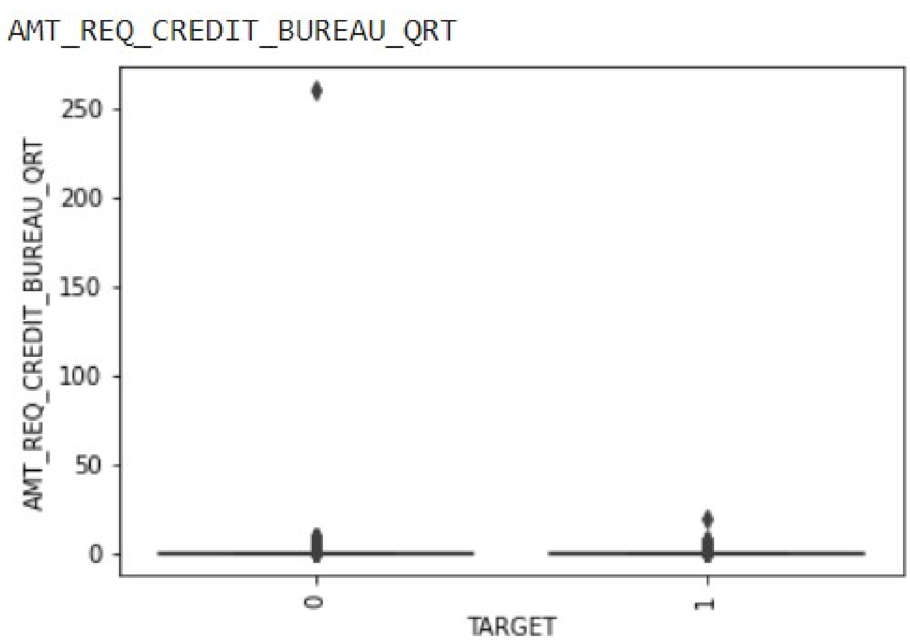
iv. DAYS_EMPOLYED

Has outliers, so we removed rows with values larger than 50000



v. AMT_REQ_CREDIT_BUREAU_QRT

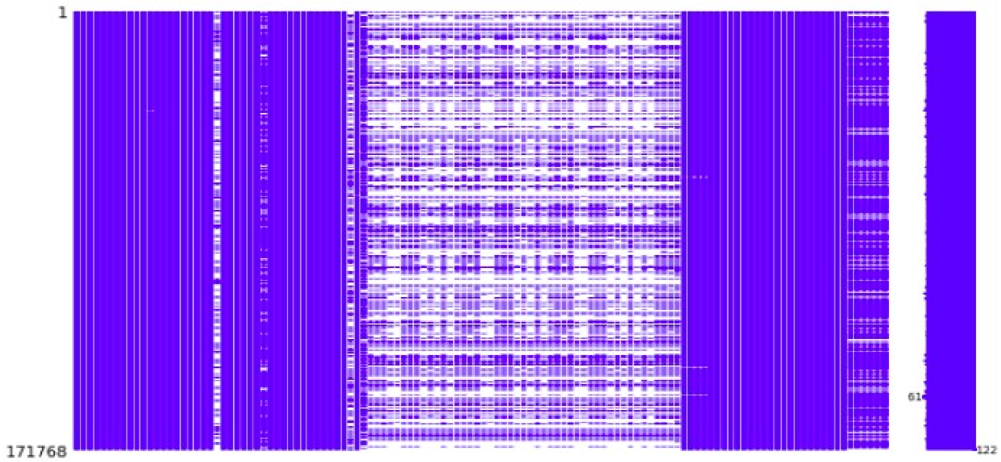
Has outliers, so we removed rows with values larger than 40



After this Step the dataset has shape of (245384, 122)

c. *Check & visualize Nan values*

Visualize Nan values using Missingno library



d. *Fill Nan Values*

Recommended 1: Drop columns with Nan values larger than 50%

Columns decreased from 122 to 73 columns.

Recommended 2: Fill columns with Nan values less than 13%

Drop remaining rows with Nan values

The output dataset after this step has the shape of (83997, 77)

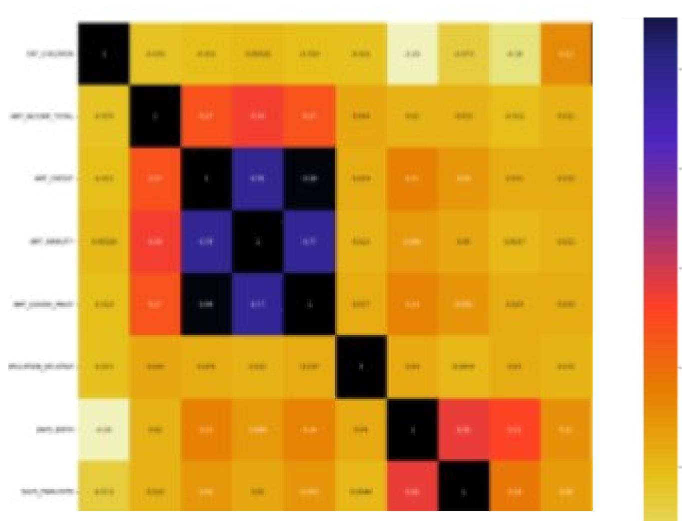
after removing SK_ID_CURR column which is not useful.

2. Bivariate Analysis

a. *Continuous Vs Continuous*

Compute correlation between continuous features to remove dependent features.

Slice of heatmap



We found the following columns highly correlated to each other, over 0.85;

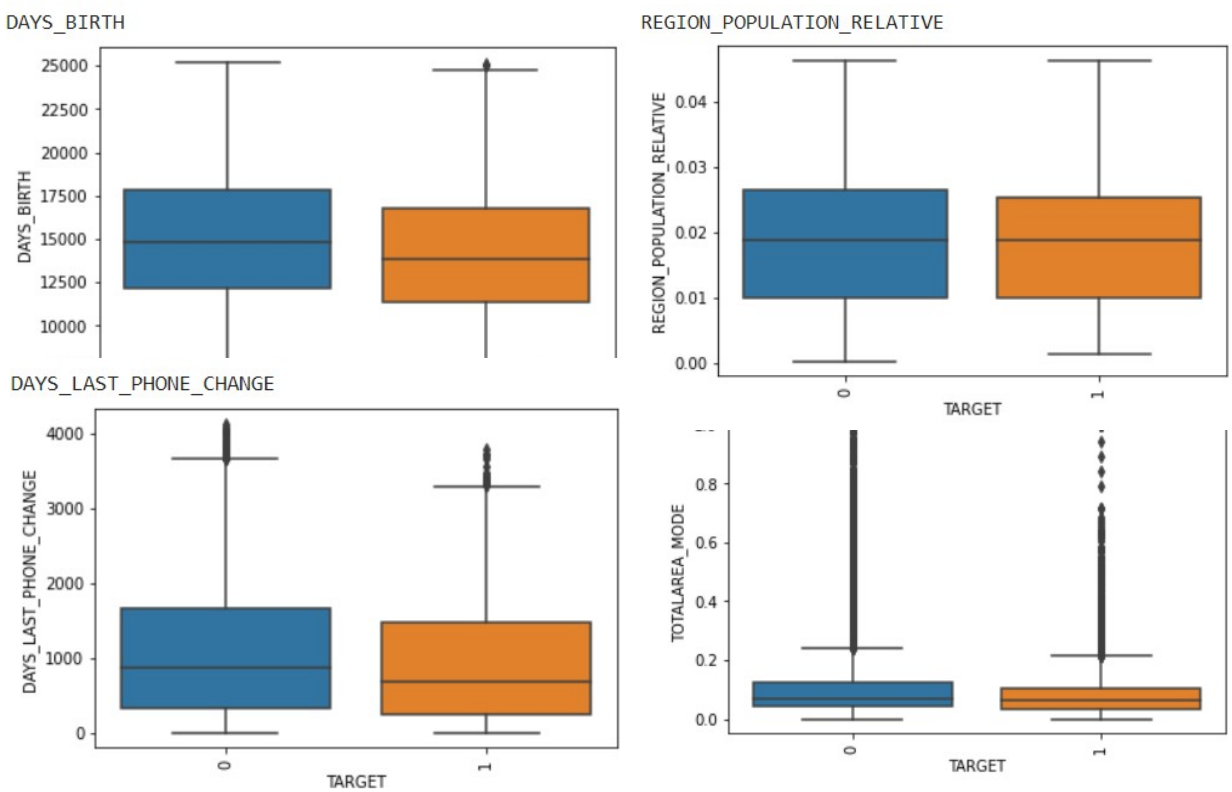
CNT_CHILDREN	VS	CNT_FAM_MEMBERS
AMT_CREDIT	VS	AMT_GOODS_PRICE
REGION_RATING_CLIENT	VS	REGION_RATING_CLIENT_W_CITY
YEARS_BEGINEXPLUATATION_AVG	VS	YEARS_BEGINEXPLUATATION_MODE
YEARS_BEGINEXPLUATATION_AVG	VS	YEARS_BEGINEXPLUATATION_MEDI
YEARS_BEGINEXPLUATATION_MODE	VS	YEARS_BEGINEXPLUATATION_MEDI
OBS_30_CNT_SOCIAL_CIRCLE	VS	OBS_60_CNT_SOCIAL_CIRCLE
DEF_30_CNT_SOCIAL_CIRCLE	VS	DEF_60_CNT_SOCIAL_CIRCLE

We Removed Unique columns from left side.

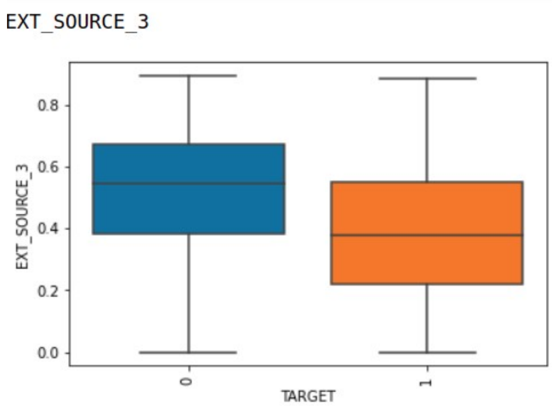
b. *Continuous Vs Output (Categorical)*

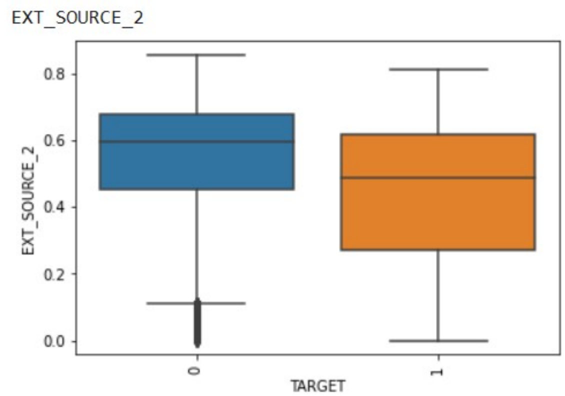
Draw box plot for continuous features Vs output target to and check if distribution changes for each class. Remove features that have same distribution for both classes

*Examples for Features that are **NOT** correlated to response feature*



*Examples for Features that **ARE** correlated to response feature*





After This step the dataset has a shape of (83997, 58)

c. *Binary Categorical Vs binary categorical*

To find correlation between binary categorical features we used PEARSON'R method. Remove dependent features with correlation r larger than 0.85 and p-value less than 0.05.

We found the following columns correlated

```
FLAG_DOCUMENT_7 Vs FLAG_DOCUMENT_13
FLAG_DOCUMENT_2 Vs FLAG_DOCUMENT_13
FLAG_DOCUMENT_6 Vs FLAG_DOCUMENT_13
```

We removed whole three columns on the left side.

After This step the dataset has a shape of (83997, 55)

d. *Binary Features Vs Output*

We used same method (PEARSON'R) to remove features with correlation r less than 0.04 with output or p-value less than 0.05.

The only remaining binary feature after this step was CODE GENDER

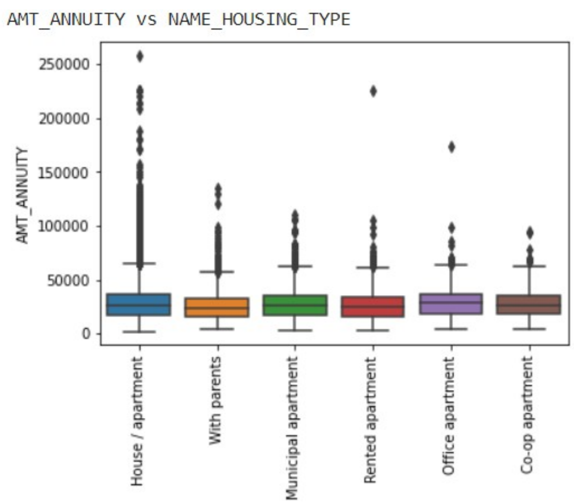
After This step the dataset has a shape of (83997, 22)

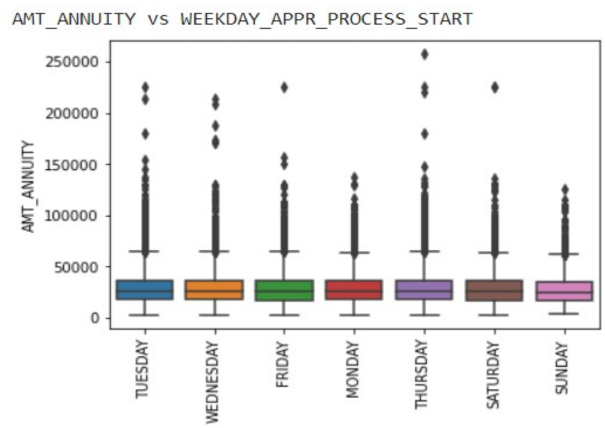
e. *Multiple Categorical Vs Continuous*

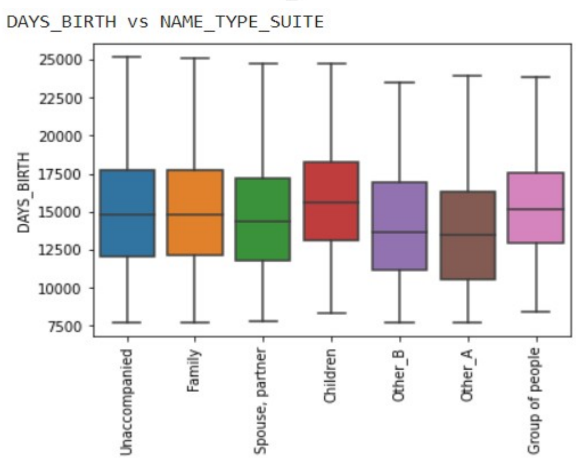
Draw box plot for continuous features Vs output target to and check if distribution change for each class. Remove features that have different distribution for all classes. Which indicates that each class has a different range of values (features are dependent).

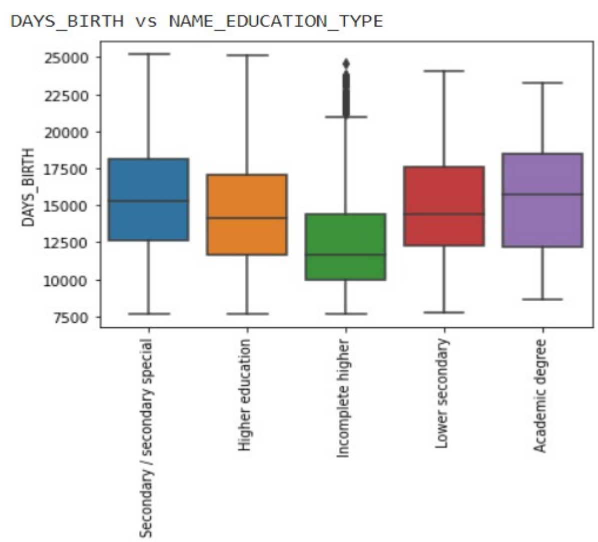
From The output figures we didn't find any features that are highly correlated where there were always large intersection between class ranges.

Examples for output figures









f. Multiple Categorical Vs Categorical

For this test we used Chi-Square test to calculate the correlation between multi-categorical features. Remove dependent features with correlation larger than 0.85 and features that have correlation with response feature of value less than 0.05.

	NAME_TYPE_SUITE	NAME_INCOME_TYPE	NAME_EDUCATION_TYPE	NAME_FAMILY_STATUS
NAME_TYPE_SUITE	1.000000	0.000000	0.019312	0.058615
NAME_INCOME_TYPE	0.000000	1.000000	0.072726	0.019701
NAME_EDUCATION_TYPE	0.019312	0.072726	1.000000	0.042783
NAME_FAMILY_STATUS	0.058615	0.019701	0.042783	1.000000

Cross section from output matrix

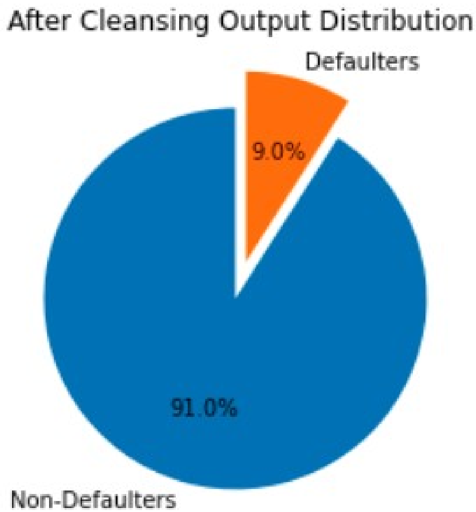
We found that no features are highly correlated to each other. And we removed features that have correlation with response features less than 0.05. The following features were removed.

- NAME_TYPE_SUITE
- NAME_INCOME_TYPE
- NAME_FAMILY_STATUS
- NAME_HOUSING_TYPE
- WEEKDAY_APPR_PROCESS_START'

After This step the dataset has a shape of (83997, 17)

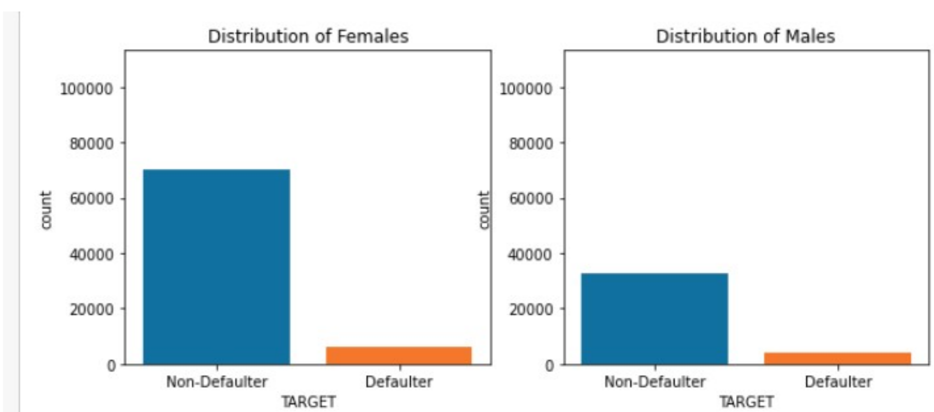
3. Univariate Analysis

a. Distribution of defaulter and non-defaulters



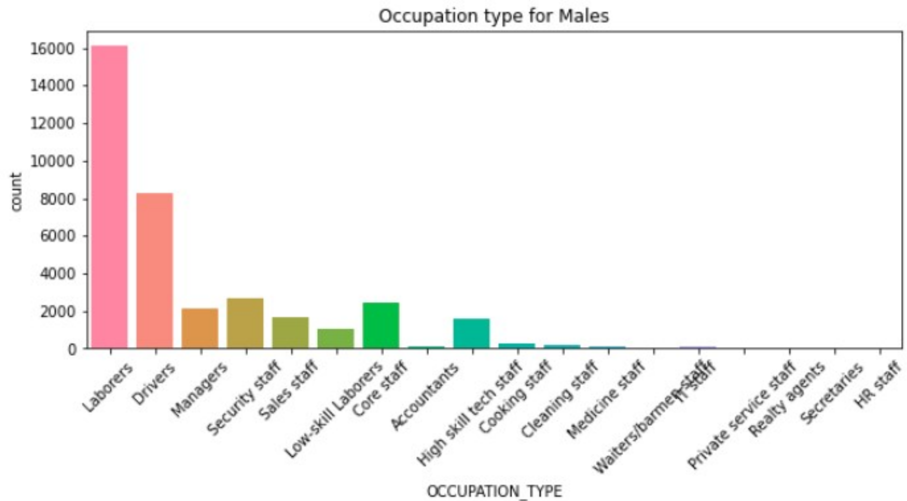
Insight: Our dataset has 91% non-Defaulters and 9% Defaulters so our dataset is imbalanced

b. Distribution of Applicants gender



Insight: the number of non-Defaulters for men is less than the number of defaulters for women to get better insight we better calculate percentage.
Insight: the percentage Defaulters of males is larger than females

c. Occupation Type Analysis



Insight: majority of male clients are laborers followed by drivers.

Insight: majority of female clients are Sales staff followed by Laborers, followed by Core staff.

4. Model/Classifier training

We used Logistic regression from Statsmodels library to find out the significance of features in prediction. We removed features that have p-value less than 0.05 which indicates that features are not significant for prediction.

We conclude that the most effective features for our prediction are the following;

- AMT_ANNUITY
- EXT_SOURCE_3
- NAME_EDUCATION_TYPE

Results and Evaluation

We trained the models to achieve best f1-score for model class 1 on validation set.

- ➢ The Logistic Regression model achieved on Over Sampled data the following scores

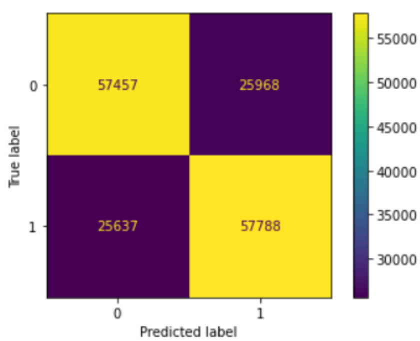
Train: accuracy = 69%, f1-score for class 1 = 69%, f1-score for class 0 = 69%

Valid: accuracy = 69%, f1-score for class 1 = 28%, f1-score for class 0 = 80%

Test: accuracy = 69%, f1-score for class 1 = 27%, f1-score for class 0 = 80%

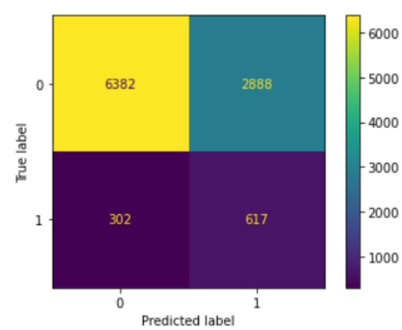
Train Set

	precision	recall	f1-score	support
0	0.69	0.69	0.69	83425
1	0.69	0.69	0.69	83425
accuracy			0.69	166850
macro avg	0.69	0.69	0.69	166850
weighted avg	0.69	0.69	0.69	166850



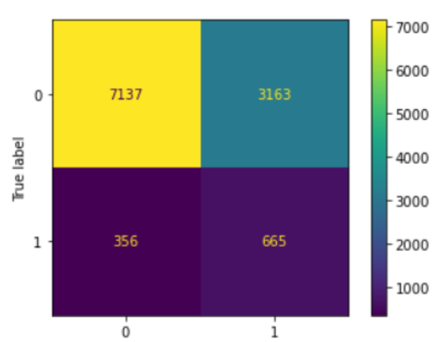
Valid Set

	precision	recall	f1-score	support
0	0.95	0.69	0.80	9270
1	0.18	0.67	0.28	919
accuracy			0.69	10189
macro avg	0.57	0.68	0.54	10189
weighted avg	0.88	0.69	0.75	10189



Test Set

	precision	recall	f1-score	support
0	0.95	0.69	0.80	10300
1	0.17	0.65	0.27	1021
accuracy			0.69	11321
macro avg	0.56	0.67	0.54	11321
weighted avg	0.88	0.69	0.75	11321



➢ Stochastic Logistic Regression on Over Sampled data using **MapReduce**

Train: accuracy = 61%, f1-score for class 1 = 66%, f1-score for class 0 = 54%

Valid: accuracy = 48%, f1-score for class 1 = 20%, f1-score for class 0 = 61%

Test: accuracy = 48%, f1-score for class 1 = 20%, f1-score for class 0 = 81%

Other Trials

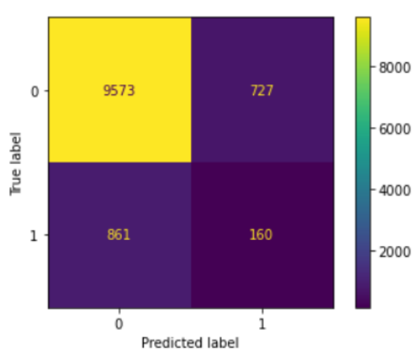
We also tried the following classifiers on imbalanced data, under sampled data, over sampled data: we will show the confusion matrix

1. Naive Bayes

Imbalanced Data

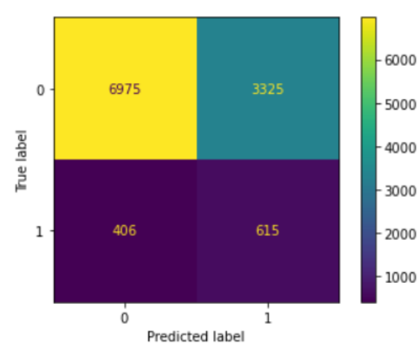
Test Accuracy:0.8597297058563731

	precision	recall	f1-score	support
0	0.92	0.93	0.92	10300
1	0.18	0.16	0.17	1021
accuracy			0.86	11321
macro avg	0.55	0.54	0.55	11321
weighted avg	0.85	0.86	0.86	11321



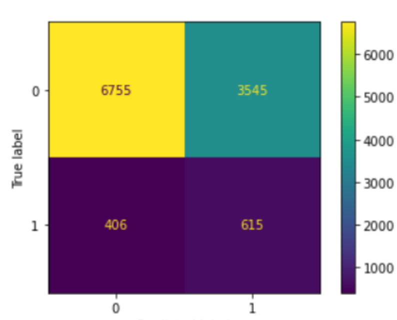
UnderSampled

	precision	recall	f1-score	support
0	0.94	0.68	0.79	10300
1	0.16	0.60	0.25	1021
accuracy			0.67	11321
macro avg	0.55	0.64	0.52	11321
weighted avg	0.87	0.67	0.74	11321



OverSampled

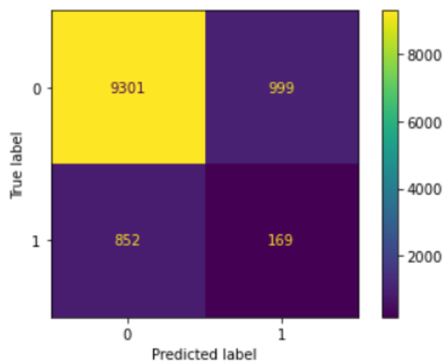
	precision	recall	f1-score	support
0	0.94	0.66	0.77	10300
1	0.15	0.60	0.24	1021
accuracy			0.65	11321
macro avg	0.55	0.63	0.51	11321
weighted avg	0.87	0.65	0.73	11321



2. Decision Tree

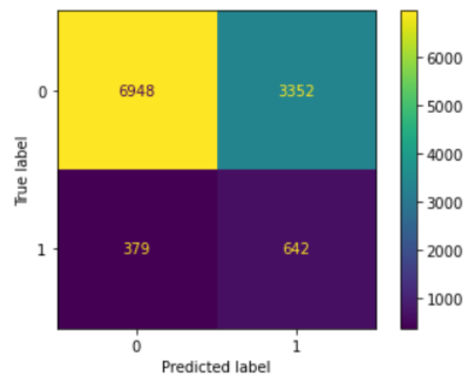
Imbalanced

	precision	recall	f1-score	support
0	0.92	0.90	0.91	10300
1	0.14	0.17	0.15	1021
accuracy			0.84	11321
macro avg	0.53	0.53	0.53	11321
weighted avg	0.85	0.84	0.84	11321



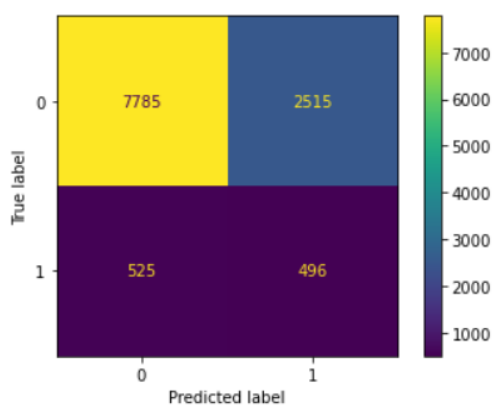
UnderSampled

	precision	recall	f1-score	support
0	0.95	0.67	0.79	10300
1	0.16	0.63	0.26	1021
accuracy			0.67	11321
macro avg	0.55	0.65	0.52	11321
weighted avg	0.88	0.67	0.74	11321



OverSampled

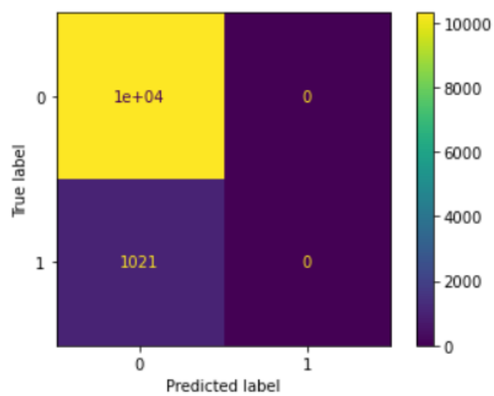
	precision	recall	f1-score	support
0	0.94	0.76	0.84	10300
1	0.16	0.49	0.25	1021
accuracy			0.73	11321
macro avg	0.55	0.62	0.54	11321
weighted avg	0.87	0.73	0.78	11321



3. KNN

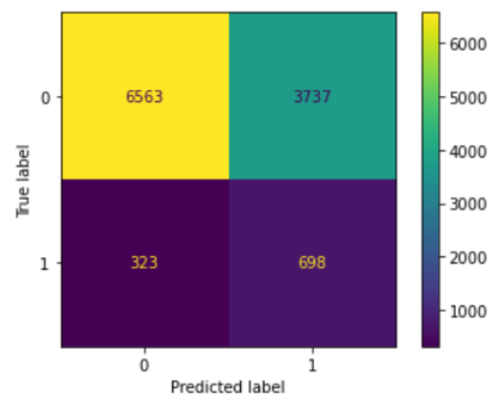
Imbalanced Data

	precision	recall	f1-score	support
0	0.91	1.00	0.95	10300
1	0.00	0.00	0.00	1021
accuracy			0.91	11321
macro avg	0.45	0.50	0.48	11321
weighted avg	0.83	0.91	0.87	11321



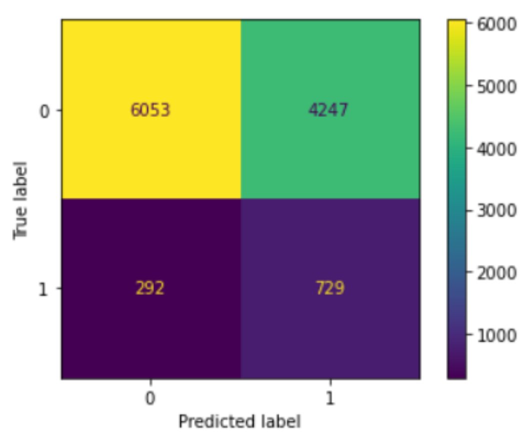
UnderSampled

	precision	recall	f1-score	support
0	0.95	0.64	0.76	10300
1	0.16	0.68	0.26	1021
accuracy			0.64	11321
macro avg	0.56	0.66	0.51	11321
weighted avg	0.88	0.64	0.72	11321



OverSampled

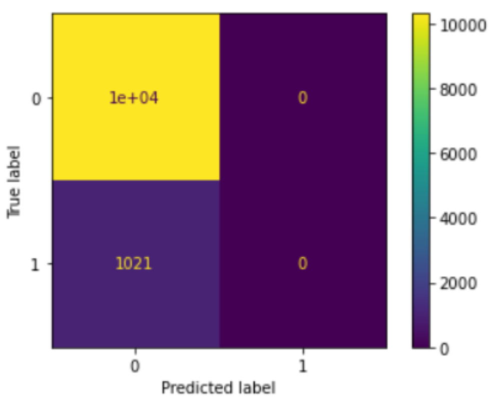
	precision	recall	f1-score	support
0	0.95	0.59	0.73	10300
1	0.15	0.71	0.24	1021
accuracy			0.60	11321
macro avg	0.55	0.65	0.49	11321
weighted avg	0.88	0.60	0.68	11321



4. Support Vector Machine

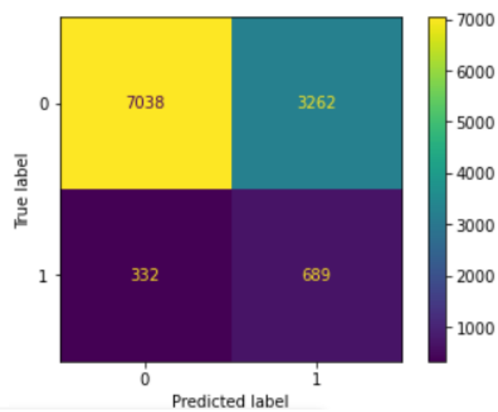
Imbalanced Data

	precision	recall	f1-score	support
0	0.91	1.00	0.95	10300
1	0.00	0.00	0.00	1021
accuracy			0.91	11321
macro avg	0.45	0.50	0.48	11321
weighted avg	0.83	0.91	0.87	11321



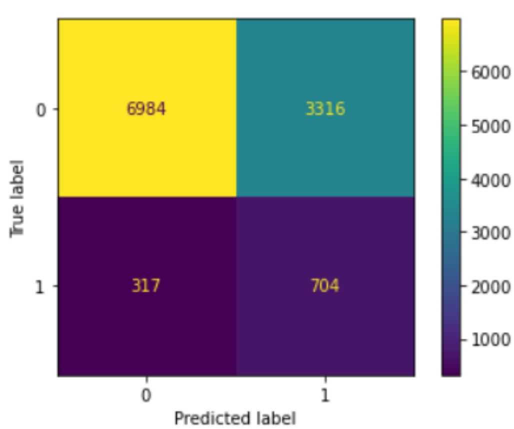
UnderSampled

	precision	recall	f1-score	support
0	0.95	0.68	0.80	10300
1	0.17	0.67	0.28	1021
accuracy			0.68	11321
macro avg	0.56	0.68	0.54	11321
weighted avg	0.88	0.68	0.75	11321



OverSampled

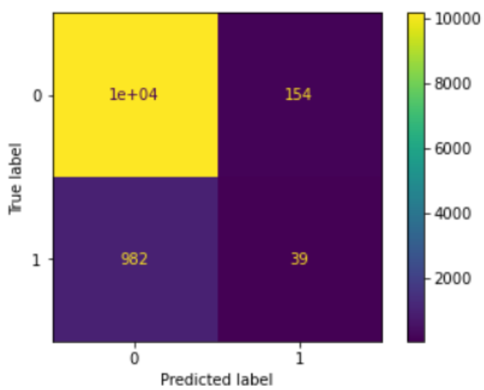
	precision	recall	f1-score	support
0	0.96	0.68	0.79	10300
1	0.18	0.69	0.28	1021
accuracy			0.68	11321
macro avg	0.57	0.68	0.54	11321
weighted avg	0.89	0.68	0.75	11321



5. Random Forest

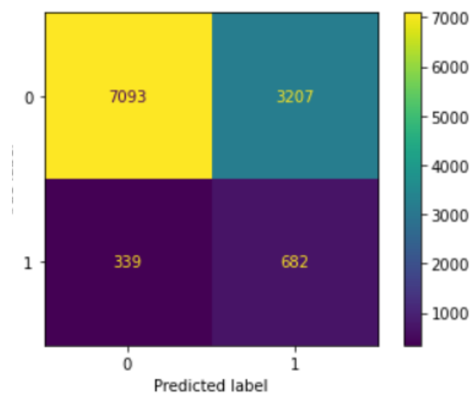
Imbalanced Data

	precision	recall	f1-score	support
0	0.91	0.99	0.95	10300
1	0.20	0.04	0.06	1021
accuracy			0.90	11321
macro avg	0.56	0.51	0.51	11321
weighted avg	0.85	0.90	0.87	11321



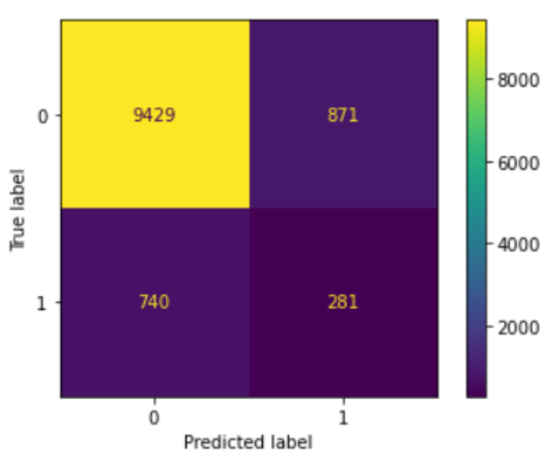
UnderSampled

	precision	recall	f1-score	support
0	0.95	0.69	0.80	10300
1	0.18	0.67	0.28	1021
accuracy			0.69	11321
macro avg	0.56	0.68	0.54	11321
ighted avg	0.88	0.69	0.75	11321



OverSampled

	precision	recall	f1-score	support
0	0.93	0.92	0.92	10300
1	0.24	0.28	0.26	1021
accuracy			0.86	11321
macro avg	0.59	0.60	0.59	11321
weighted avg	0.87	0.86	0.86	11321



Over Sampling and Under Sampling achieved almost similar results, imbalanced data achieved the worst f1-score for class 1 where it was equal to zero.

Best Model Based on F1-Score is **Support Vector Machine**
Which achieved the following scores on Over-Sampled Data
Train: accuracy = 68%, f1-score for class 1 = 28%, f1-score for class 0 = 80%

Future Work: We can work on collecting more data of class 1 to achieve some sort of balance which will be useful for our prediction.