



FINAL REPORT

Kaushal Jaya Navaraju, Shubham Nilesh Palande, Sindhu Sundararajan, Swetha Lakshmana

Perumal

College of Professional Studies, Northeastern University

ALY6040: Data Mining Applications

PROFESSOR: Kasun Samarasinghe

February 13, 2023

INTRODUCTION

The Final Project dataset is a Loan Application dataset with a real-world business scenario. The dataset is further examined to gain a better understanding of risk analysis in banking and finance services. This information represents a real-world loan structure that is used to reduce the risk of losing money when lending to customers. The number of rows in the Loan Application Dataset is 307511, and the number of columns is 122, as shown in the figure 1.

☞

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL
0	100002	1	Cash loans	M	N	Y	0	202500.0
1	100003	0	Cash loans	F	N	N	0	270000.0
2	100004	0	Revolving loans	M	Y	Y	0	67500.0
3	100006	0	Cash loans	F	N	Y	0	135000.0
4	100007	0	Cash loans	M	N	Y	0	121500.0
...
307506	456251	0	Cash loans	M	N	N	0	157500.0
307507	456252	0	Cash loans	F	N	Y	0	72000.0
307508	456253	0	Cash loans	F	N	Y	0	153000.0
307509	456254	1	Cash loans	F	N	Y	0	171000.0
307510	456255	0	Cash loans	F	N	N	0	157500.0

307511 rows x 122 columns

FIGURE 1

DATA EXPLORATION

Figure 2 illustrates the descriptive statistics of the loan application dataset.

☞

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL
0	100002	1	Cash loans	M	N	Y	0	202500.0
1	100003	0	Cash loans	F	N	N	0	270000.0
2	100004	0	Revolving loans	M	Y	Y	0	67500.0
3	100006	0	Cash loans	F	N	Y	0	135000.0
4	100007	0	Cash loans	M	N	Y	0	121500.0
...
307506	456251	0	Cash loans	M	N	N	0	157500.0
307507	456252	0	Cash loans	F	N	Y	0	72000.0
307508	456253	0	Cash loans	F	N	Y	0	153000.0
307509	456254	1	Cash loans	F	N	Y	0	171000.0
307510	456255	0	Cash loans	F	N	N	0	157500.0

307511 rows x 122 columns

FIGURE 2

Figure 3 depicts the information about the dataset. It can be seen from the figure that there are 65 float characters, 41 integers, and 16 object datatypes in the loan application dataset.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 307511 entries, 0 to 307510
Columns: 122 entries, SK_ID_CURR to AMT_REQ_CREDIT_BUREAU_YEAR
dtypes: float64(65), int64(41), object(16)
memory usage: 286.2+ MB
```

FIGURE 3

DATA CLEANING

In order to check and drop the duplicate values, `drop_duplicates()` function is used. It is clear that the number of rows is still the same and hence, it can be concluded that there are no duplicate values in our dataset.

```
df = df.drop_duplicates()
df
```

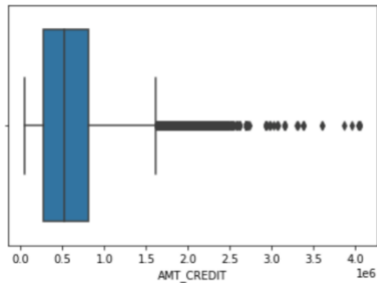
	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL
0	100002	1	Cash loans	M	N	Y	0	202500.0
1	100003	0	Cash loans	F	N	N	0	270000.0
2	100004	0	Revolving loans	M	Y	Y	0	67500.0
3	100006	0	Cash loans	F	N	Y	0	135000.0
4	100007	0	Cash loans	M	N	Y	0	121500.0
...
307506	456251	0	Cash loans	M	N	N	0	157500.0
307507	456252	0	Cash loans	F	N	Y	0	72000.0
307508	456253	0	Cash loans	F	N	Y	0	153000.0
307509	456254	1	Cash loans	F	N	Y	0	171000.0
307510	456255	0	Cash loans	F	N	N	0	157500.0

307511 rows x 122 columns

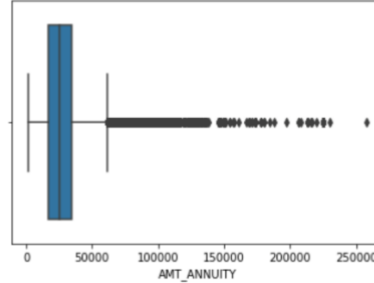
FIGURE 4

To check if there are any outliers in the dataset, box plot is used. Figure 5 illustrates the box plot for `amt_credit` attribute and it is clear from the figure that there are many outliers in the amount credited column. Similarly, the box plot for `amt_annuity` is shown in Figure 6 and it can be seen that `amt_annuity` also has outliers.

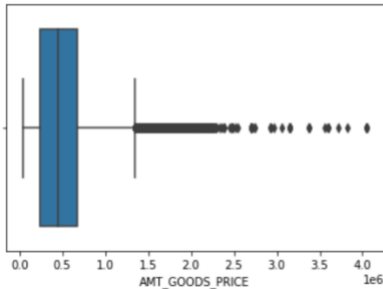
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f14ce0dde50>
```

**FIGURE 5**

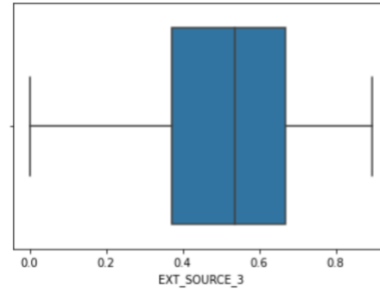
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f14ca69d700>
```

**FIGURE 6**

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f14cdfd31f0>
```

**FIGURE 7**

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f14ca30f550>
```

**FIGURE 8**

It can be inferred from Figure 7 that the amount good price attribute has outliers and the box plot depicted in Figure 8 shows that the attribute `ext_source_3` does not have any outliers.

From Figure 9, it is evident that the loan application dataset has some NULL values. Hence, the data is said to be unclean.

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f14ca30f550>
```

SK_ID_CURR	0
TARGET	0
NAME_CONTRACT_TYPE	0
CODE_GENDER	0
FLAG_OWN_CAR	0
...	
AMT_REQ_CREDIT_BUREAU_DAY	41519
AMT_REQ_CREDIT_BUREAU_WEEK	41519
AMT_REQ_CREDIT_BUREAU_MON	41519
AMT_REQ_CREDIT_BUREAU_QRT	41519
AMT_REQ_CREDIT_BUREAU_YEAR	41519
Length: 122, dtype: int64	

FIGURE 9

Figure 10 illustrates the percentage of NULL values in each attribute in the dataset. The attributes that have a percentage of NULL values greater than 20% are dropped.

☐➔

	Column Name	Total_Number_of_Nulls	Missing_Percent
0	COMMONAREA_MEDI	214865	69.872297
1	COMMONAREA_AVG	214865	69.872297
2	COMMONAREA_MODE	214865	69.872297
3	NONLIVINGAPARTMENTS_MODE	213514	69.432963
4	NONLIVINGAPARTMENTS_AVG	213514	69.432963
...
117	NAME_HOUSING_TYPE	0	0.000000
118	NAME_FAMILY_STATUS	0	0.000000
119	NAME_EDUCATION_TYPE	0	0.000000
120	NAME_INCOME_TYPE	0	0.000000
121	SK_ID_CURR	0	0.000000

122 rows x 3 columns

FIGURE 10

After dropping the attributes with high NULL values, the number of columns got reduced from 122 to 73 as seen in Figure 11.

☐➔

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL
0	100002	1	Cash loans	M	N	Y	0	202500.0
1	100003	0	Cash loans	F	N	N	0	270000.0
2	100004	0	Revolving loans	M	Y	Y	0	67500.0
3	100006	0	Cash loans	F	N	Y	0	135000.0
4	100007	0	Cash loans	M	N	Y	0	121500.0
...
307506	456251	0	Cash loans	M	N	N	0	157500.0
307507	456252	0	Cash loans	F	N	Y	0	72000.0
307508	456253	0	Cash loans	F	N	Y	0	153000.0
307509	456254	1	Cash loans	F	N	Y	0	171000.0
307510	456255	0	Cash loans	F	N	N	0	157500.0

307511 rows x 73 columns


FIGURE 11

The attributes 'NAME_TYPE_SUITE', 'OBS_30_CNT_SOCIAL_CIRCLE', 'DEF_30_CNT_SOCIAL_CIRCLE', 'OBS_60_CNT_SOCIAL_CIRCLE', 'DEF_60_CNT_SOCIAL_CIRCLE' are not required for the further data analysis as they do not align with our target variable. After removing these columns, the total attributes became 68.

The NULL values in the attributes 'AMT_ANNUITY', 'AMT_GOODS_PRICE', 'DAYS_LAST_PHONE_CHANGE' are discarded as the minimum value and the maximum value has a huge difference and imputing mean value to these NULL values won't be biased.

The NULL values of the attributes 'EXT_SOURCE_2' and 'EXT_SOURCE_3' are replaced with the mean values of the respective variables. But, for the attributes AMT_REQ_CREDIT_BUREAU_HOUR, AMT_REQ_CREDIT_BUREAU_DAY, AMT_REQ_CREDIT_BUREAU_DAY, AMT_REQ_CREDIT_BUREAU_MON, AMT_REQ_CREDIT_BUREAU_QRT, and AMT_REQ_CREDIT_BUREAU_YEAR, there is a huge difference between the maximum value and the mean value. Hence, the NULL values in these attributes are replaced by the mode values. The attribute OCCUPATION_TYPE has around 9633 NULL values and it is a character data type. So, these NULL values are replaced by the term 'UNKNOWN'.

Figure 12 depicts the final checking of NULL values in the loan application dataset and there are no more NULL values. It can now be concluded that the dataset is clean.



SK_ID_CURR	0
TARGET	0
NAME_CONTRACT_TYPE	0
CODE_GENDER	0
FLAG_OWN_CAR	0
FLAG_OWN_REALTY	0
CNT_CHILDREN	0
AMT_INCOME_TOTAL	0
AMT_CREDIT	0
AMT_ANNUITY	0
AMT_GOODS_PRICE	0
NAME_INCOME_TYPE	0
NAME_EDUCATION_TYPE	0
NAME_FAMILY_STATUS	0
NAME_HOUSING_TYPE	0
REGION_POPULATION_RELATIVE	0
DAYS_BIRTH	0
DAYS_EMPLOYED	0
DAYS_REGISTRATION	0
DAYS_ID_PUBLISH	0
FLAG_MOBIL	0

FIGURE 12

STANDARDIZING THE VALUES

In standardizing, the high value columns are segregated under the right type of data. In the loan application dataset, the columns AMT_INCOME_TOTAL, AMT_CREDIT, AMT_GOODS_PRICE are considered as these are high value numeric columns. Now, the values

of these columns are placed in categorical bin columns. Appropriate bins are created for income columns in terms of Lakhs. This can be seen in Figures 13 and 14.

```

1-2 L      50.73
2-3 L      21.21
0-1 L      20.73
3-4 L       4.78
4-5 L       1.74
5-6 L       0.36
6-7 L       0.28
8-9 L       0.10
7-8 L       0.05
9-10 L      0.01
10 L Above  0.01
Name: INCOME_RANGE, dtype: float64

```

FIGURE 13

```

count      307511.000000
mean        1.687979
std         2.371231
min         0.256500
25%         1.125000
50%         1.471500
75%         2.025000
max         1170.000000
Name: AMT_INCOME_TOTAL, dtype: float64

```

FIGURE 14

DATA ANALYSIS

Figure 15 is a box plot that illustrates the outliers in the dataset. The columns that had the highest difference between maximum value and its 75th percentile value are chosen to find the outliers for them. The columns that have few outliers are AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE, and CNT_CHILDREN.

Columns with the highest number of Outliers are AMT_INCOME_TOTAL (the graph indicates that a few select people have higher income compared to the rest of the members). The column, DAYS_EMPLOYED have outlier values at 350000 days, which is a humongous number of days, which indicates that there is mistake while collecting data or some bogus information was provided. The column, DAYS_BIRTH doesn't have any outlier indicating that it is most reliable inform among the following values.

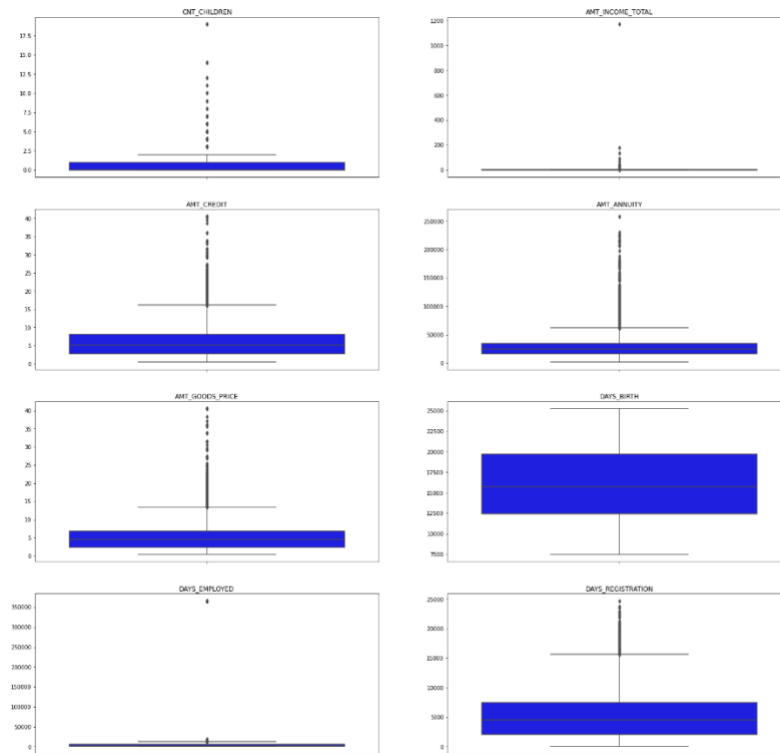
**FIGURE 15**

Figure 16 depicts the number of loan applications from both the genders. It can be inferred from the Figure that female took twice loan as men accounting to nearly 202251 and male contributed to around 104965 applications.

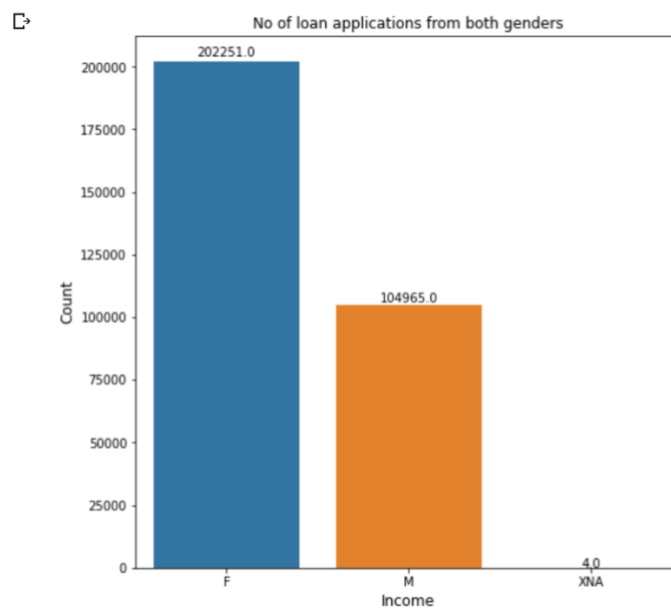
**FIGURE 16**

Figure 17 is a pie chart of the target variable i.e., the actual value of the repayer and the defaulters. It is clear from the chart that there are 92% of repayers and only 8% defaulters in the target variable. So, only a very small portion of the population had issues in repaying the loans.

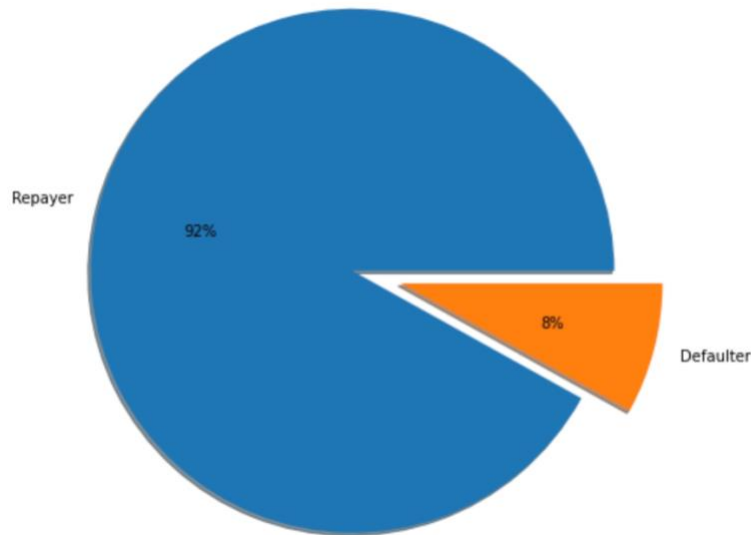


FIGURE 17

Figure 18 is a pair plot we can clearly observe that there is a good increasing trend or strong correlation between the variables AMT_ANNUIITY vs AMT_CREDIT and AMT_CREDIT vs AMTG_GOODS_PRICE. These variables can be perfect for explaining the interdependence and how these play as a factor in a customer either becoming a "Repayer" or "Defaulter".

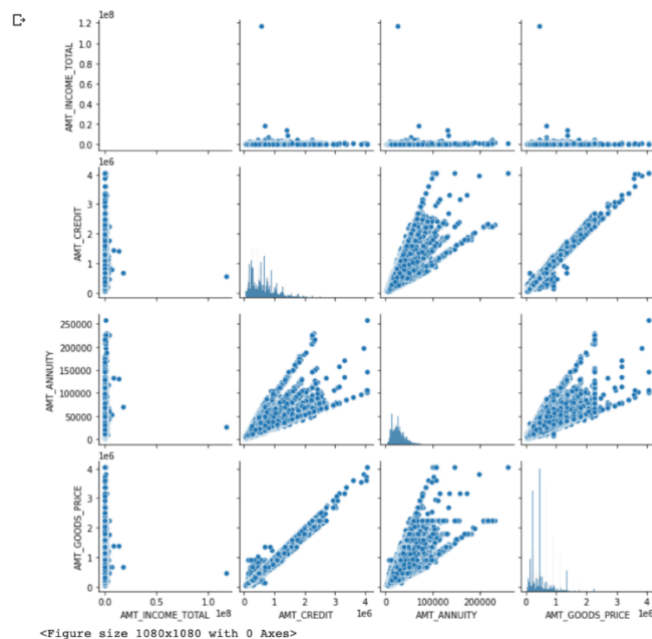


FIGURE 18

CATEGORICAL VARIABLE ANALYSIS

Important function for univariate analysis is performed. Now, a function is created for plotting variables to univariate analysis. This function will create two plots as shown in Figure 19.

Plot of the given column in relation to the TARGET column2.

Finding the percentage of defaulters within each column of the given datasets.

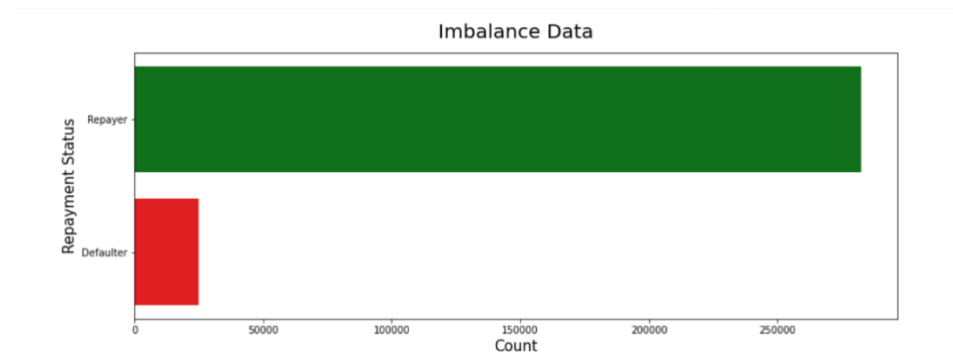


FIGURE 19

Around 90% of the loans are cash loans, which indicate that the loan applicant is in the need of liquid cash. Only 10% of the loans are revolving loans. Solid 8 % of the cash loan recipients turn out to be defaulters. Whereas in the case of revolving loans, 5.5 % of the loan recipients turn out to be defaulters.

It is clearly seen that most of the cash loan and revolving loan recipients repay their loans. But in terms of sheer numbers the defaulters are more in the case of cash loans. This is illustrated in Figure 20.

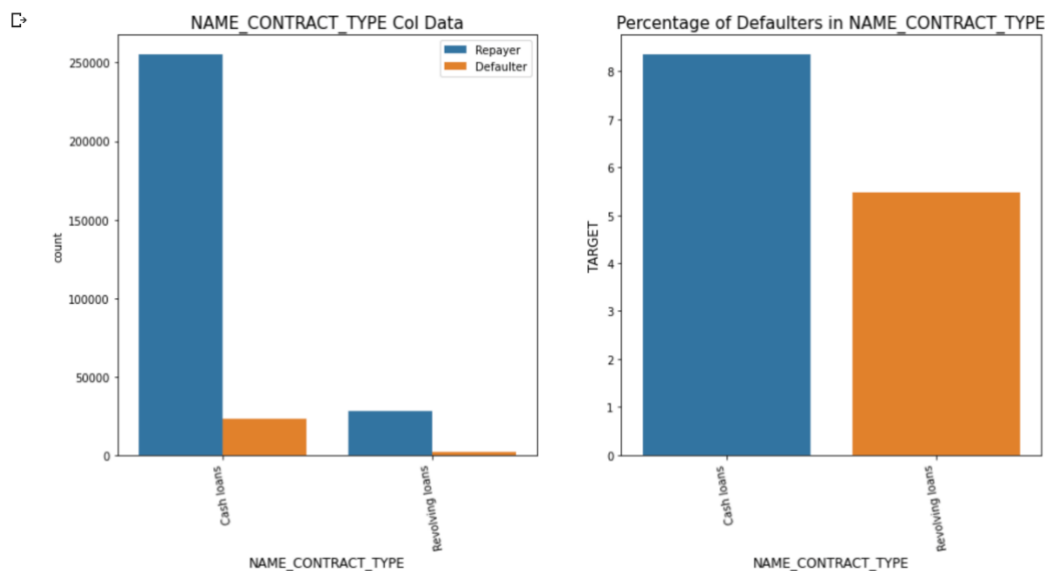


FIGURE 20

From the plot shown in Figure 21, it is clear that most of the loan applications are female. But, when it comes to the defaulter percentage, it can be seen that male clients tend to not return their loans. 10% of the male clients turn out to be defaulters but in case of the female clients, the defaulter percentage is as low as 7%.

This clearly indicates that most the female clients are more trust worthy clients to the bank and this in turn would impact the decisions matrix. Its more likely that female clients would have their loan approvals faster and the overall process would be seemingly smooth.

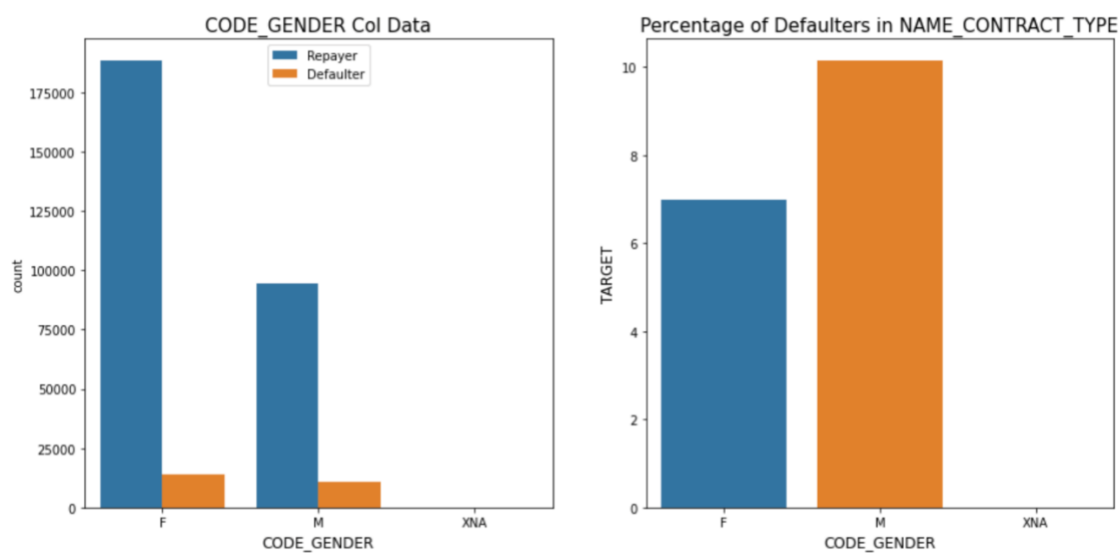


FIGURE 21

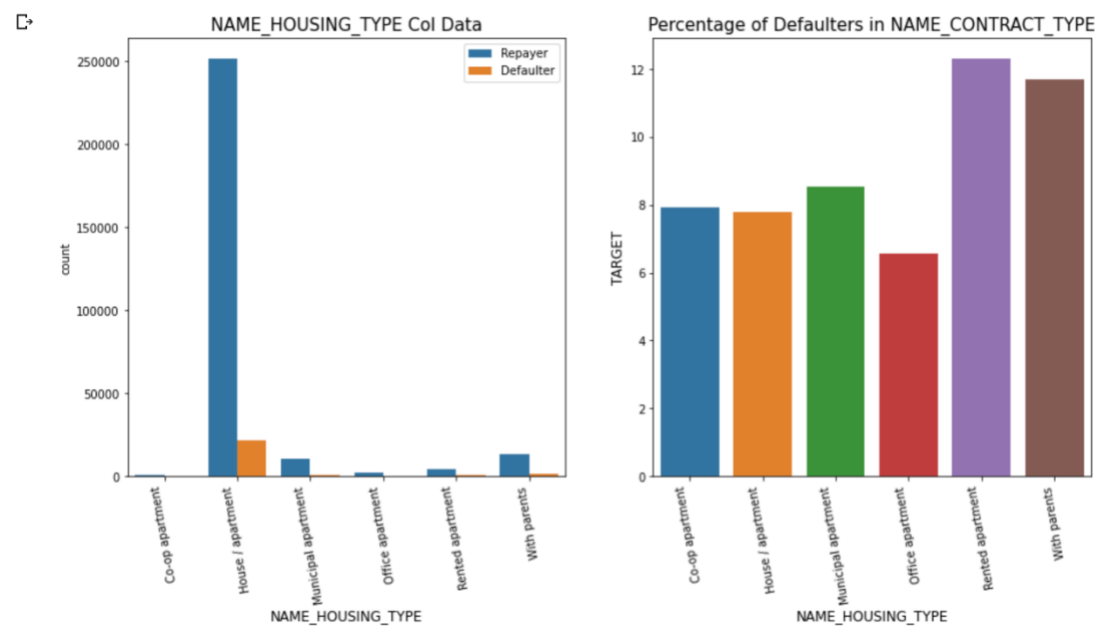


FIGURE 22

Most of the loan applicants live in independent house or apartment. People living in co-op apartment don't tend to apply for loans in general.

Coming to the defaulter percentage, people living in office apartments tend to default the least. only 6.5 % of the loan applicants default in their case.

Majority of the defaulters fall under the rented apartment category (12% applicants) and applicants living with their parents (around 11%).

But in terms of sheer numbers, majority of defaulters fall under the house and apartment category, which can be clearly seen in Figure 22.

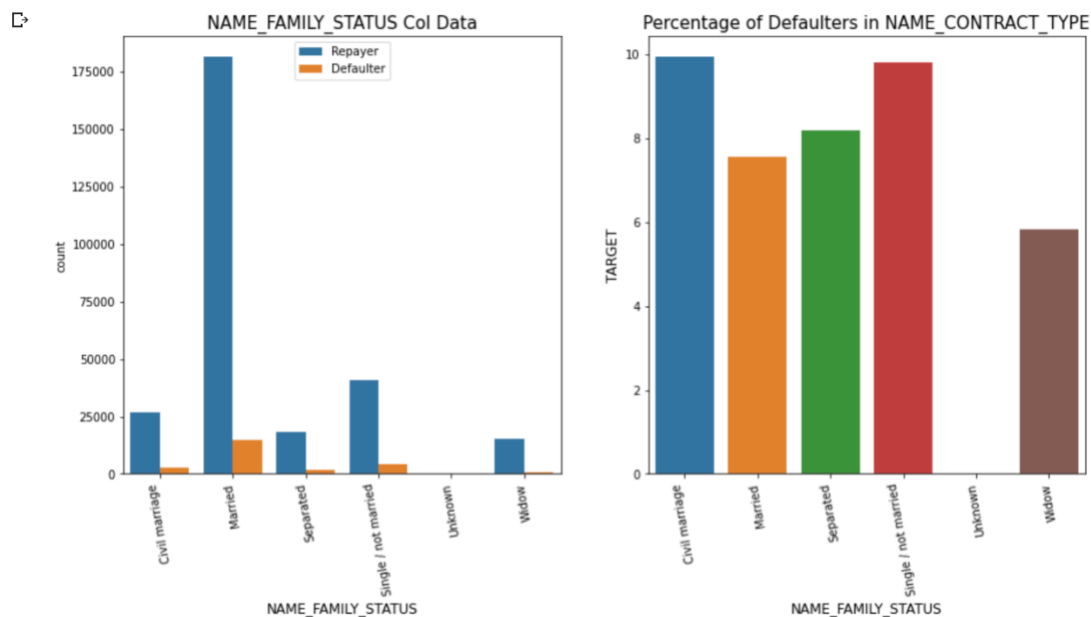


FIGURE 23

From the plot shown in Figure 23, it is clear that most of the loan applications have own property of their own.

However, in terms of the defaulter percent, approximately 8% of the applicants in both parties' default. But the applicants without property slightly default more than the other one.

The impact of educational qualification on loan repayment is shown in Figure 24.

People with secondary educational qualification are higher education are the major applicants for loan.

But when it comes to defaulter percentage, applicants with lower secondary degree are more tend to default which is around 12%.

Majority of the higher education applicants tend to return the loan.

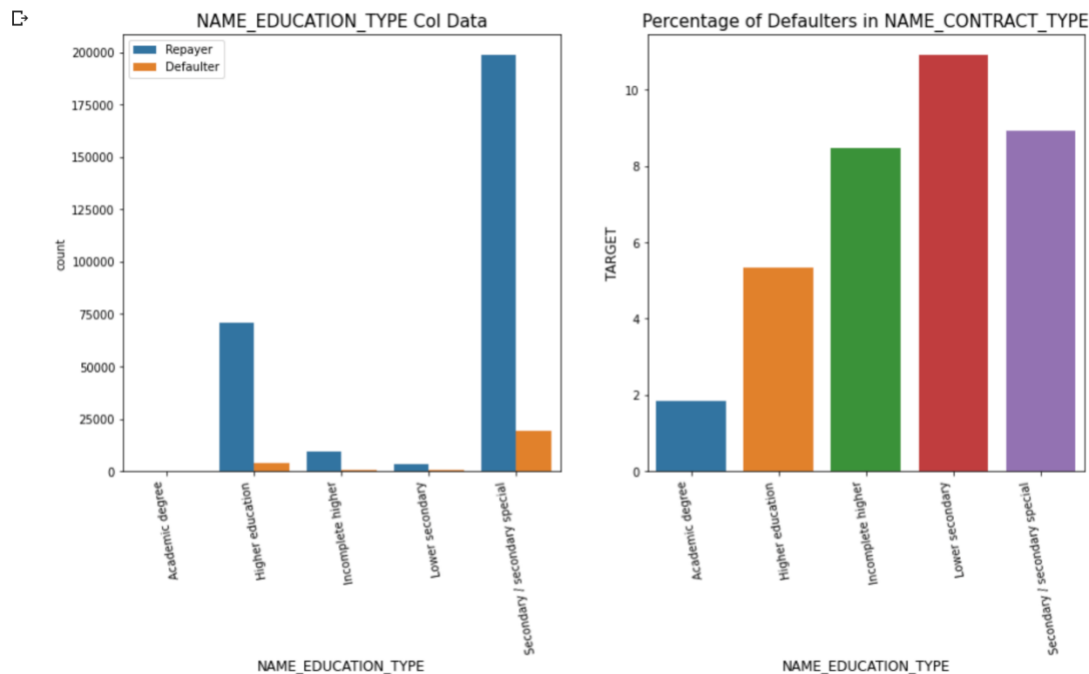


FIGURE 24

From the plot shown in Figure 25, it can be inferred that the maximum repayers are in working category and the secondary repayers are commercial associates and pensioner. Where as in defaulters, the people in maternity leave are the highest defaulters and the second highest are the unemployed people.

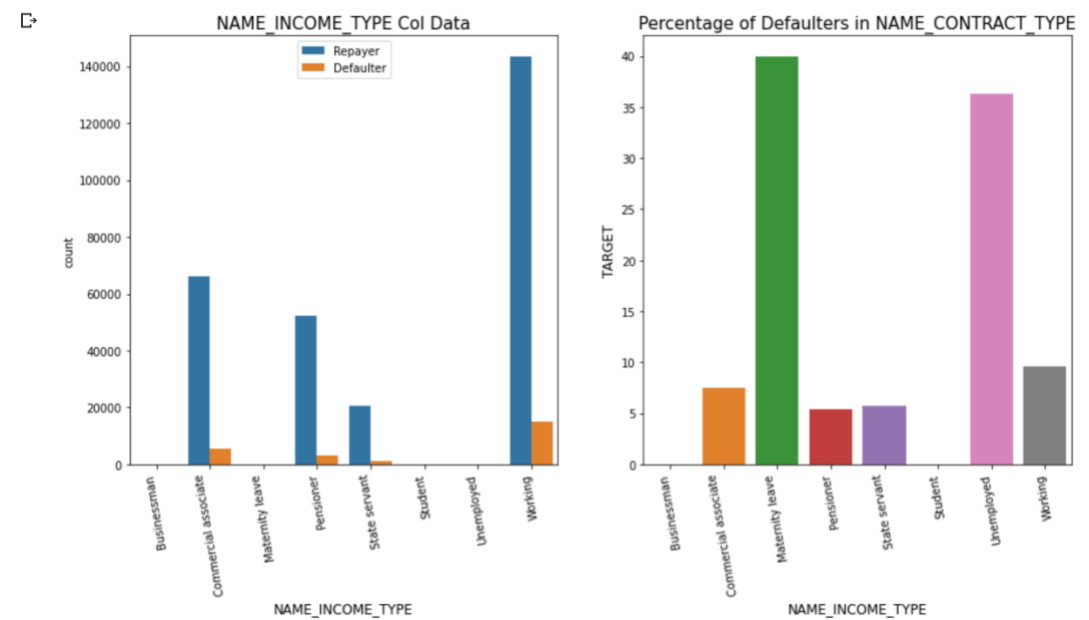


FIGURE 25

NUMERICAL VARIABLE ANALYSIS

In numerical variable analysis, bisecting the app_data dataframe based on Target value 0 and 1 for correlation and other analysis is done. This is shown in Figure 26.

```
Index(['SK_ID_CURR', 'TARGET', 'NAME_CONTRACT_TYPE', 'CODE_GENDER',
      'FLAG_OWN_REALTY', 'CNT_CHILDREN', 'AMT_INCOME_TOTAL', 'AMT_CREDIT',
      'AMT_ANNUITY', 'AMT_GOODS_PRICE', 'NAME_TYPE_SUITE', 'NAME_INCOME_TYPE',
      'NAME_EDUCATION_TYPE', 'NAME_FAMILY_STATUS', 'NAME_HOUSING_TYPE',
      'REGION_POPULATION_RELATIVE', 'DAYS_BIRTH', 'DAYS_EMPLOYED',
      'DAYS_REGISTRATION', 'DAYS_ID_PUBLISH', 'FLAG_MOBIL', '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', 'LIVE_REGION_NOT_WORK_REGION',
      'REG_CITY_NOT_LIVE_CITY', 'REG_CITY_NOT_WORK_CITY',
      'LIVE_CITY_NOT_WORK_CITY', 'ORGANIZATION_TYPE',
      'OBS_30_CNT_SOCIAL_CIRCLE', 'DEF_30_CNT_SOCIAL_CIRCLE',
      'OBS_60_CNT_SOCIAL_CIRCLE', 'DEF_60_CNT_SOCIAL_CIRCLE',
      'DAYS_LAST_PHONE_CHANGE', 'FLAG_DOCUMENT_3',
      'AMT_REQ_CREDIT_BUREAU_HOUR', 'AMT_REQ_CREDIT_BUREAU_DAY',
      'AMT_REQ_CREDIT_BUREAU_WEEK', 'AMT_REQ_CREDIT_BUREAU_MON',
      'AMT_REQ_CREDIT_BUREAU_QRT', 'AMT_REQ_CREDIT_BUREAU_YEAR',
      'INCOME_RANGE', 'CREDIT_RANGE', 'GOODS_PRICE_RANGE', 'AGE', 'AGE_GROUP',
      'YEARS_EMPLOYED', 'EMPLOYEMENT_YEARS'],
      dtype='object')
```

FIGURE 26

The correlation plot for the repayers is illustrated in Figure 27. It can be seen that the credit amount is highly correlated with "Goods Price Amount", "Loan Annuity", and "Total Income". It is also clear that repayers have high correlation in number of days employed.

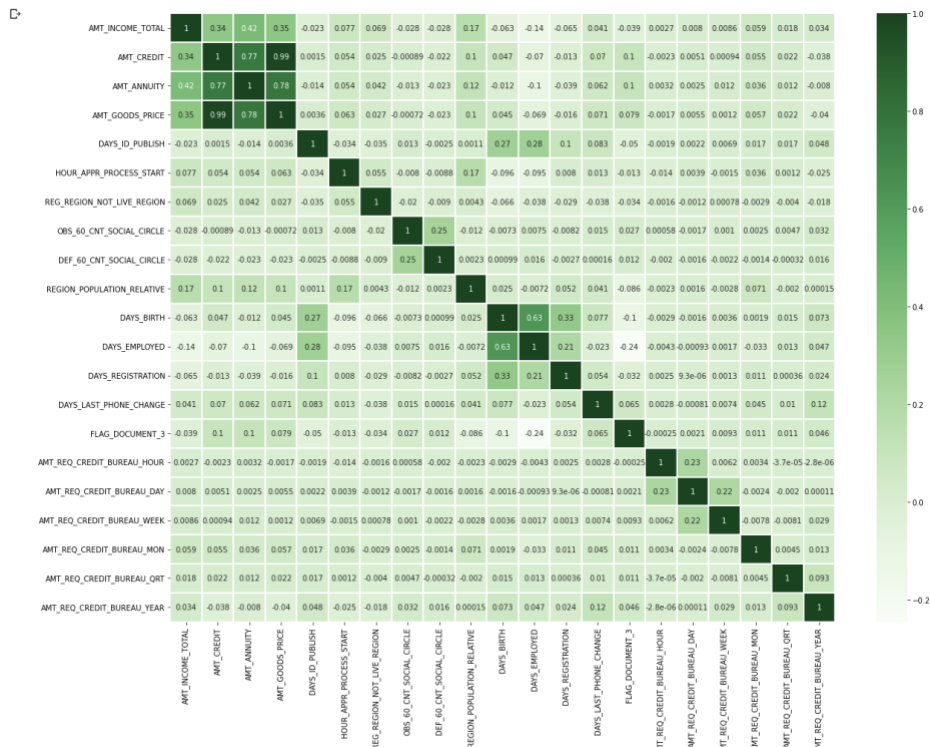


FIGURE 27

The correlation plot for defaulters is illustrated in Figure 28.

There is a high correlation between the goods price and Credit amount, which in turn infers that most of the applicants take loan to purchase consumer goods.

The correlation between credit amount and loan annuity is less among the defaulter (0.75) when compared to that of repayers (0.77).

Similarly, even for the employed days, the repayers exhibit more correlation than defaulters.

The correlation between credit amount and total income sees a major difference, as the repayers are seen to have fairly average correlation around (0.34), whereas for defaulter (0.03), we can notice a sharp drop.

The correlation between days birth column and number of childrens column has dropped to 0.259 in defaulters when compared to that of the repayers 0.33.

There is a slight increase in defaulted to observed count in social circle among defaulters(0.264) when compared to repayers(0.254)

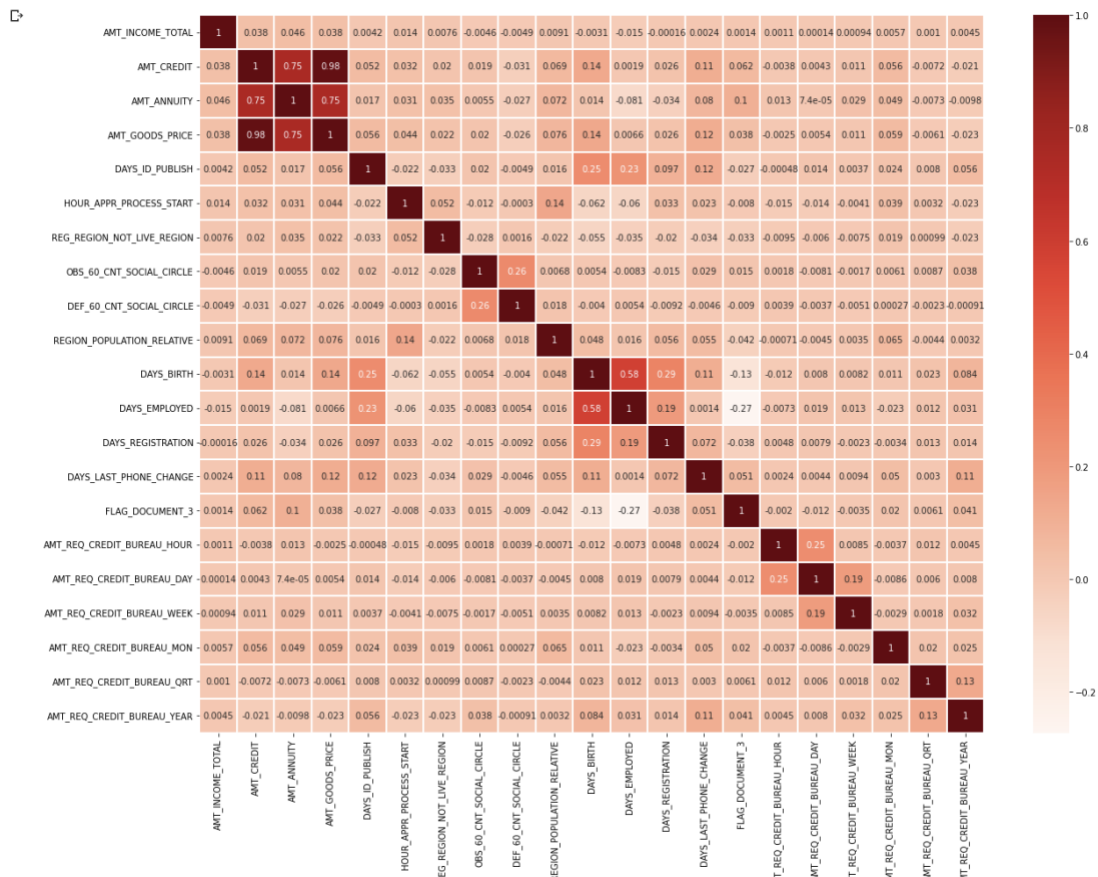


FIGURE 28

NUMERICAL UNIVARIATE ANALYSIS

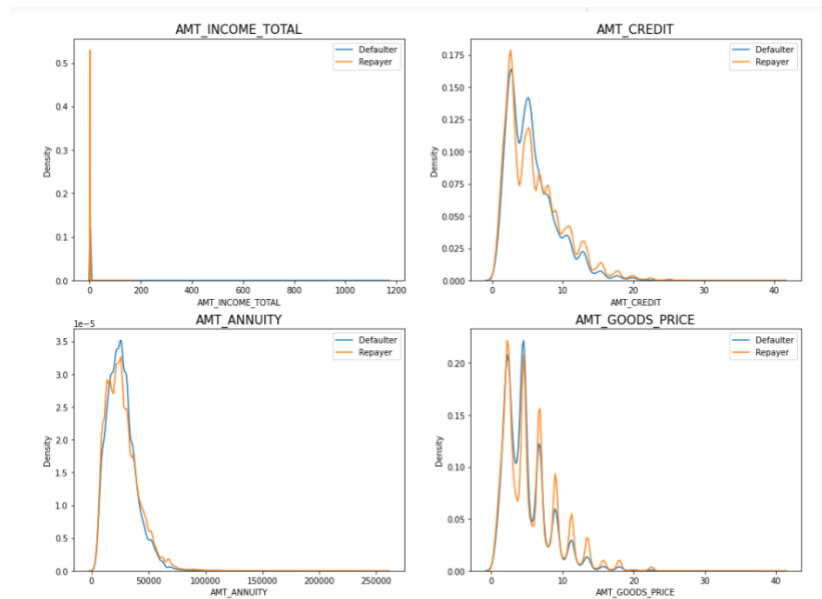


FIGURE 29

Figure 29 depicts the numerical univariate analysis.

Most number of loans range between 2 - 10 lakhs for the goods price. Most people pay annuity below 50K for the credit loan. Highest amount of Credit amount loans are for the amount 2-5 lakhs and majorly ranges from 2-10 lakhs bracket.

NUMERICAL BIVARIATE ANALYSIS

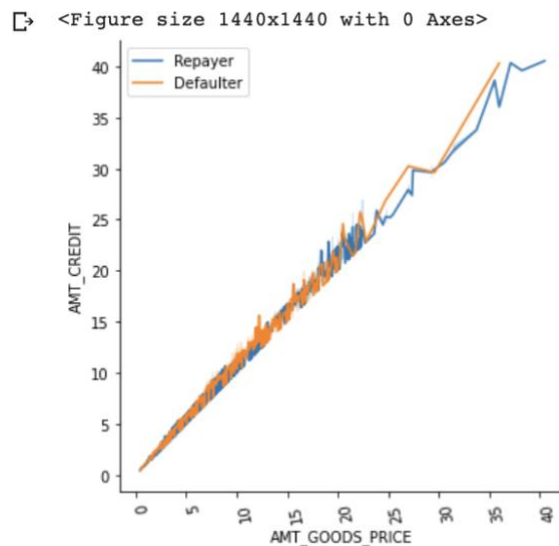


FIGURE 30

Figure 30 shows the numeric bivariate analysis for the dataset.

The amount of repayers and defaulters are continuously seem to raise with the increase in the loan amount.

However, it can be seen that the defaulter percentage seems to steadily raise but post 30 lakhs, there is spike in the defaulter percentage and the repayers percentage sees a small downfall.

DECISION TREE

Decision tree is used to break down complicated data to more manageable components and to fine the key parameter that is suitable for the dataset. From the decision tree shown in Figure 31, it can be inferred that the parameter AGE is the most effective at illustrating several applications in the dataset. The AGE decision node serves as the foundation for most classifications that can be applied to the dataset.

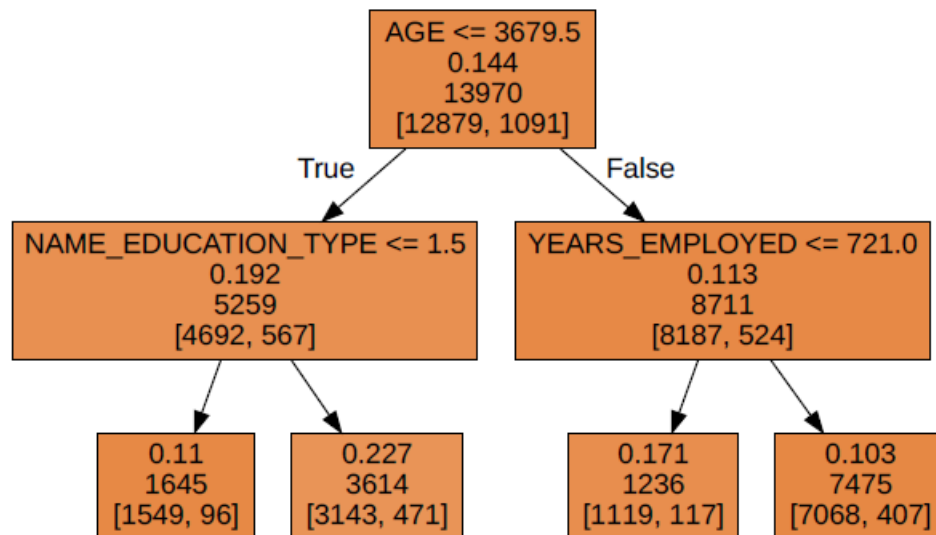


FIGURE 31

The attributes that might have a significant influence on the loan application dataset can be figured out by looking at the bar graph shown in Figure 32. It is clear from the bar plot that the value AGE is crucial compared to the other parameters. Moreover, the decision tree shown in Figure 18 makes it clear that the AGE is the major decision node.



FIGURE 32

The classification report of the decision tree can be seen in Figure 33.

Classification report from Decision Tree:				
	precision	recall	f1-score	support
0	0.92	1.00	0.96	2688
1	0.00	0.00	0.00	224
accuracy			0.92	2912
macro avg	0.46	0.50	0.48	2912
weighted avg	0.85	0.92	0.89	2912

FIGURE 33

An accuracy test was conducted on the decision tree. It can be inferred from the below output that the accuracy of the decision tree is 92.31%, which indicates that the model is a good fit.

Test Accuracy: 92.31 Percent

A confusion matrix depicting the graphical representation of relationship of two variables namely Repayer and Defaulter can be seen in Figure 34. The confusion matrix plots their values in relation to predictions and actual values. It can be inferred from the plot that the defaulters are also repayers and since there are no false positive and true negative values which indicates that the model is a good fit.

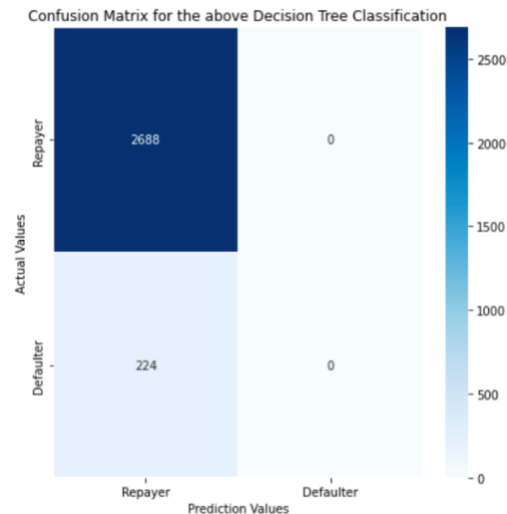


FIGURE 34

LOGISTIC REGRESSION

Logistic model predicts the likelihood of an event occurring by combining its log-odds with one or more independent factors in a linear fashion. The classification report of the logistic regression can be seen in Figure 35.

Classification report from Logistic Regression:				
	precision	recall	f1-score	support
0	0.92	1.00	0.96	2688
1	1.00	0.00	0.01	224
accuracy			0.92	2912
macro avg	0.96	0.50	0.48	2912
weighted avg	0.93	0.92	0.89	2912

FIGURE 35

From the test accuracy conducted for logistic regression, it can be inferred that the accuracy is 92.34%. Hence, it can be concluded that the model is a good fit.

Test Accuracy: 92.34 Percent

Figure 36 is a confusion matrix showing the confusion matrix for the logistic regression. It is clear from the matrix that there is no false positive value and the true negative value is 1. So, the model is an ideal fit.

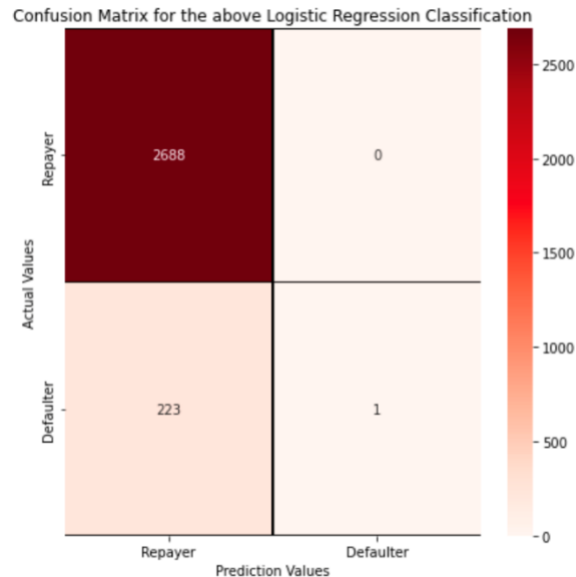


FIGURE 36

A function is created for calculating the co-efficient of determination. The adjusted R square value for the train and the test dataset can be seen in Figure 37.

```
In [79]: print(adj_r2(x_train, y_train, log_reg))
```

```
0.9194165743033891
```

```
In [80]: print(adj_r2(x_test, y_test, log_reg))
```

```
0.92125629801142
```

FIGURE 37

RANDOM FOREST CLASSIFICATION

The random forest classification algorithm creates many randomly chosen decision trees from the data using ensemble learning techniques and the decision tree architecture, averaging the results to get a new result that frequently produces accurate predictions or classifications. The classification report of the random forest classification can be seen in Figure 38.

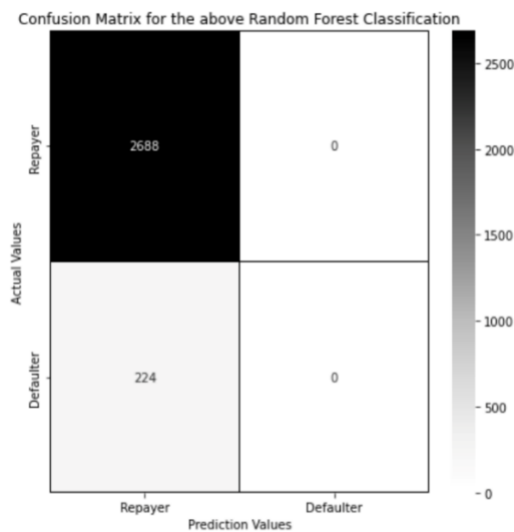
Classification report from Random Forest :				
	precision	recall	f1-score	support
0	0.92	1.00	0.96	2688
1	0.00	0.00	0.00	224
accuracy			0.92	2912
macro avg	0.46	0.50	0.48	2912
weighted avg	0.85	0.92	0.89	2912

FIGURE 38

The accuracy test is conducted for the random forest classification. It can be seen that the accuracy is 92.31%. This shows that the model is a good fit.

Test Accuracy: 92.31 Percent

Figure 39 depicts the confusion matrix for random forest classification. It can be inferred from the confusion matrix that there are no false positive and true negative values. So, it can be concluded that the model is an ideal fit.

**FIGURE 39**

SUPPORT VECTOR MACHINE

Support vector machine helps to identify a hyperplane in an N-dimensional space that clearly divides the data points into categories. The classification report of the support vector machine can be seen in Figure 40.

Classification report from SVM :

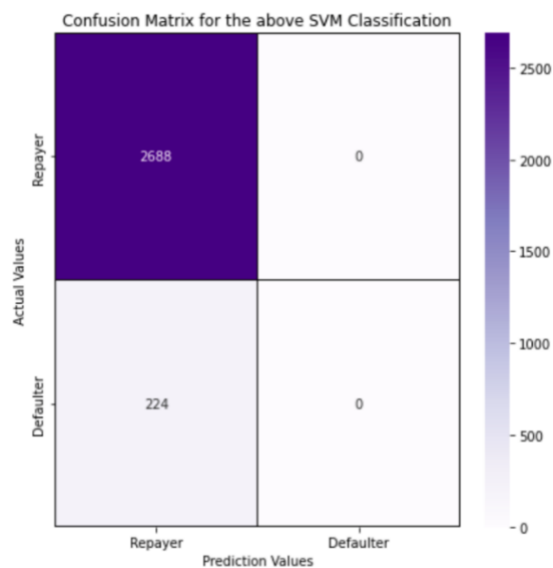
	precision	recall	f1-score	support
0	0.92	1.00	0.96	2688
1	0.00	0.00	0.00	224
accuracy			0.92	2912
macro avg	0.46	0.50	0.48	2912
weighted avg	0.85	0.92	0.89	2912

FIGURE 40

The test accuracy is conducted for support vector machine. It can be seen that the accuracy is 92.31%. This shows that the model is a good fit.

Test Accuracy: 92.31 Percent

It can be inferred from the confusion matrix for support vector machine from Figure 41 that there is no false positive and true negative values. So, the model is said to be an ideal fit.

**FIGURE 41**

K-NEAREST NEIGHBORS (KNN)

KNN models are simply the technical application of a widely held belief that entities with similar properties are likely to be, well, similar. The classification report of the K-nearest neighbors(KNN) can be seen in Figure 42.

Classification report from KNN :

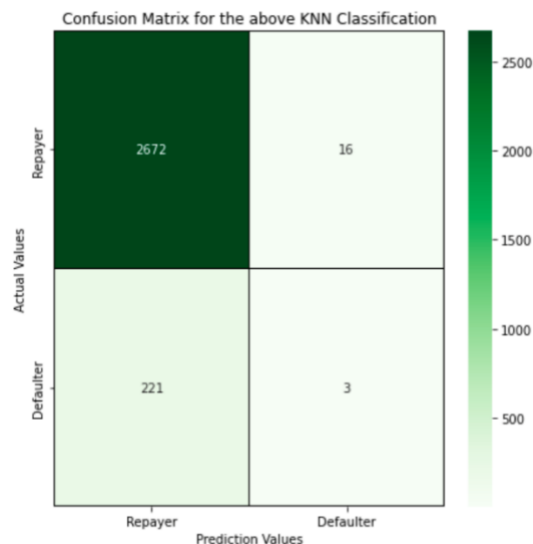
	precision	recall	f1-score	support
0	0.92	0.99	0.96	2688
1	0.16	0.01	0.02	224
accuracy			0.92	2912
macro avg	0.54	0.50	0.49	2912
weighted avg	0.86	0.92	0.89	2912

FIGURE 42

From the test accuracy conducted for KNN, it is clear that the accuracy is 92.1%. This shows that the model is a good fit.

Fitting KNN Value: 9
Test Accuracy: 92.1 Percent

From the confusion matrix for support vector machine illustrated in Figure 43, it can be seen that there are 16 false positive values and 3 true negative values which affects the accuracy of the model. This is why the model is not 100% accurate.

**FIGURE 43**

In order to check whether the trained model is predicting the values correctly, the test values are predicted. It can be inferred from Figure 44 that the predicted and test values are not matching, indicating that the model is not good at predicting the values.

ROC CURVE

The ROC (Receiver Operating Characteristics) curve, gives a value that indicates the Sensitivity, Specificity of a model designed. As the curve is not too close to the left corner of the graph. It indicates that the model is not a perfect fit. Indicating that it is not the most viable solution. This is shown in Figure 47.

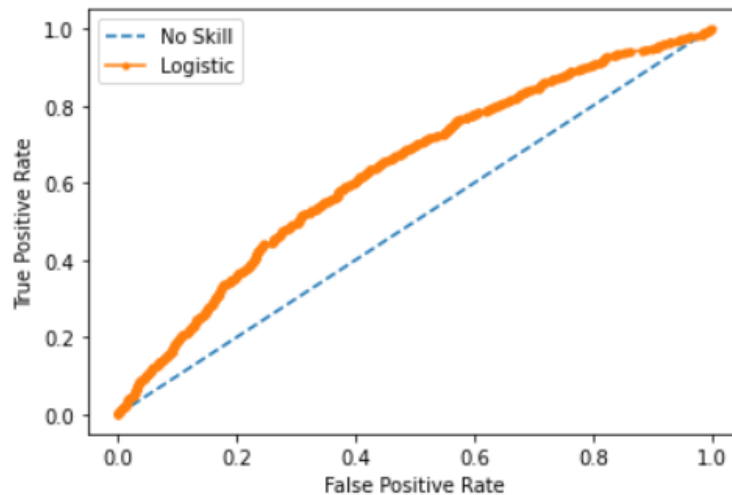


FIGURE 47

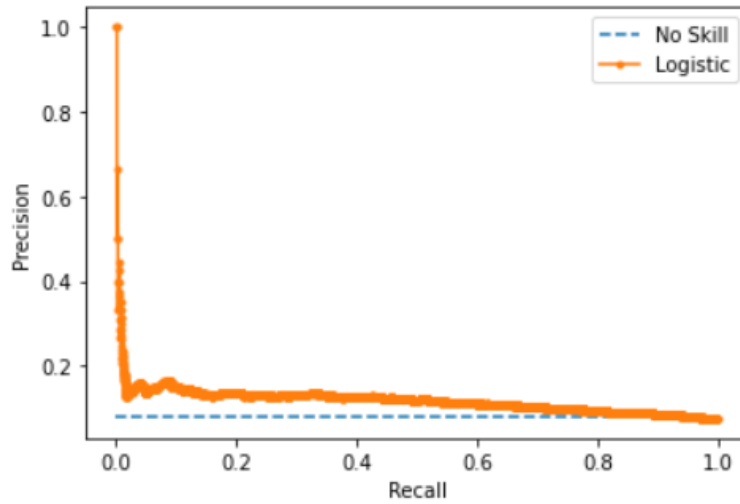
From Figure 48, it can be seen that the AUC value for the above ROC curve is "0.678", showing that it is the best model to predict the values.

```
No Skill ROC AUC 0.502
Logistic ROC AUC 0.634
```

FIGURE 48

PR (PRECISION-RECALL) CURVE

Precision-recall curve shows tradeoff that happens between precision and recall for different thresholds. In this curve as the precision value increases the recall opportunities diminish. From the above graph we can see that the precision of the values are very less and the recall features are more, again proving that it is not a good fit. This is illustrated in Figure 49.

**FIGURE 49**

From the figure 50, AUC values for the PR curve we can see that the no skill AUC values are around "0.119", where the logistic AUC value is "0.167", indicating that it is better than the no skill value. However, the model still proves to be inefficient

```
No Skill PR AUC: 0.120
Logistic PR AUC: 0.120
```

FIGURE 50

SUMMARY

The banking industry has been at the pinnacle of various businesses' growth and has helped improve living standards. The banking industry has contributed to developing particular countries' economic standards through the following activities. But one of the biggest challenges banks face nowadays is identifying the right potential customers. Though a majority of the customer's turn out to be repayers, they are constantly susceptible to frequent losses due to defaulters. These defaulters are hard to identify, as they initially make payments to the concerned bank, and as the months go, they stop making payments or make payments late. This kind of activity often leads the bank to losses. So, in this case, the study we are going to propose is a way in which the banks can identify and make informed decisions using previously acquired customer data.

Using the previously available data from the bank authorities, we will perform an Exploratory data analysis to understand the structure and datatypes we are dealing with for our study. During this, we will extract and isolate the parameters which can help us, further the study. Also, after EDA we used marplot and seaborn libraries to visualize the data effectively. Through appropriate visualization methods, we were able to get a pictorial representation different data that are present in the dataset. The correlation between various data columns were found using heatmaps. Then we applied various statistical concepts such as univariate, bivariate analysis and implemented decision tree and to pinpoint the parameter that carries high coherence with the objective. Later, by implementing machine learning algorithms such as KNN, Random Forest, and SVM techniques, we designed a model that can learn and predict the values, which in our case is to identify whether the applicant is a repayer or a defaulter. To solve classification and regression problems, we are implementing "XGB" gradient-boosting machine. Finally, using the gini index and Lorentz curve, we can that how varied the teen data set is in our case the gini index is close to 0.93. Indicating that the dataset having a good inequality thus helping us to design a good prediction model.

REFERENCES

- Loan Application Dataset*. (n.d.). Loan Application Dataset | Kaggle. Retrieved February 5, 2023, from <https://datasets/ramakrushnamohapatra/loan-application-dataset>
- seaborn.boxplot* — *seaborn 0.12.2 documentation*. (n.d.). Seaborn.Boxplot — Seaborn 0.12.2 Documentation. Retrieved February 5, 2023, from <https://seaborn.pydata.org/generated/seaborn.boxplot.html>
- seaborn.pairplot* — *seaborn 0.12.2 documentation*. (n.d.). Seaborn.Pairplot — Seaborn 0.12.2 Documentation. Retrieved February 5, 2023, from <https://seaborn.pydata.org/generated/seaborn.pairplot.html>
- Implementing Gradient Boosting Regression in Python* | *Paperspace Blog*. (2019, December 13). Paperspace Blog. Retrieved January 22, 2023, from <https://blog.paperspace.com/implementing-gradient-boosting-regression-python/>
- Beheshti, N. (2022, March 2). *Random Forest Regression*. Medium. Retrieved January 22, 2023, from <https://towardsdatascience.com/random-forest-regression-5f605132d19d>
- Gandhi, R. (2018, July 5). *Support Vector Machine — Introduction to Machine Learning Algorithms*. Medium. Retrieved January 22, 2023, from <https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47>