CREDIT EDA CASE STUDY

SATADHRITI CHAKRABARTY

PRASASTI CHOUDHURY

DS C17 Group 2

PROBLEM STATEMENT

- Apply EDA in credit business for banking and financial services to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default and in turn to minimise the loss resulting from the rejection of good loans.

BUSINESS CONTEXT

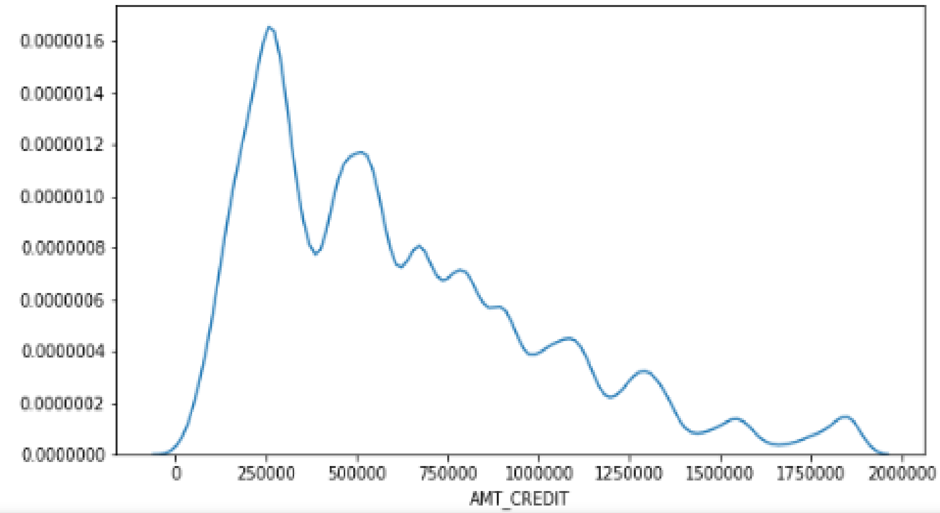
- Credit approval has a cost of assessment for a financial company apart from the various risks that arise during the lifetime of the credit.
- Studying a customer's credit history has a pivotal role in minimizing the loss or consequentially, maximizing the profit from the company's point of view.
- The analysis presented here attempts to ease this decision making for the company and to ensure that the business gets a *bang for their buck!*

ANALYSIS APPROACH

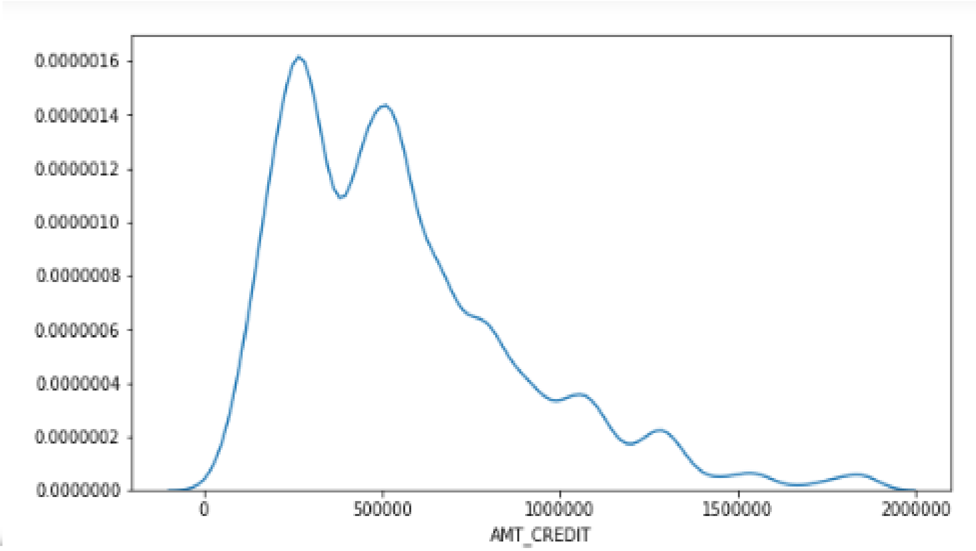
- The "application_data.csv" file was imported and the missing values were analysed and the ways to impute the missing values were reported in the python file(as markdown text).
- Handling of outliers and binning of continuous variables were done
- Imbalance percentage was checked and the data was divided into subsets for Target 0 and Target 1.
- Univariate and Bivariate analysis were performed for continuous and categorical variables.
- The data was then merged with the "previous_application.csv" file and the univariate and bivariate analysis were done for both Target 0 and Target 1.

CUSTOMER PROFILING

UNIVARIATE ANALYSIS OF CONTINUOUS VARIABLE



TARGET 0

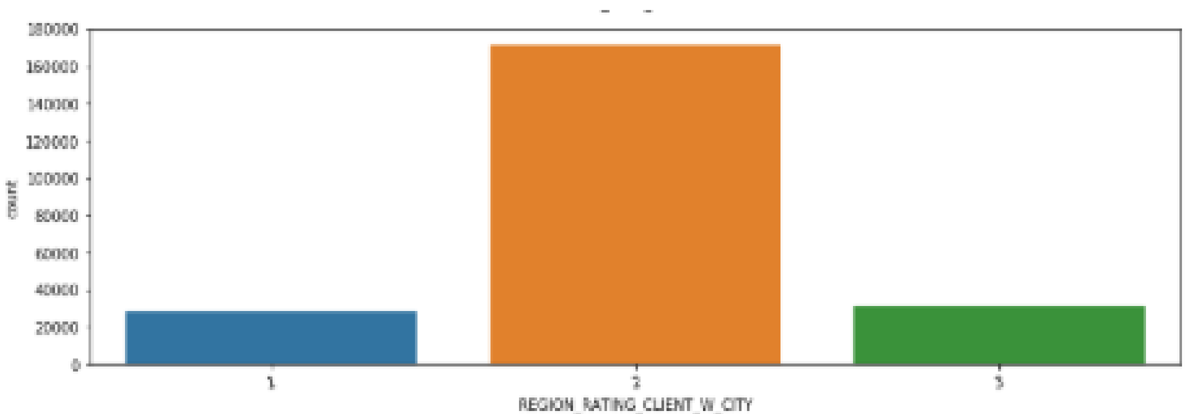


TARGET 1

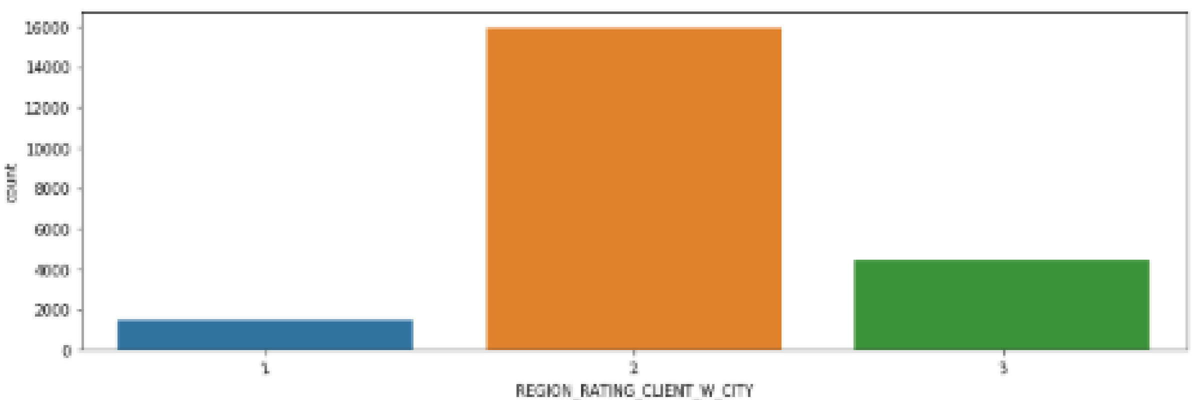
- From the "AMT_CREDIT" graph, it is visible that the people taking a loan amount of 10lakhs or less are more likely to make default.

CUSTOMER PROFILING

UNIVARIATE ANALYSIS OF CATEGORICAL VARIABLE



TARGET 0



TARGET 1

➢ From "REGION_RATING_CLIENT_W_CITY" graph it is visible that people belonging to the Tier-3 city are most likely to make a default.

CUSTOMER PROFILING

BIVARIATE ANALYSIS OF CONTINUOUS-CONTINUOUS

	VAR1	VAR2	Correlation
96	AMT_CREDIT	AMT_GOODS_PRICE	0.981005
13	CNT_FAM_MEMBERS	CNT_CHILDREN	0.893278
135	AMT_ANNUITY	AMT_GOODS_PRICE	0.766939
137	AMT_ANNUITY	AMT_CREDIT	0.760092
136	AMT_ANNUITY	AMT_INCOME_TOTAL	0.473806
83	AMT_INCOME_TOTAL	AMT_GOODS_PRICE	0.407687
97	AMT_CREDIT	AMT_INCOME_TOTAL	0.398691
165	DAYS_BIRTH	DAYS_EMPLOYED	0.352337
167	DAYS_BIRTH	DAYS_REGISTRATION	0.298951
156	DAYS_BIRTH	CNT_CHILDREN	0.242430

TARGET 0

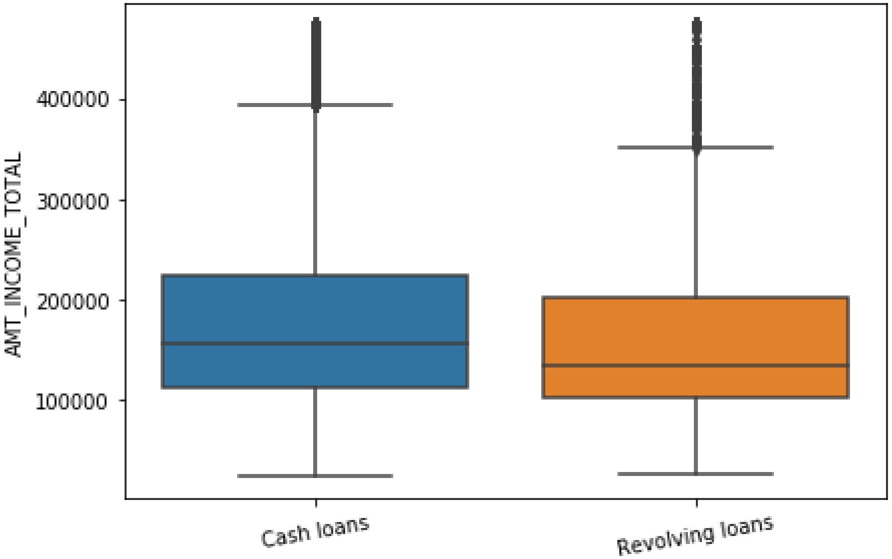
	VAR1	VAR2	Correlation
96	AMT_CREDIT	AMT_GOODS_PRICE	0.978765
13	CNT_FAM_MEMBERS	CNT_CHILDREN	0.893829
135	AMT_ANNUITY	AMT_GOODS_PRICE	0.749379
137	AMT_ANNUITY	AMT_CREDIT	0.748359
136	AMT_ANNUITY	AMT_INCOME_TOTAL	0.424363
83	AMT_INCOME_TOTAL	AMT_GOODS_PRICE	0.358797
97	AMT_CREDIT	AMT_INCOME_TOTAL	0.353931
165	DAYS_BIRTH	DAYS_EMPLOYED	0.307018
167	DAYS_BIRTH	DAYS_REGISTRATION	0.241202
163	DAYS_BIRTH	AMT_CREDIT	0.190989

TARGET 1

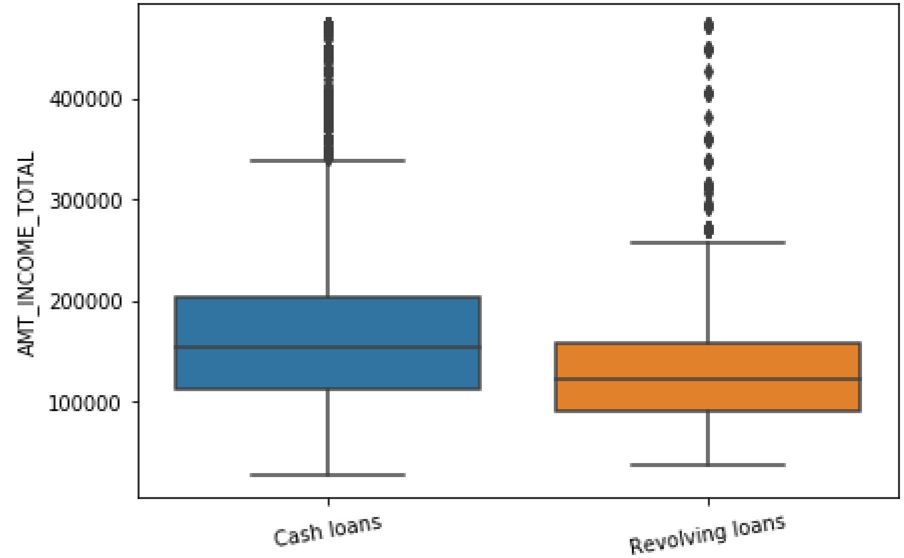
- The top 9 correlation chart of defaulters and non-defaulters are almost same both with respect to the type of variables and correlation coefficient which proves that the variables are independent of defaulting. If we inspect more into the correlation aspect, the top two variables represent very strong correlations and an increase in one variable would show similar increase in the other and vice versa. The next three variables are mildly correlated and can be considered depending on the context. Values less than 0.5 are not to be considered for correlation anyhow. Only the 10th combination is different.

CUSTOMER PROFILING

BIVARIATE ANALYSIS OF CONTINUOUS-CATEGORICAL VARIABLE



TARGET 0

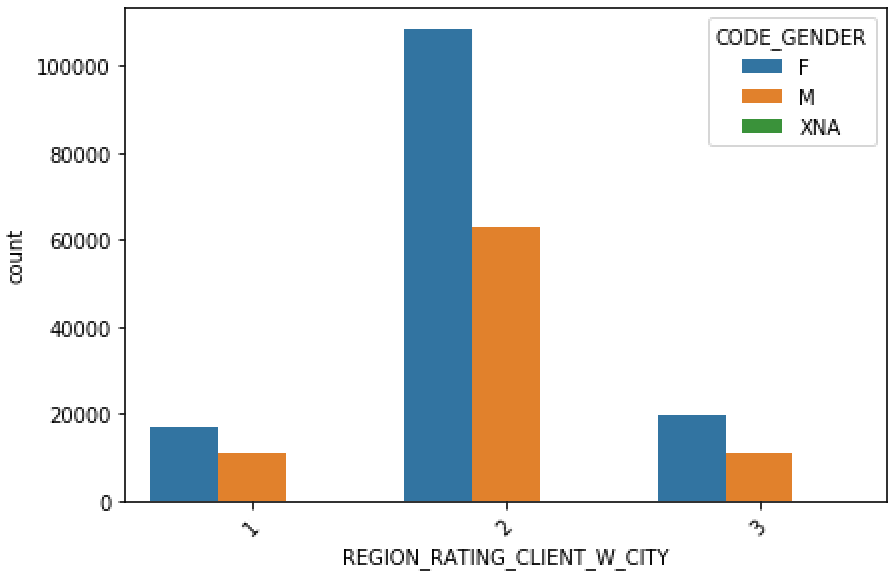


TARGET 1

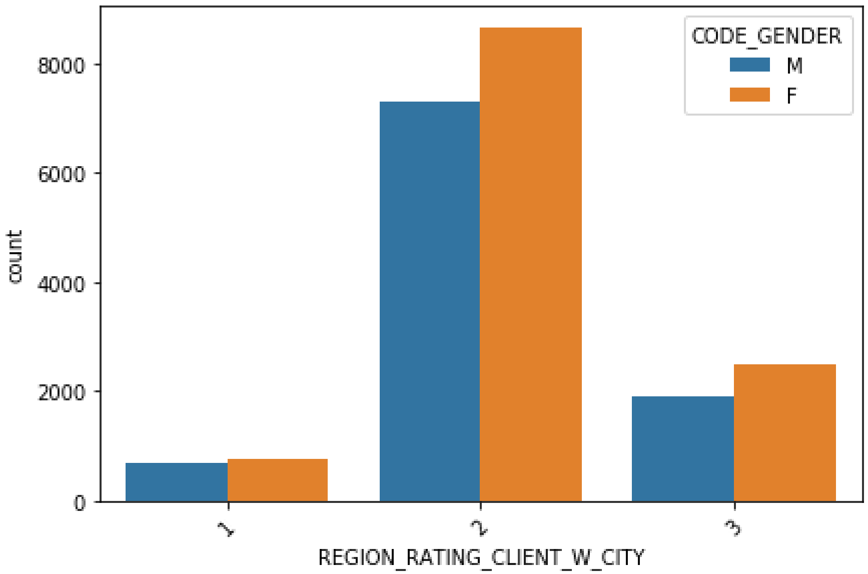
➢The box plot trends are different for defaulters and non-defaulters. For non defaulters, the box plot inter quartile range show a greater variability, and right skewed, which essentially means that people earning greater than the median salary are more in no. as compared to people earning lesser than the median salary. For defaulters the inter quartile region seems to be balanced. The logic seems to be right as per us since there are more people who earn more than the median salary. The chances of defaulting goes down.

CUSTOMER PROFILING

BIVARIATE ANALYSIS OF CATEGORICAL-CATEGORICAL VARIABLE



TARGET 0

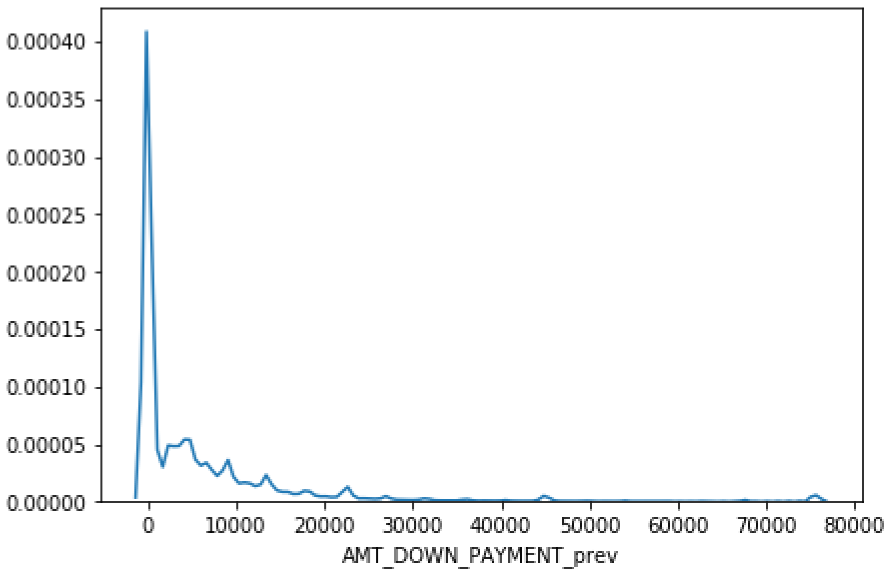


TARGET 1

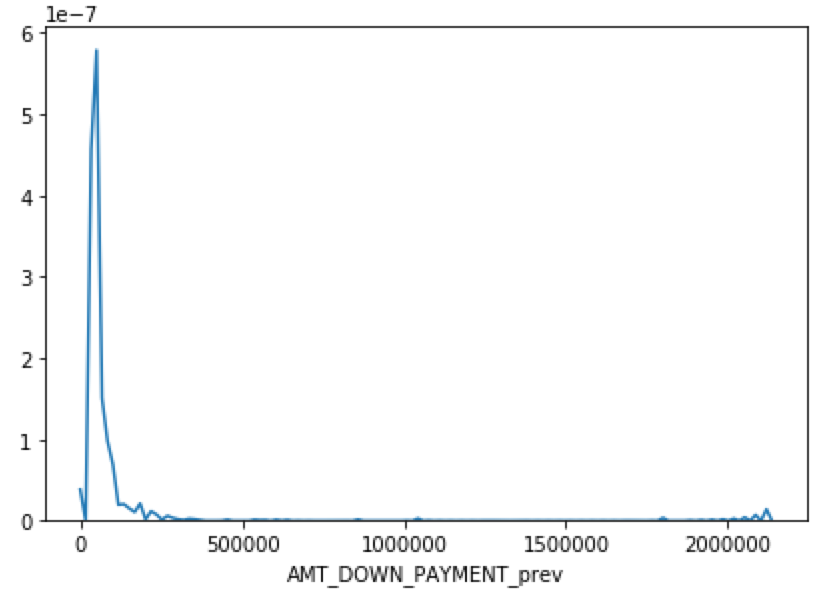
- In all the Rated cities, female are less likely to default, and if seen Tier-2 city females are not likely to default at all.

CUSTOMER PROFILING

UNIVARIATE ANALYSIS OF CONTINUOUS VARIABLE OF **MERGED DATA**



MERGED TARGET 0

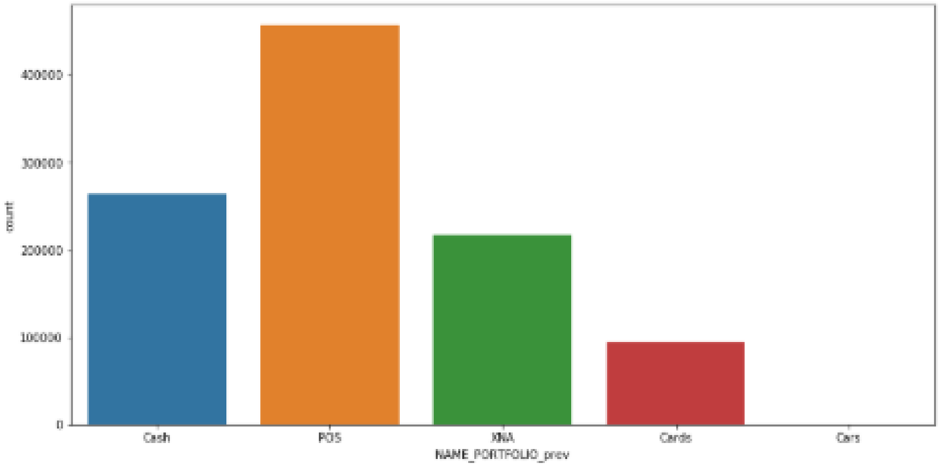


MERGED TARGET 1

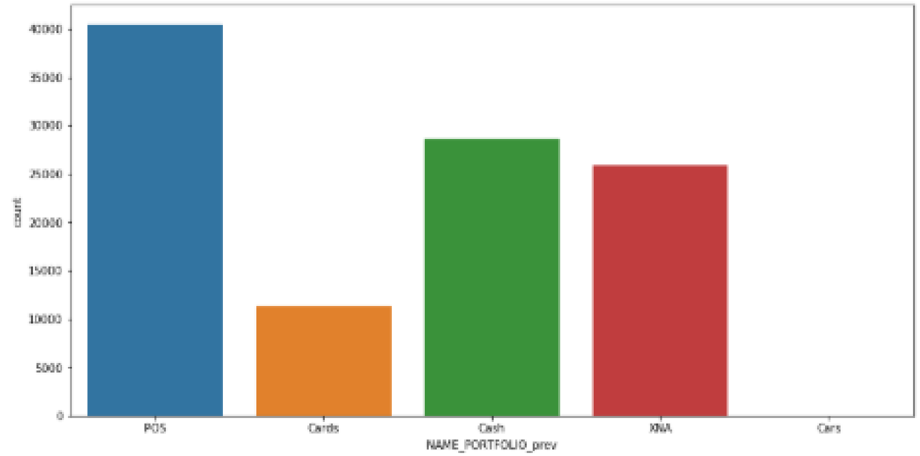
➢From "AMT_DOWN_PAYMENT_prev" graph it is visible that people making more down payment in the previous applications are likely to default.

CUSTOMER PROFILING

UNIVARIATE ANALYSIS OF CONTINUOUS VARIABLE OF MERGED DATA



MERGED TARGET 0



MERGED TARGET 1

➢ From "NAME_PORTFOLIO_prev" graph, it is visible that the previous application for cash and cards is more for the defaulters.

CUSTOMER PROFILING

BIVARIATE ANALYSIS OF CONTINUOUS-CONTINUOUS VARIABLE OF MERGED DATA

	VAR1	VAR2	Correlation
14	AMT_CREDIT_prev	AMT_APPLICATION_prev	0.967011
8	AMT_APPLICATION_prev	AMT_DOWN_PAYMENT_prev	0.356535
12	AMT_CREDIT_prev	AMT_DOWN_PAYMENT_prev	0.223065
13	AMT_CREDIT_prev	RATE_INTEREST_PRIMARY_prev	0.152969
9	AMT_APPLICATION_prev	RATE_INTEREST_PRIMARY_prev	0.138287
4	RATE_INTEREST_PRIMARY_prev	AMT_DOWN_PAYMENT_prev	0.016289

MERGED TARGET 0

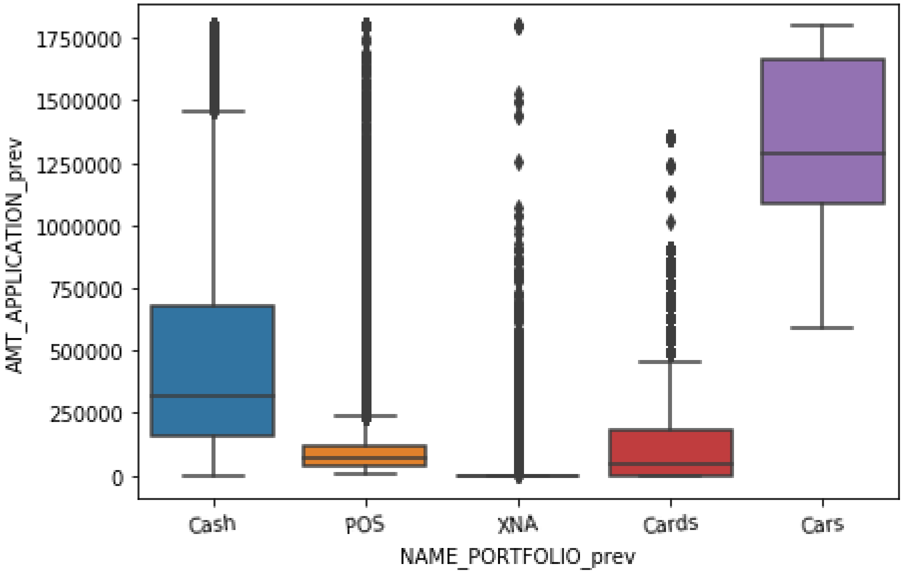
	VAR1	VAR2	Correlation
14	AMT_CREDIT_prev	AMT_APPLICATION_prev	0.966235
8	AMT_APPLICATION_prev	AMT_DOWN_PAYMENT_prev	0.394201
12	AMT_CREDIT_prev	AMT_DOWN_PAYMENT_prev	0.281301
13	AMT_CREDIT_prev	RATE_INTEREST_PRIMARY_prev	0.155668
9	AMT_APPLICATION_prev	RATE_INTEREST_PRIMARY_prev	0.121843
4	RATE_INTEREST_PRIMARY_prev	AMT_DOWN_PAYMENT_prev	0.000922

MERGED TARGET 1

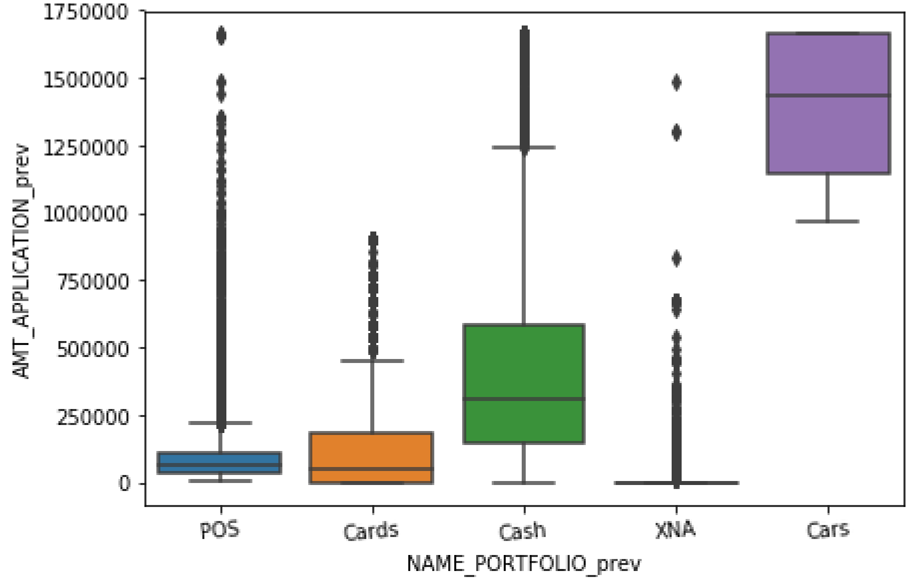
- Comparing the two tables, it can be seen that for both defaulters and non defaulters the correlation between the amount asked in the application by the customer and the amount sanctioned by the credit agency is the highest, which means that the change in one variable will highly affect the change in the other. All others in the table can be neglected as the correlation value is very low.

CUSTOMER PROFILING

BIVARIATE ANALYSIS OF CONTINUOUS-CATEGORICAL VARIABLE OF MERGED DATA



MERGED TARGET 0

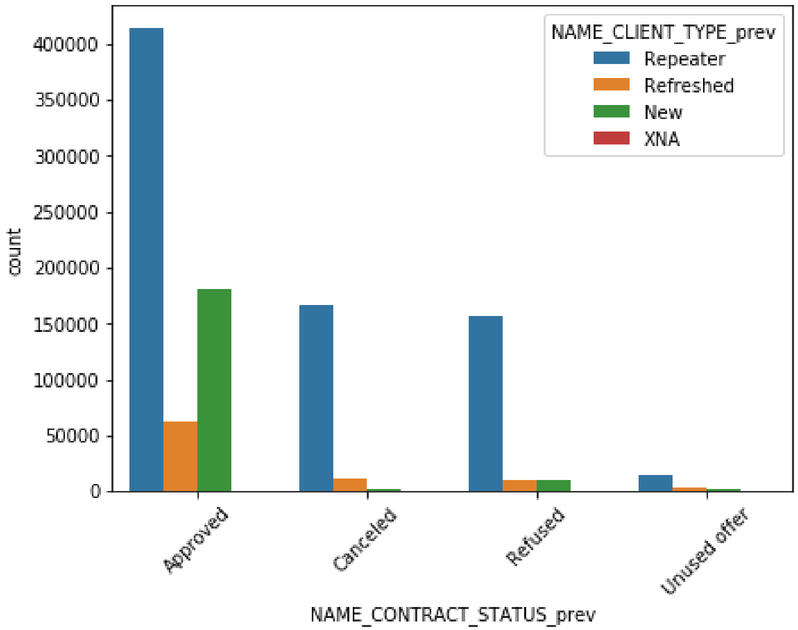


MERGED TARGET 1

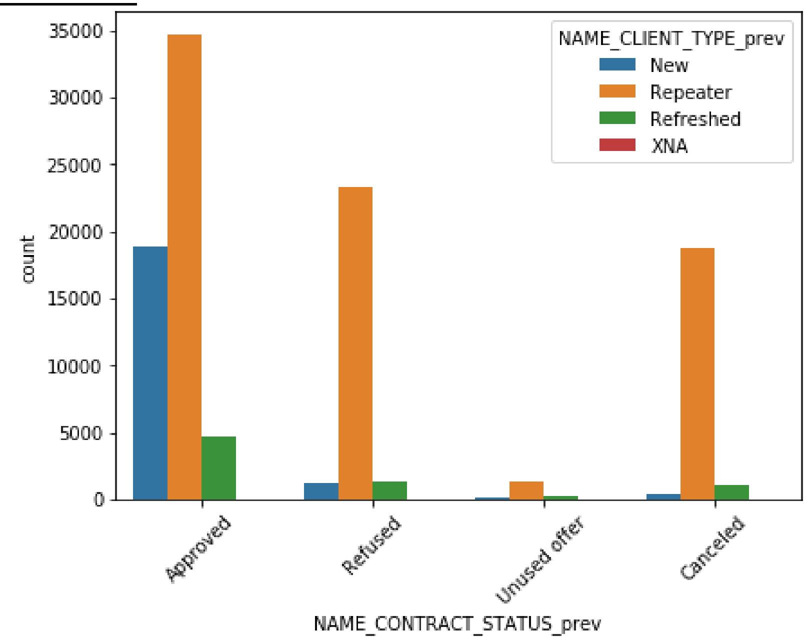
➢ For all the categories of portfolio it is same in both the cases except for the portfolio of car category. Defaulters having cars in the portfolio category have a higher median value and minimum value and it is also visible that they have applied for loan amount lesser than the median value in the application.

CUSTOMER PROFILING

BIVARIATE ANALYSIS OF CATEGORICAL-CATEGORICAL VARIABLE OF MERGED DATA



MERGED TARGET 0



MERGED TARGET 1

➢ People whose previous application have been 'Refused' or 'Cancelled' and were Repeaters are more likely to default. People whose previous application have been 'Approved' and were New customers are more likely to default.

MAJOR RECOMMENDATIONS

While extending or cancelling a loan application, **credit history of the applicant** is of utmost importance**. *Even if a credit application for a customer has been approved in the past, a customer may default***. So the **customer demographics** play a prime role here. To avoid a default, major recommendations for the business are to verify the following types of customers:

- Male customers in Tier 3 cities and applying for loans of less than 10 lakhs
- Customers making more down payment in the previous loan applications
- Customers who applied for cards in the previous applications
- Customers who were 'Repeaters' and were 'Refused' in previous applications

INFERENCE

```
TARGET   NAME_CONTRACT_STATUS_prev
0        Approved                         0.576803
         Canceled                         0.159182
         Refused                          0.154019
         Unused offer                     0.016725
1        Approved                         0.051135
         Canceled                         0.017790
         Refused                          0.022787
         Unused offer                     0.001559
Name: NAME_CONTRACT_STATUS_prev, dtype: float64
```

- **15.4% of the sample who were refused loans previously turned out to be Non-Defaulters, where as 5.1% of the sample who were approved loans previously turned out to be defaulters.**
- **This is the loss to the financial company that can result from *refusal of good loans* (15.4%) and *approval of bad loans*(5.1%). So as a company our objective should be to reduce these losses as much as possible.**

THANK YOU!