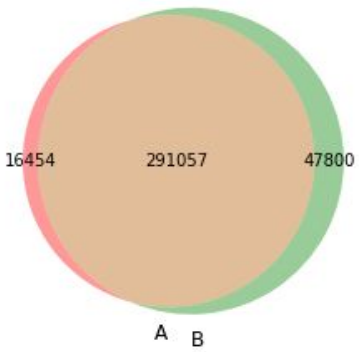
EDA CASE STUDY

Please Note

- ONLY the important/few of the insights are added to the presentation, all the analysis performed has been explained in the notebook in detail.
- Analysis on the Numerical features has been done by Binning these features by user defined functions for further analysis and plotting.
- Features of data sets have also been analysed by segmenting both the data sheets into ‘Defaulters’ & ‘Non Defaulters’ (TARGET 1/0) category. User defined variables would also be explained in the notebook for better clarity.
- New features/variable created would be explicitly briefed in the notebook, all predefined features have been understood as per the ‘Column_description’ data sheet.

Data Quality Summary

- 307511 unique current applications
- 8.8 % of defaulters - High class imbalance!
- Not all the records in current applications have a corresponding history.
- Hence, for the analysis of previous application data, we will only consider current applicants data

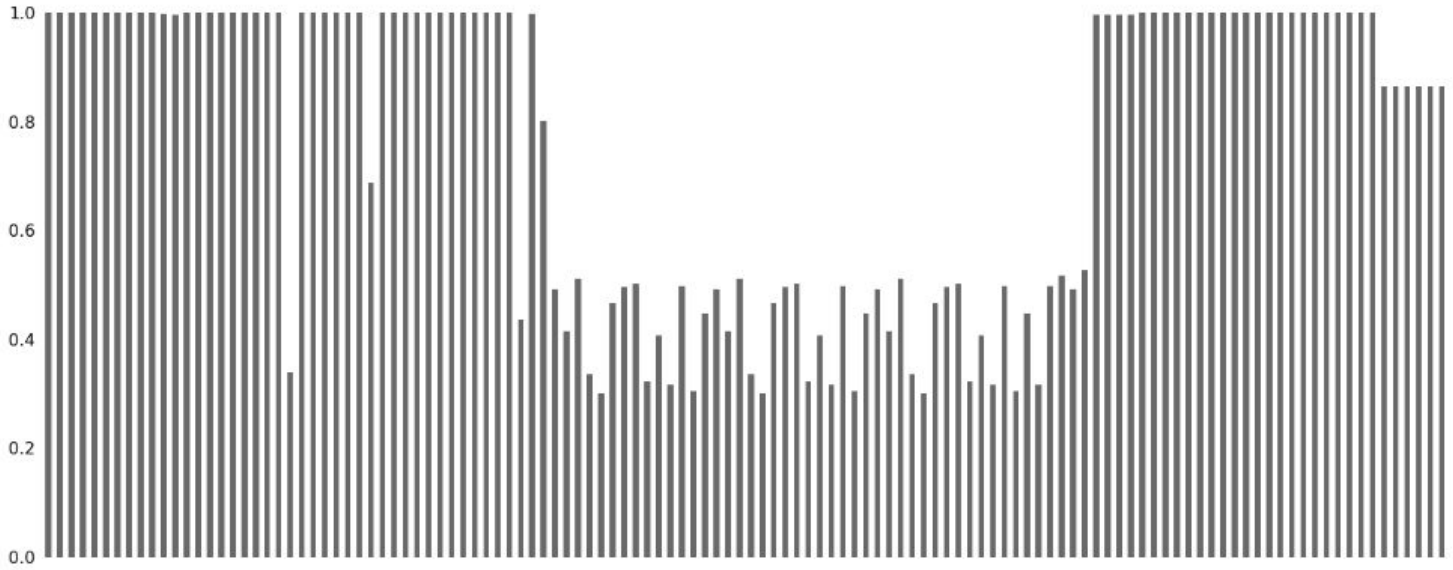


Data Types Check:

- For Previous application : All the data types of the columns are appropriate to the context and do not need any further manipulation except SK_ID_PREV, NFLAG_INSURED_ON_APPROVAL and few other flag columns
- For Current application : As the SK_ID_CURR and TARGET will be treated as a categorical data and not numerical, converting the same.

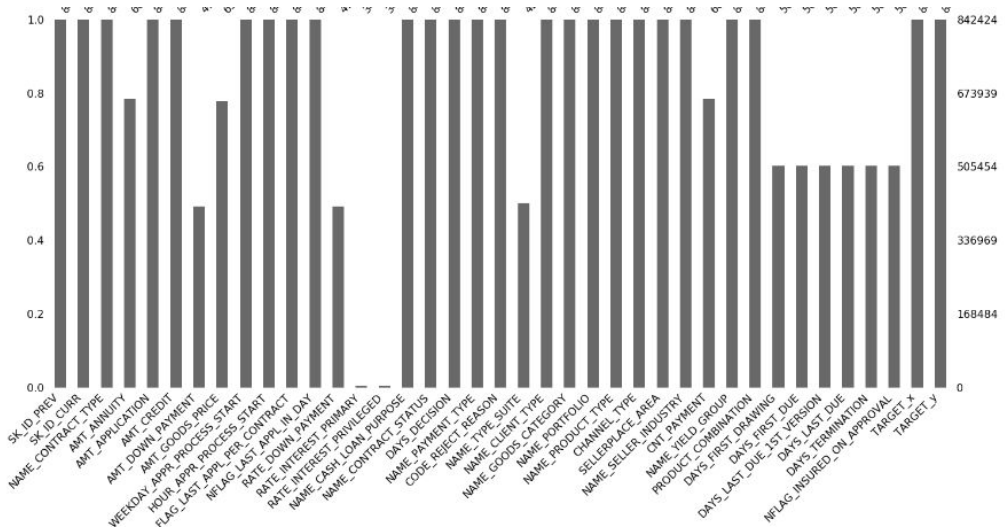
Overview on Missing Values

- Current Application
  - We ideally should remove the columns which has more than 70% values
  - OWN_CAR_AGE Column has 66% missing values and the remaining can be imputed with zero as it can be assumed that blank cells are people with no car.
  - EXT_SOURCE_1 column has 56% missing value and can be imputed with value 0, because we can assume that there is no value from that source.
  - LANDAREA_MODE column has 60% missing value which can be imputed by taking the mean of the remaining values.
  - HOUSETYPE_MODE column has 51% missing value and can be imputed with mode of the remaining values.



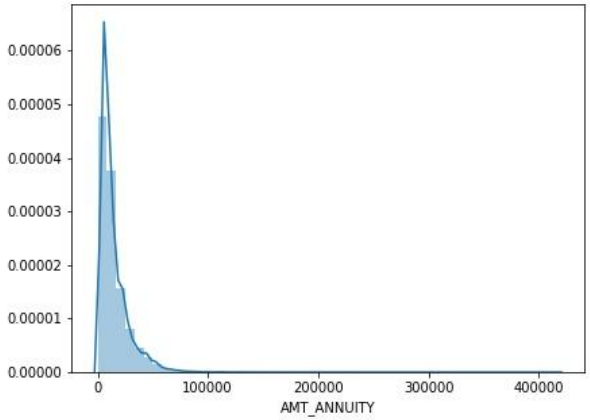
Overview on Missing Values

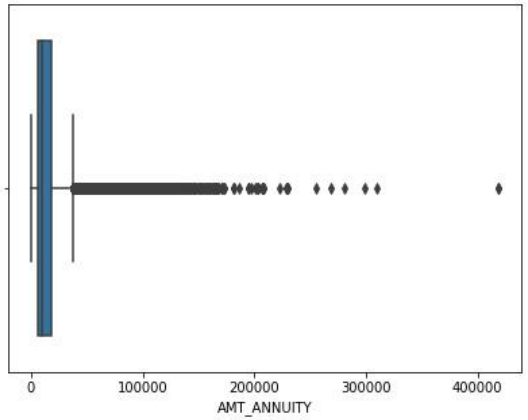
- Previous Application
  - NAME_TYPE_SUITE is missing for 49% of the rows and this is a categorical variable which can be ideally imputed with NA instead of imputing with a mode value as 49% is a high value.
  - Variables DAYS_FIRST_DRAWING, DAYS_FIRST_DUE, DAYS_LAST_DUE_1ST_VERSION, DAYS_LAST_DUE, DAYS_TERMINATION have around 40% missing values and can be imputed with mode as they seem to have discrete distributions and imputing them with mean would not be appropriate
  - NFLAG_INSURED_ON_APPROVAL can be imputed with 0 meaning that the loan was not insured if there is not information provided on this column
  - CNT_PAYMENT can be imputed with the mean value as it will give an idea of the most common term that is opted for. It is not required to impute it with median as the data does not seem to have few extreme values as such in this column



Outliers Analysis

- Current Application:
  - Outliers can be removed in few of the columns: 'AMT_INCOME_TOTAL','AMT_CREDIT' 'REGION_POPULATION_RELATIVE','DAYS_EMPLOYED'
- Previous Application:
  - In this use case, there are possibilities of having valid and rare extreme values in columns like AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE etc. So in such cases, an outlier treatment would simply mean loss of data, So we keep them as is.
  - But in case of columns like HOUR_APPR_PROCESS_START, RATE_DOWN_PAYMENT, DAYS_FIRST_DUE, DAYS_LAST_DUE, DAYS_FIRST_DUE, DAYS_LAST_DUE_1ST_VERSION, DAYS_TERMINATION there are extreme values which can be regarded as invalid cases and hence discarded or we can also replace those instances with such invalid values with the mean/median.





Please note: Distribution plots for the other variables mentioned in this analysis can be found in the distributions folder

Outliers Analysis contd.

- For e.g., in HOUR_APPR_PROCESS_START, it is impossible for this process to start in any non working hours of the facility and we see a lot of such cases where the hour of the application process is before 8 AM and seems mostly implausible.
- In case of DAYS_LAST_DUE_1ST_VERSION, and other relative dates, there are few instances with values of the range 365243 which is an impossible case. It's is not possible to have a previous application for the same customer which dates back to (365243 / 365) = 1000 years old!
- We will cap the instances in columns DAYS_FIRST_DUE & DAYS_LAST_DUE_1ST_VERSION which have invalid values as the number of such instances seem to be less. We can remove the column DAYS_FIRST_DRAWING as it mostly seems to have invalid value

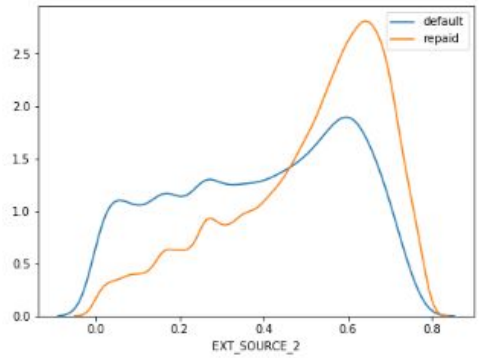
Feature Creation

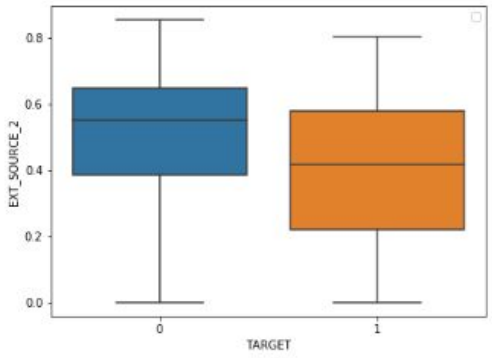
Following features have been created:

1. credit_application_ratio: ratio of the loan amount actually credited VS the amount the customer actually applied for
2. goods_price_credit_ratio: ratio of the cost of the goods against which the loan is sanctioned
3. annuity_credit_ratio: precentage of the annuity amount that the customer pays periodically against the loan amount
4. approved_credit_flag: flag depicting if the loan was approved

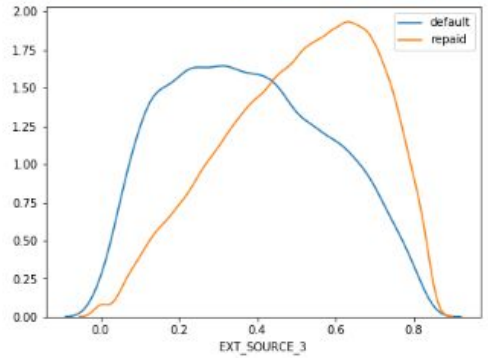
Univariate Analysis

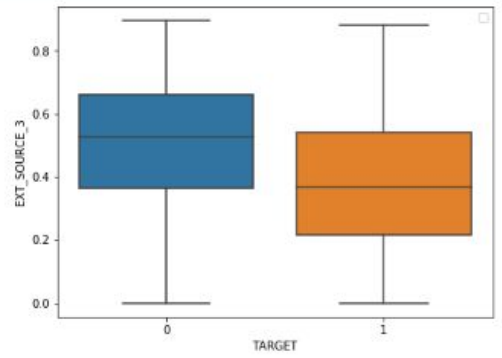
- The distribution of the values is very similar in the 2 cases of the target values except in cases in EXT_SOURCE_2 & EXT_SOURCE_3 which do mean that these 3 variables are able to differentiate between the defaulters & non defaulters
- We do see a slight difference in the median of the CNT_PAYMENT column. It depicts that there a slightly higher chance that the customer will default in the loan of the CNT_PAYMENT is higher

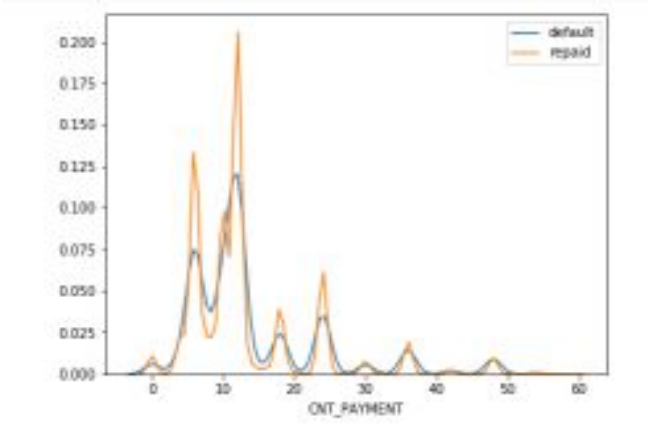


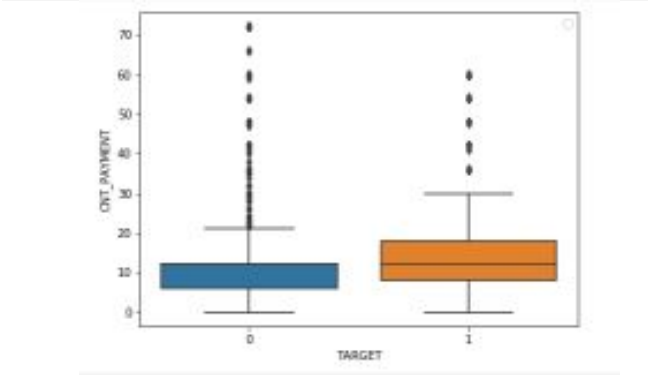



```
No handles with labels found to put in legend.
```



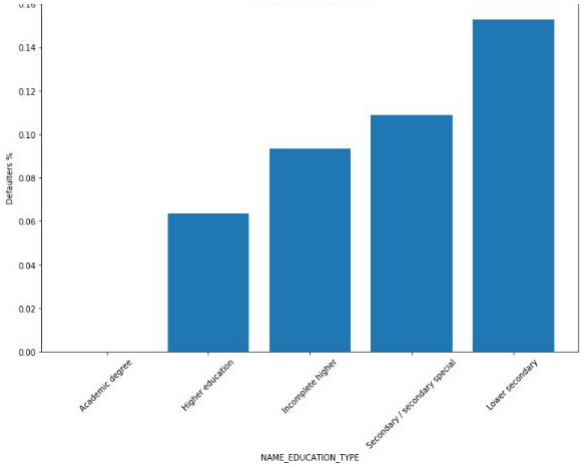


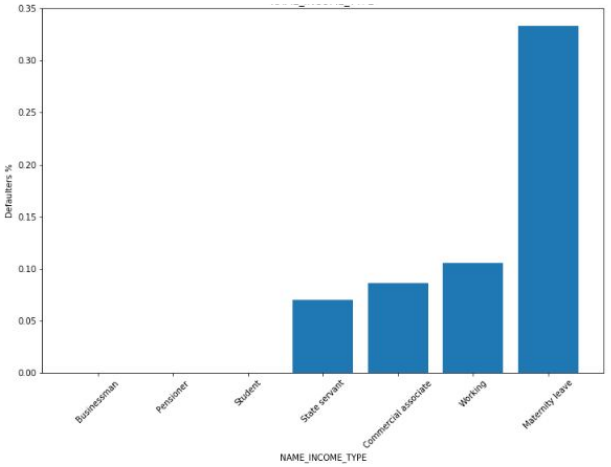


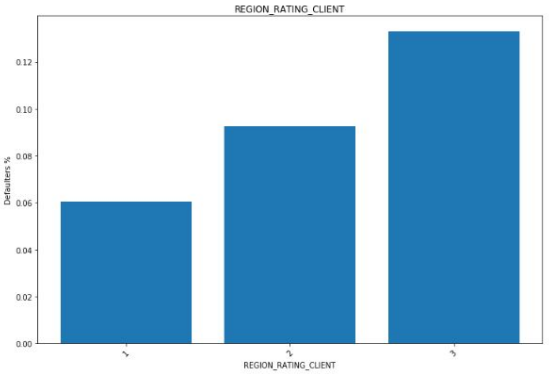


Univariate Analysis

- Around 10 % of the current applicants with Lower secondary education were defaulter, highest amongst the education type.
- Amongst the employment category 'Maternity leave' and 'Unemployed' showed the highest defaulter rates of 40% and 35% respectively.
- Amongst the region rating, region rating of 3 had the maximum % of defaulter, >10 % * There is a significantly high number of defaulters in the income group of 300000 - 650000
- Loans with annuity < 60,000 have a relatively big chunk of the defaulters

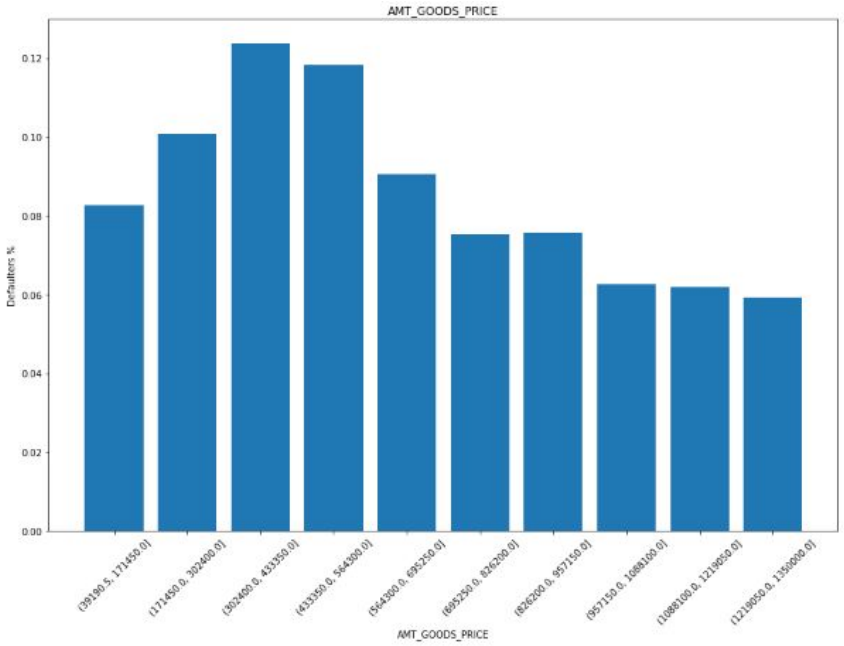


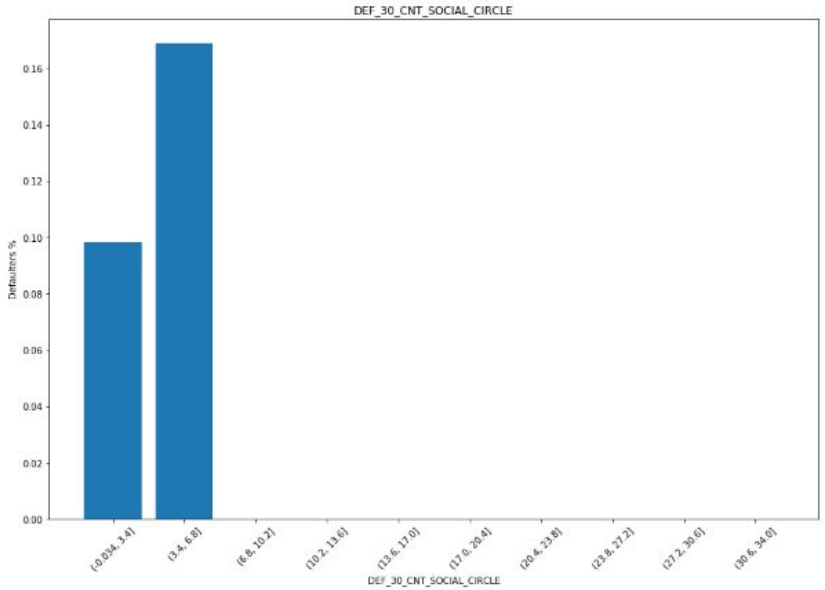




Univariate Analysis

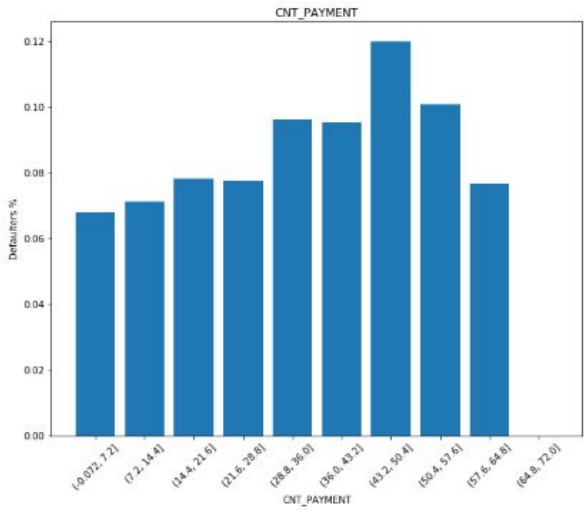
- Loans sanctioned against good priced 3,00,000 - 5,50,000 have a relatively high default rate
- Customers with 3-7 people defaulting in their social circle also have a high chance of defaulting

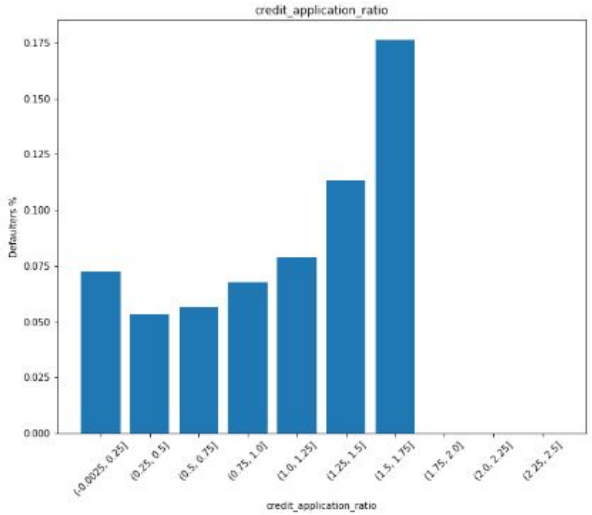


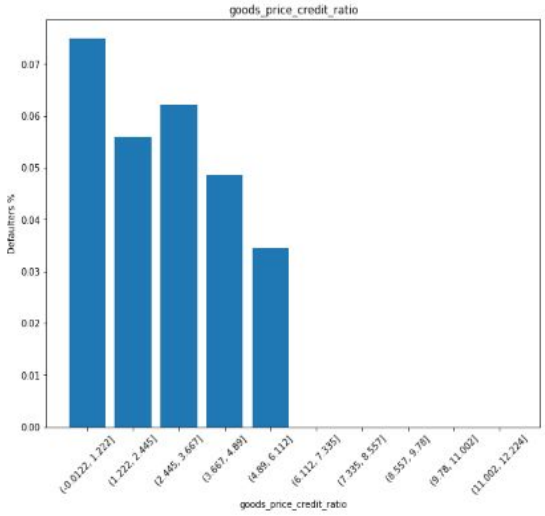


Univariate Analysis

- As the annuity_credit_ratio increases, that is, when the annuity amount with respect to the credit amount increases, the proportion of defaulters decreases.
- We can observe here that typically, any loans with terms between 28-57 have relatively higher default ratios
- We can observe that whenever credit_application_ratio is higher than 1, the default rates significantly go up. That means that if a loan amount that is sanctioned is higher than the amount the customer actually applied for, there are higher chances that they might default
- Here we can see that as the good price to credit amount ratio increases, the percentage of defaulters reduces. That is, if the goods prices is very close to the credit amount or lesser than the credit amount, then there is a relatively higher chances of default, rather than the cases where the goods price is much higher than that of the credit amount







Correlation Analysis (Previous Application)

- The correlations of Target =1 and Target =0 are almost identical

	index	variable	value	key
1682	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998229	OBS_30_CNT_SOCIAL_CIRCLE_OBS_60_CNT_SOCIAL_CIRCLE
122	AMT_GOODS_PRICE	AMT_CREDIT	0.976398	AMT_CREDIT_AMT_GOODS_PRICE
1021	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.951423	REGION_RATING_CLIENT_REGION_RATING_CLIENT_W_CITY
16	CNT_FAM_MEMBERS	CNT_CHILDREN	0.895167	CNT_CHILDREN_CNT_FAM_MEMBERS
1742	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.869188	DEF_30_CNT_SOCIAL_CIRCLE_DEF_60_CNT_SOCIAL_CIRCLE
1261	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.832163	LIVE_REGION_NOT_WORK_REGION_REG_REGION_NOT_WOR...
1441	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.752657	LIVE_CITY_NOT_WORK_CITY_REG_CITY_NOT_WORK_CITY
181	AMT_GOODS_PRICE	AMT_ANNUITY	0.724443	AMT_ANNUITY_AMT_GOODS_PRICE
121	AMT_ANNUITY	AMT_CREDIT	0.724127	AMT_ANNUITY_AMT_CREDIT
2045	FLAG_DOCUMENT_8	FLAG_DOCUMENT_3	-0.626313	FLAG_DOCUMENT_3_FLAG_DOCUMENT_8

TARGET =1

	index	variable	value	key
1682	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998502	OBS_30_CNT_SOCIAL_CIRCLE_OBS_60_CNT_SOCIAL_CIRCLE
122	AMT_GOODS_PRICE	AMT_CREDIT	0.980563	AMT_CREDIT_AMT_GOODS_PRICE
1021	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.932369	REGION_RATING_CLIENT_REGION_RATING_CLIENT_W_CITY
16	CNT_FAM_MEMBERS	CNT_CHILDREN	0.895137	CNT_CHILDREN_CNT_FAM_MEMBERS
1742	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.867117	DEF_30_CNT_SOCIAL_CIRCLE_DEF_60_CNT_SOCIAL_CIRCLE
1261	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.837126	LIVE_REGION_NOT_WORK_REGION_REG_REGION_NOT_WOR...
1441	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.800954	LIVE_CITY_NOT_WORK_CITY_REG_CITY_NOT_WORK_CITY
181	AMT_GOODS_PRICE	AMT_ANNUITY	0.729398	AMT_ANNUITY_AMT_GOODS_PRICE
121	AMT_ANNUITY	AMT_CREDIT	0.727299	AMT_ANNUITY_AMT_CREDIT
2045	FLAG_DOCUMENT_8	FLAG_DOCUMENT_3	-0.551639	FLAG_DOCUMENT_3_FLAG_DOCUMENT_8

TARGET =0

Top 10 Correlation of Current application

Correlation Analysis (Previous Application)

	index	variable	value	key
21	AMT_GOODS_PRICE	AMT_APPLICATION	1.000000	AMT_APPLICATION_AMT_GOODS_PRICE
146	DAYS_FIRST_DUE	DAYS_DECISION	0.996114	DAYS_DECISION_DAYS_FIRST_DUE
38	AMT_GOODS_PRICE	AMT_CREDIT	0.989992	AMT_CREDIT_AMT_GOODS_PRICE
19	AMT_CREDIT	AMT_APPLICATION	0.976193	AMT_APPLICATION_AMT_CREDIT
183	DAYS_TERMINATION	DAYS_FIRST_DUE	0.961373	DAYS_FIRST_DUE_DAYS_TERMINATION
149	DAYS_TERMINATION	DAYS_DECISION	0.960481	DAYS_DECISION_DAYS_TERMINATION
200	DAYS_TERMINATION	DAYS_LAST_DUE_1ST_VERSION	0.912940	DAYS_LAST_DUE_1ST_VERSION_DAYS_TERMINATION
253	goods_price_credit_ratio	credit_application_ratio	-0.900242	credit_application_ratio_goods_price_credit_ratio
100	goods_price_credit_ratio	RATE_DOWN_PAYMENT	0.894009	RATE_DOWN_PAYMENT_goods_price_credit_ratio
181	DAYS_LAST_DUE_1ST_VERSION	DAYS_FIRST_DUE	0.878710	DAYS_FIRST_DUE_DAYS_LAST_DUE_1ST_VERSION

Defaulters

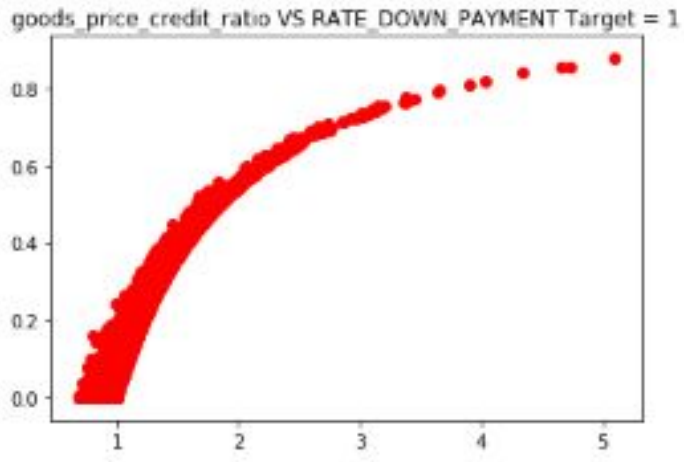
	index	variable	value	key
21	AMT_GOODS_PRICE	AMT_APPLICATION	1.000000	AMT_APPLICATION_AMT_GOODS_PRICE
146	DAYS_FIRST_DUE	DAYS_DECISION	0.996815	DAYS_DECISION_DAYS_FIRST_DUE
38	AMT_GOODS_PRICE	AMT_CREDIT	0.991589	AMT_CREDIT_AMT_GOODS_PRICE
19	AMT_CREDIT	AMT_APPLICATION	0.975601	AMT_APPLICATION_AMT_CREDIT
183	DAYS_TERMINATION	DAYS_FIRST_DUE	0.960005	DAYS_FIRST_DUE_DAYS_TERMINATION
149	DAYS_TERMINATION	DAYS_DECISION	0.959445	DAYS_DECISION_DAYS_TERMINATION
200	DAYS_TERMINATION	DAYS_LAST_DUE_1ST_VERSION	0.923407	DAYS_LAST_DUE_1ST_VERSION_DAYS_TERMINATION
100	goods_price_credit_ratio	RATE_DOWN_PAYMENT	0.893099	RATE_DOWN_PAYMENT_goods_price_credit_ratio
99	credit_application_ratio	RATE_DOWN_PAYMENT	-0.889696	RATE_DOWN_PAYMENT_credit_application_ratio
253	goods_price_credit_ratio	credit_application_ratio	-0.887009	credit_application_ratio_goods_price_credit_ratio

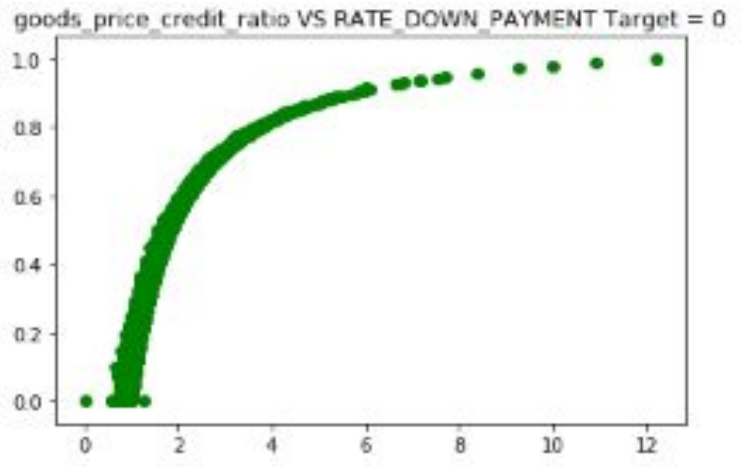
Non-Defaulters

Top 10 Correlation of Previous application

Correlation Analysis (Previous Application)

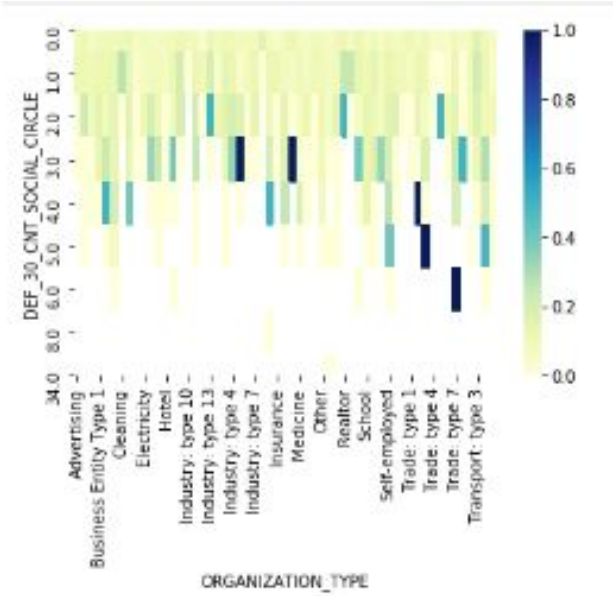
- As we can clearly see above, a lot of the amount related metrics like AMT_ANNUITY, AMT_CREDIT, AMT_APPLICATION, AMT_ANNUITY, and the features created from the same are correlated. The obvious reason for this is that these amounts are related to each other. Definitely, the loan amount credited will be dependant on the loan amount which the customer applied for and the annuity or the down payment(if any) will also be proportionate to the same. Also, the metrics derived from these variables will also be highly correlated.
- Similarly, all the relative dates related features are also highly correlated amongst each other
- Although, some of the below correlated pairs give some important inferences:
  - goods_price_credit_ratio, RATE_DOWN_PAYMENT 0.895156
  - credit_application_ratio RATE_DOWN_PAYMENT -0.893378
- Also, we can see that there are slight variations in the correlation values in the 2 different data sets (defaulters and non defaulters). For example, the pair goods_price_credit_ratio, RATE_DOWN_PAYMENT is slightly more correlated in the case of non defaulters which can also be seen in the graphs.

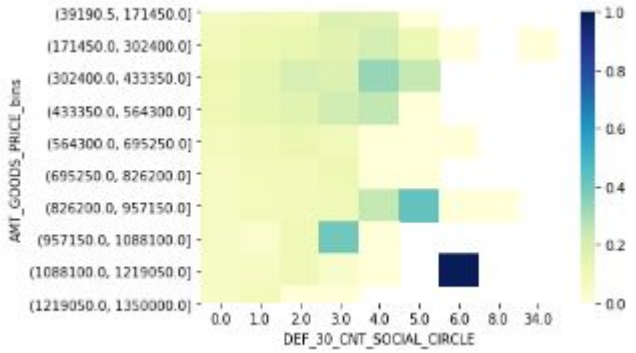


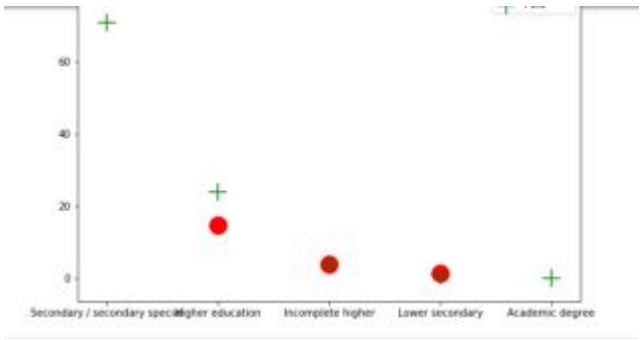


Bivariate Analysis

- Low skill labours with region rating = 3 are prone to being a defaulter
- Customers belonging to Industry type 4 and have around 3 people in their social circle defaulting up to 30 days are also a high risk customer
- Customers with 6 people defaulting in their social circle and the price of the goods being in the range 1088100-1219050 have an alarmingly high default rate compared to the rest of the segments

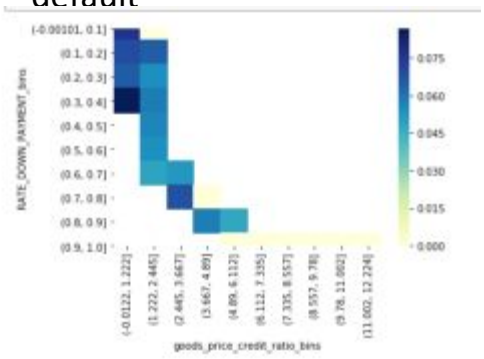


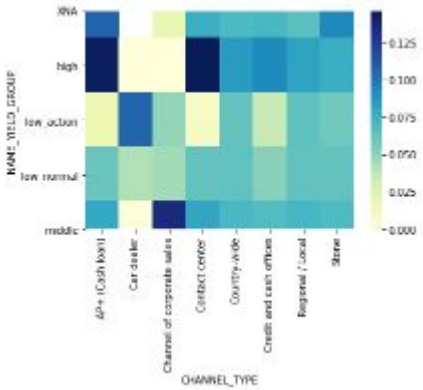


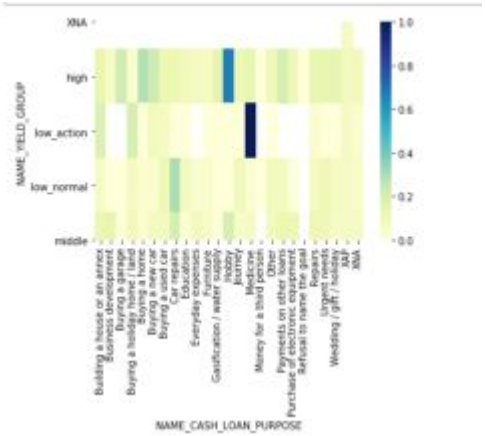


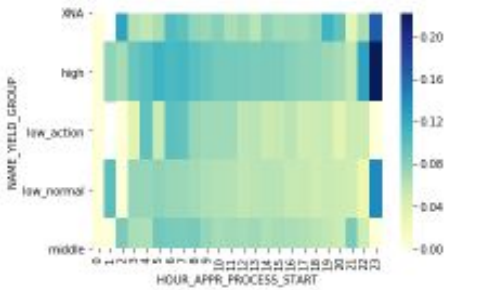
Bivariate Analysis

- The following pairs of loan attributes seem to be high in frequency
  - Medicine & low_action yield
  - Hobby and high yield
- High yield loans sanctioned through contact center or AP+(cash loans) have a high ratio and needs to be put in check
- The loans applied at odd hours - 11 PM and of high yield or low_normal yield have a high default rate
- Loans with low goods_price_credit_ratio and 0.3-0.4 downpayment % have a slightly higher chances of default









Conclusions

Below are some of the important insights:

1. EXT_SOURCE_2 & EXT_SOURCE_3 scores depict the behaviour of the defaulters to an extent
2. The number of people in the social circle of the customer who have defaulted proves to be an important factor
3. The distribution of the term of payment (CNT_PAYMENT) of the defaulters has slightly shifted to the higher end
4. Loans sanctioned against good priced 3,00,000 - 5,50,000 have a relatively high default rate
5. Around 10 % of the current applicants with Lower secondary education were defaulter, highest amongst the education type.
6. As the annuity_credit_ratio increases, that is, when the annuity amount with respect to the credit amount increases, the proportion of defaulters decreases.
7. Here we can see that as the good price to credit amount ratio increases, the percentage of defaulters reduces. That is, if the goods prices is very close to the credit amount or lesser than the credit amount, then there is a relatively higher chances of default, rather than the cases where the goods price is much higher than that of the credit amount
8. Customers with 6 people defaulting in their social circle and the price of the goods being in the range 1088100-1219050 have an alarmingly high default rate compared to the rest of the segments
9. High yield loans sanctioned through contact center or AP+(cash loans) have a high ratio and needs to be put in check
10. Low skill labours with region rating = 3 are prone to being a defaulter