Credit EDA Assignment

- Khantil Shah

Case Study Introduction

- This case study aims at making us understand the risk of losing money while lending. With data analytics and EDA we can get to know what are the risks and identify them so that an informed decision can be made. EDA in real scenario.
- Objectives-
  - EDA to understand how consumer attributes and loan attributes influence the tendency of default.
  - The company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.

Approach for the analysis

- For explorating analysis below mentioned steps were undertaken:
  - Importing libraries
  - Importing modules
  - Reading the datasets
  - Understanding the datasets (variables, columns, etc)
  - Describing the data – understanding features and datatype in both the dataframes
  - Shape of the dataset
  - Data Cleaning
    - Percentage of missing values in both datasets
    - Identifying missing values
    - Dropping missing values
    - Filling missing values
  - Exploring Numeric values
  - Merging required columns from both datasets
  - Handling the outliers
  - Checking Data Imbalance
  - Univariate/Bivariate analysis
  - Identifying value wise defaulter percentage
  - Correlation between Numeric features in previous application data
  - Analysis of categorical values
  - Observations and concluding remarks

Understanding the datasets (variables, columns, etc)

- application_df has 307511 rows and 122 columns
  - Out of which 121 has features and 1 is a target variable
    - 65 features are float64
    - 41 are integar
    - 16 are object datatype

- prev_app_df has 1670214 rows and 37 columns
  - Out of which
    - 15 features are float64
    - 6 are integars
    - 16 are object datatype

- Out of all SK_ID_CURR is the common and unique identifier and which will be later on used to merge in both dataframes

Numeric values extraction in prev_app_df

- Given below are the numeric values columns
  - 'AMT_ANNUITY', 'AMT_APPLICATION', 'AMT_CREDIT', 'AMT_DOWN_PAYMENT', 'AMT_GOODS_PRICE', 'RATE_DOWN_PAYMENT', 'CNT_PAYMENT', 'DAYS_FIRST_DRAWING', 'DAYS_FIRST_DUE', 'DAYS_LAST_DUE_1ST_VERSION', 'DAYS_LAST_DUE', 'DAYS_TERMINATION', 'NFLAG_INSURED_ON_APPROVAL’
  - Total 13 columns with float64 with nonnull counts.

Filling missing values with Median

- Checking columns in application.df have less null %
- AMT_ANNUITY has missing values and hence lets try and impute it by filling missing values by Median
- Same is done for
  - AMT_DOWN_PAYMENT
  - AMT_GOODS_PRICE
  - RATE_DOWN_PAYMENT
  - CNT_PAYMENT
  - DAYS_FIRST_DRAWING
  - DAYS_FIRST_DUE
  - DAYS_LAST_DUE_1ST_VERSION
  - DAYS_LAST_DUE
  - DAYS_TERMINATION
  - NFLAG_INSURED_ON_APPROVAL

Data Dropping

- In dataset prev_app_df there are missing values
  - Dropping RATE_INTEREST_PRIMARY, RATE_INTEREST_PRIVILEGED as they have over 90% of the data values missing
  - PRODUCT_COMBINATION and AMT_CREDIT won’t affect the analysis if dropped as the missing value % is very low. So it won’t be dropped.
  - So the columns will be reduced from 37 to 35 and rows 1670214 to 1669867
- In dataset application_df there are missing values
  - Dropping "NAME_TYPE_SUITE", "OWN_CAR_AGE", "OCCUPATION_TYPE", "EXT_SOURCE_1", "EXT_SOURCE_2", "EXT_SOURCE_3", "APARTMENTS_AVG", "BASEMENTAREA_AVG", "YEARS_BEGINEXPLUATATION_AVG", "YEARS_BUILD_AVG", "COMMONAREA_MODE", "ELEVATORS_MODE", "ENTRANCES_MODE", "FLOORSMAX_MODE", "FLOORSMIN_MODE", "LANDAREA_MODE", "LIVINGAPARTMENTS_MODE", "LIVINGAREA_MODE", "NONLIVINGAPARTMENTS_MODE", "NONLIVINGAREA_MODE", "APARTMENTS_MEDI", "BASEMENTAREA_MEDI", "BASEMENTAREA_MEDI", "YEARS_BEGINEXPLUATATION_MEDI", "YEARS_BEGINEXPLUATATION_MEDI", "YEARS_BUILD_MEDI", "COMMONAREA_MEDI", "ELEVATORS_MEDI", "ENTRANCES_MEDI", "FLOORSMAX_MEDI", "FLOORSMIN_MEDI", "LANDAREA_MEDI", "LIVINGAPARTMENTS_MEDI", "NONLIVINGAREA_MEDI", "FONDKAPREMONT_MODE", "HOUSETYPE_MODE", "TOTALAREA_MODE", "WALLSMATERIAL_MODE", "EMERGENCYSTATE_MODE", "COMMONAREA_AVG", "ELEVATORS_AVG", "ENTRANCES_AVG", "FLOORSMAX_AVG", "FLOORSMIN_AVG", "LANDAREA_AVG", "LIVINGAPARTMENTS_AVG", "NONLIVINGAREA_AVG", "APARTMENTS_MODE", "YEARS_BEGINEXPLUATATION_MODE", "YEARS_BUILD_MODE", "LIVINGAREA_MEDI", "NONLIVINGAPARTMENTS_MEDI“LIVINGAREA_AVG", "NONLIVINGAPARTMENTS_AVG", "BASEMENTAREA_MODE" to make dataset smaller and also the missing values are over 40%.
  - So the columns will be reduced to 69 from 122 earlier and rows will remain same at 307511.

Categorical columns with XNA

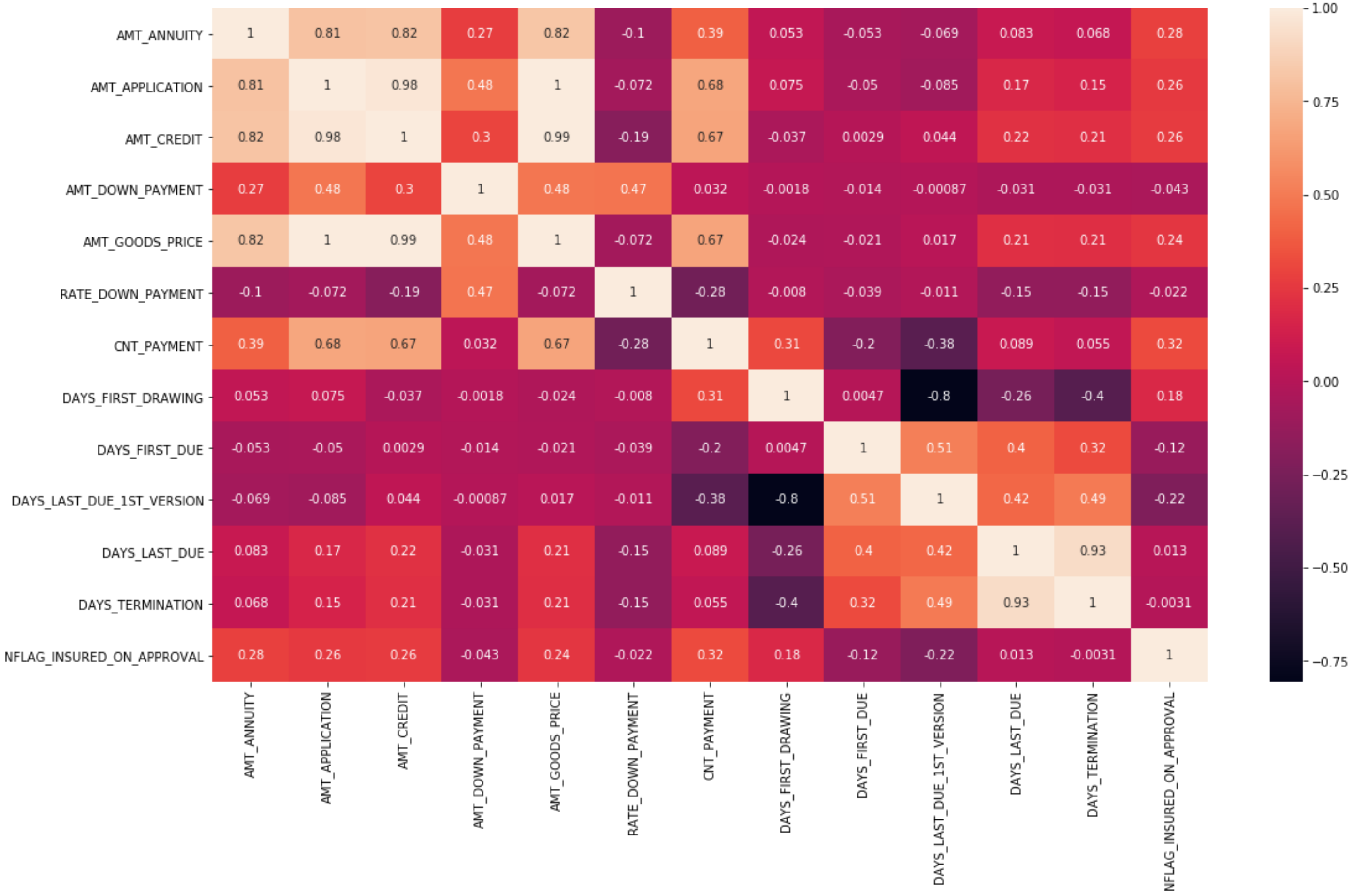
- Gender Column
- Organization Column

Merging relevant columns from 2 data frames

- Adding SK_ID_CURR & TARGET from application_df to prev_app_df
- Creating a new data frame NewDataMerge
    - With 35 columns & 1430155 having 17 float64, 2 int64, and 16 object features

10.46 is the data imbalance ratio.

Checking correlation of Numerica values in prev_app_df dataset



	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE	RATE_DOWN_PAYMENT	CNT_PAYMENT	DAYS_FIRST_DRAWING	DAYS_FIRST_DUE	DAYS_LAST_DUE_1ST_VERSION	DAYS_LAST_DUE	DAYS_TERMINATION	NFLAG_INSURED_ON_APPROVAL
AMT_ANNUITY	1	0.81	0.82	0.27	0.82	-0.1	0.39	0.053	-0.053	-0.069	0.083	0.068	0.28
AMT_APPLICATION	0.81	1	0.98	0.48	1	-0.072	0.68	0.075	-0.05	-0.085	0.17	0.15	0.26
AMT_CREDIT	0.82	0.98	1	0.3	0.99	-0.19	0.67	-0.037	0.0029	0.044	0.22	0.21	0.26
AMT_DOWN_PAYMENT	0.27	0.48	0.3	1	0.48	0.47	0.032	-0.0018	-0.014	-0.00087	-0.031	-0.031	-0.043
AMT_GOODS_PRICE	0.82	1	0.99	0.48	1	-0.072	0.67	-0.024	-0.021	0.017	0.21	0.21	0.24
RATE_DOWN_PAYMENT	-0.1	-0.072	-0.19	0.47	-0.072	1	-0.28	-0.008	-0.039	-0.011	-0.15	-0.15	-0.022
CNT_PAYMENT	0.39	0.68	0.67	0.032	0.67	-0.28	1	0.31	-0.2	-0.38	0.089	0.055	0.32
DAYS_FIRST_DRAWING	0.053	0.075	-0.037	-0.0018	-0.024	-0.008	0.31	1	0.0047	-0.8	-0.26	-0.4	0.18
DAYS_FIRST_DUE	-0.053	-0.05	0.0029	-0.014	-0.021	-0.039	-0.2	0.0047	1	0.51	0.4	0.32	-0.12
DAYS_LAST_DUE_1ST_VERSION	-0.069	-0.085	0.044	-0.00087	0.017	-0.011	-0.38	-0.8	0.51	1	0.42	0.49	-0.22
DAYS_LAST_DUE	0.083	0.17	0.22	-0.031	0.21	-0.15	0.089	-0.26	0.4	0.42	1	0.93	0.013
DAYS_TERMINATION	0.068	0.15	0.21	-0.031	0.21	-0.15	0.055	-0.4	0.32	0.49	0.93	1	-0.0031
NFLAG_INSURED_ON_APPROVAL	0.28	0.26	0.26	-0.043	0.24	-0.022	0.32	0.18	-0.12	-0.22	0.013	-0.0031	1

What we can infer from this heatmap is that¶

1. High Correlation between

a) AMT_ANNUITY,AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE b) DAYS_LAST_DUE, DAYS_TERMINATION
c) AMT_GOODS_PRICE, AMT_CREDIT

2. High Negative Correlation between

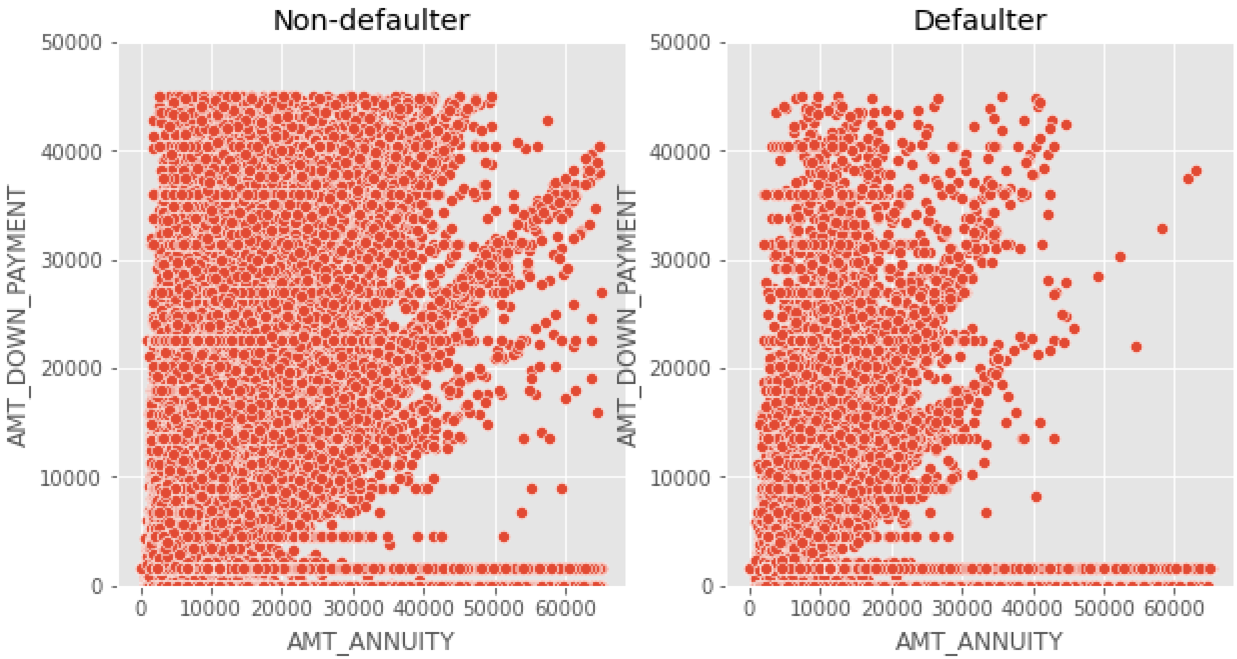
a) DAYS_FIRST_DRAWING, DAYS_LAST_DUE_1ST_VERSION

3. Somewhat Correlated

a) CNT_PAYMENT, AMT_GOODS_PRICE, AMT_CREDIT, AMT_APPLICATION

Analysis of Numeric Features of Previous Application Data

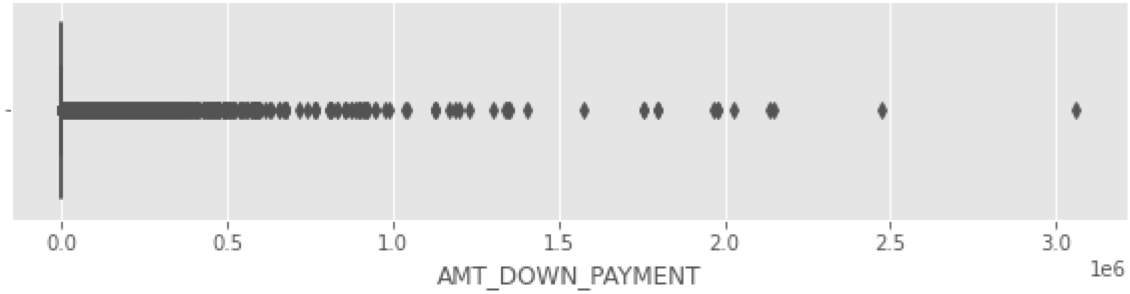
Numeric features of previous app data

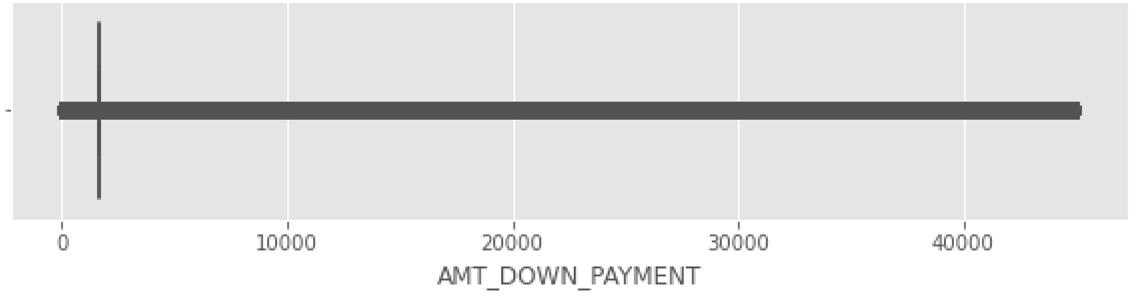


The number of defaulters are less for larger amount of annuity of previous application.
Where there is higher down payment, defaulter cases are less.

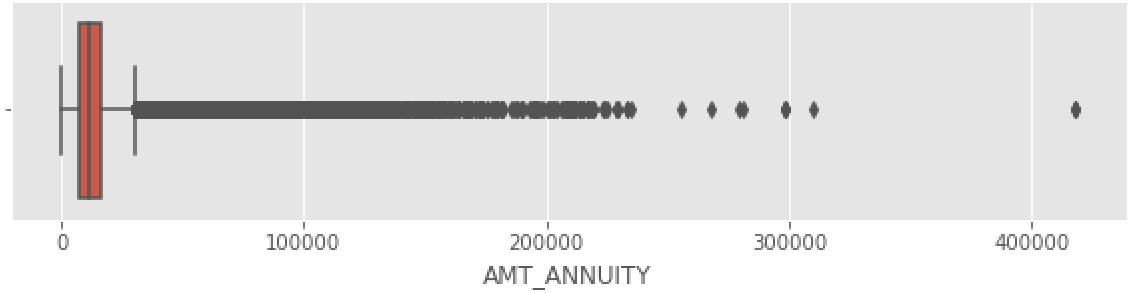
Handling Outliers

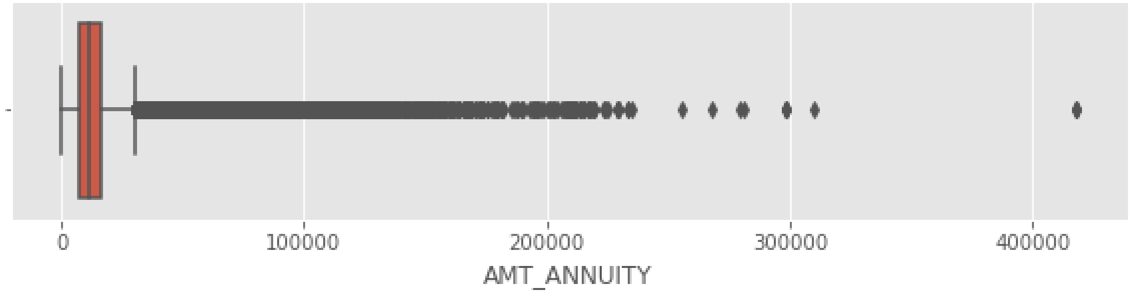
1. AMT_DOWN_PAYMENT





2. AMT_Annuity

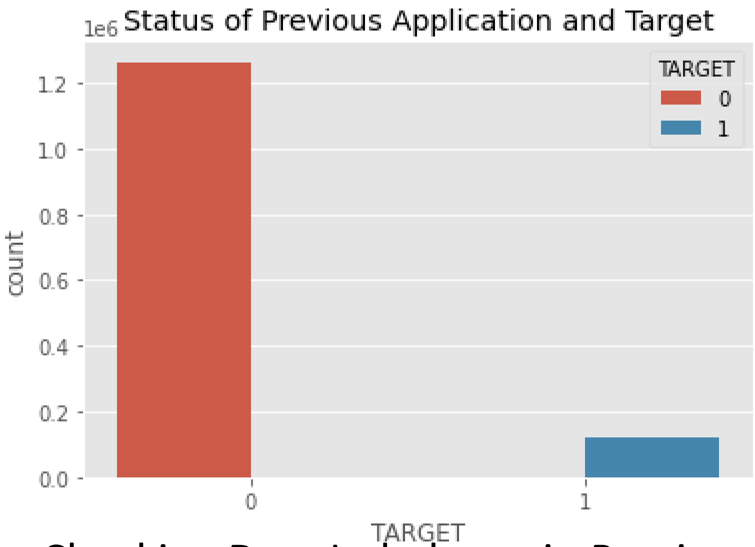




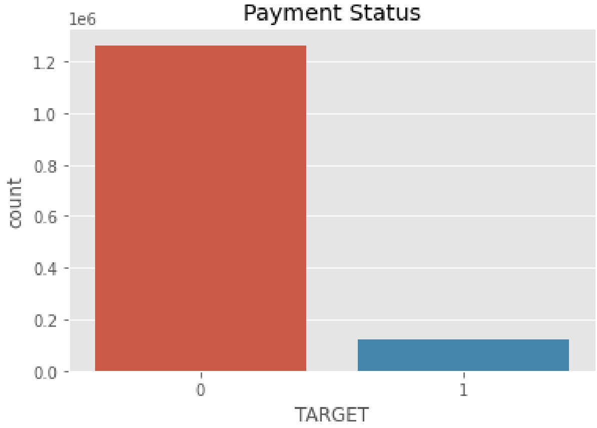
% of applicants

- % of applicants approved previously that defaulted in current loan ; 7.665149896341809
- % of applicants refused previously but paid current loan ; 87.88653859839336

The applicants whose previous loans were approved are more likely to pay current loan in time, than the applicants whose previous loans were rejected.

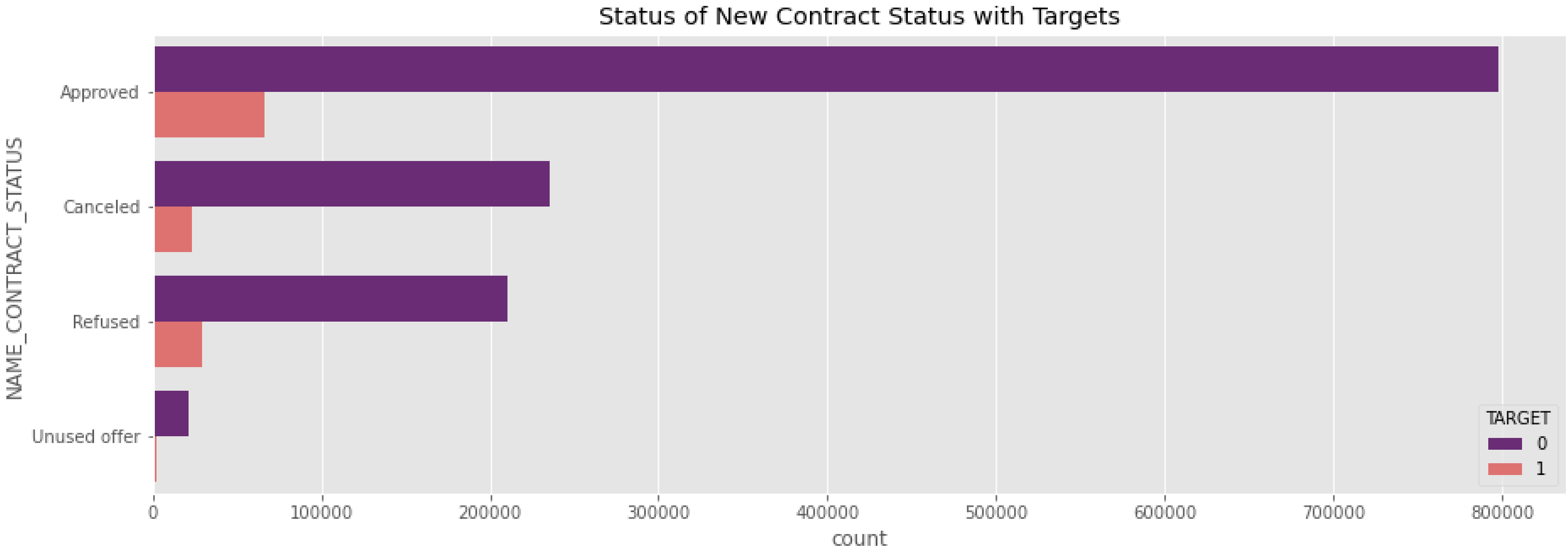


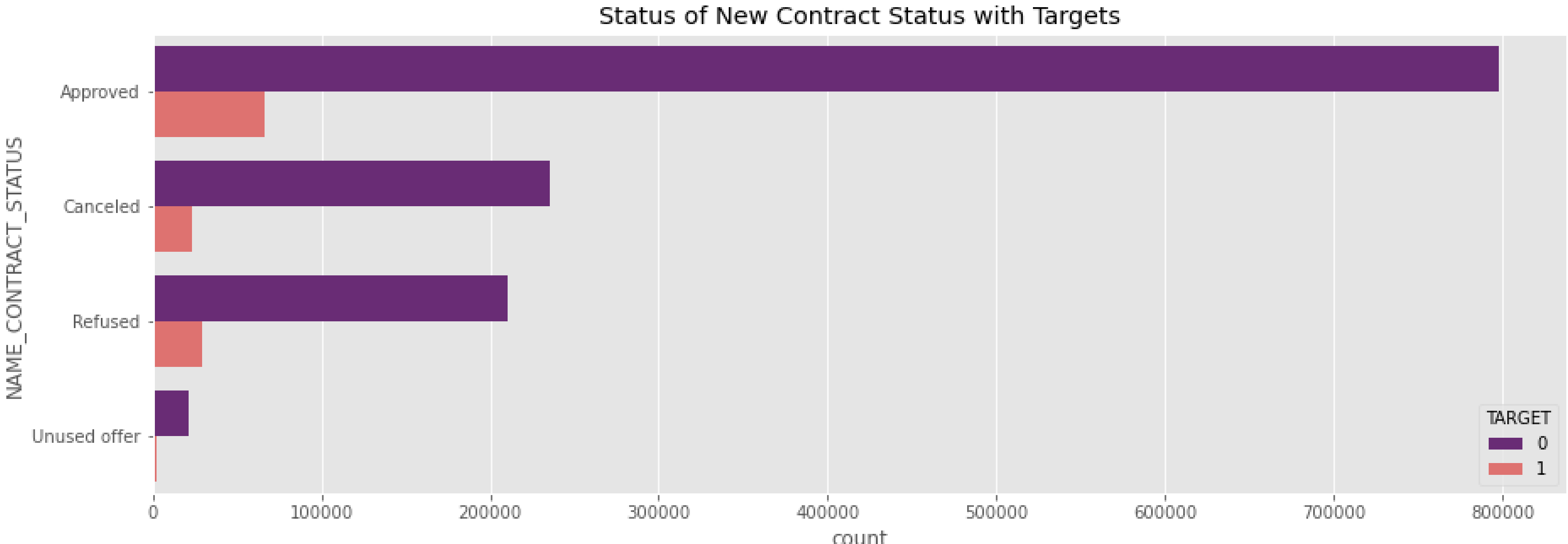
Checking Data Imbalance in Previous Application Data



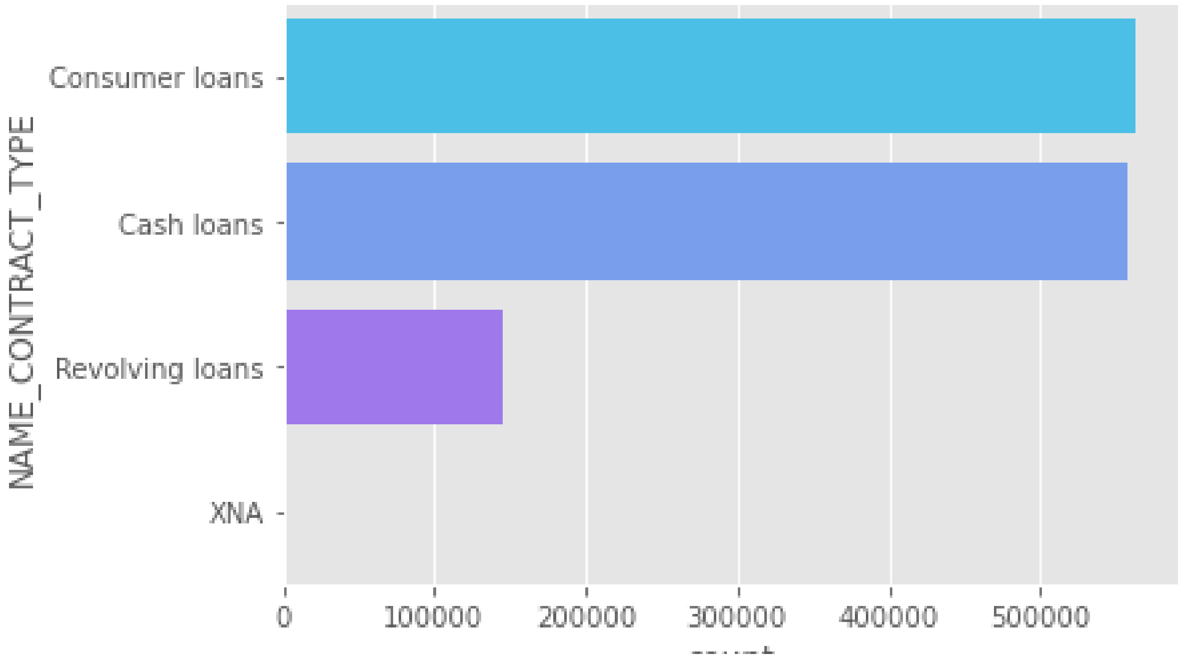
This data is highly imbalanced given that in total population the number of defaulters are very few.

Graphs

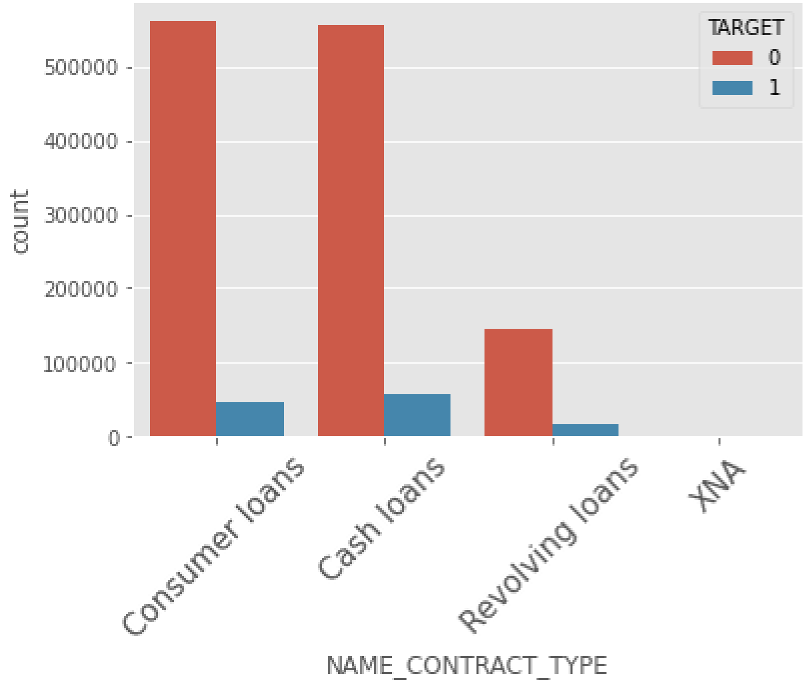

Status of New Contract Status with Targets
NAME_CONTRACT_STATUS
Approved
Canceled
Refused
Unused offer
0
100000
200000
300000
400000
500000
600000
700000
800000
count
TARGET
0
1

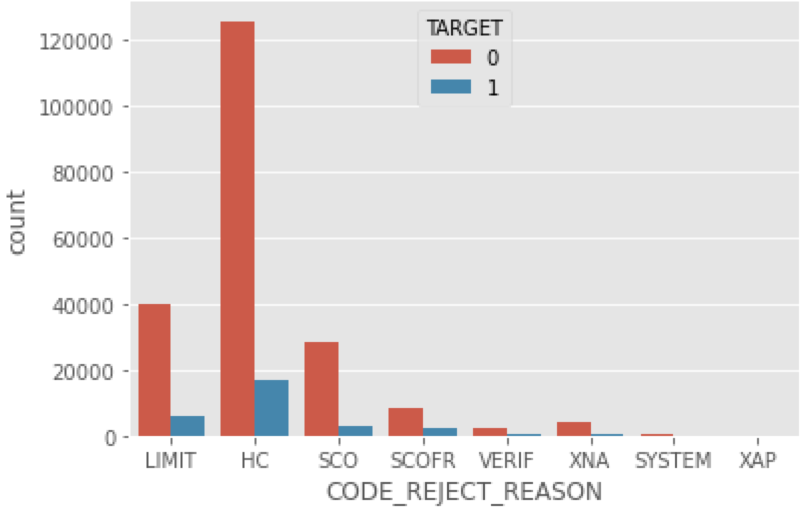

Status of New Contract Status with Targets
NAME_CONTRACT_STATUS
Approved
Canceled
Refused
Unused offer
0
100000
200000
300000
400000
500000
600000
700000
800000
count
TARGET
0
1

Analysing Categorial features of prev_app_df



Consumer loans are the highest applied number of loans followed by Cash loans





The most common reasons for rejection are HC, Limit and SCO.

Concluding remarks

- Insurance and Vehicles applicants has the highest % of default cases
- Default rate is highest for the Card segment(10%)
- Those who walked in are the highest amongst defaulted cases
- 13% loan applicant defaulted for AP+ (Cash Loan)
- Auto technology in the seller category has the highest defaulting rate
- Tourism has lowest defaulting rate in seller category
- Default are higher where name is not known or details are not known
- Dataset is highly imbalanced
- There are certain columns in both datasets which are highly correlated
- The Bank should focus more on Tourism, Fitness, Medical Supplies, Clothing, Furniture, Education, etc as the default rates are very low.
- Bank should focus less on Cash loan or Contact center and more on Corporate Sales and Car dealer which has lower default rates