EDA Case Study

Ankit Tambe

BUSINESS Understanding

We have data from the banks database and we have to help them make decision of weather a customer gets a loan or not.

Two types of risks are associated with the bank's decision:

A. If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company

B. If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

EDA Strategy

To avoid these risks, we first need to find out the factors leading to –

1. Loan Default
2. On time payment

Mitigation

Risk A – We will know what factors lead to a good customer

Risk B- We will know the factors leading to a default and hence avoid such customers

Data Understanding

Data Understanding

1. 'application_data.csv' contains all the information of the client at the time of application. The data is about whether a client has payment difficulties. This is file from which we will find out the different factors leading to a default
2. 'previous_application.csv' contains information about the client's previous loan data. It contains the data on whether the previous application had been Approved, Cancelled, Refused or Unused offer . This files tells us which previous loans were Accepted, Rejected, Cancelled and Unused.

Data Cleaning

Data Cleaning- Dealing with Null Values

1. Keeping the threshold of 40%, all columns that have more than 40% null values are deleted leaving 73 columns from the 122 we started with
2. Dropping some unwanted columns we are left with 28 columns

Data Cleaning- Categorical Columns

1. The NAME_FAMILY_STATUS column has Unknown category for null values. We need to replace these with pn.nan for better processing. Similar function done for "XNA' in the CODE_GENDER column
2. As we see, 'OCCUPATION_TYPE' has more than 31% null values. Also Occupation is an important column for our analysis. Therefore imputing these values will affect the analysis heavily. Hence it is better if create a new category call it unknown and replace all the NAs with 'Unknown’
3. For NAME_TYPE_SUITE,NAME_FAMILY_STATUS and CODE_GENDER the no of null values is quite less, also the column is not that significant in our analysis. Therefore we can impute it with the mode

Data Cleaning- Numerical Columns

1. The percentage of missing values in these columns is not that significant, also the column with the most missing values - AMT_REQ_CREDIT_BUREAU_YEAR is not that important for the analysis. Hence we can impute them.
2. For normalized data, the mean and the median are either same or very close. For data that has outlies, the mean is heavily affected. Hence it is safe to impute missing values with median instead of mean. This will give us a good approximation even if a certain column has outliers
3. 'AMT_REQ_CREDIT_BUREAU_YEAR','AMT_ANNUITY','YEARS_LAST_PHONE_CHANGE','YEARS_EMPLOYED','CNT_CHILDREN','AMT_INCOME_TOTAL' are imputed with flooring and capping using IQR techniques.
4. Bins created for 'AGE', 'AMT_GOODS_PRICE', 'AMT_ANNUITY', 'AMT_CREDIT', 'AMT_INCOME_TOTAL'
5. Outlier decisions explained in detail in the python file

Data Analysis

Univariate Analysis - Numerical

1. The category that is actually showing a huge variation in defaulters and non defaulters is 'YEARS_EMPLOYED' – 0-7 years,
2. The lesser the job experience, more are the defaults.
3. For AMT_CREDIT between 250000 and approximately 650000, there are more clients with Payment difficulties. For AMT_CREDIT > 750000 , there are more clients with On-Time Payments
4. For YEARS_BIRTH between 20 and 40, there are more clients with Payment difficulties. Conversely, for YEARS_BIRTH > 40 , there are more clients with On-Time Payments
5. For CNT_CHILDREN 0 (those with no children), there are lots of clients with On-Time Payments. For CNT_CHILDREN with 1 OR 2 (those with 1 or 2 children), there are few more clients with On-Time Payments
6. Based on AMT_INCOME_TOTAL, for clients with Payment difficulties, the distribution resembles a normal distribution approximately. But for clients with On-Time Payments, there are erratic spikes in the distribution which doesn't give any valid observations

Univariate Analysis - Categorical

1. Clearly not defaulters - IT Staff, Student, Businessman, where goods > 2400000

2. Clearly Defaulters - Labourers, Higher Eduication, Pentioner,

Multivariate Analysis - Categorical

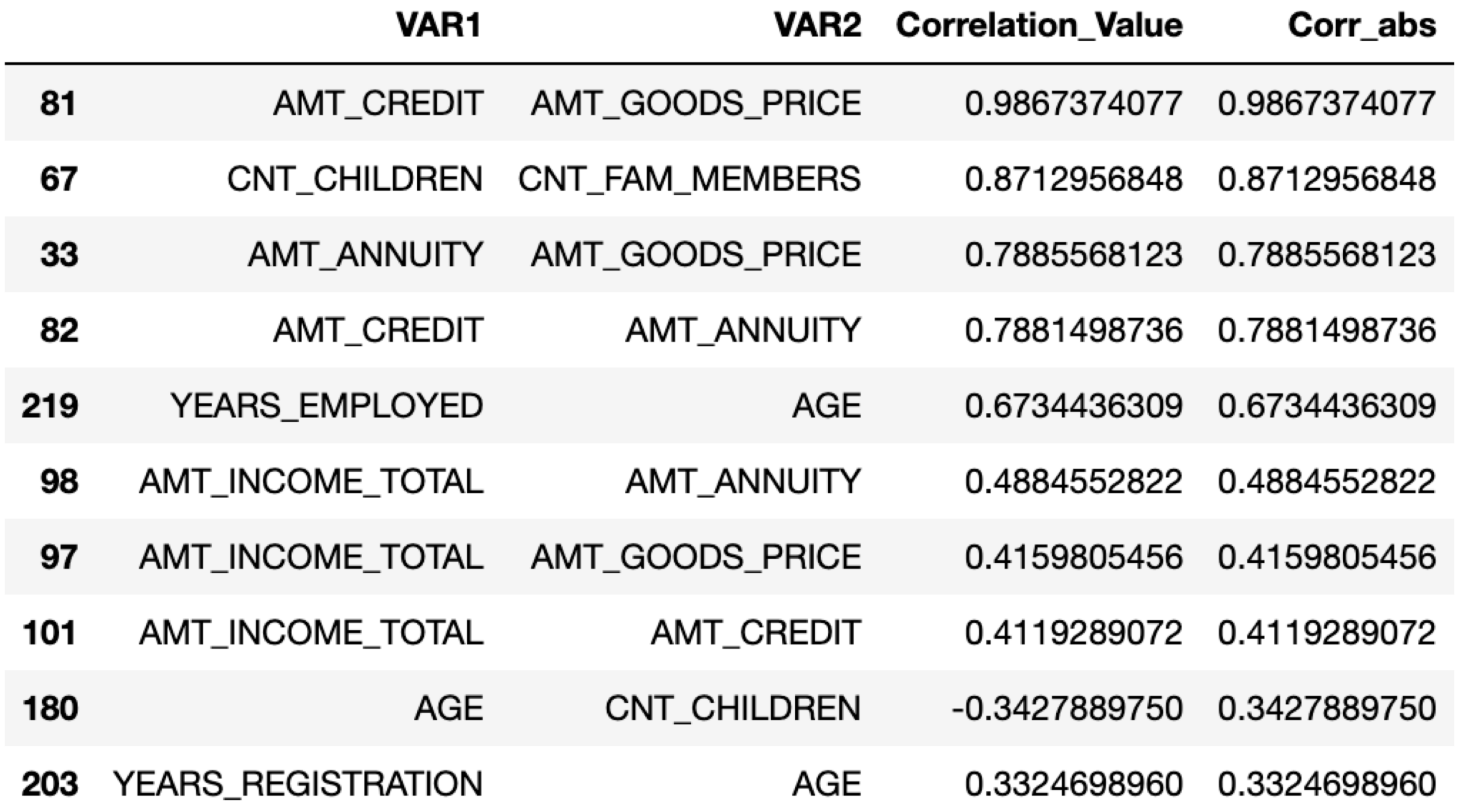
Observations

1. Strong Rate of Default - 0.9-1
    1. Unemployed - 18-27, Unemployed - 27-36, Unemployed - 63-73, Maternity Leave all valid age groups
2. High Rate of Default - 0.7-0.9
    1. Unemployed - Married Married - Maternity Leave Unemployed - Male
3. Medium Rate of Dwfault - 0.5 -0.7
    1. Unemployed ,low skill labourers , 63-73
4. Weak Rate of default - 0.2-0.5
    1. Lower Secondary , 18-27 low skill labourers , 63-73 low skill labourers, Lower Secondary low skill labourers, Single Not Married Low Skill labourers, Male

Unemployed People, People on Maternity Leave, Low Skilled Labourers - are features that the bank should avoid while providing a loan

Multivariate Analysis - Numerical

TOP 10 correlations for both Target types are the same



	VAR1	VAR2	Correlation_Value	Corr_abs
81	AMT_CREDIT	AMT_GOODS_PRICE	0.9867374077	0.9867374077
67	CNT_CHILDREN	CNT_FAM_MEMBERS	0.8712956848	0.8712956848
33	AMT_ANNUITY	AMT_GOODS_PRICE	0.7885568123	0.7885568123
82	AMT_CREDIT	AMT_ANNUITY	0.7881498736	0.7881498736
219	YEARS_EMPLOYED	AGE	0.6734436309	0.6734436309
98	AMT_INCOME_TOTAL	AMT_ANNUITY	0.4884552822	0.4884552822
97	AMT_INCOME_TOTAL	AMT_GOODS_PRICE	0.4159805456	0.4159805456
101	AMT_INCOME_TOTAL	AMT_CREDIT	0.4119289072	0.4119289072
180	AGE	CNT_CHILDREN	-0.3427889750	0.3427889750
203	YEARS_REGISTRATION	AGE	0.3324698960	0.3324698960

FINAL SUGGESTIONS to Bank

1. Following are the features to look to avoid –
   - Work Experience- The lesser the job experience, more are the defaults. Target clients with more than 20 years of employment
   - Income Type - People on Maternity Leave,
   - Occupation - Low Skilled Labourers, Unemployed People
2. Following are the features to look for –
   - Target clients with more than 20 years of empoyment
   - Student, Businessman
   - Pensioners also seem a good bet