Credit EDA Case Study

Business Use Case

- Identify patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. In other words, there is need to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.

Solution

After doing extensive univariate, and bivariate analysis, on all the columns of the provided dataset, we came across following patterns:

Results of Univariate analysis of categorical columns:

- If the credit amount of the client loan is in the range (43239-749394), then there is high probability that client will have payment difficulties.
- Client whose Loan Annuity is in the range (43141-70582) are less likely to default on their loan.

- Client's whose permanent address does not match contact address are more likely to default.
- Client’s who stay in the region rated as 2^{nd} and 3^{rd} are highly likely to default on their payments.
- Client’s whose permanent address does not match with work address are more prone to default.
- More percentage of males default as compared to females.
- Under column - How many days before the application the person started current employment- A lot of clients have put 365243 number. Further investigation is needed to find the root cause.

Results of Univariate analysis of numerical columns:

- If ‘EXT_SOURCE_2’, ‘EXT_SOURCE_3’ i.e. Normalized score from external data sources is less, of clients is less, then clients are likely to have payment related issues.
- If number of enquiries to Credit Bureau about the client one day year (excluding last 3 months before application) are more then 3, then there is higher possibility of default.

Results of Bivariate analysis:

- If value of ‘AMT_REQ_CREDIT_BUREAU_QRT’ (Credit Bureau enquiries) oscillates between (0.00 - 0.25), and the AMT_INCOME_TOTAL (Income of client) is in range (707130.0, 877500.0], then client is more likely to default on payment.

- If AMT_CREDIT (Credit amount of the loan) is in range (43239.015, 397197.0], and CNT_FAMILY_Members (Count of Family members) > 3 then there is more probability that client will default.
- If AMT_REQ_CREDIT_BUREAU_YEAR is greater then 3, and AMT_CREDIT in range (749394.0, 1101591.0], then client is less likely to default.
- If Count of children > 0, and DAYS_REGISTRATION (How many days before the application did client change his registration) is in range – (9331.8 - 6221.2], then client is more likely to default.
- If CNT_FAMILY_Members > 2, and NAME_HOUSING_TYPE is Municipal apartment, then client is more likely to default.

- If AMT_REQ_CREDIT_BUREAU_YEAR (Credit bureau enquiries) > 5, and NAME_HOUSING_TYPE is Office apartment, then client is likely to default.

Numeric variable Correlations

- EXT_SOURCE_3 positively correlates with AMT_REQ_CREDIT_BUREAU_HOUR, AMT_REQ_CREDIT_BUREAU_DAY, AMT_REQ_CREDIT_BUREAU_WEEK when client faces payment difficulties, otherwise, it negatively correlates.
- AMT_REQ_CREDIT_BUREAU_YEAR negatively correlates with AMT_REQ_CREDIT_BUREAU_HOUR,AMT_REQ_CREDIT_BUREAU_DAY and AMT_REQ_CREDIT_BUREAU_WEEK when client faces difficulty, otherwise it correlates positively.