CREDIT EDA ASSIGNMENT

By Anju Mary Samuel

What & Why?

- To identify loan applicants who are likely to default in the repayments

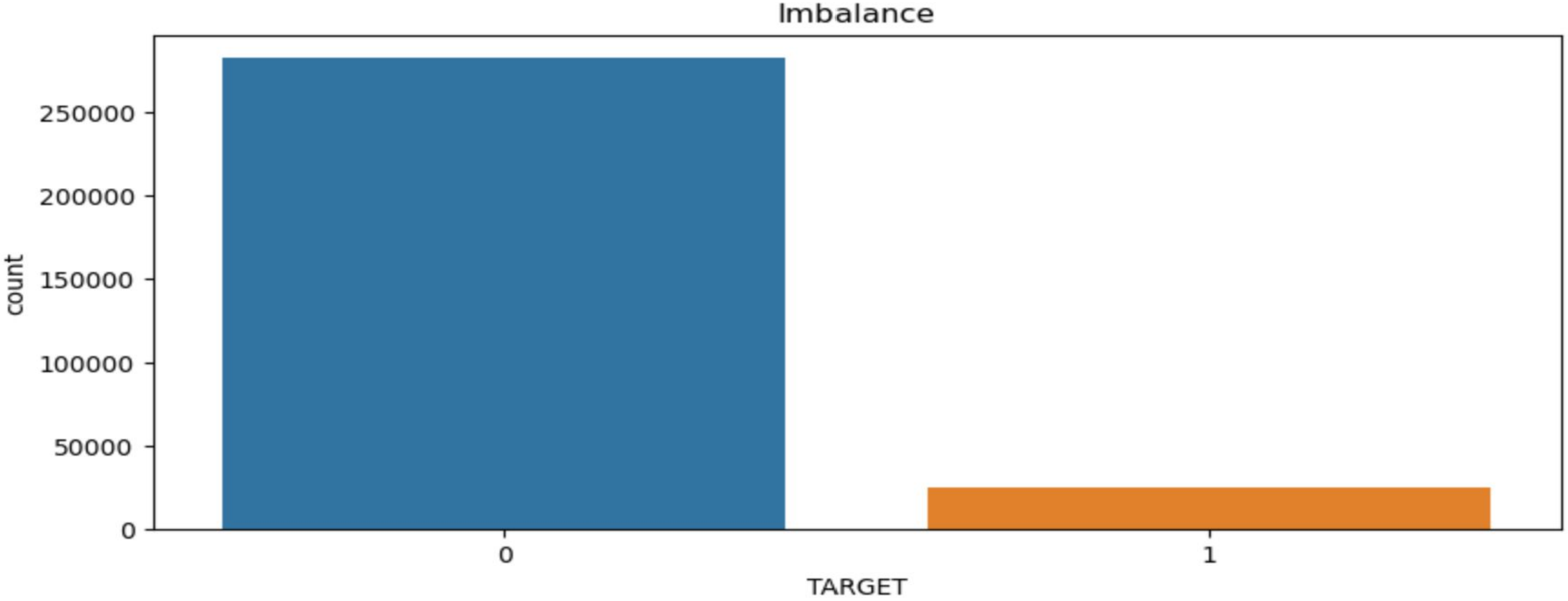
- To provide insights to the bank to aid its decision & risk management strategy

How?

- ❑ Analyze data provided by bank
    - ❑ Identify target and relevant columns
- ❑ Clean data
    - ❑ Identify missing values, impute / drop as required
    - ❑ Standardize relevant data
    - ❑ Fix invalid values
    - ❑ Identify outliers
- ❑ Identify relationships between relevant data
    - ❑ Univariate analysis
    - ❑ Multivariate analysis
    - ❑ Bivariate analysis

Analyze data provided by bank

Data imbalance: for every defaulter there are 11 repayers

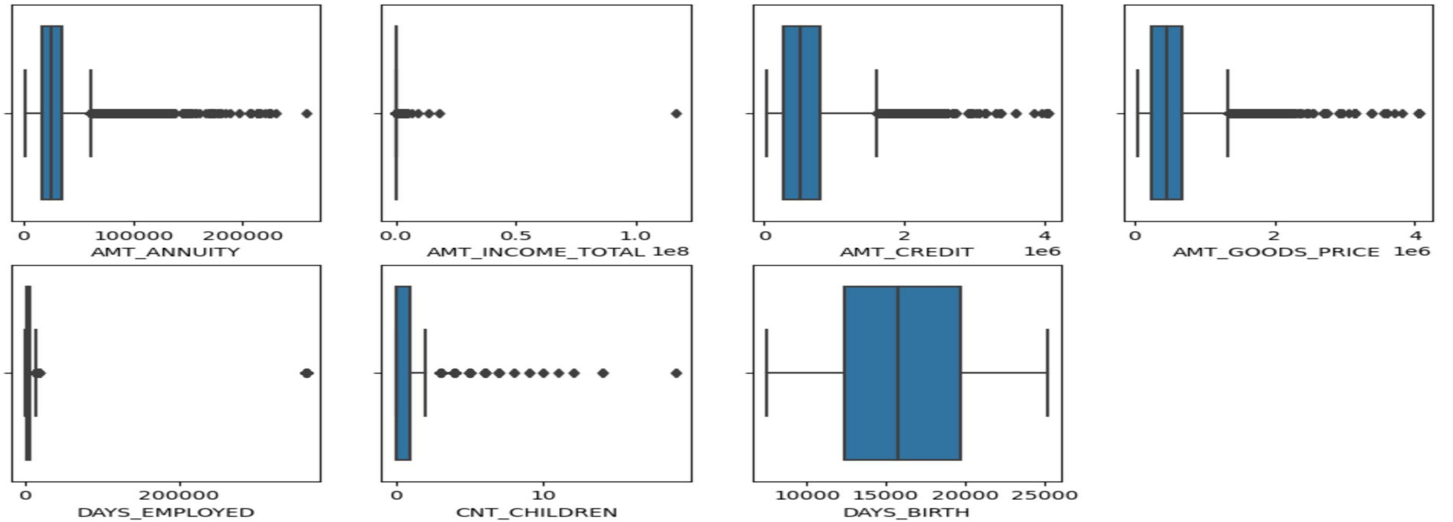


Data Cleaning - part 1

- ❑ Identify Columns with null values
  - ❑ Drop columns having > 40% null values as this will skew the analysis results if left as is; ELEVATORS_MODE, ENTRANCES_MODE, FLOORSMAX_MODE etc, ~ 50 columns were dropped.
  - ❑ Impute missing values for data that are relevant and have significant data missing; OCCUPATION_TYPE has ~ 30% missing data - impute with new value so as to not skew other occupation types.
  - ❑ Check correlation of other columns with TARGET and drop them if no correlation exists; AMT_REQ_CREDIT_BUREAU_X, FLAG_DOCUMENT_X were dropped
  - ❑ Drop irrelevant columns; FLAG_PHONE, FLAG_EMAIL, FLAG_EMAIL were also dropped
- ❑ Fix invalid values
  - ❑ CODE_GENDER had an invalid value ‘XNA’ and was replaced with the mode.
  - ❑ DAYS_BIRTH, DAYS_EMPLOYED etc had negative values -> converted to +ve
- ❑ Standardize relevant data
  - ❑ DAYS_BIRTH, DAYS_EMPLOYED were converted in terms of ‘year’

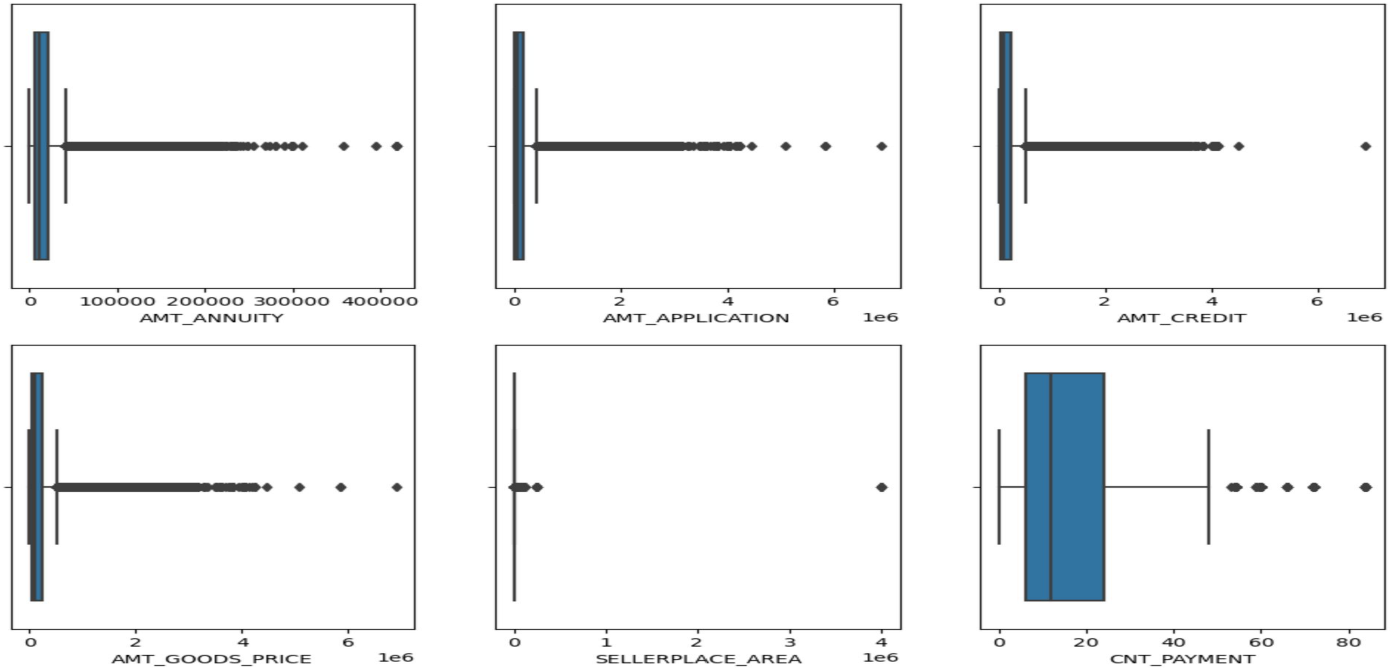
Data Cleaning - part 2

- Identify outliers in application data
  - AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE,CNT_CHILDREN have some number of outliers.
  - AMT_INCOME_TOTAL has huge number of outliers which indicate that few of the loan applicants have high income when compared to the others.
  - DAYS_BIRTH has no outliers which means the data available is reliable.
  - DAYS_EMPLOYED has outlier values around 350000(days) which is around 958 years which is impossible and hence this has to be incorrect entry.



Data Cleaning - part 3

- Identify outliers in previous data
  - AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE, SELLERPLACE_AREA have huge number of outliers.
  - CNT_PAYMENT has few outlier values.



Univariate Analysis - Insights

- Women are likely to repay than men - Women take significantly higher number of loans but default much lesser.
- Revolving loans are defaulted more - they are much lesser in number than cash loans but have significant number of default cases compared to cash loans.
- Civil marriage couples and Single/unmarried are likely to default more.
- Applicants who have not completed their higher education are likely to default
- Low-skill labourers have the highest chance of defaulting
- People who live in region with rating 3 have very high default rating.
- Unemployed and maternity leave applicants are the highest defaulters.

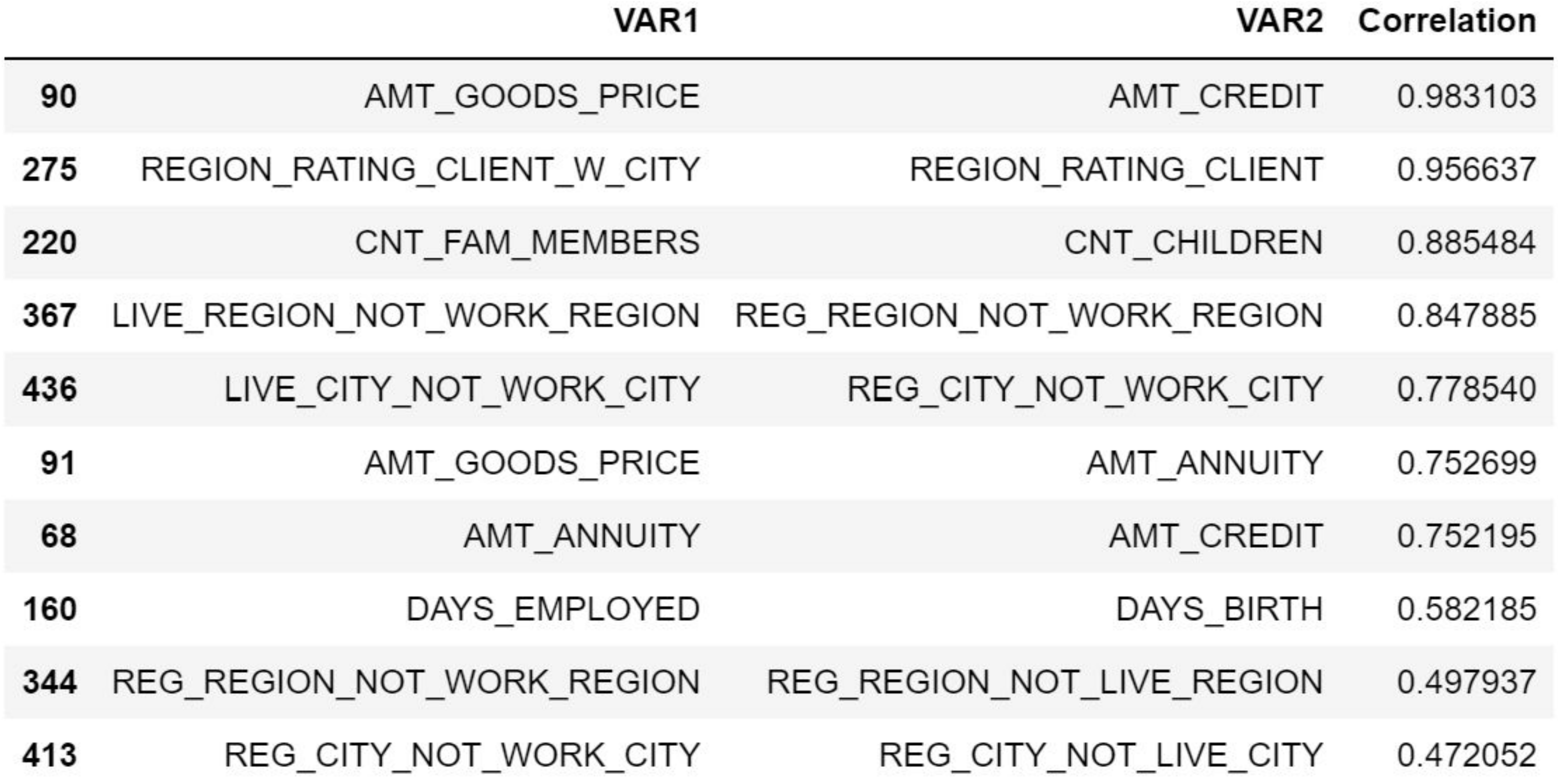
Univariate Analysis - Insights (contd.)

- People in age group 20-30 are most likely to default and > 50 are least likely to default
- People in income range < 300k have a high likelihood of defaulting
- People who get loan for 300-600k tend to default more than others.

Bivariate & Multivariate Analysis - insights

- People with higher education have higher income
- People with difficulty to pay have higher credit amount as compared to their income.
- Very high correlation between AMT_GOODS_PRICE & AMT_CREDIT -> as AMT_GOODS_PRICE increases, so does AMT_CREDIT
- Applicants for whom previous loans were refused, have had no difficulty repaying their current loan on time
- Significant number of loans were refused for repairs

Top 10 correlation factors



	VAR1	VAR2	Correlation
90	AMT_GOODS_PRICE	AMT_CREDIT	0.983103
275	REGION_RATING_CLIENT_W_CITY	REGION_RATING_CLIENT	0.956637
220	CNT_FAM_MEMBERS	CNT_CHILDREN	0.885484
367	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.847885
436	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.778540
91	AMT_GOODS_PRICE	AMT_ANNUITY	0.752699
68	AMT_ANNUITY	AMT_CREDIT	0.752195
160	DAYS_EMPLOYED	DAYS_BIRTH	0.582185
344	REG_REGION_NOT_WORK_REGION	REG_REGION_NOT_LIVE_REGION	0.497937
413	REG_CITY_NOT_WORK_CITY	REG_CITY_NOT_LIVE_CITY	0.472052