Analysis of fraud detection in bank data using EDA

Problem statement

- A certain bank wants to find a way to provide loan for the applicants to run its business .
- But they facing difficulty in finding the client as :
  1. providing loan to a client who is not likely to repay loan will lead to financial loss.
  2. Denying loan to a client who is likely to repay loan will lead to business loss.
- To find a solution, we are given a three datasets which contains the current application and previous application data of the clients and description of the columns.
- We are required to perform the EDA and use the data visualization techniques on these data to find the clients who are most likely to repay the loan and who are not.

Assumptions

- Since some the data are missing in the data , we are removing columns which has higher percentage missing data as imputing values to them may affect the analysis.
- For the data where the missing values are less, we are imputing them with median value of the remaining data for continuous data and mode value of the remaining data for categorical data.

Approach and methodology

UNDERSTANDING THE DATA

- The primary step involves understanding the data, here we are given a dataset containing column description to understand the meaning of the data and their importance

IMPORTING THE DATA

- Now we load the data into the platform to check its shape, datatypes and other information.

HANDING MISSING DATA

- Generally the data(column) containing more than 30% (depending on total number of columns) of the missing values are removed.
- For the columns containing lesser missing values, are replaced median value for continuous data and mode value for categorical data.

Approach and methodology

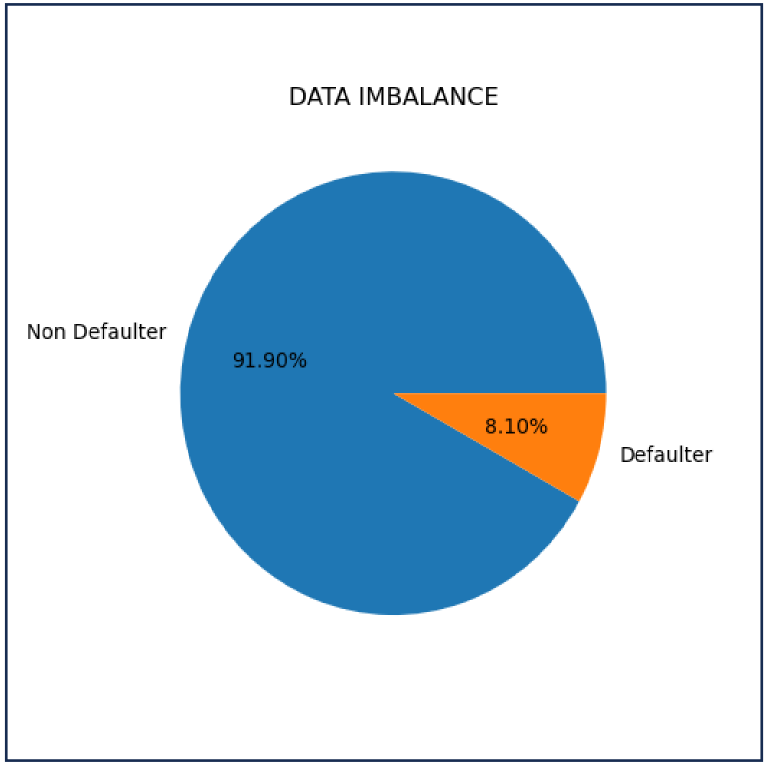
OUTLIERS

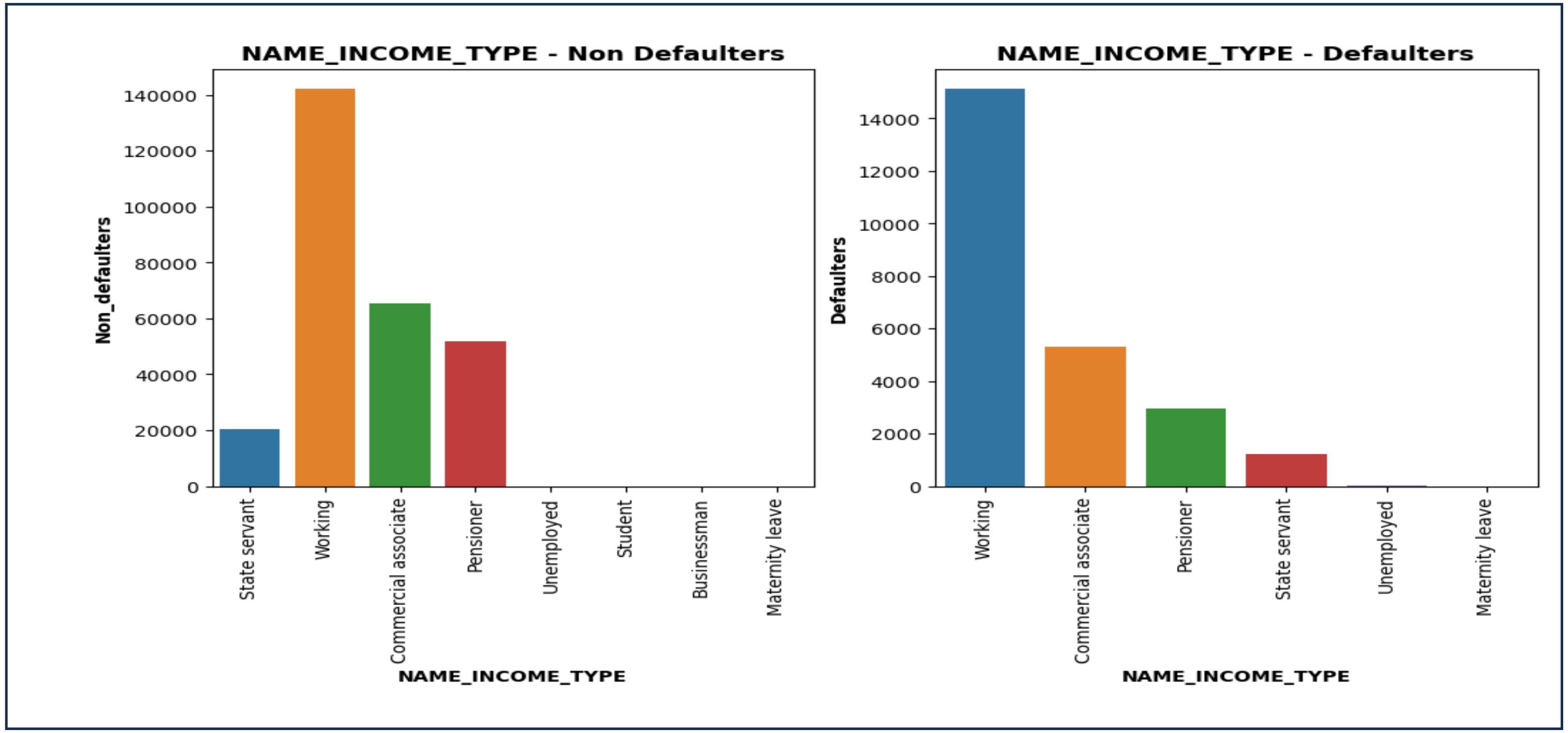
- After handling the missing data, we use describe method or boxplot to visually identify the potential outliers which are later verified further for data correctness.

VISUAL ANALYSIS

- Here we use various sophisticated visual techniques such as UNIVARIATE, BIVARIATE AND MULTIVARIATE ANALYSIS for identifying the important feature, trends and correlation among the data
- The above information is used to find the valuable insights and applied in favour of the target variable
- Such examples can be seen in next slide.

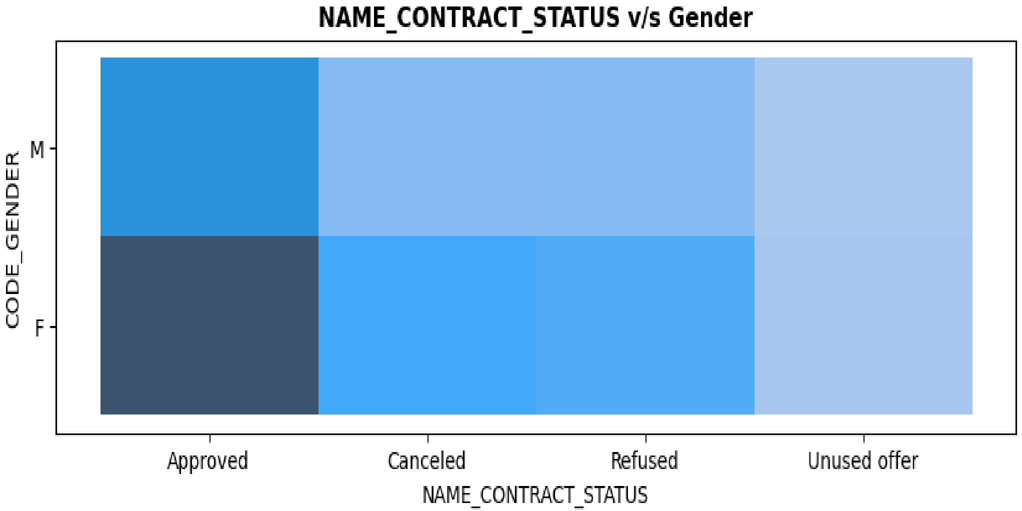
Graphs and Insights - Univariate

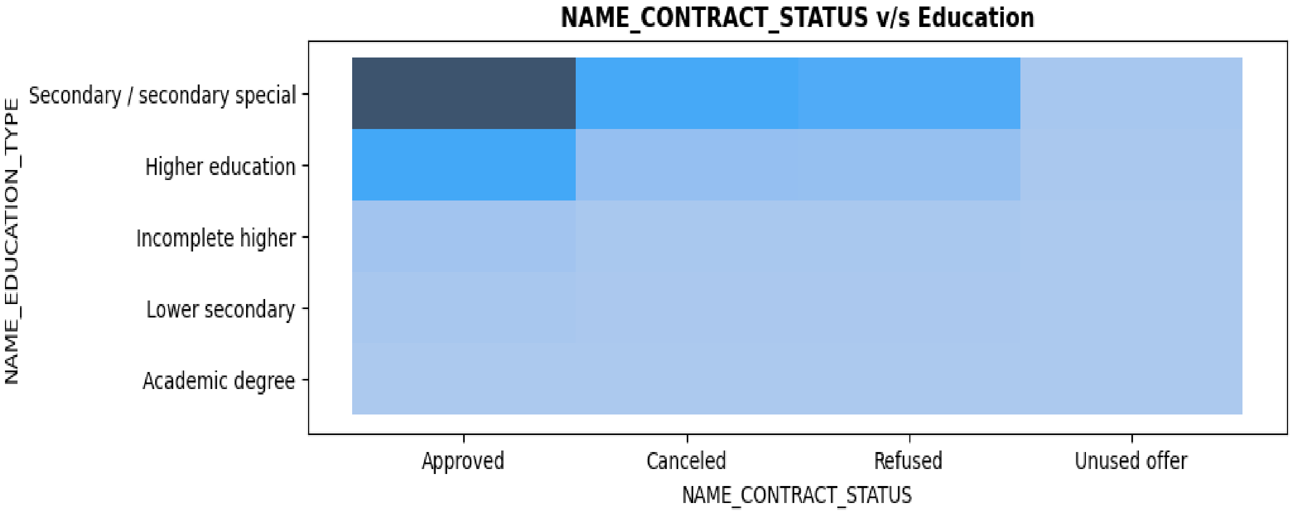




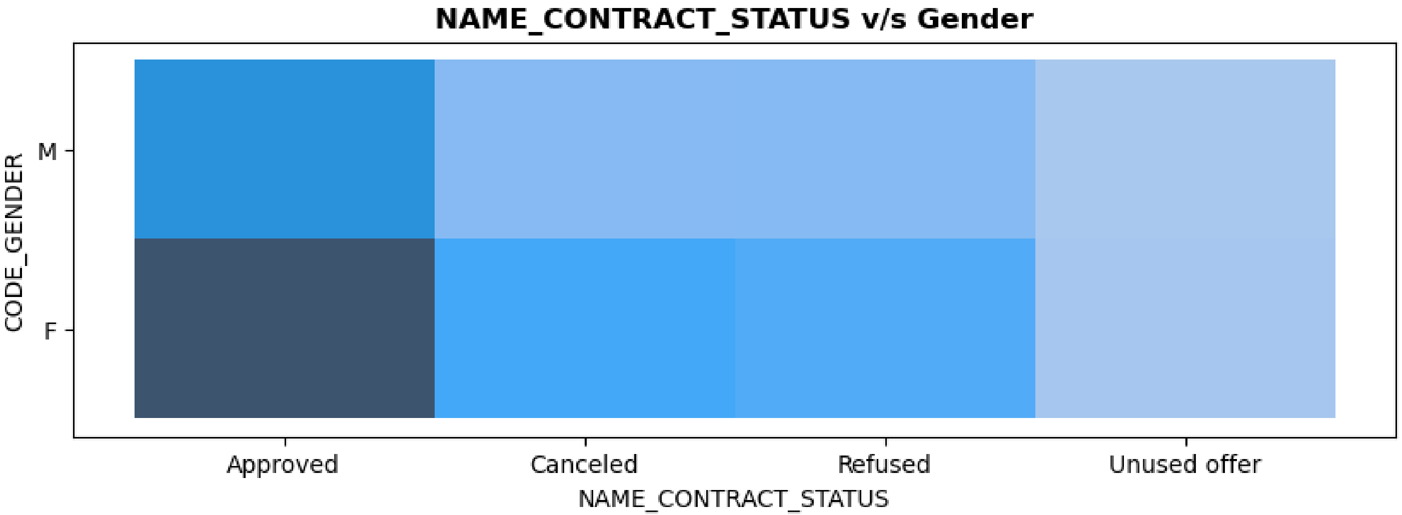
- Here we can see that the data imbalance ratio is 11.35, which is a favorable condition for the bank as it indicated that the defaulter are less compared to non defaulters.
- In the second graph we can see that the working class people are applying more for the loan.

Graphs and Insights - Bivariate





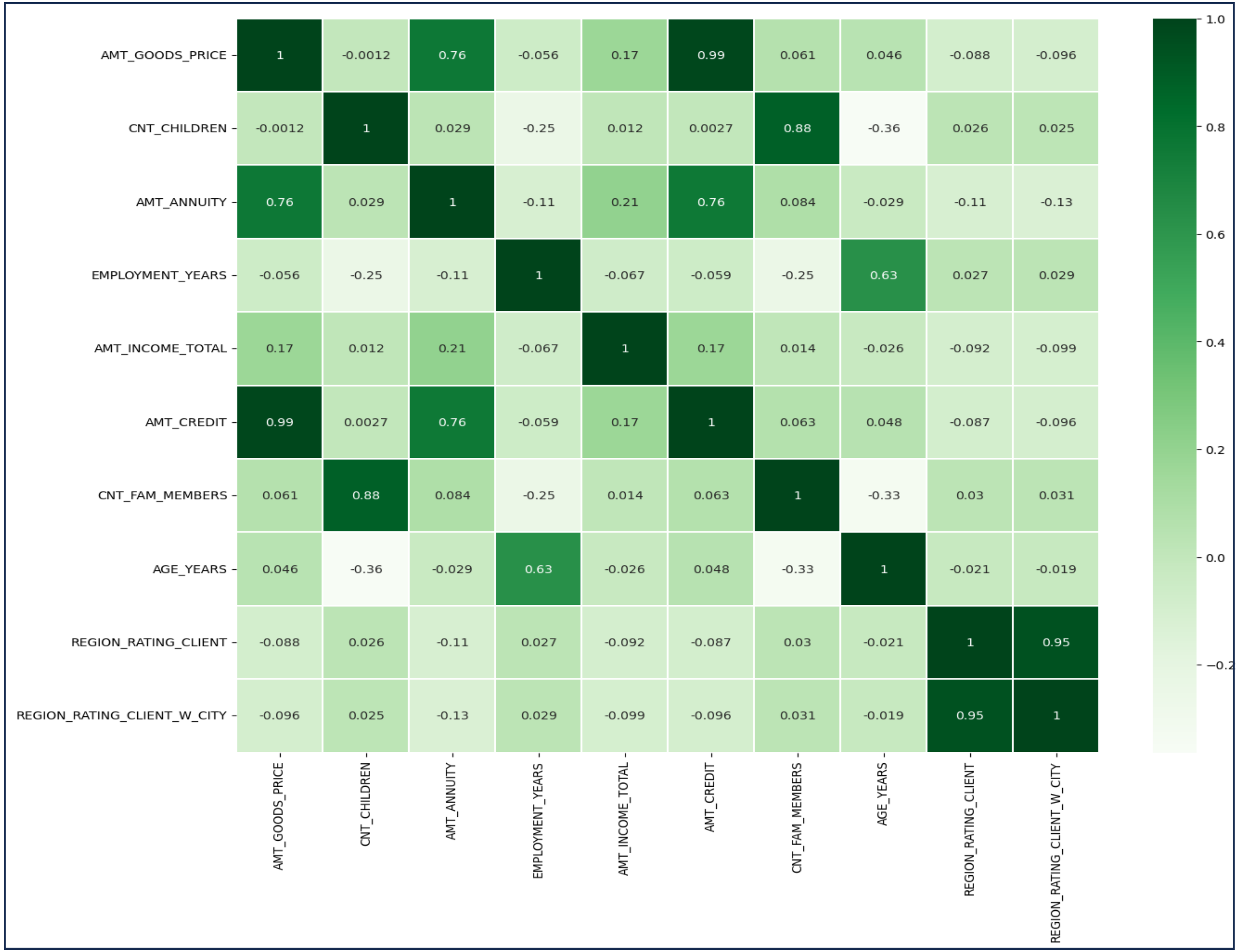
Here we can see that the approval rate is high repeater applicant and in applicant with secondary education



- Approval rate is higher in females than males

Graphs and Insights - Multivariate

- AMT_GOODS_PRICE and AMT_CREDIT have the highest correlation
- AMT CREDIT and AMT_ANNUITY also has high correlation



	AMT_GOODS_PRICE	CNT_CHILDREN	AMT_ANNUITY	EMPLOYMENT_YEARS	AMT_INCOME_TOTAL	AMT_CREDIT	CNT_FAM_MEMBERS	AGE_YEARS	REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY
AMT_GOODS_PRICE	1	-0.0012	0.76	-0.056	0.17	0.99	0.061	0.046	-0.088	-0.096
CNT_CHILDREN	-0.0012	1	0.029	-0.25	0.012	0.0027	0.88	-0.36	0.026	0.025
AMT_ANNUITY	0.76	0.029	1	-0.11	0.21	0.76	0.084	-0.029	-0.11	-0.13
EMPLOYMENT_YEARS	-0.056	-0.25	-0.11	1	-0.067	-0.059	-0.25	0.63	0.027	0.029
AMT_INCOME_TOTAL	0.17	0.012	0.21	-0.067	1	0.17	0.014	-0.026	-0.092	-0.099
AMT_CREDIT	0.99	0.0027	0.76	-0.059	0.17	1	0.063	0.048	-0.087	-0.096
CNT_FAM_MEMBERS	0.061	0.88	0.084	-0.25	0.014	0.063	1	-0.33	0.03	0.031
AGE_YEARS	0.046	-0.36	-0.029	0.63	-0.026	0.048	-0.33	1	-0.021	-0.019
REGION_RATING_CLIENT	-0.088	0.026	-0.11	0.027	-0.092	-0.087	0.03	-0.021	1	0.95
REGION_RATING_CLIENT_W_CITY	-0.096	0.025	-0.13	0.029	-0.099	-0.096	0.031	-0.019	0.95	1

Recommendations and Insights

- We have observed that the applicant who have secondary education are most likely to repay the loan, hence the bank can lend loan to them.
- Banks can offer the loans to repeater as they already have a good credit history and likely to repay the loan
- Students and businessmen have almost negligible defaulters, hence bank can offer the loan for them as well
- Banks can avoid clients with academic degree and medium level income as the defaulters are more than non defaulters in these class

THANK YOU