Exploratory Data Analysis

Case Study : Bank Loan Analysis

Domain : Risk Analytics in Banking

Submission by	:	**Anmol Mehta**
Batch	:	**DS C 36**

Business Problem & Objective

- This case study aims to give us an idea of applying EDA in a real business scenario. In this case study, we develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimise the risk of losing money while lending to customers.
- This case study aims to identify patterns which indicate if a client has difficulty paying their instalments 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. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study. The company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

Data Loading, Processing & Cleansing

This dataset has 3 files as explained below:

- ***'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.

- ***'previous_application.csv'***

  contains information about the client's previous loan data. It contains the data whether the previous application had been Approved, Cancelled, Refused or Unused offer.

- ***'columns_description.csv'***

  is data dictionary which describes the meaning of the variables.

Application Data was our primary Dataset.

- **Data quality & inspection checks** were performed on the data.
- Data was checked for **Null/missing values** and columns were dropped where the Null percentage **threshold** breached **50%.**
- **"EXT_SOURCE" & "FLAG_DOC"** were dropped due to irrelevance to our analysis.
- Data was checked for **duplicates**.
- Datatypes for some columns were converted to **Numeric** for analysis.

Data Imputation

- For some columns like AMT_REQ_CREDIT_BUREAU_MON, AMT_REQ_CREDIT_BUREAU_WEEK, etc. values were **imputed with 0** because of their **median, min, IQR** values being 0.
- Values for **days** for all columns were converted from **negative** (for some cases) into **positive**.
- 'XNA' values were imputed with the **Median** & **Null** for Gender & Organization Type columns respectively.

Binning of the Data

- For some columns like **Income Range, Credit Range**, etc. values were **Binned** into the following ranges :

  Very low, Low, Medium, High & Very High
- For column **Work experience**, values were binned into:

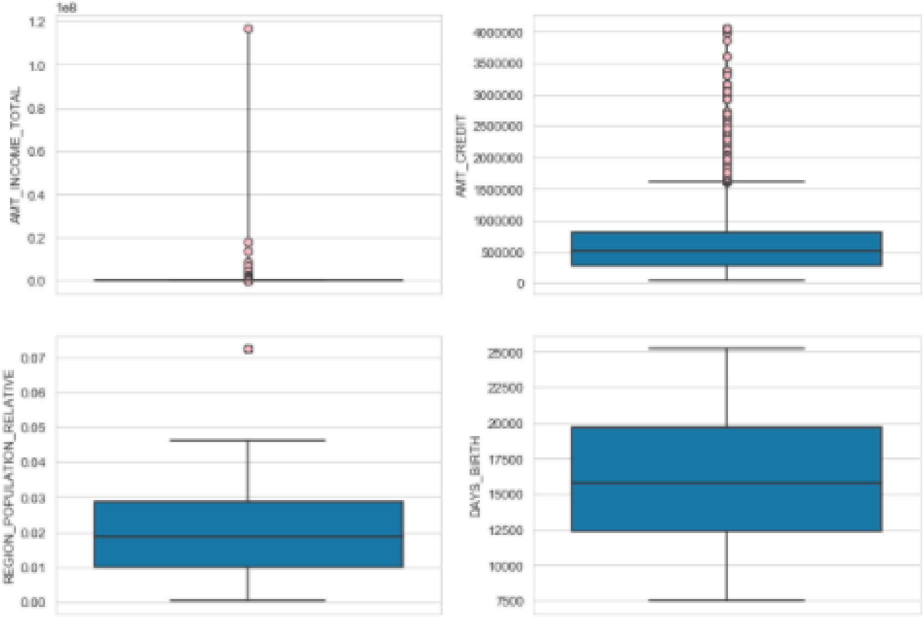
  'Not Experinced','Fresher', 'Entry Level', 'Mid Level', 'Senior Level', 'Higher Senior Level'.
- For column **Days of birth**, values were binned into:

  'Very_Young', 'Young', 'Middle_Age', 'Senior_Citizen'

Outlier Detection & Treatment

- ***Box Plots***

  Plotted boxplots to visualise distribution of the data through their quartiles & plotting outliers outside the whiskers.



(Fig 1.) Columns like Income Total, Credit amount, etc.

- After analysing all the plots below columns were found to have outliers:

  AMT_INCOME_TOTAL,AMT_CREDIT,INCOME_CREDIT_RT,LTV_RT,EMPLOYED_YEAR,CNT_FAM_MEMBERS, DEF_30_CNT_SOCIAL_CIRCLE, AMT_REQ_CREDIT_BUREAU_*, AMT_ANNUITY, DAYS_EMPLOYED, DAYS_REGISTRATION

Outlier Detection & Treatment

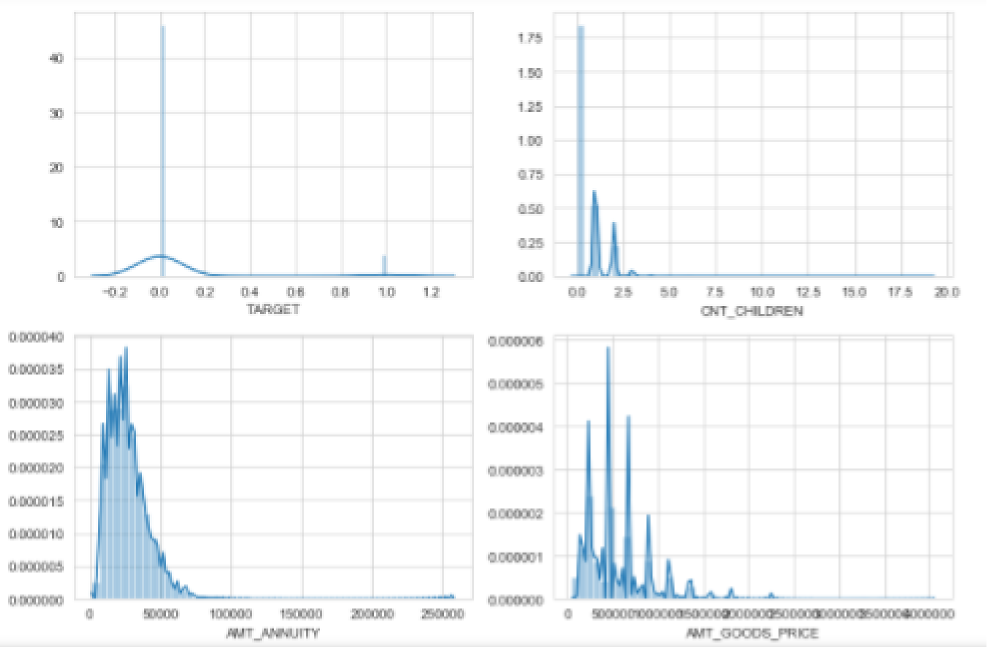
- ***Percentile Distribution***

  For the columns with outliers, we calculated the percentile distributions of the data for 25th, 50th, 75th, 95th and 100th percentile.

  We found that for columns like **Income Total , Amount Annuity, count of children**, etc. there was a drastic shift in figures from 95th to 100th percentile, suggesting a very high possibility of those values being outliers.

- ***Histograms***

  We plotted the frequency distribution plots to get understand the representation of the data.

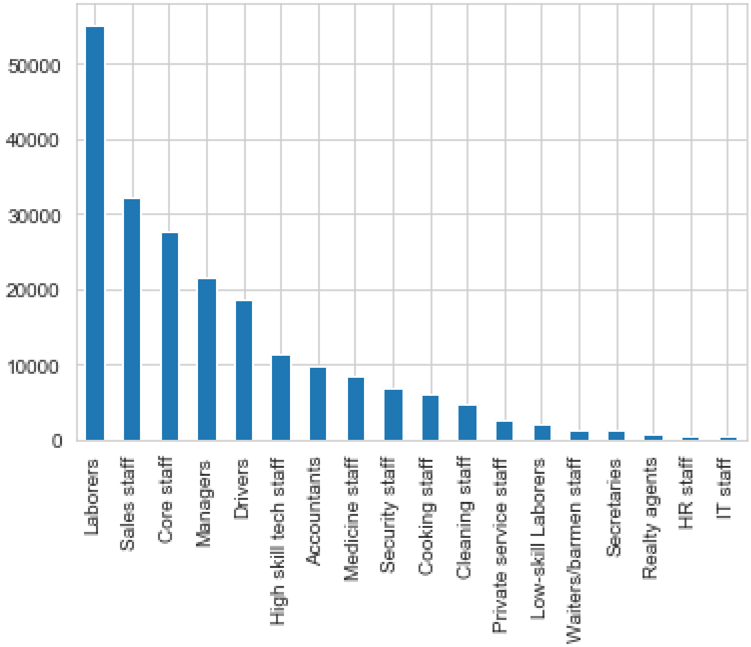


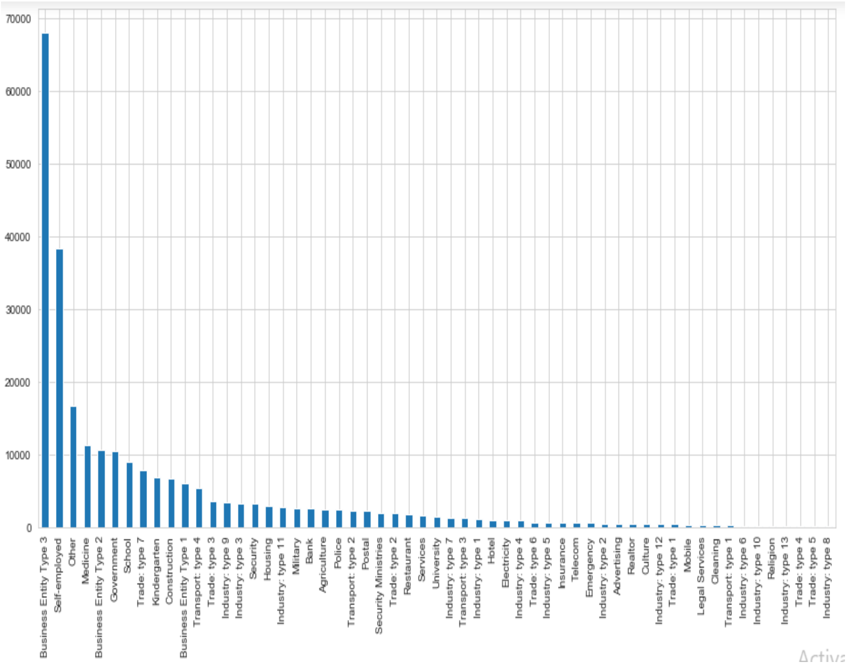
(Fig 2.) Columns like Amount Annuity, count of children etc.

Univariate Analysis (Application data)

- ***Bar Plots***

  We plotted bar graphs to check the categorical representation of the data for some columns.





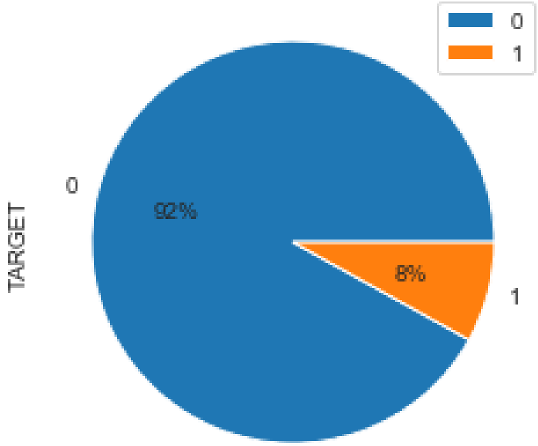
(Fig 3.) Occupation Type : We observed that Labourers, Sales Staff & Core staff applies the most for loans, where as IT Staff applied the least.

(Fig 4.) Organization Type : We observed that Business Entities & Self-Employed people applies the most for loan applications.

Data Imbalance & Segmentation

- ***Pie Plot***

  We plotted pie charts to check the numerical proportion & balance of the data for Target Column (used for Loan & Non-Loan payment difficulties).



(Fig 5.) Target: We observed a very high imbalance of data.

Due to the nature of imbalance in the data for Target column, we segmented the whole dataset into two sections for further analysis:

- **Target = 0 (Non – Loan Payment Difficulties)**
- **Target = 1 (Loan Payment Difficulties)**

Segmented Univariate Analysis

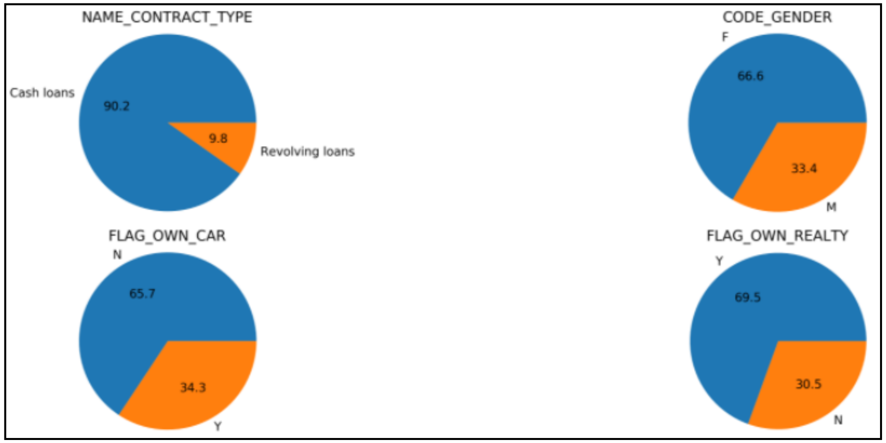
(Categorical Variables)

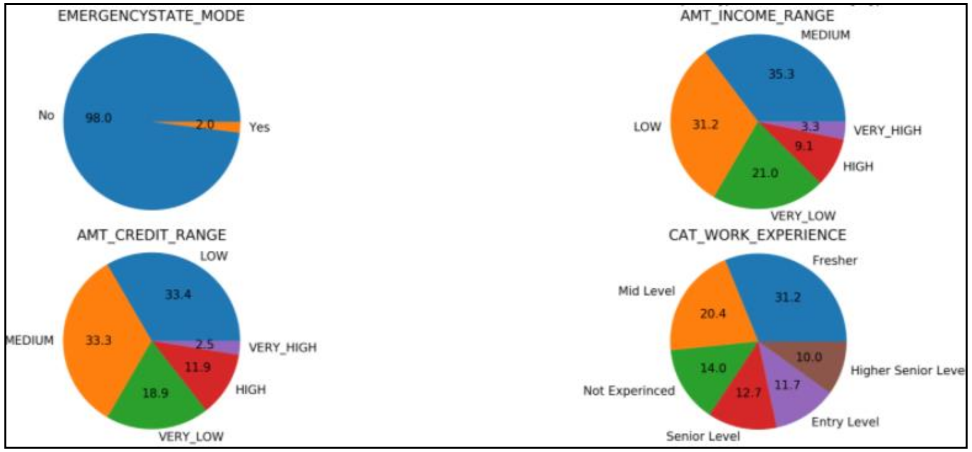
- ***Pie Plot***

  We plotted pie charts to check the numerical proportion & balance of the data for:

Target = 0

Target = 1





(Fig 6.) Columns like Contract Type, Gender, Flag Own Car, etc. for Target = 0

(Fig 7.) Columns like Income Range, Credit Range, Work Experience, etc. for Target = 1

Segmented Univariate Analysis

(Categorical Variables)

- ***Key Takeaways & Observations:***

1. **Contract Type** : Applicants prefer Cash loans in both the categories. although there was a decrease in payment difficulties for applicants that prefer Revolving Loans.

2. **Gender** : Females come out to be holding a majority in terms of loan applications, in both categories (Payment and non payment difficulties). Although there was a notable rise in Male Payment difficulties from Non Payment difficulties.

3**. Income Type** : Working class makes the majority (more than 50%) in both the categories. There was a notable decrease in Pensioners that were having Payment difficulties compared to Non-Payment difficulties.

4. **Education** : Secondary education is prevalent in the loan applications in both categories, but there was an increase in Payment difficulties compared to Non Payment difficulties. Whereas there was a decrease in Payment difficulties for people with higher education.

5**. Family Status** : Married folks make up the majority of the loan applications, though there was a decrease in Payment difficulties for Married & divorced applicants, and there was an increase Payment difficulties for Single/Unmarried & Civil marraige applicants from Non-payment difficulties.

6. **Type of house** : Applicants with House/Appt makes up the majority. Though, people who live with their parents show an increase in Payment difficulties compared to non payment difficulties.

7. **Income Range** : Applicants with Low incomes showed a notable increase in Payment difficulties compared to non payment difficulties.

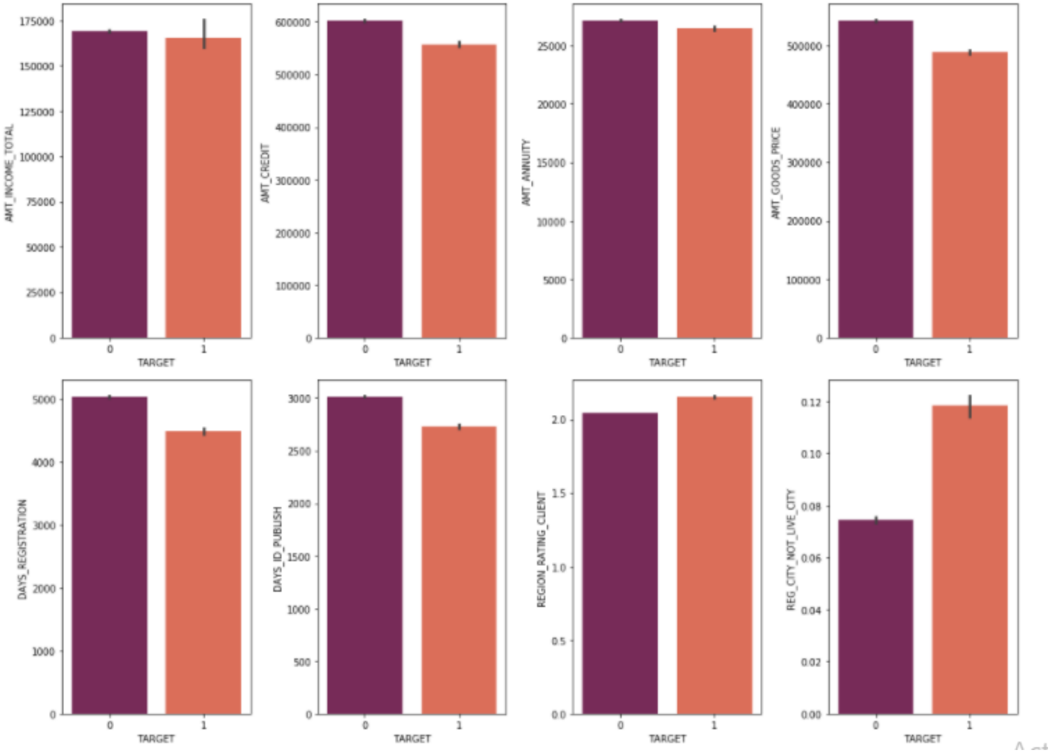
8. **Age Range** : Young Loan applicants saw an increase in Payment difficulties compared to non payment difficulties.

Segmented Univariate Analysis

(Numerical Variables)

- ***Bar Plots***

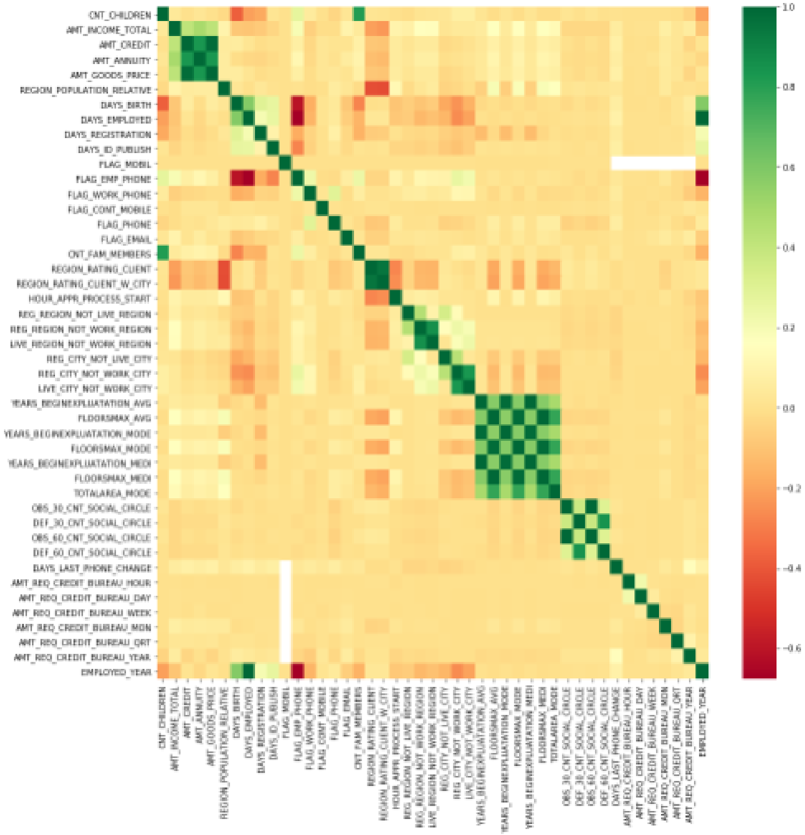
  We plotted bar charts to compare Target variable across various numerical categories in the data:



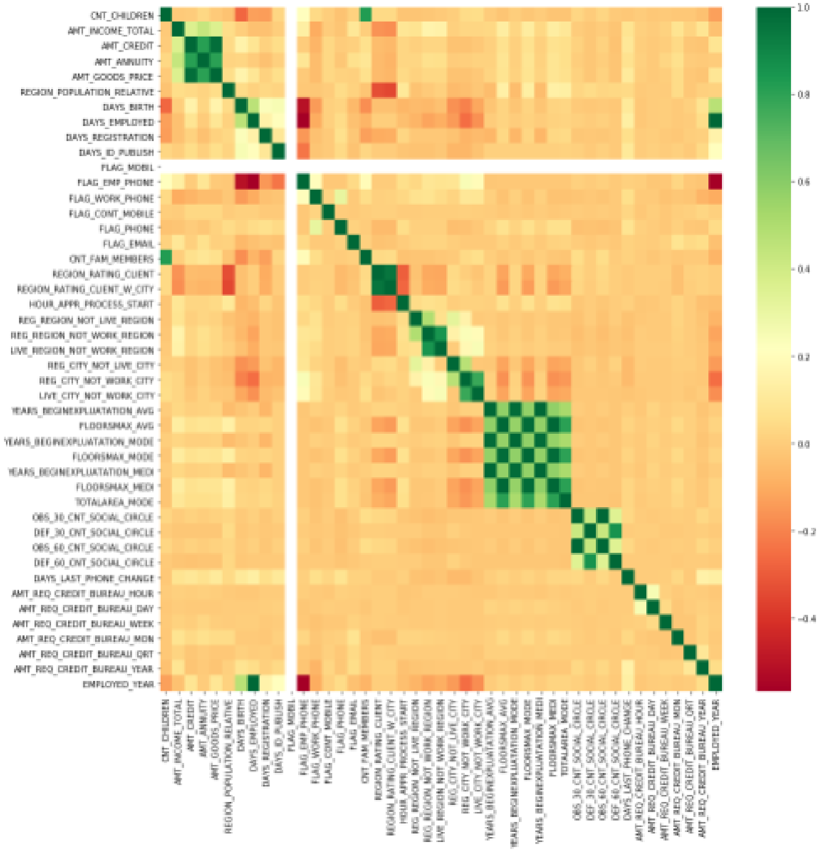
(Fig 8.) Columns like Income, Credit amount, Region rating, etc. for both Target = 0 & 1

- **Observations** : There was a notable rise in Payment difficulties of applicants whose Live & Registered Cities are different than their Work Cities.

Segmented Correlation of variables



(Fig 9.) Correlation Matrix for Target = 0



(Fig 10.) Correlation Matrix for Target = 1

Observations :

There is a **high correlation** between **credit amount** and **goods price**. Some **deviancies** are noticeable in the correlation for **credit amount** & **Income** for both the targets.

Bivariate Analysis
(Categorical v/s Numerical variables)

- **Analysis of Education v/s Credit Amount (basis different Family statuses)**

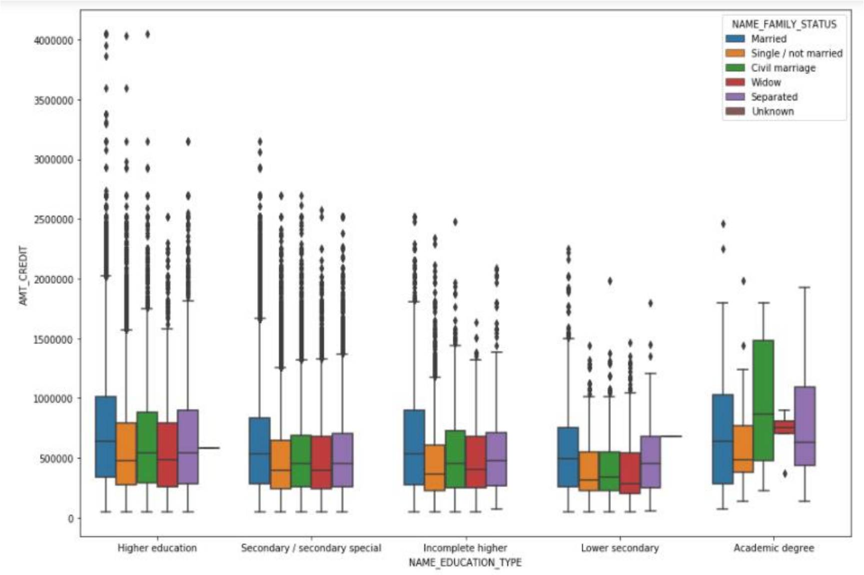


Fig 11. Segmented Box-Whisker for **Target = 0**

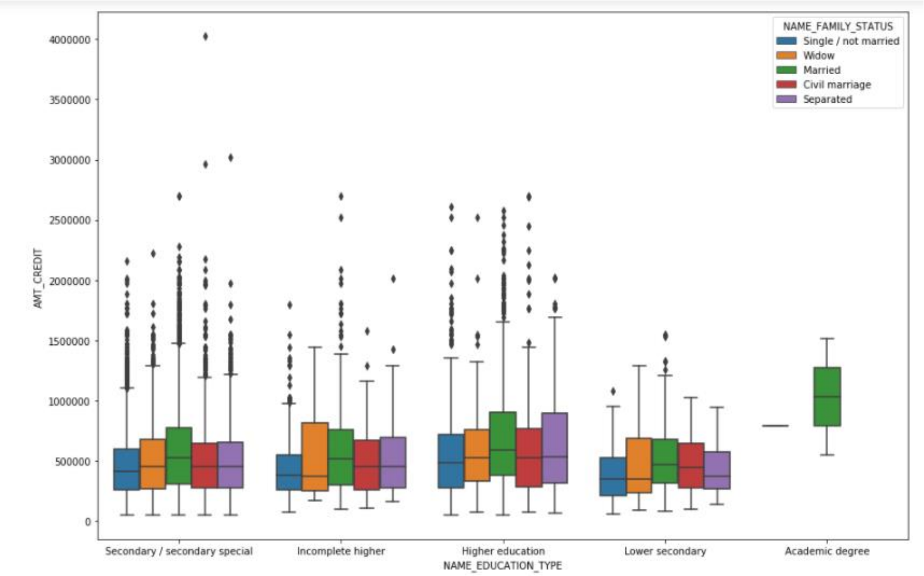


Fig 12. Segmented Box-Whisker for **Target = 1**

- **Observations** :
  The applicants with status of **'civil marriage', 'marriage' and 'separated**' of **Academic degree education** are getting maximum credit sanctioned compared to others. Majority of the outliers are from **'Higher education**' and **'Secondary'. Married** folks made the majority of the records with Loan difficulties with **Academic Degree**.

Bivariate Analysis

(Categorical v/s Numerical variables)

- **Analysis of Income Range v/s Credit Amount (basis different Family statuses)**

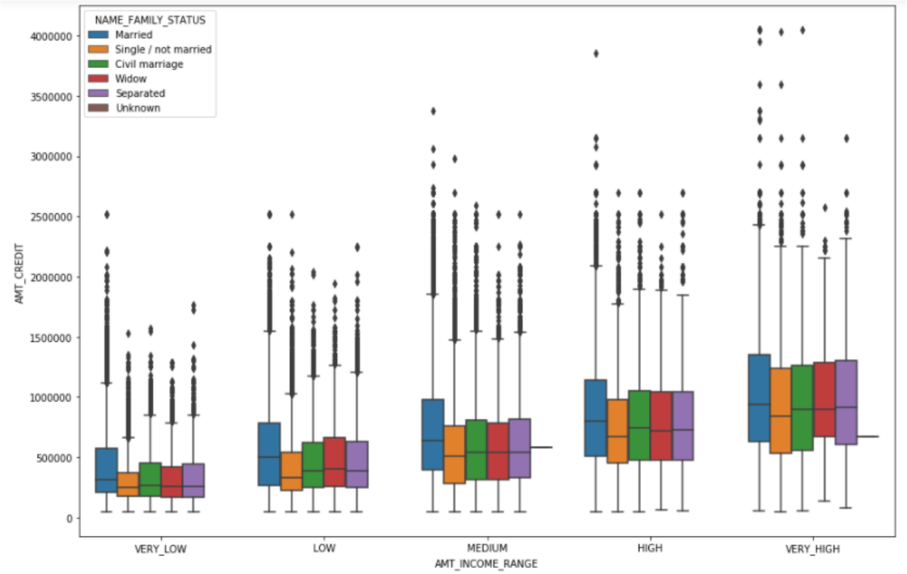


Fig 13. Segmented Box-Whisker for **Target = 0**

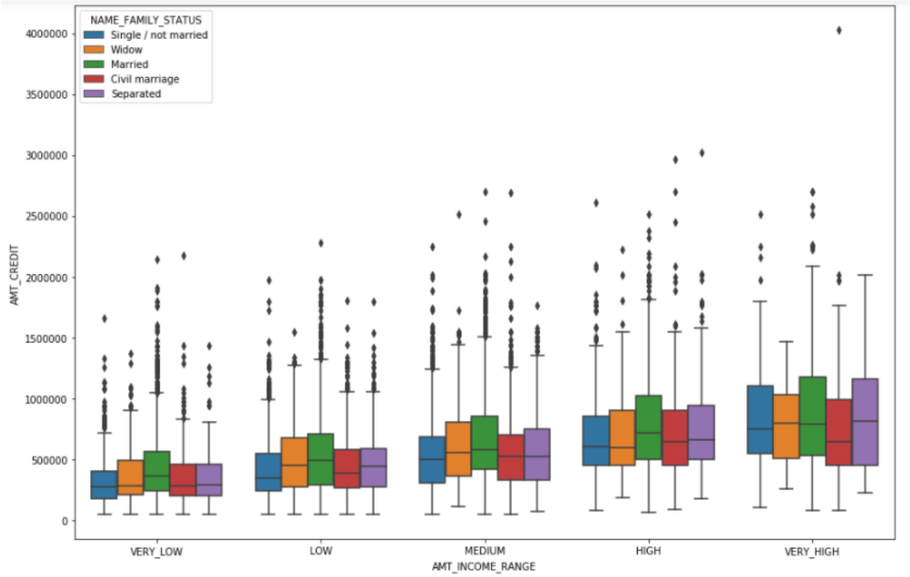


Fig 14. Segmented Box-Whisker for **Target = 1**

- **Observations** :
  Single, Separated & Married applicants with **"Very high" income range** have relatively higher credit sanctioned compared to other categories.

Bivariate Analysis
(Categorical v/s Categorical variables)

Realty Owned Flag v/s Gender

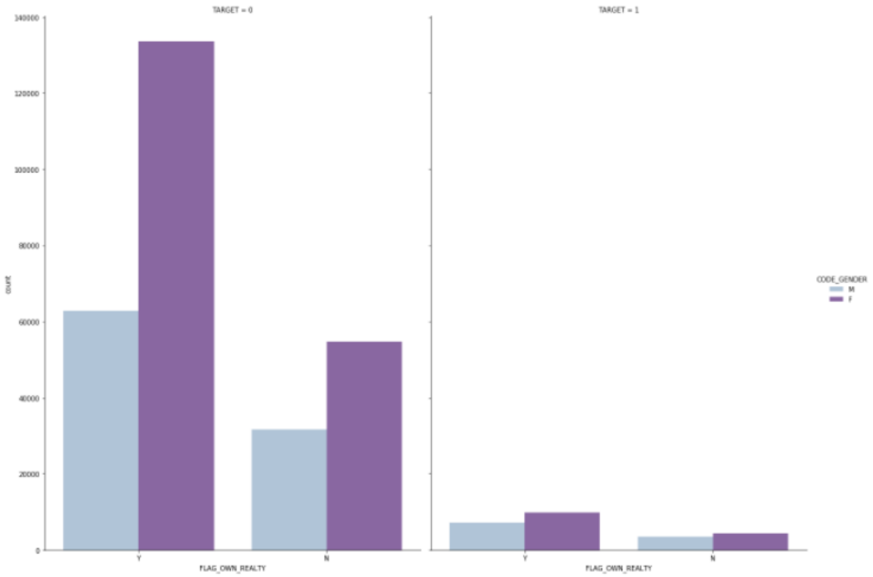


Fig 15. Bar Plots for Flag Owned Realty for **Target = 0 & 1**

Education Type v/s Family Status

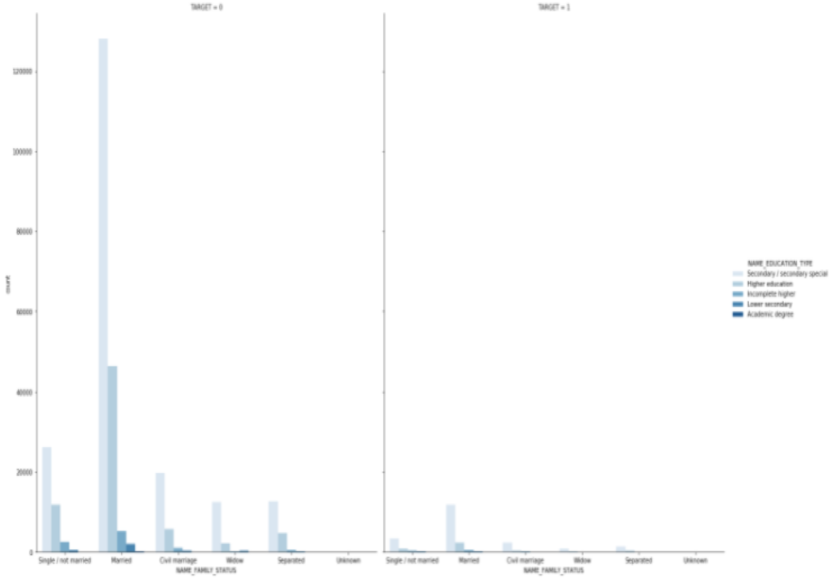
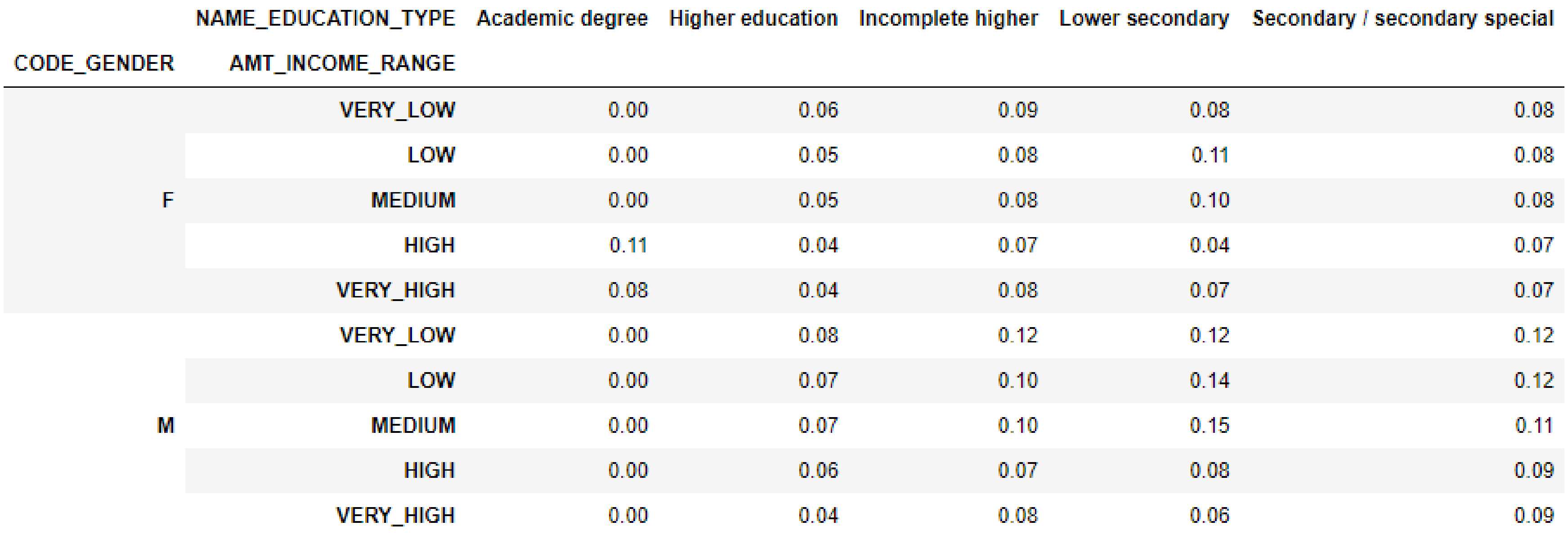


Fig 16. Bar Plots for Family Status for **Target = 0 & 1**

- **Observations** :
  Females were the major **Realty Owners** in applications for **Non-loan Payment difficulties**.
  Married applicants with **Secondary education** levels were the major loan applicants in both the **Target** categories.

Multivariate Analysis
(using Pivot Tables)

Education Type v/s Gender v/s Income Range



	NAME_EDUCATION_TYPE	Academic degree	Higher education	Incomplete higher	Lower secondary	Secondary / secondary special
CODE_GENDER	AMT_INCOME_RANGE					
F	VERY_LOW	0.00	0.06	0.09	0.08	0.08
	LOW	0.00	0.05	0.08	0.11	0.08
	MEDIUM	0.00	0.05	0.08	0.10	0.08
	HIGH	0.11	0.04	0.07	0.04	0.07
	VERY_HIGH	0.08	0.04	0.08	0.07	0.07
M	VERY_LOW	0.00	0.08	0.12	0.12	0.12
	LOW	0.00	0.07	0.10	0.14	0.12
	MEDIUM	0.00	0.07	0.10	0.15	0.11
	HIGH	0.00	0.06	0.07	0.08	0.09
	VERY_HIGH	0.00	0.04	0.08	0.06	0.09

Fig 17. Pivot Table for Education Type, Gender & Income Range for Mean Values of Target.

- **Observations** :
  Female category: Applicants with **LOW income** and **ACADEMIC DEGREE** education have **maximum** number of Loan-Payment Difficulties
  Male category: Applicants with **MEDIUM income** and **LOWER SECONDARY** education have **maximum** number of Loan-Payment Difficulties

Bivariate Analysis

(Numerical v/s Numerical variables)

- ***Pair Plots***
  We plotted Pair Plots, which are a collection of Scatter & density plots for comparison amongst numerical variables. We performed this for both Target = 0 & 1.

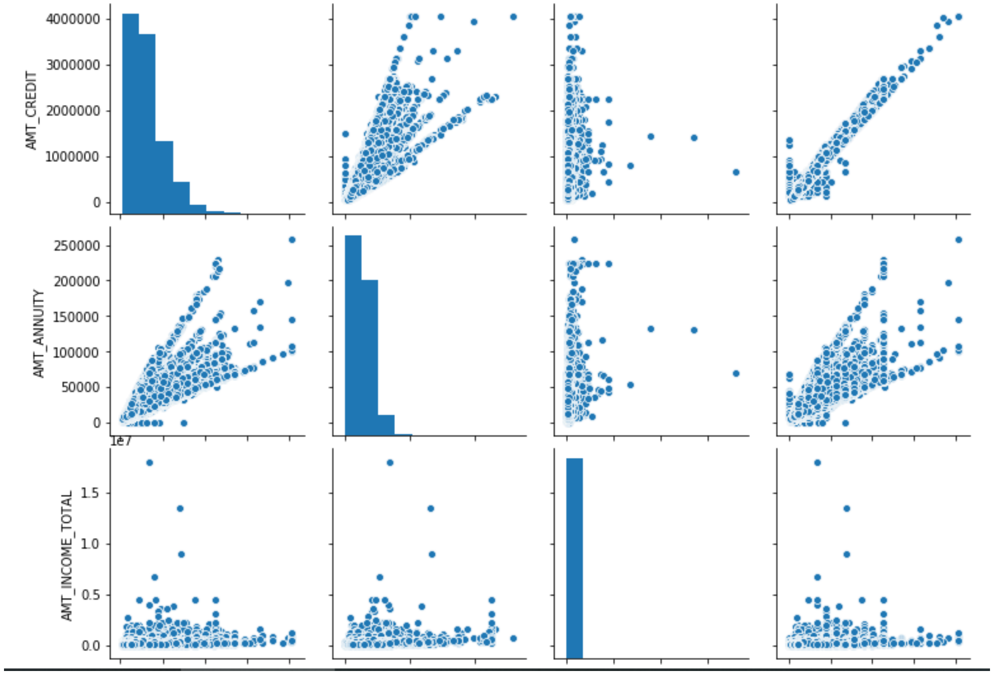


Fig 18. Sample of Pair Plots across some numerical variables for **Target = 0 & 1**

- **Observations** :
  High correlation between **Credit Amount & Goods Price**.

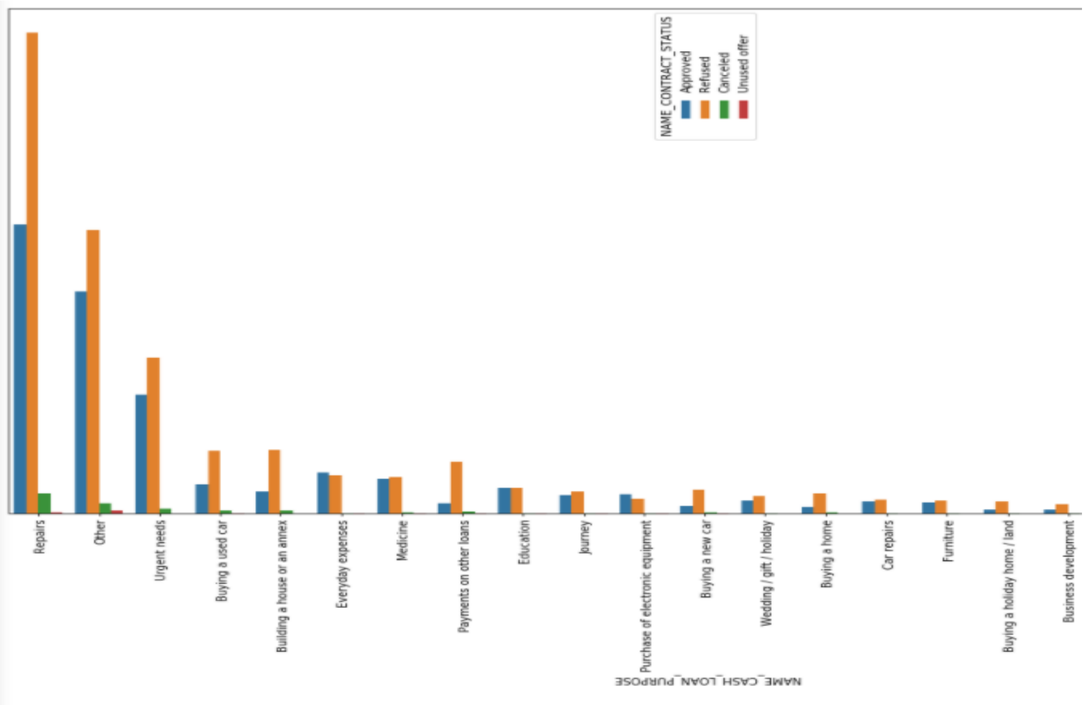
Analysis of **Previous Application Data**

- **Dimension & Data quality** checks performed.
- Pulled up **Summary Statistics** of the Previous application data.
- Treated **missing/null** & **logically wrong** values.
- Dropped duplicates.

Univariate Analysis **(Previous Application data)**

- ***Bar Plots***

Cash Loan Purpose v/s Cash Loan Status



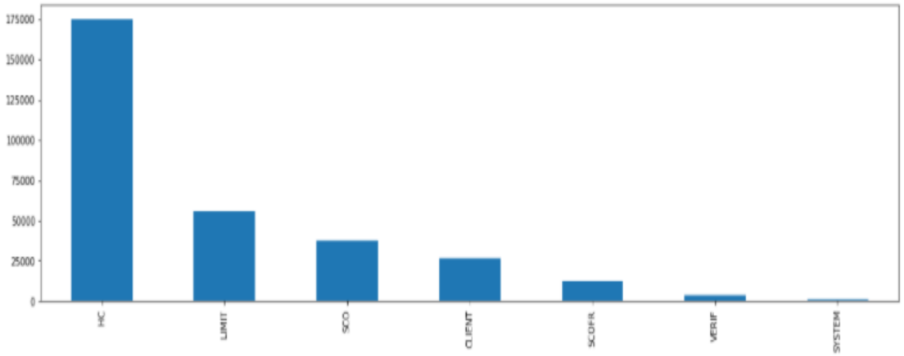
(Fig 19.) Bar Plot for Cash Loan Purpose v/s Cash Loan Status

- **Observations:**

1. **Repairs** as a Cash Loan purpose showed most loan **rejections**.
2. For purpose of **Education** the **Approval : Rejection** ratio was close to 1:1
3. **Buying a car & Paying other loans** had **higher rejection** rate.

Analysis of **Previous Application Data**

Top Rejection Reasons

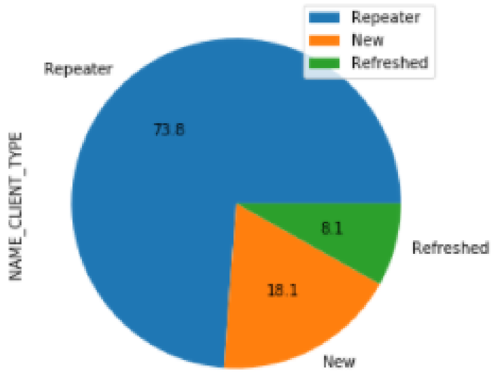


(Fig 20.) Bar Plot for Top Loan rejection reasons.

- **Observations:**

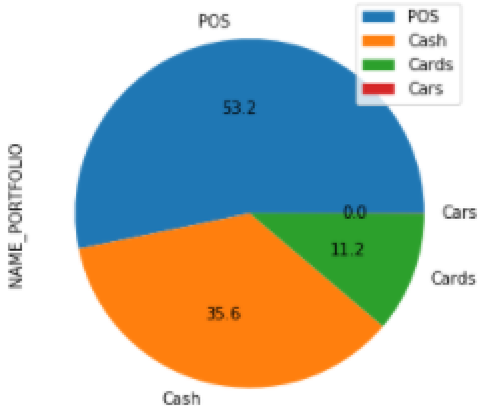
  - HC turns out to be the primary reason for loan application rejection.
  - Majority of the loan applications came from Repeater clients.
  - Majority of loan applications were for POS & Cash.

Client Type Distribution



(Fig 21.) Pie Plot for Client Type Distribution

Portfolio Distribution

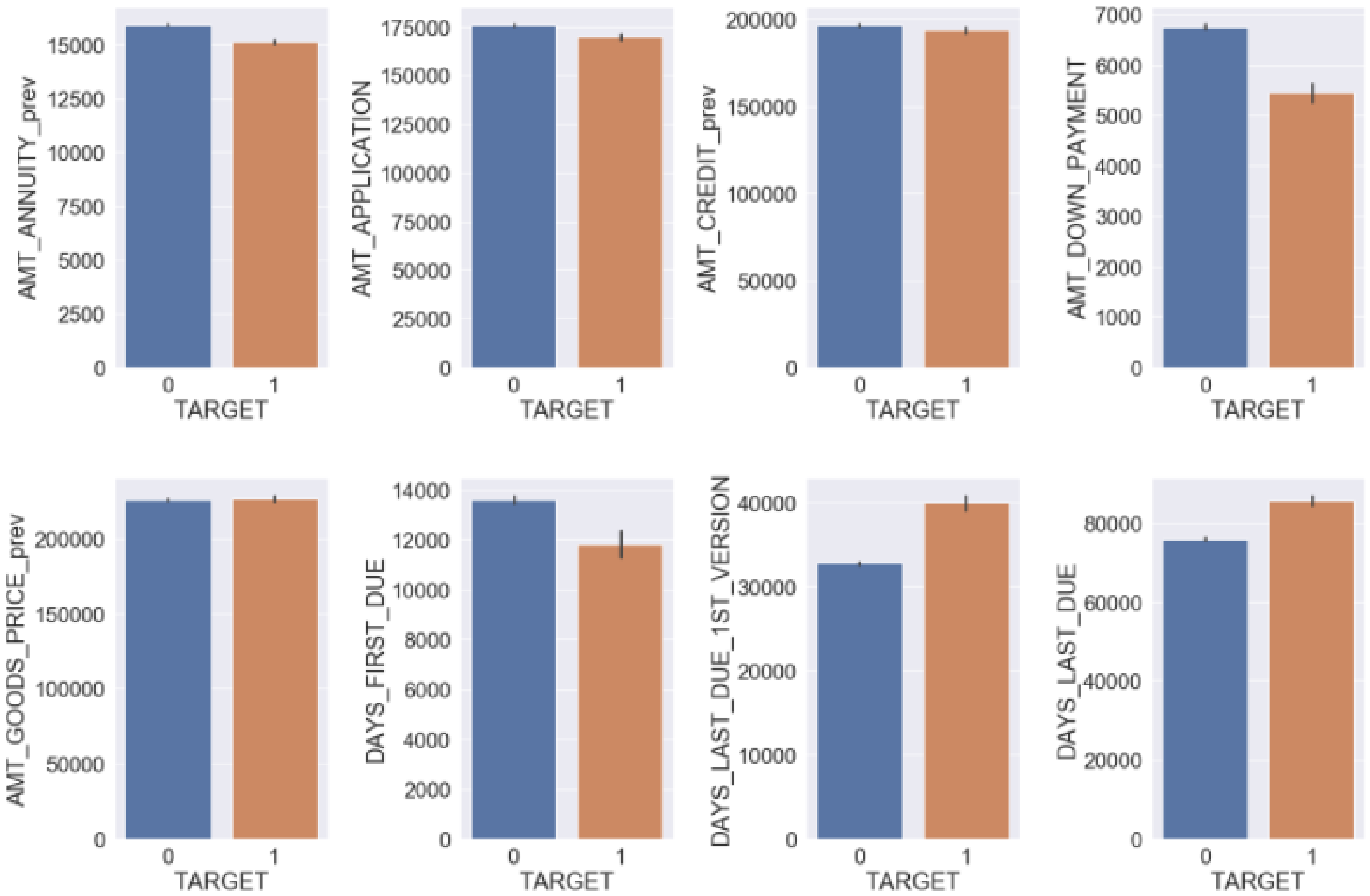


(Fig 22.) Pie Plot for Portfolio Distribution

Analysis of merged data

(Application & Prev. Application)

- ***Segmented Univariate Analysis - Numerical***

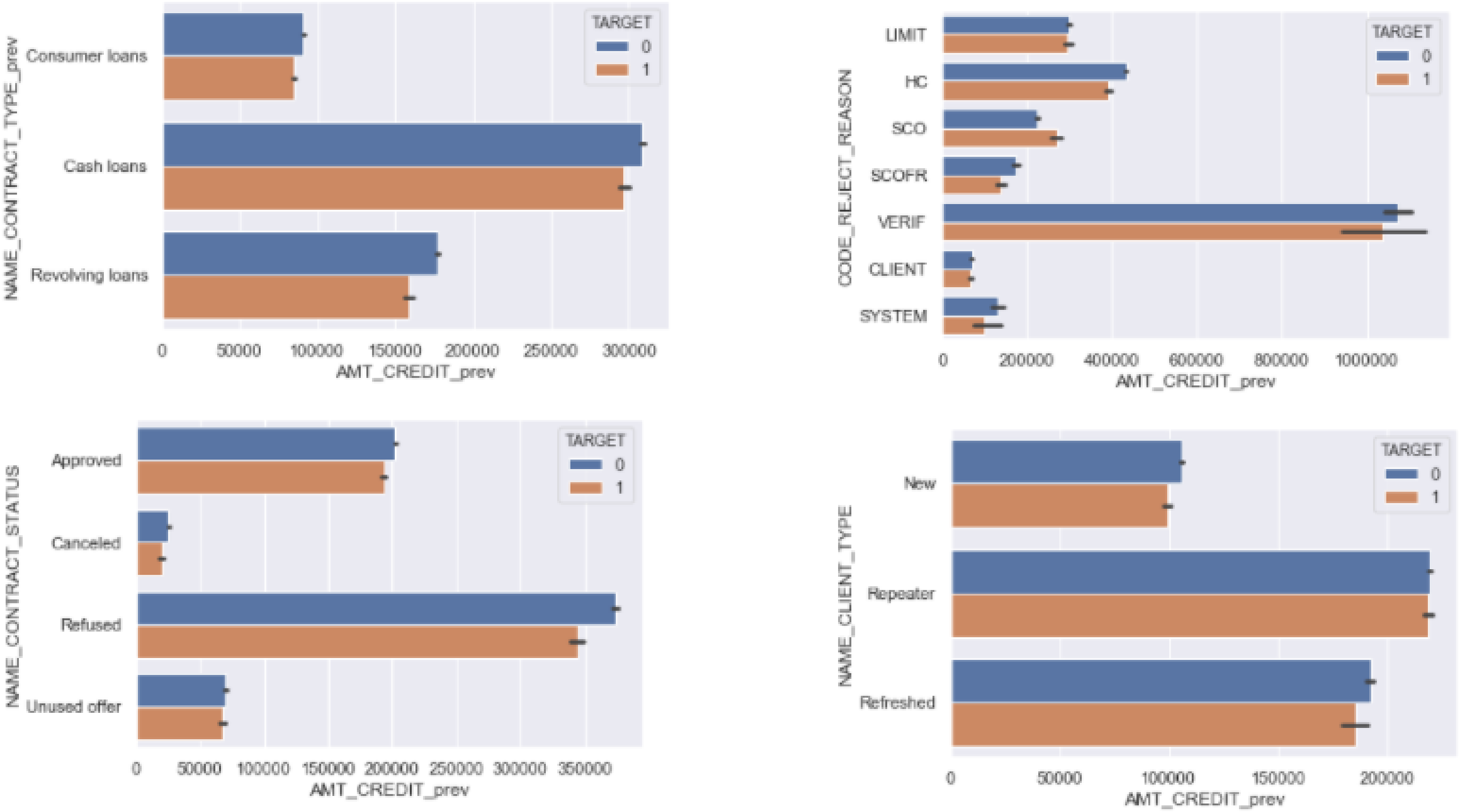


(Fig 23.) Bar Plots for both Target = 0 & 1

Analysis of merged data

(Application & Prev. Application)

- ***Segmented Univariate Analysis – Categorical***



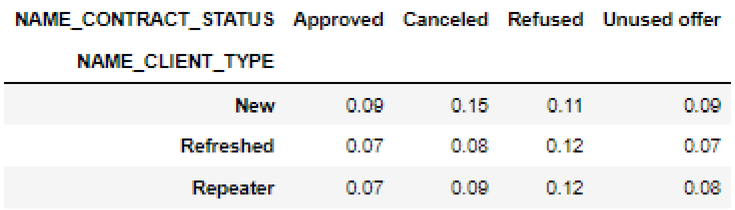
(Fig 24.) Bar Plots for both Target = 0 & 1

Multivariate Analysis of merged data

(Application & Prev. Application)

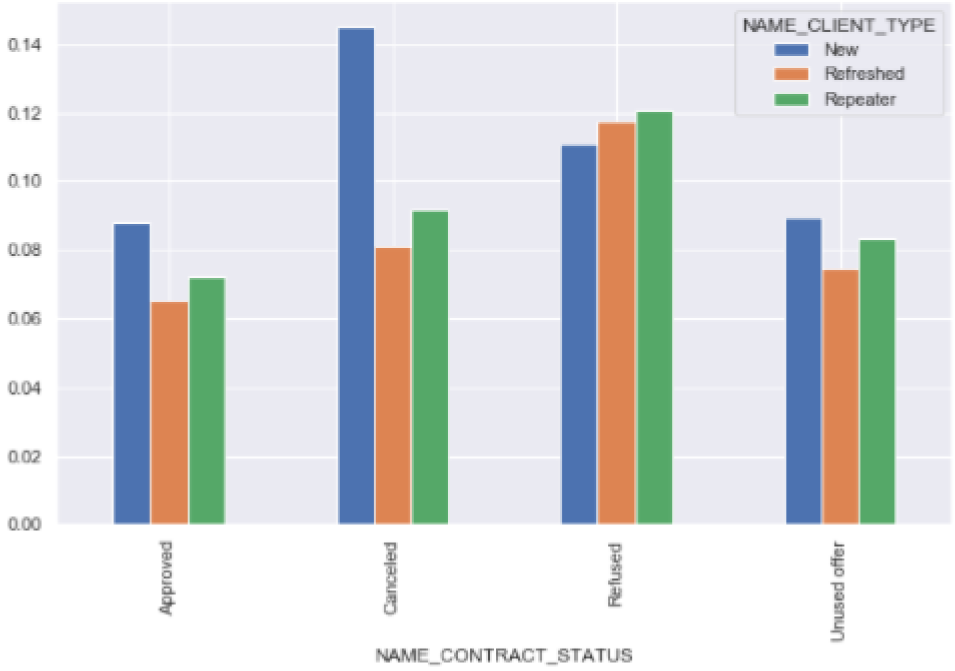
- ***Pivot Table & Bar Plots***

Client Type v/s Contract Status v/s Target



NAME_CONTRACT_STATUS	Approved	Canceled	Refused	Unused offer
NAME_CLIENT_TYPE				
New	0.09	0.15	0.11	0.09
Refreshed	0.07	0.08	0.12	0.07
Repeater	0.07	0.09	0.12	0.08

(Fig 25.) Pivot Table for Client Type & Contract Status for Mean Values of Target.



(Fig 26.) Bar Plots for Fig.25

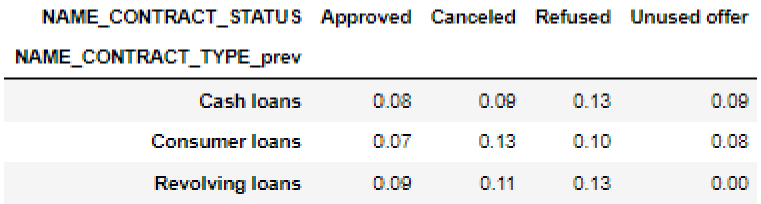
- **Observations:**
  Applicants with 'New' and 'Cancelled' previous applications have more Loan-Payment Difficulties in current application

Multivariate Analysis of merged data

(Application & Prev. Application)

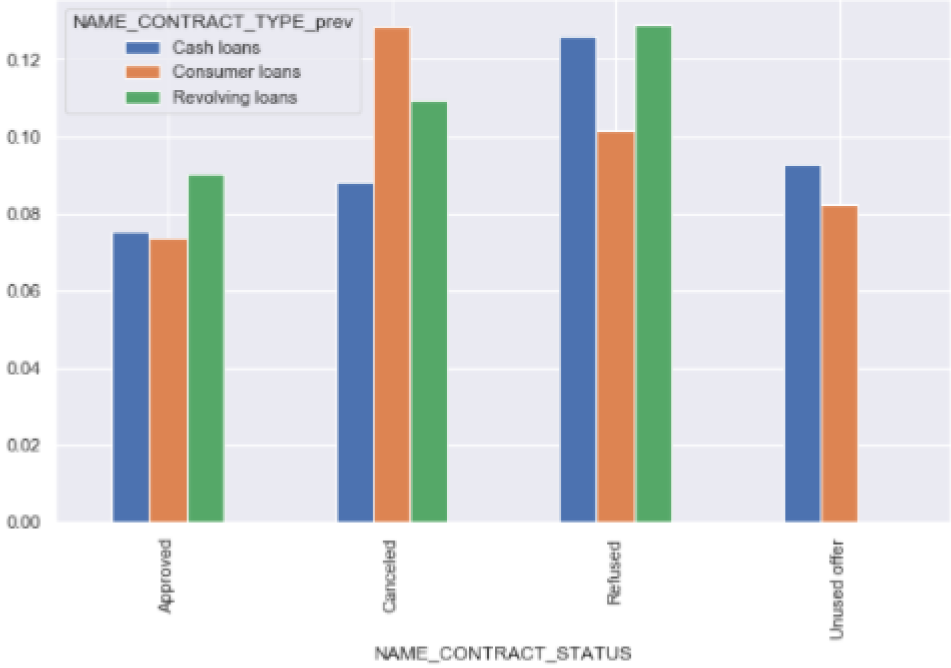
- ***Pivot Table & Bar Plots***

Contract Type v/s Contract Status v/s Target



NAME_CONTRACT_STATUS	Approved	Canceled	Refused	Unused offer
NAME_CONTRACT_TYPE_prev				
Cash loans	0.08	0.09	0.13	0.09
Consumer loans	0.07	0.13	0.10	0.08
Revolving loans	0.09	0.11	0.13	0.00

(Fig 27.) Pivot Table for Contract Type & Contract Status for Mean Values of Target.



(Fig 28.) Bar Plots for Fig.27

- **Observations:**
  Applicants with 'Revolving loans' & 'Cash' and with 'Refused' previous applications face more Loan-Payment Difficulties in current application

Recommendations

- The Client should focus more on Contract type **‘Student’ ,’pensioner’ and ‘Businessmen’** with Housing type other than ‘**Co-op apartment**’ for **successful payments**.

- The Client can afford to focus less on Income type ‘**Working**’ as they are having most number of unsuccessful payments. They should also investigate for other admin related causes or take a survey for these unsuccessful payments.

- The Client can target potential applicants from Housing type ‘**With parents’** as they have the **least** number of **unsuccessful** payments.