Credit EDA Case Study

By Gunavardhan Sattu

Business Objective

- Use data analysis to assist a loan company in making better lending choices
- We aim to avoid lending to unreliable borrowers and ensure those who can repay aren't turned away
- By looking at data, we aim to make loan approvals better while also avoiding potential losses.
- Figure out the main reasons why people might not be able to pay back loans.
- We want to find clues in the data that can help us predict if someone might not pay on time.

Approach

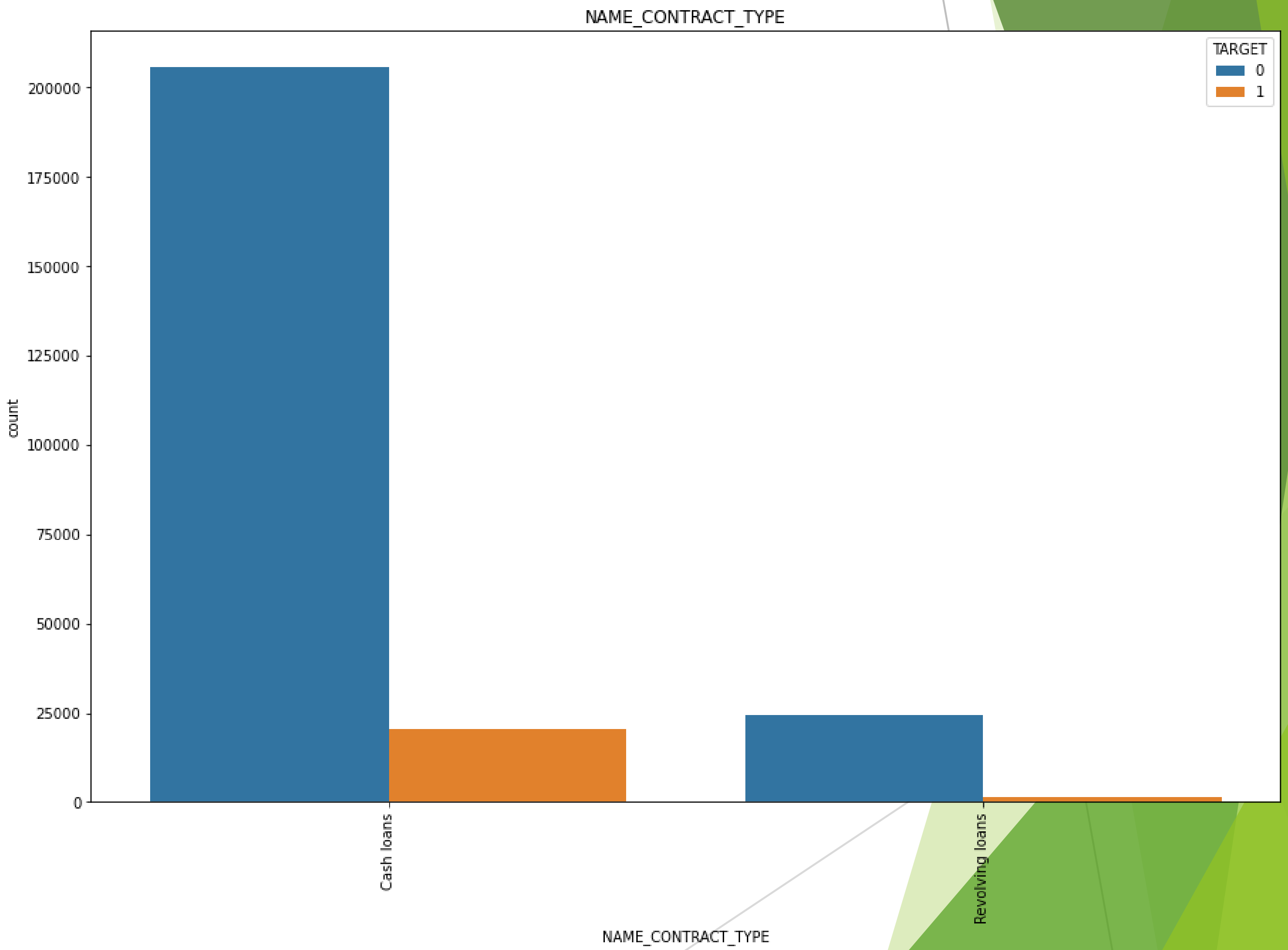
- Data Understanding
- Data Cleaning
  - Removing columns with Null values greater than 40
  - Impute/Remove missing values
  - Fixing invalid values and Filter Data
- Univariate Analysis
- Bivariate Analysis
- Multivariate Analysis

Assumptions

- Setting a threshold of 40% to eliminate columns with null values
- Filling missing values with median or mode
- In the dataset, the "TARGET" column has been divided into "Target 1" for clients with payment difficulties and "Target 0" for all other cases
- Columns containing values like "XPA," "XNA," or other non-relevant values are treated as missing data

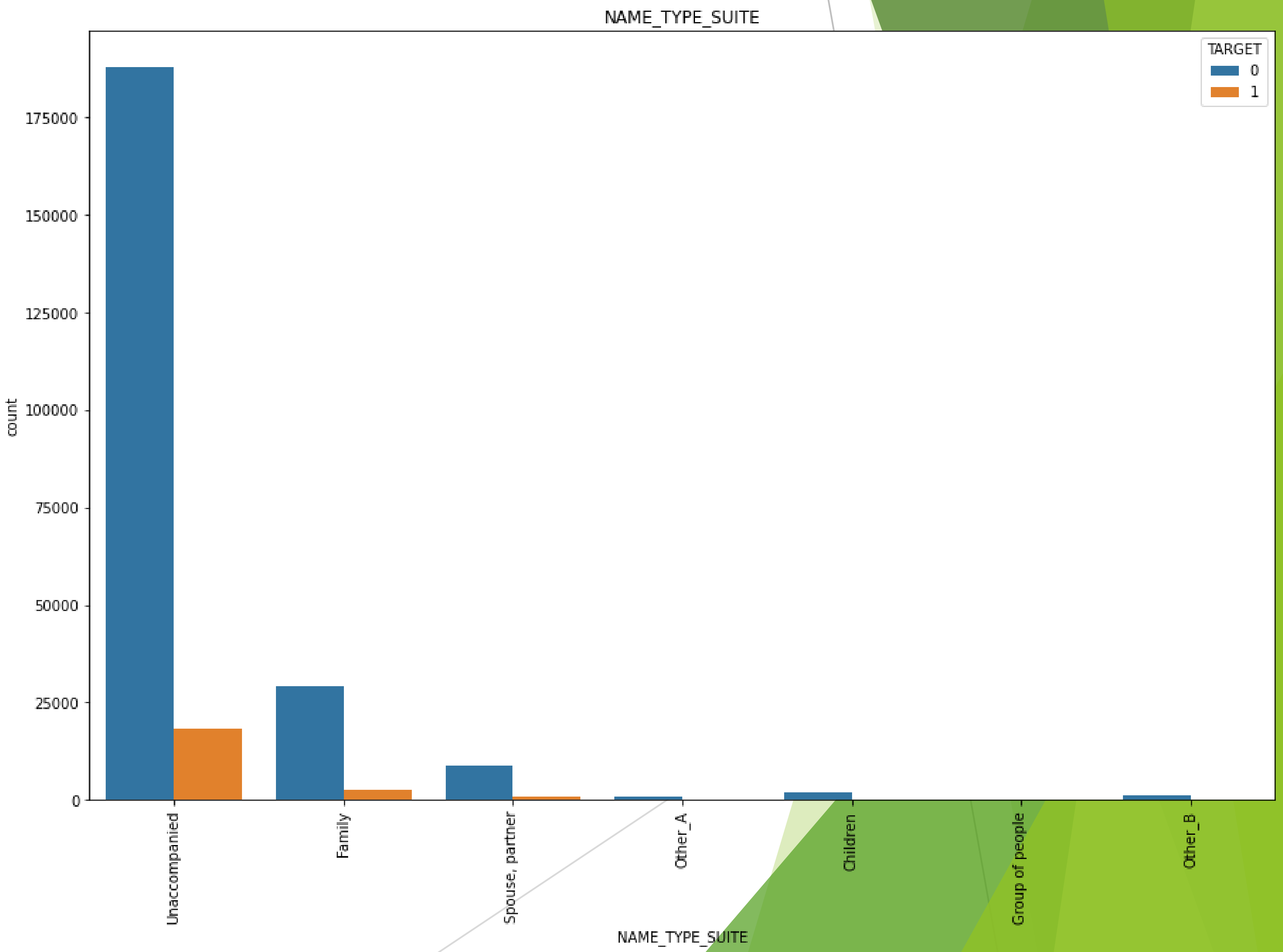
Univariate Analysis

- The majority opted for cash loans, with a higher 9% default rate, compared to a 5.65% default rate for revolving loans.



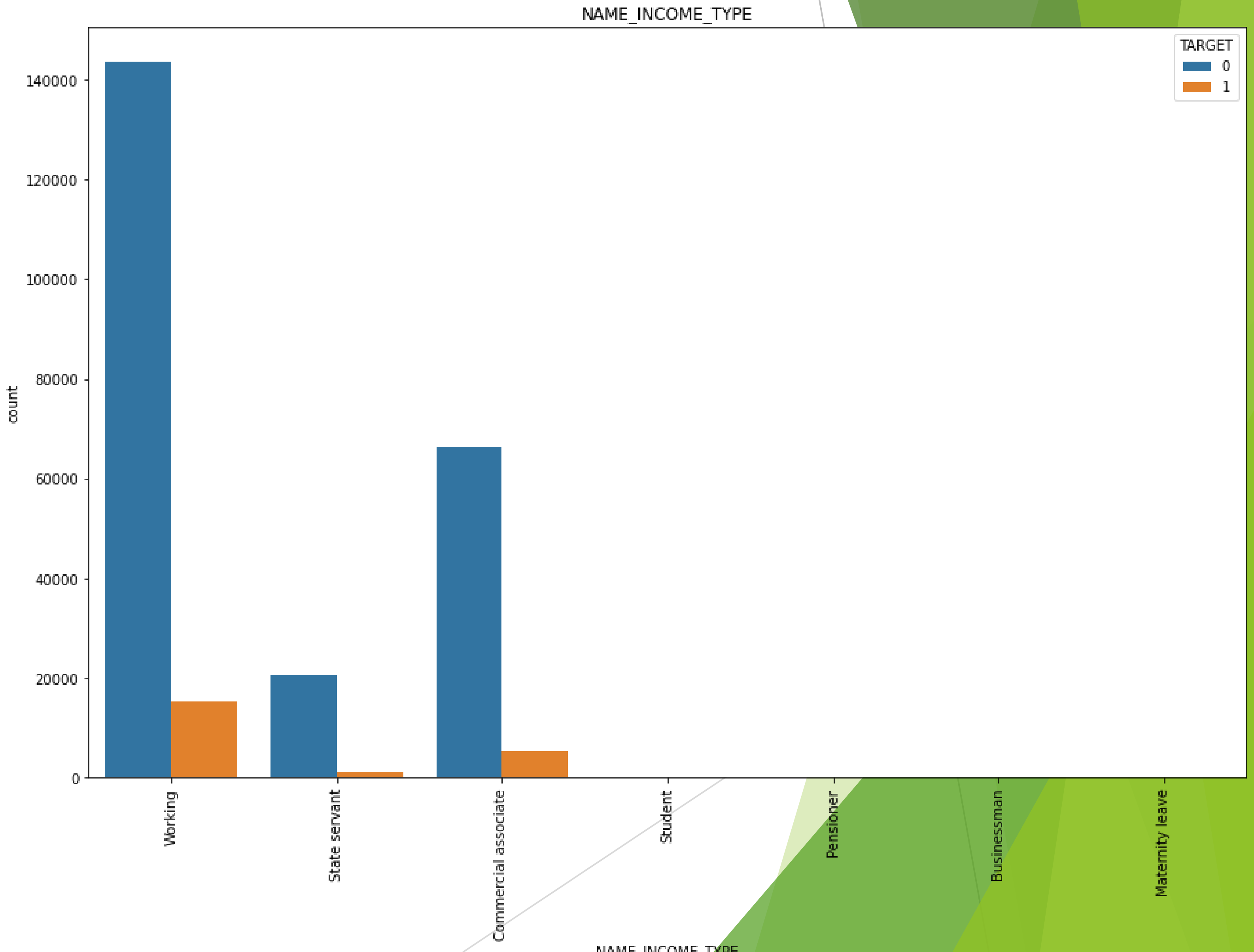
Univariate Analysis

- ❑ Accompanied customers take more loans, with an 8.5% default rate. Other_B name type is prone to default, while clients with children are less likely to default.



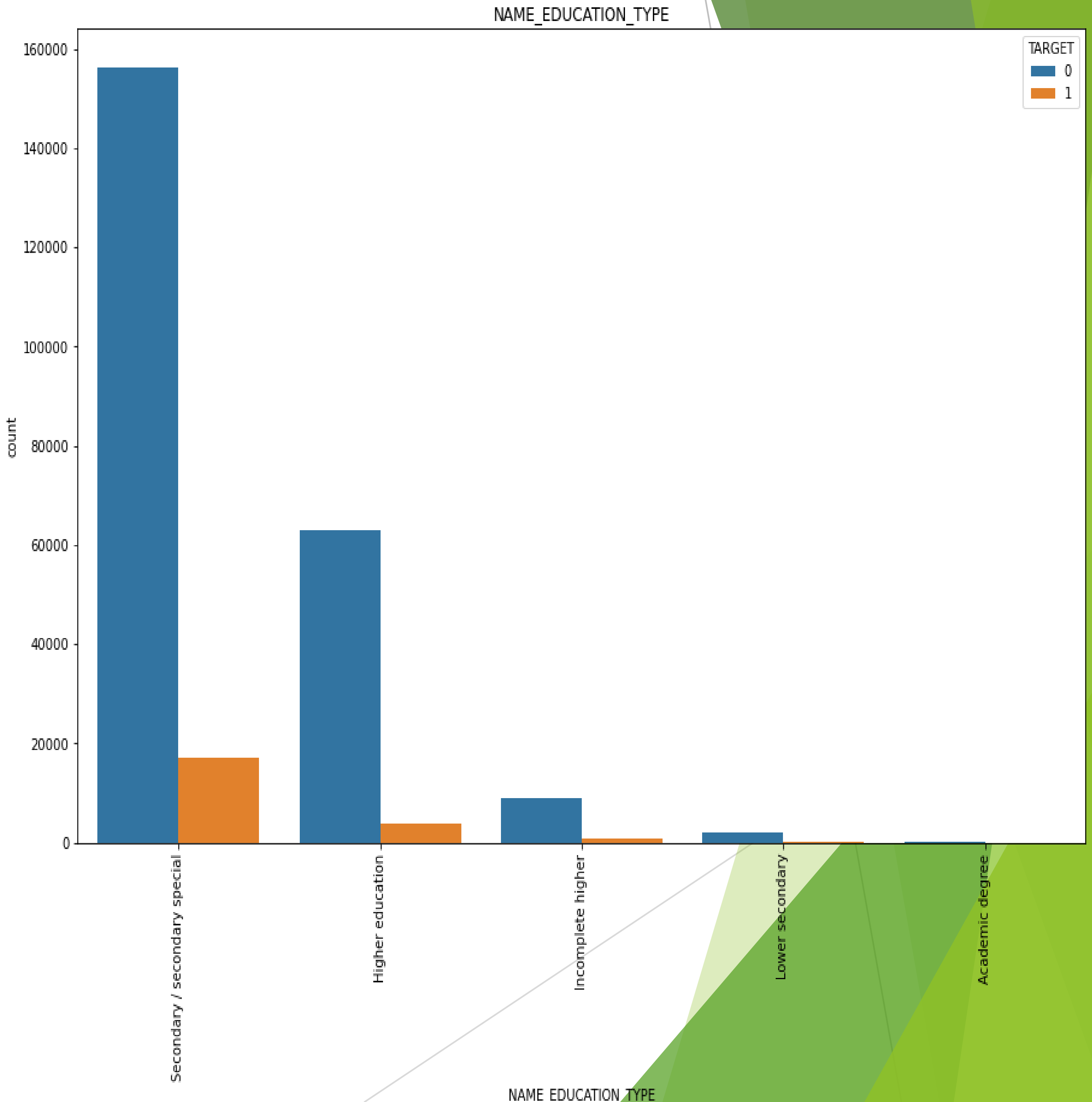
Univariate Analysis

- Working professionals have a 9.5% default rate, seeming secure with higher loan count. Maternity leave individuals show higher default rates.



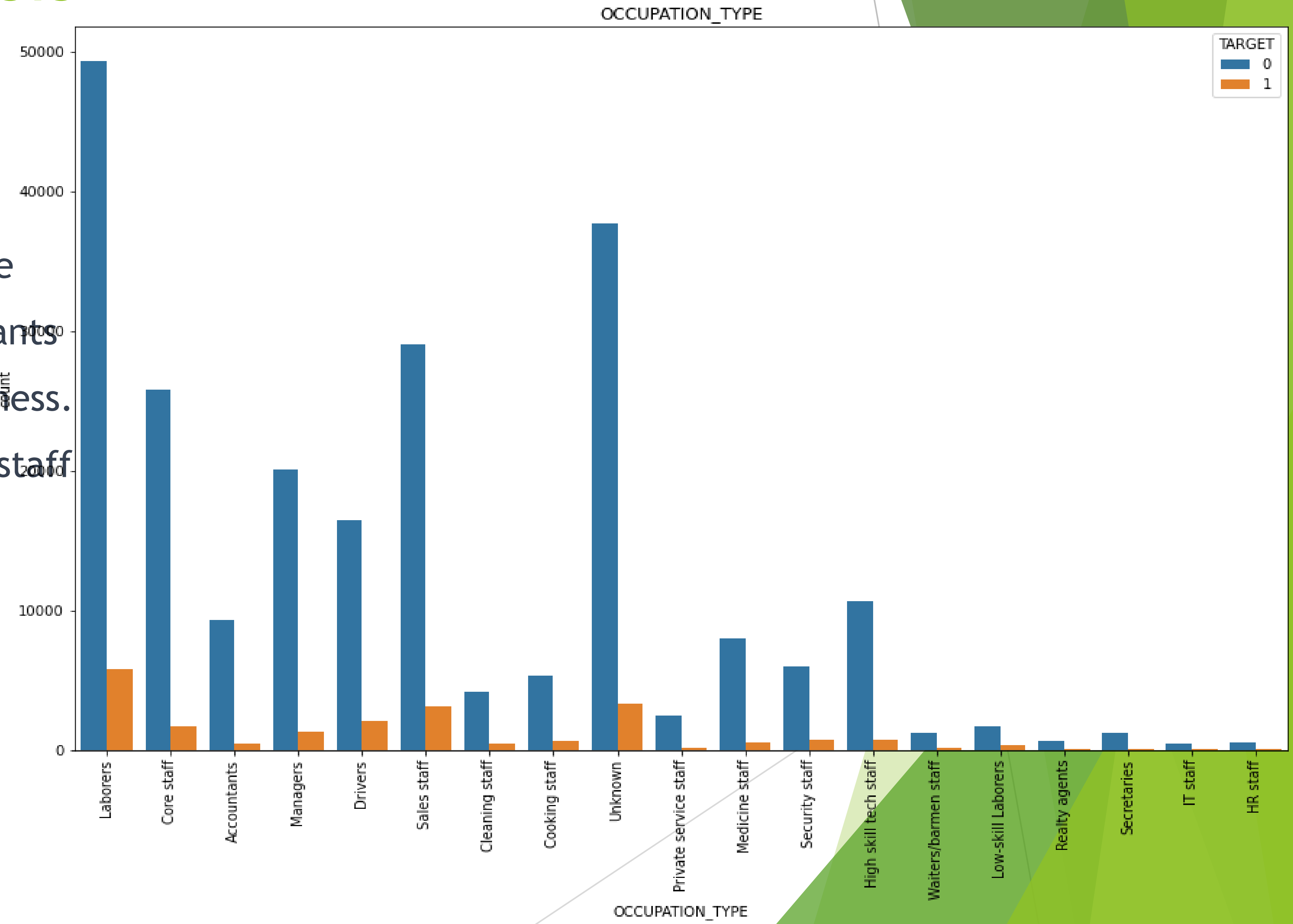
Univariate Analysis

- ❑ People who are low educated are more likely to default whereas people with higher education and academic degree are less likely to default.



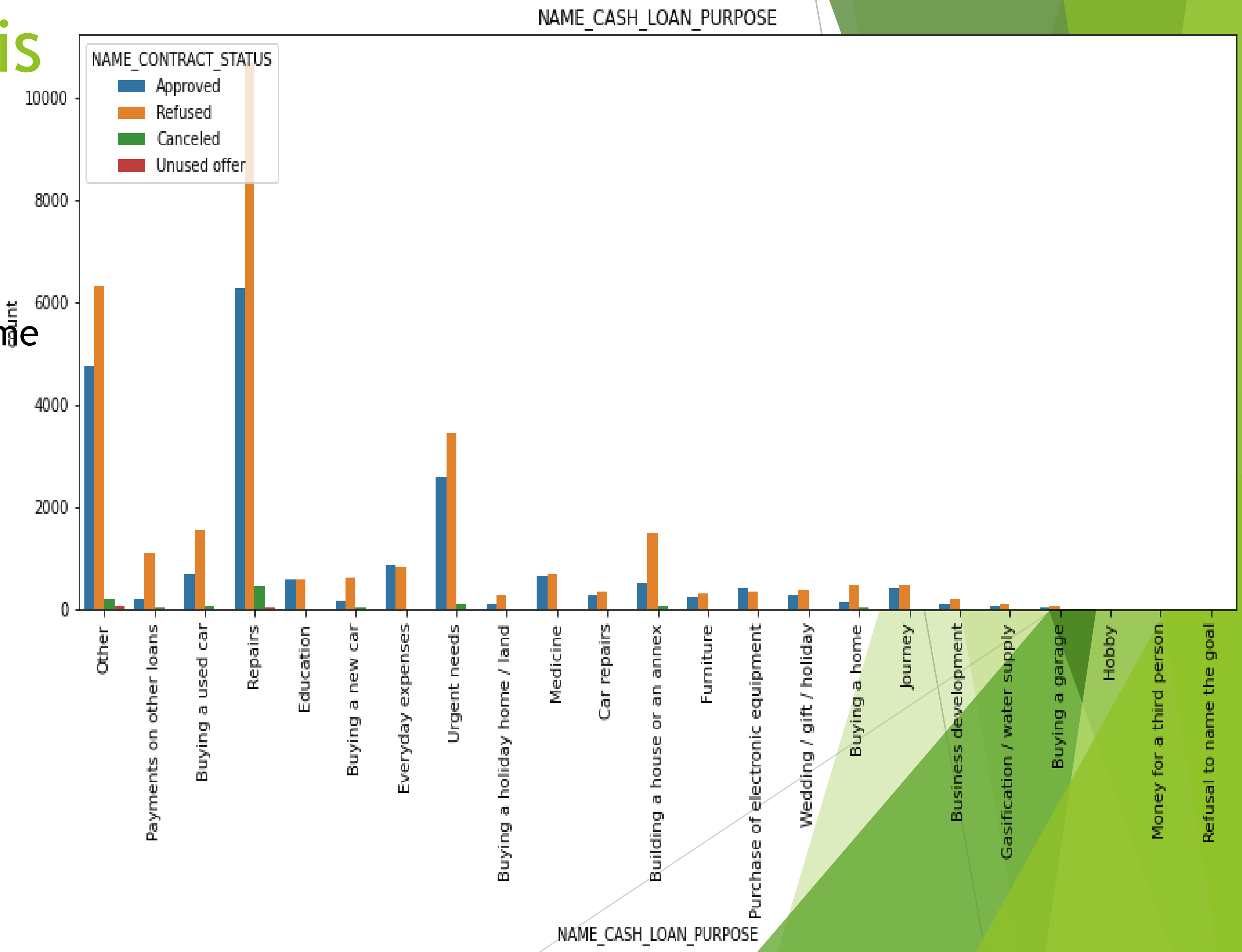
Univariate Analysis

- Low-skill workers like laborers, drivers, waiters, and barmen are likely to default, while accountants have low defaults due to awareness. Core/HR/IT/Manager/Medicine staff show low default rates.



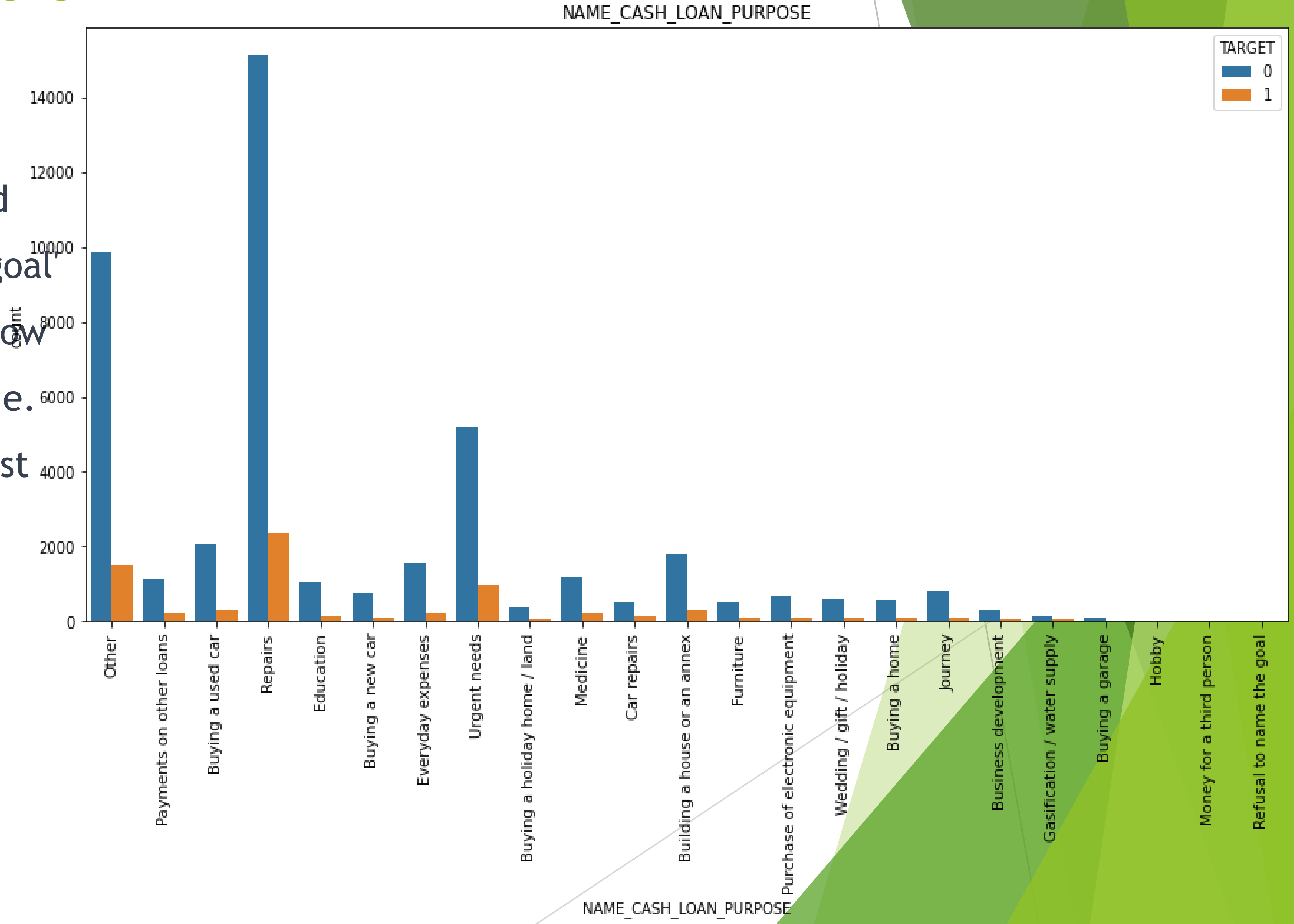
Univariate Analysis

- Most of the customers applied for the purpose of Repairs and the same type has most of the refusals and cancellations



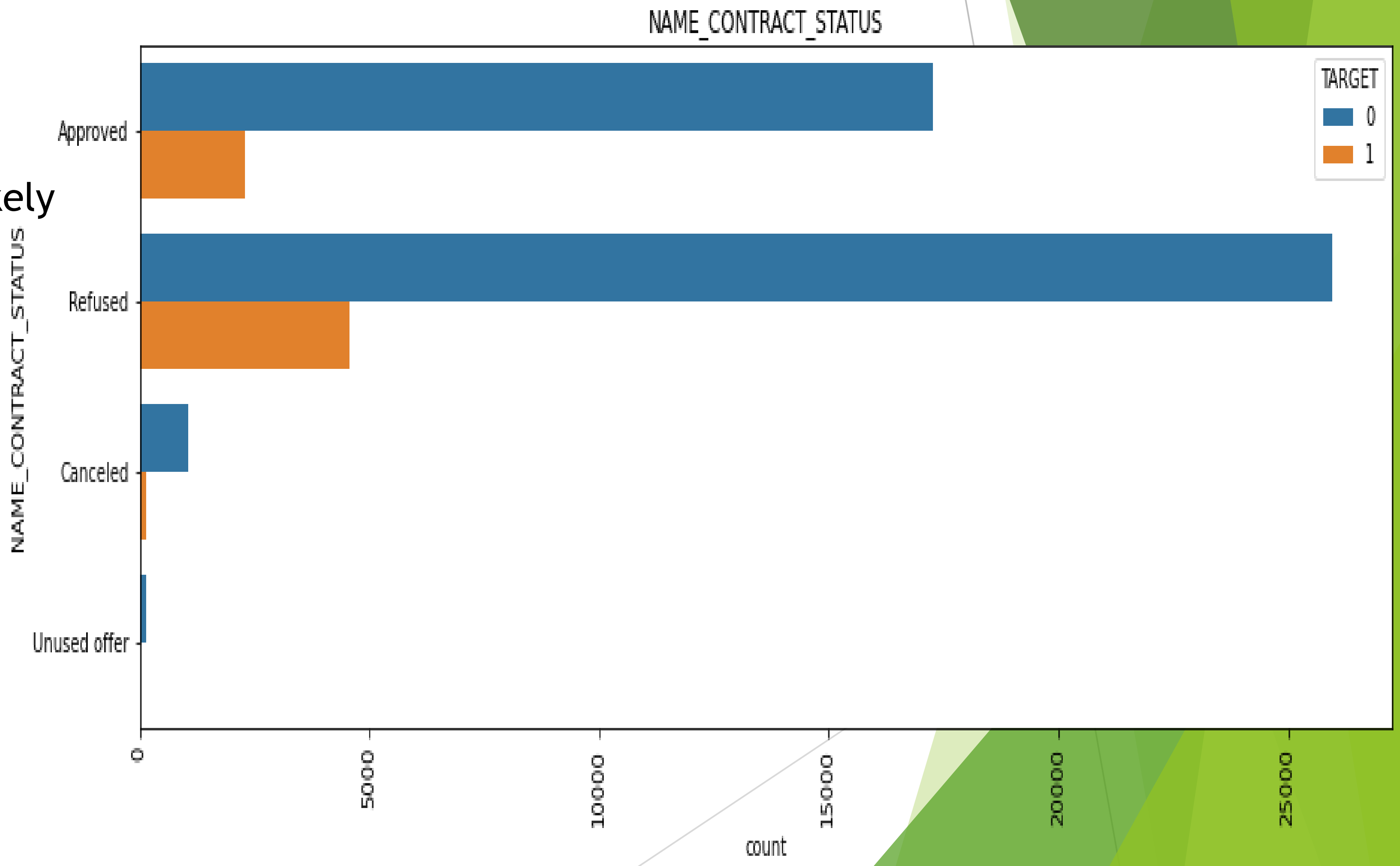
Univariate Analysis

- Repairs have high defaulters and repayers. 'Refusal to name the goal' & 'Gasification/water supply' show 25% default rate with low volume. Further analysis based on interest rates is needed for a clearer picture.



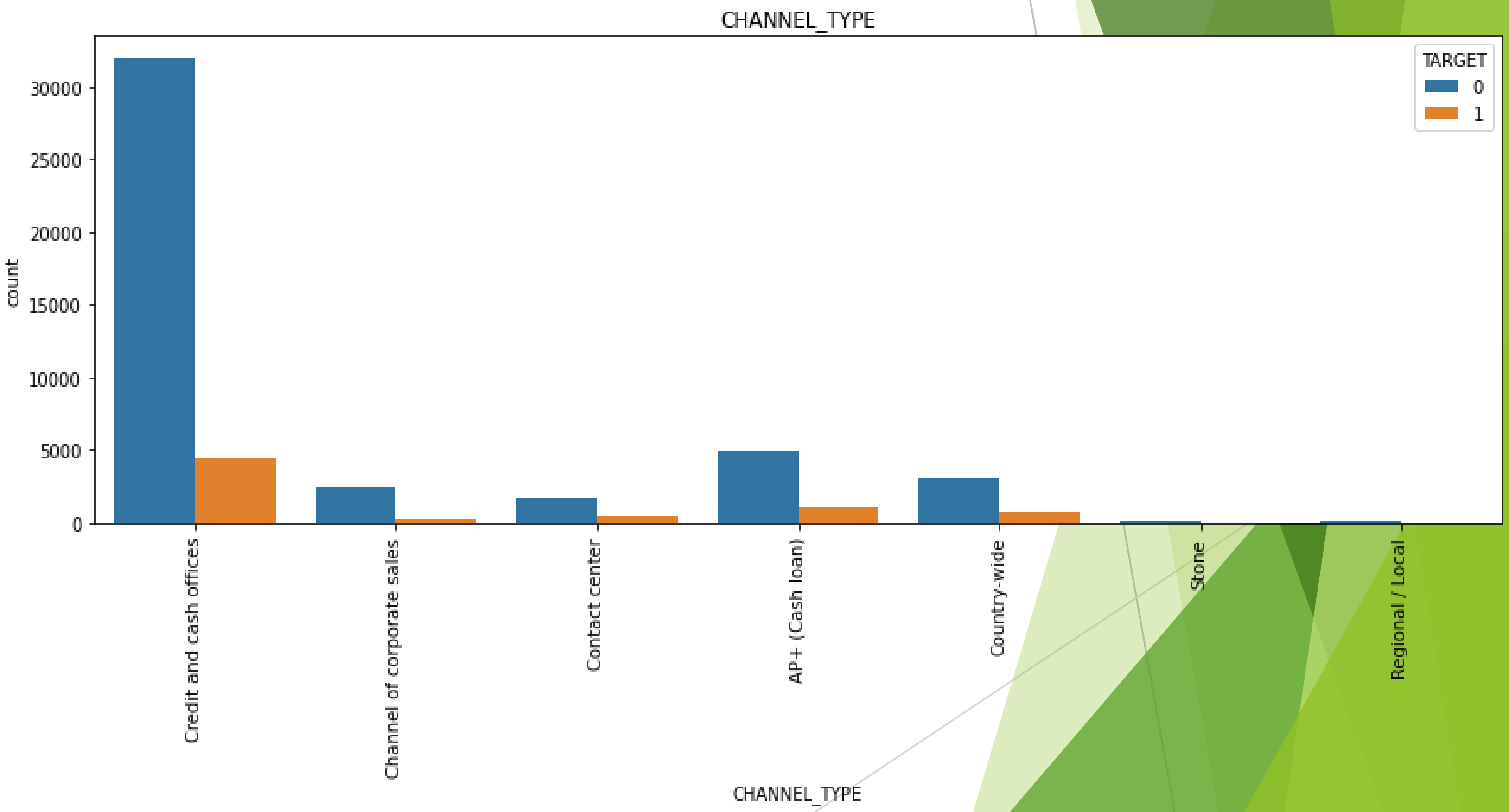
Univariate Analysis

- Applicants whose loans were previously refused are more likely to be defaulters



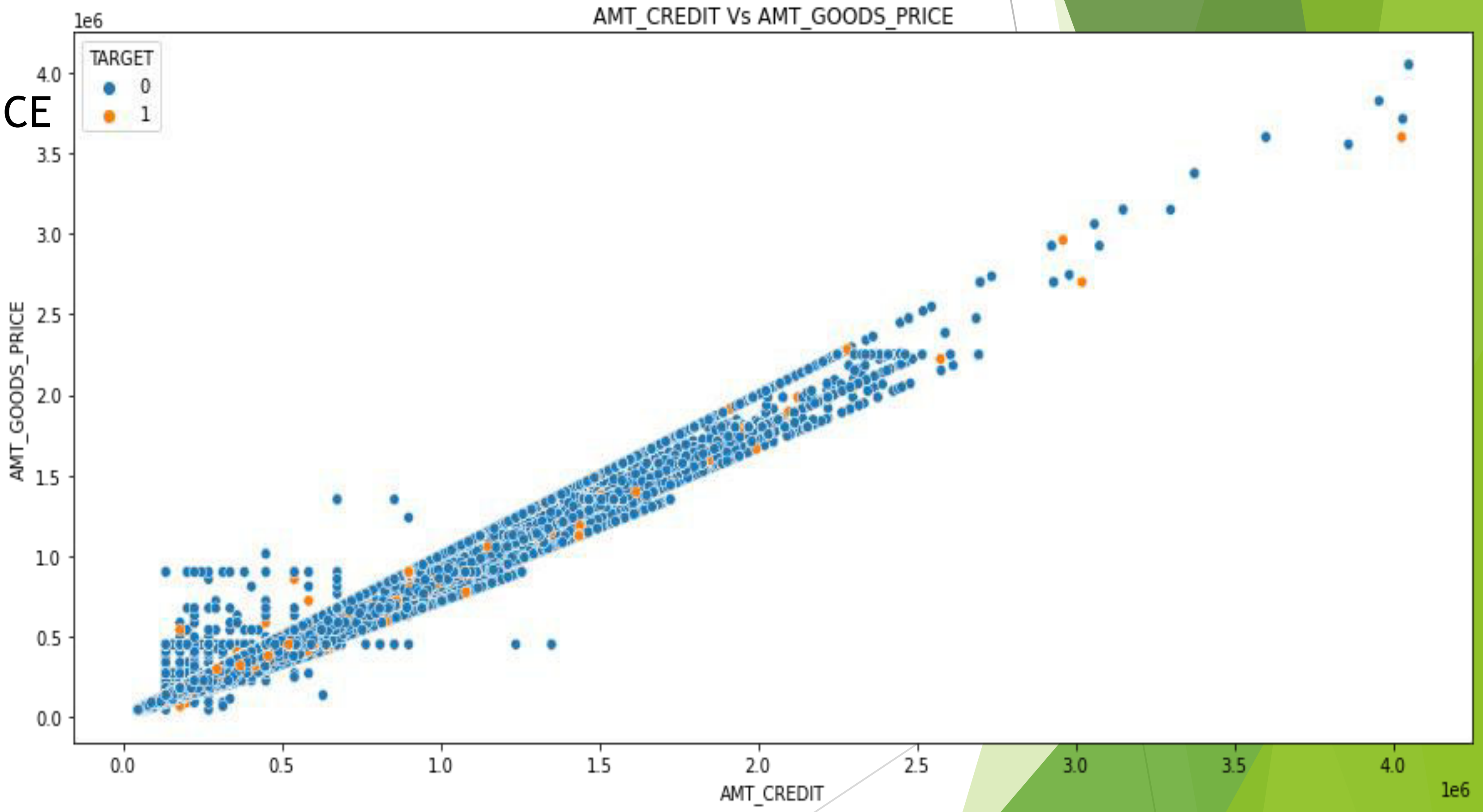
Univariate Analysis

- The most effective channels for acquiring new customers with a lower rate of defaulters are corporate sales and regional/local channels.



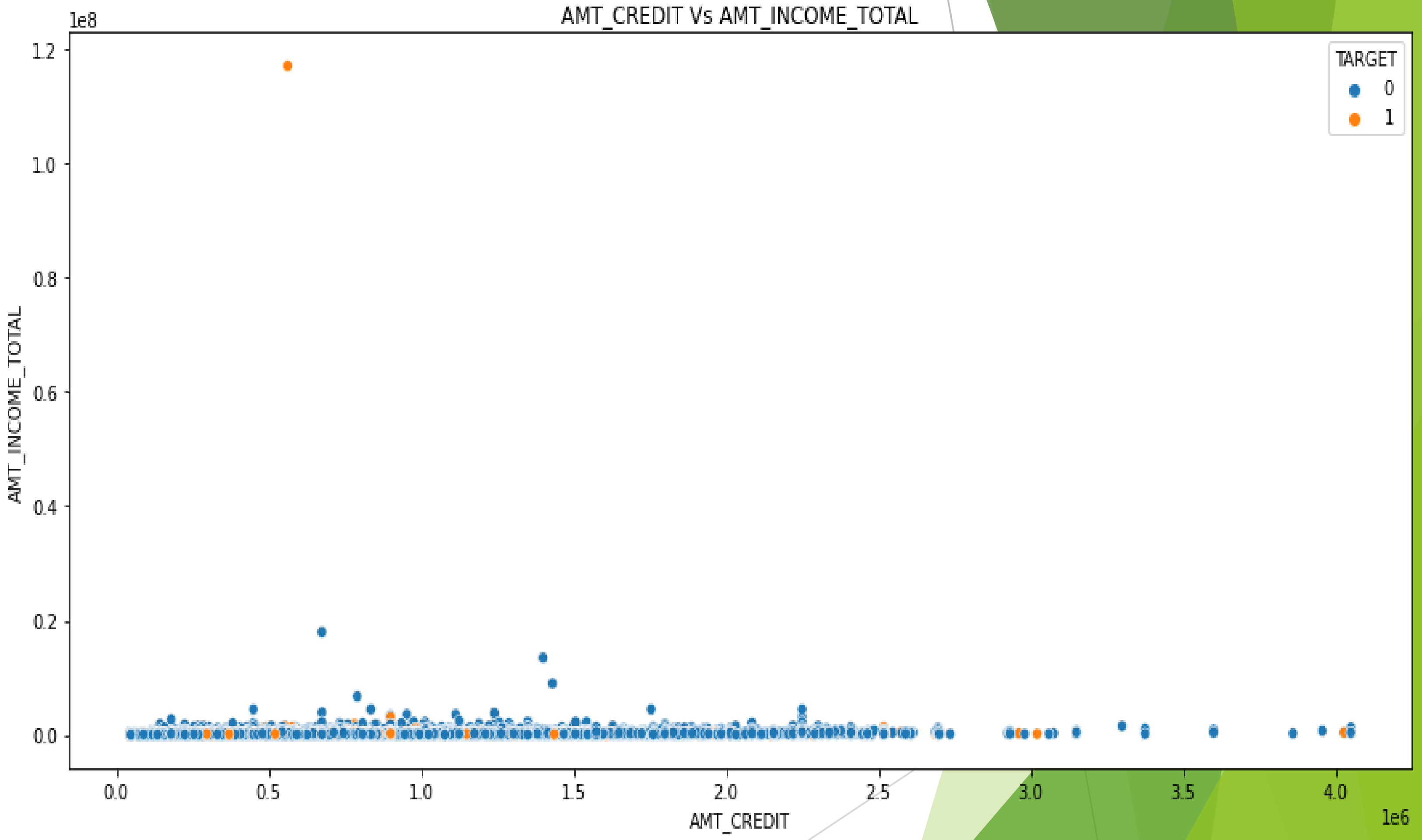
Bivariate Analysis

- ❑ AMT_CREDIT and AMT_GOODS_PRICE are linearly corelated. The defaulters got decreased as the credit amount is increasing.



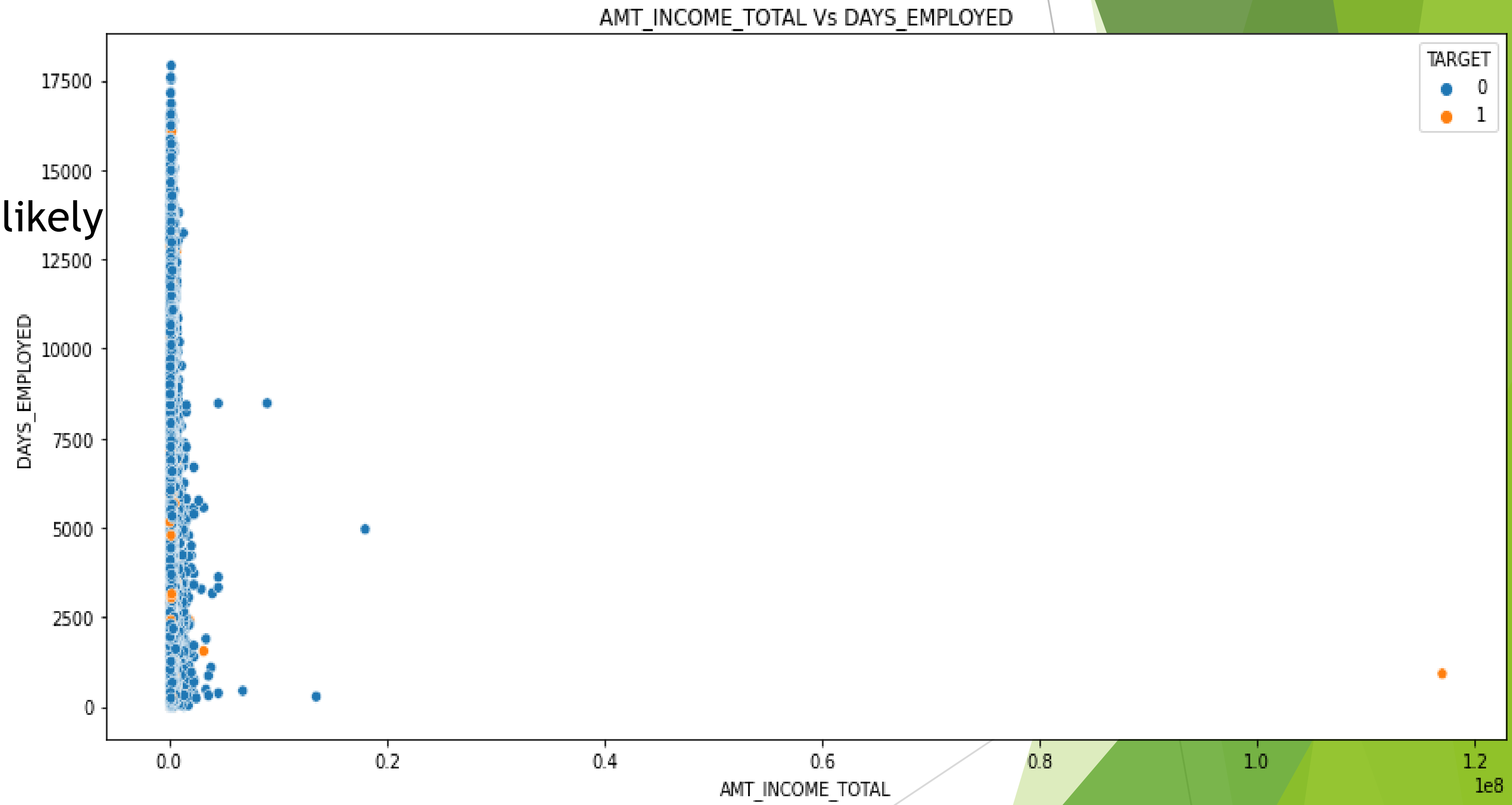
Bivariate Analysis

- ❑ Lower-income individuals take more loans; defaulters decrease as income and credit amounts rise.



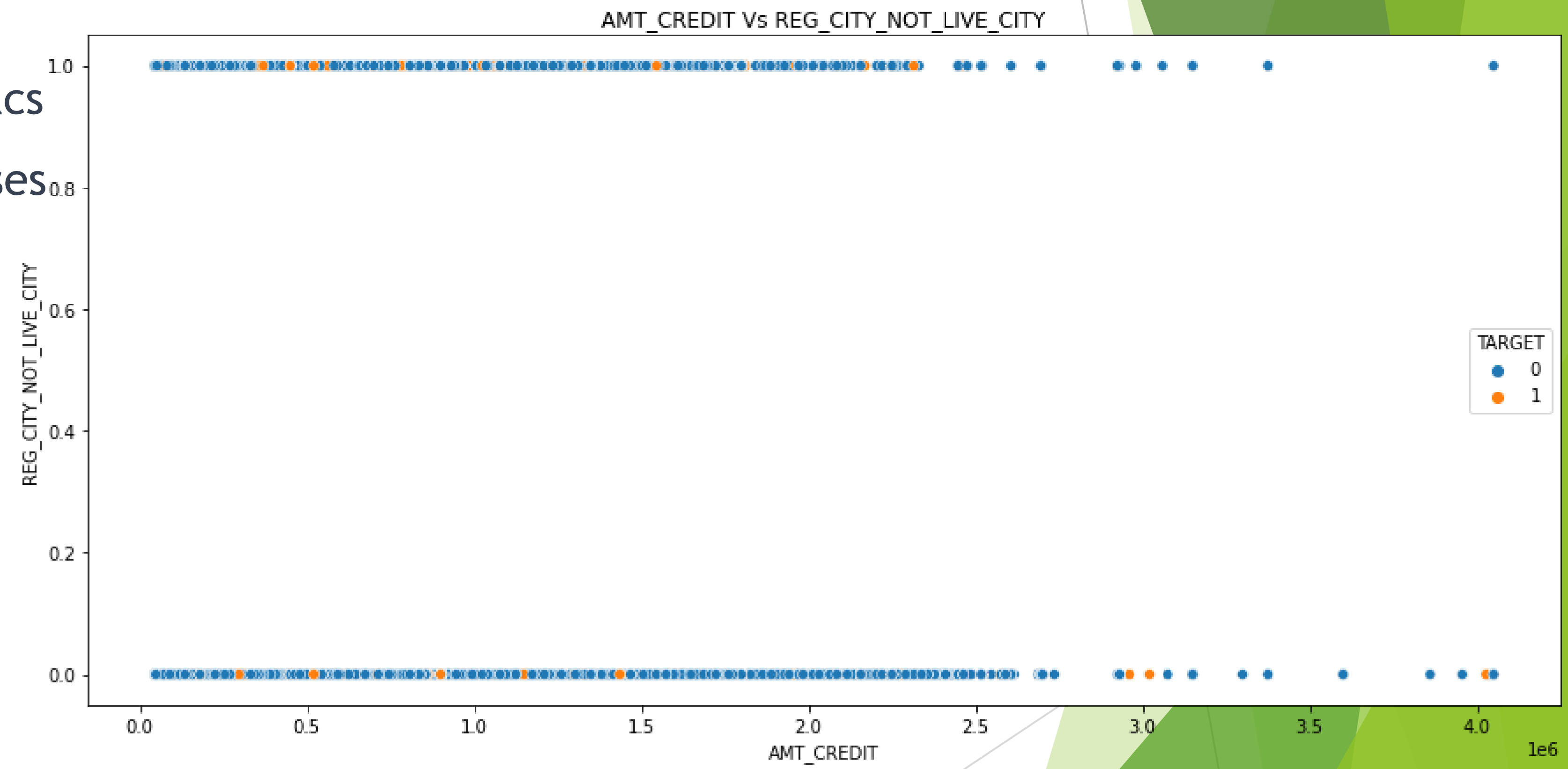
Bivariate Analysis

- Customers with extended employment duration are less likely to be defaulters.



Bivariate Analysis

- Customers with credit under 5 lacs and mismatched contact addresses have higher default likelihood

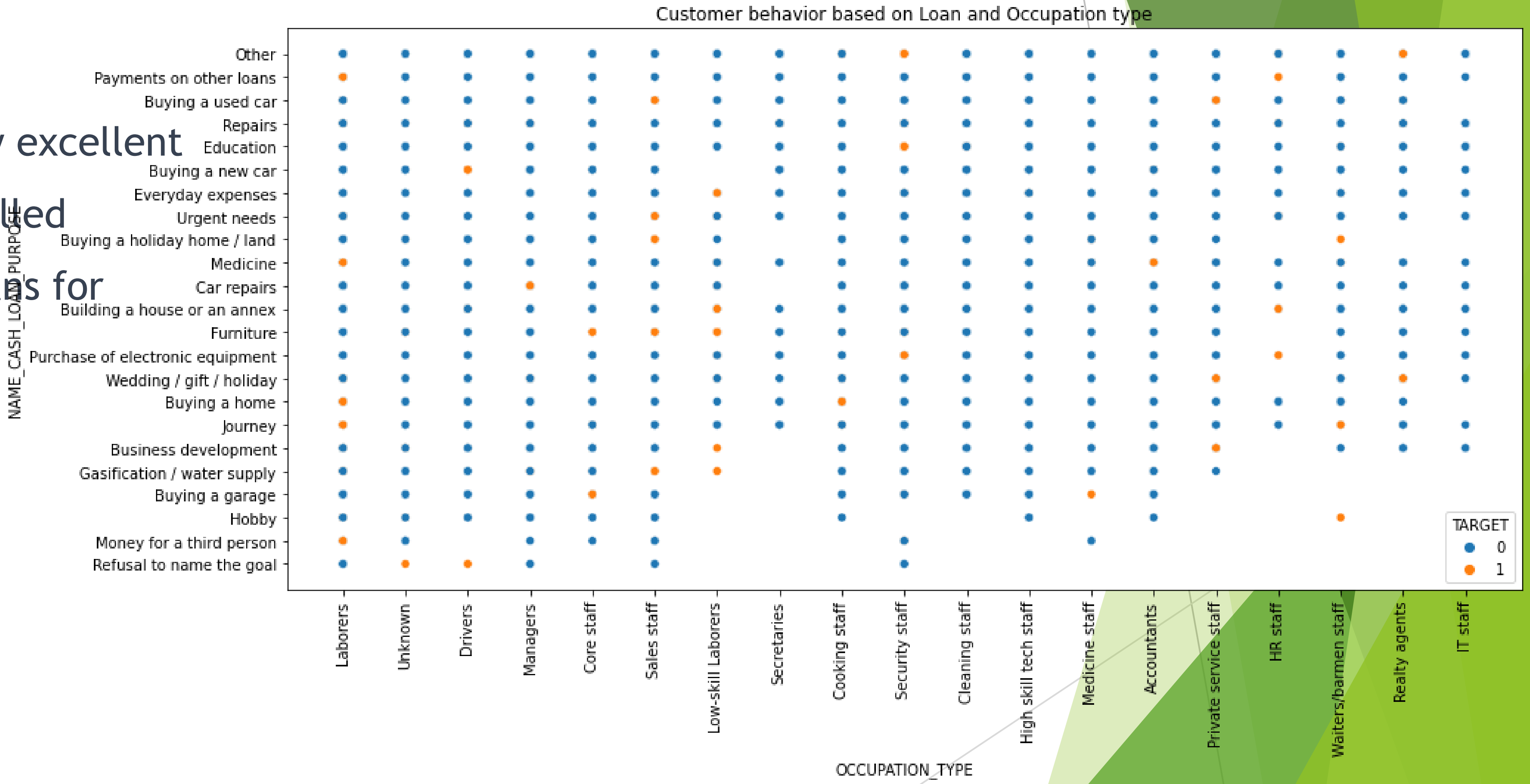


Bivariate Analysis

- Customers with medical and new car/home/garage loans have low default rates, suggesting safe lending decisions for these segments.

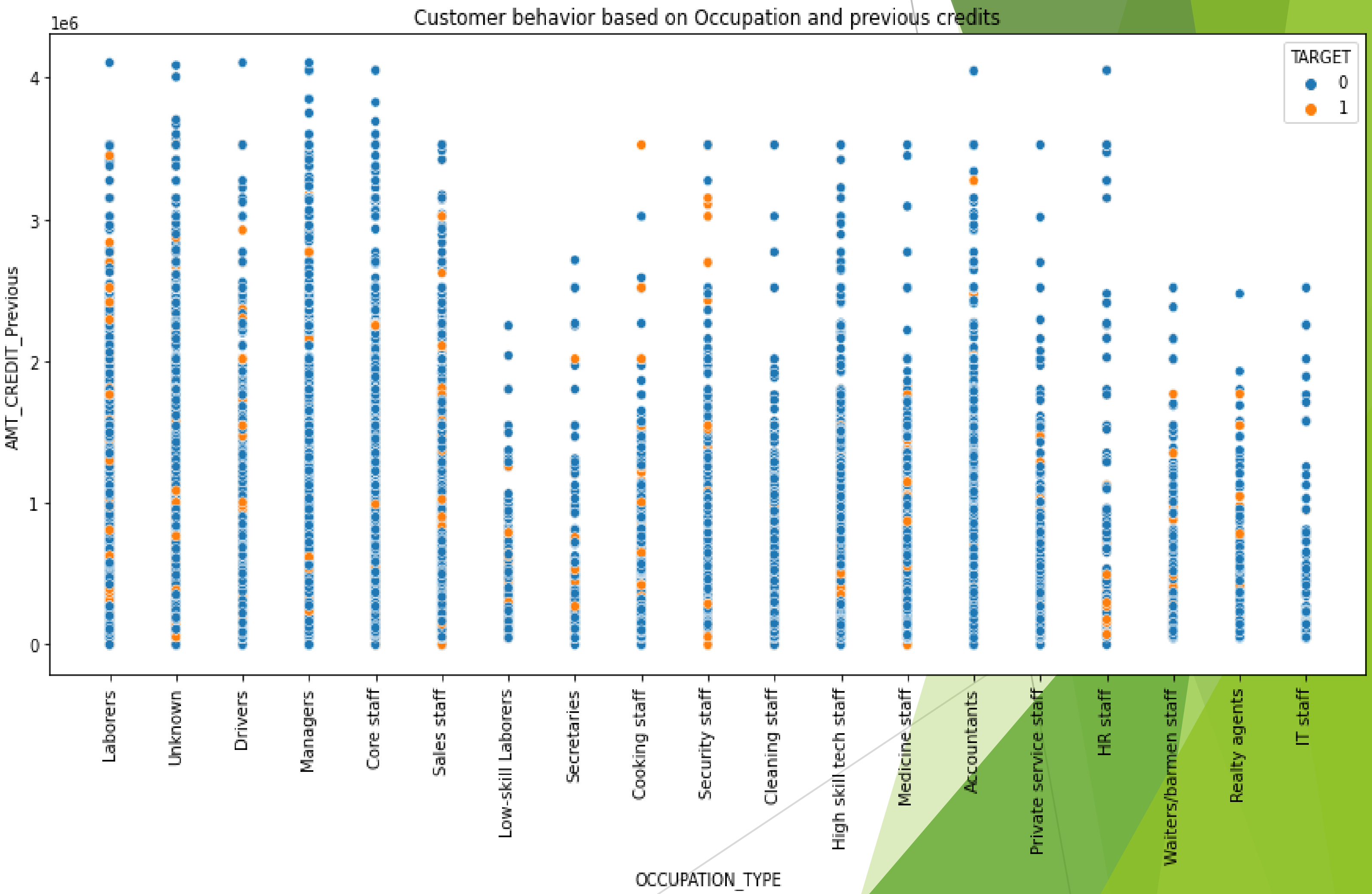
Bivariate Analysis

- ❑ Skilled tech/IT staff show excellent repayment, while low-skilled laborers struggle with loans for home/business.



Bivariate Analysis

- Higher credit amounts lead to increased default rates among laborers, drivers, and security staff.

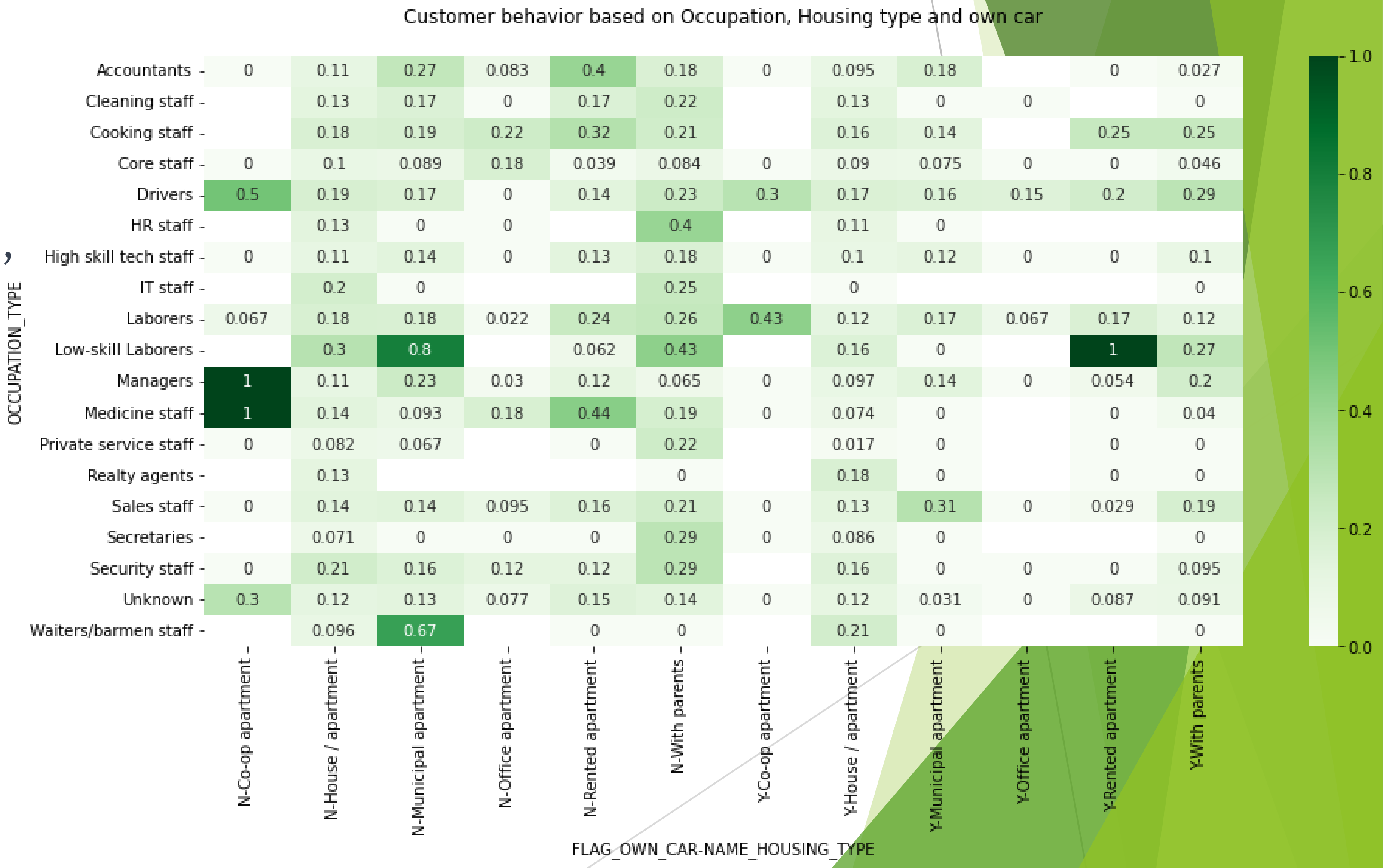


Multivariate Analysis

- Both Male and Female Customers with higher income are linked to lower defaults. Certain jobs like drivers, laborers, and security staff are struggling with payments.

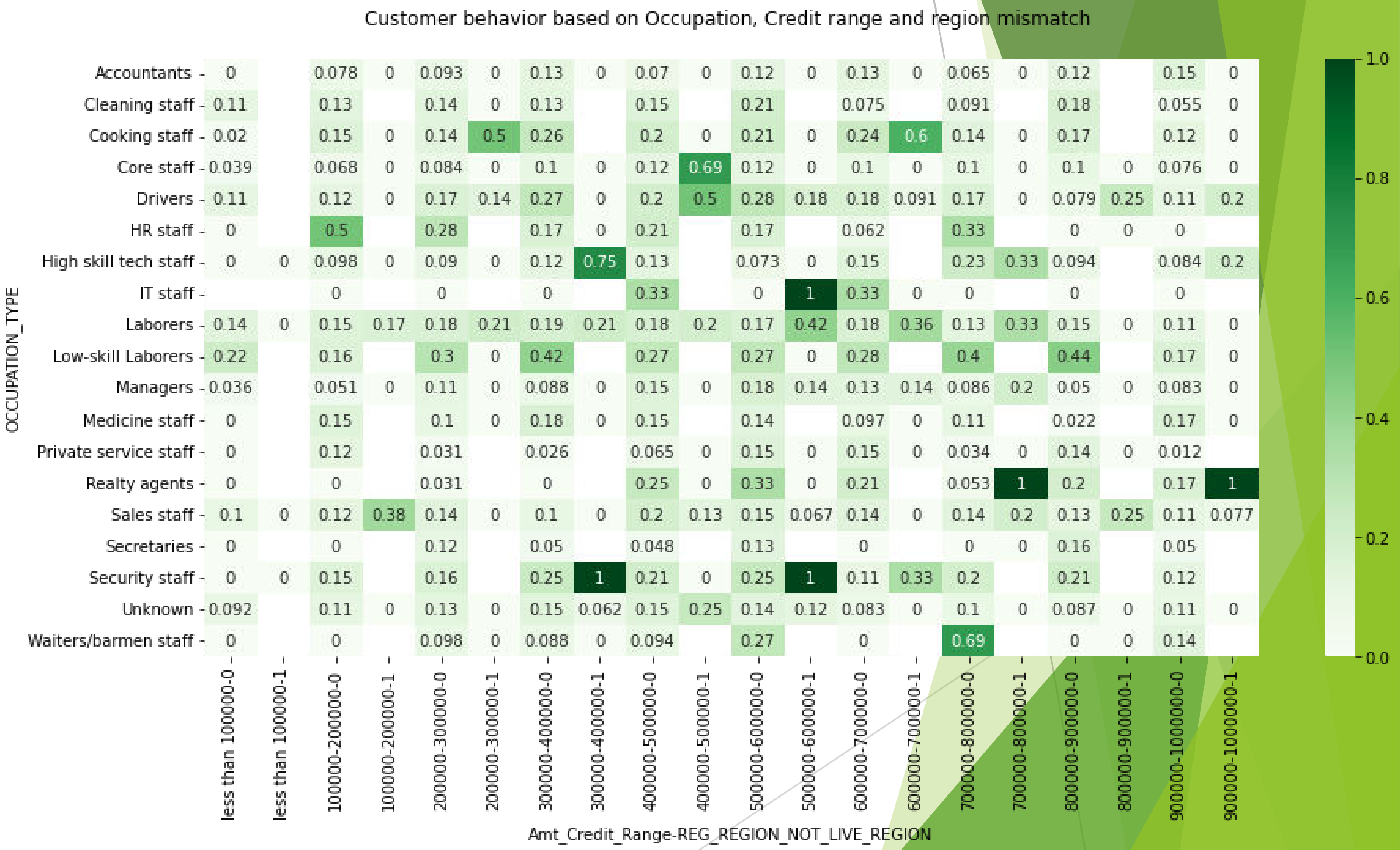
Multivariate Analysis

- Payment issues noted among specific groups: co-op apartment residents without cars (e.g., drivers, medical staff), low-skilled laborers, and barmen in municipal apartments. Caution is advised for loans to these demographics.



Multivariate Analysis

- High credit customers with mismatched permanent and contact addresses, especially among security staff and realty agents, exhibit higher default tendencies.



Findings

- Defaulters are fewer compared to non-defaulters and the data imbalance ratio is 10.55
- Few percentage of customers are having different contact and permanent addresses and they are turning out to be defaulters in cases of high credit
- Customers with cash loans tend to have higher default rates compared to revolving loans
- Customers with own cars are less likely to default may be because of financial stability
- Married individuals default less than single or civil-married counterparts and parental statuses are also impacting default rates

Findings

- Customers having decent job tenurity having good behavior those on maternity leave present higher default rates
- Lower-educated individuals tend to default more, while higher-educated clients show better repayment behavior
- Property owners have lower default rates
- Certain occupations, like low-skilled laborers and drivers, are prone to higher default rates
- Higher incomes correlate with lower defaults, but extremely high incomes again have higher default risk

Conclusion

- Considering the findings, banks should focus on customers with moderate/highly skilled occupations, stable income, and tenure. Married individuals owning properties and cars are preferred, while caution is advised with low-skilled and less educated customers. Attention is needed for accuracy in documents and matching contact regions.

Thank You