EDA Case Study

By ISHAAN BHATTACHARYA

&

KEVAL LAD

BUSINESS OBJECTIVES

- This case study aims to 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.
- 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.
- In other words, 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 utilize this knowledge for its portfolio and risk assessment.

ANALYSIS WORKFLOW

We performed the following steps in this analysis.

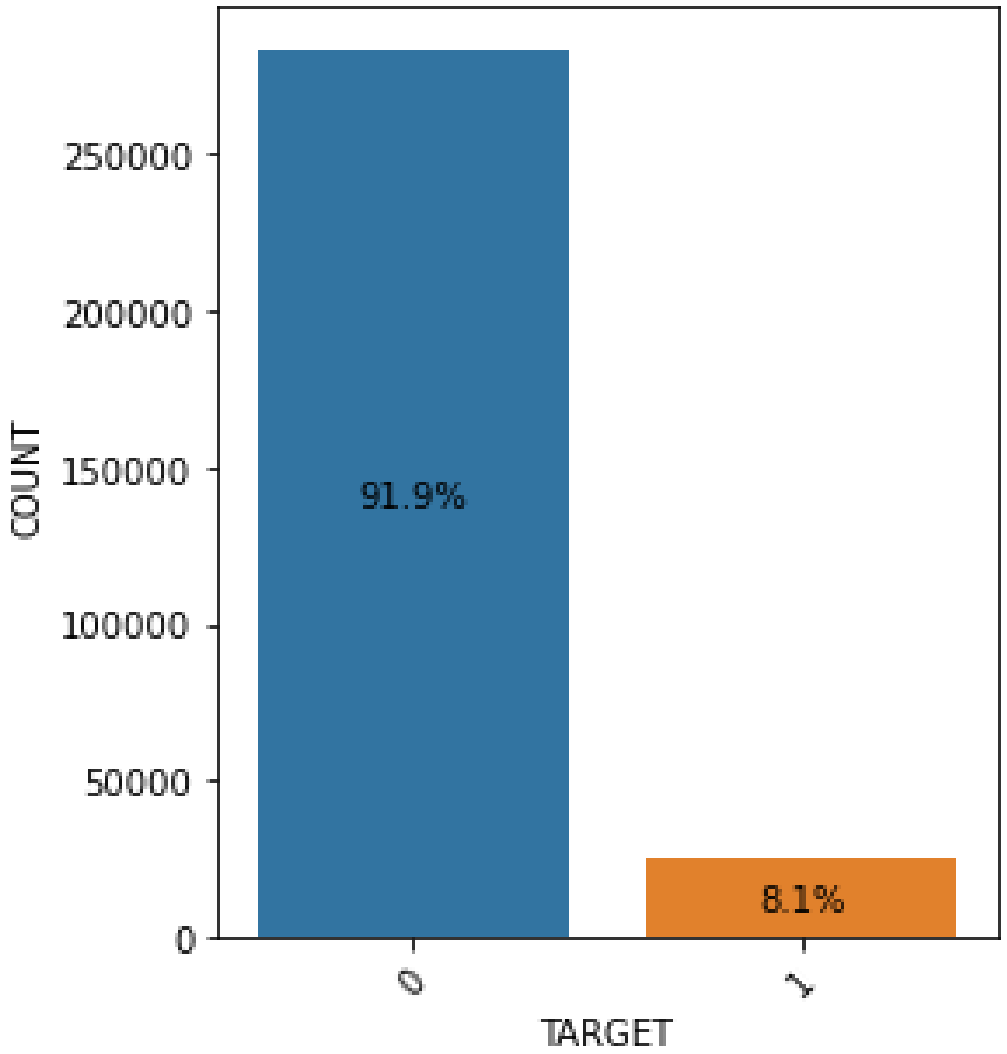
- 1. Read the “application_data.csv”.
- 2. Missing Value Treatment
- 3. Outlier Analysis
- 4. Modifying the Columns
- 5. Exploratory Data Analysis
- 6. Read the “previous_application.csv”
- 7. Missing Value Treatment
- 8. Outlier Analysis
- 9. Modifying the Columns
- 10. Merging the two datasets
- 11. Exploratory Data Analysis
- 12. Deriving the Conclusion

MISSING VALUES

- There were a number of columns with missing values in both the current application dataset and the previous application dataset.
- The columns which had 40% missing values were dropped.
- The columns which had less than 40% of missing values were imputed by the median value for numerical columns and by the mode value for the categorical columns.

DATA IMBALANCE

- There is data imbalance in the current application dataset.
- 91.9% values in the TARGET column are 0 whereas only 8.1%values are 1.
- The ratio of data imbalance is 11.34 in favor of 0.



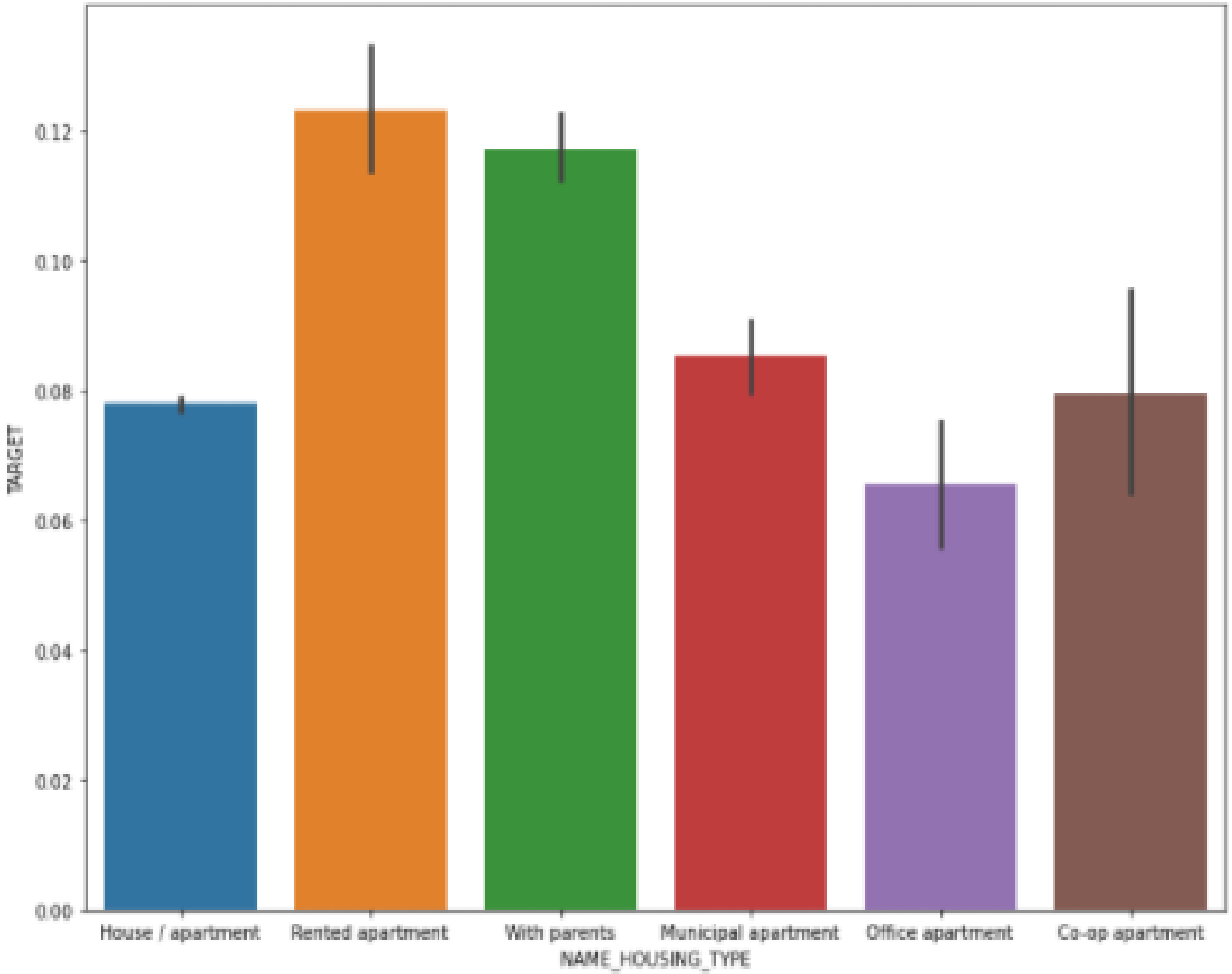
EDA INSIGHTS

- Over the next few slides we will present the most important graphs and the insights derived.
- The graphs having significant differences between different categories in a single column are the most important.
- We have divided the analysis into two parts – one for the current application dataset and another for the previous application dataset.

CURRENT APPLICATION DATASET

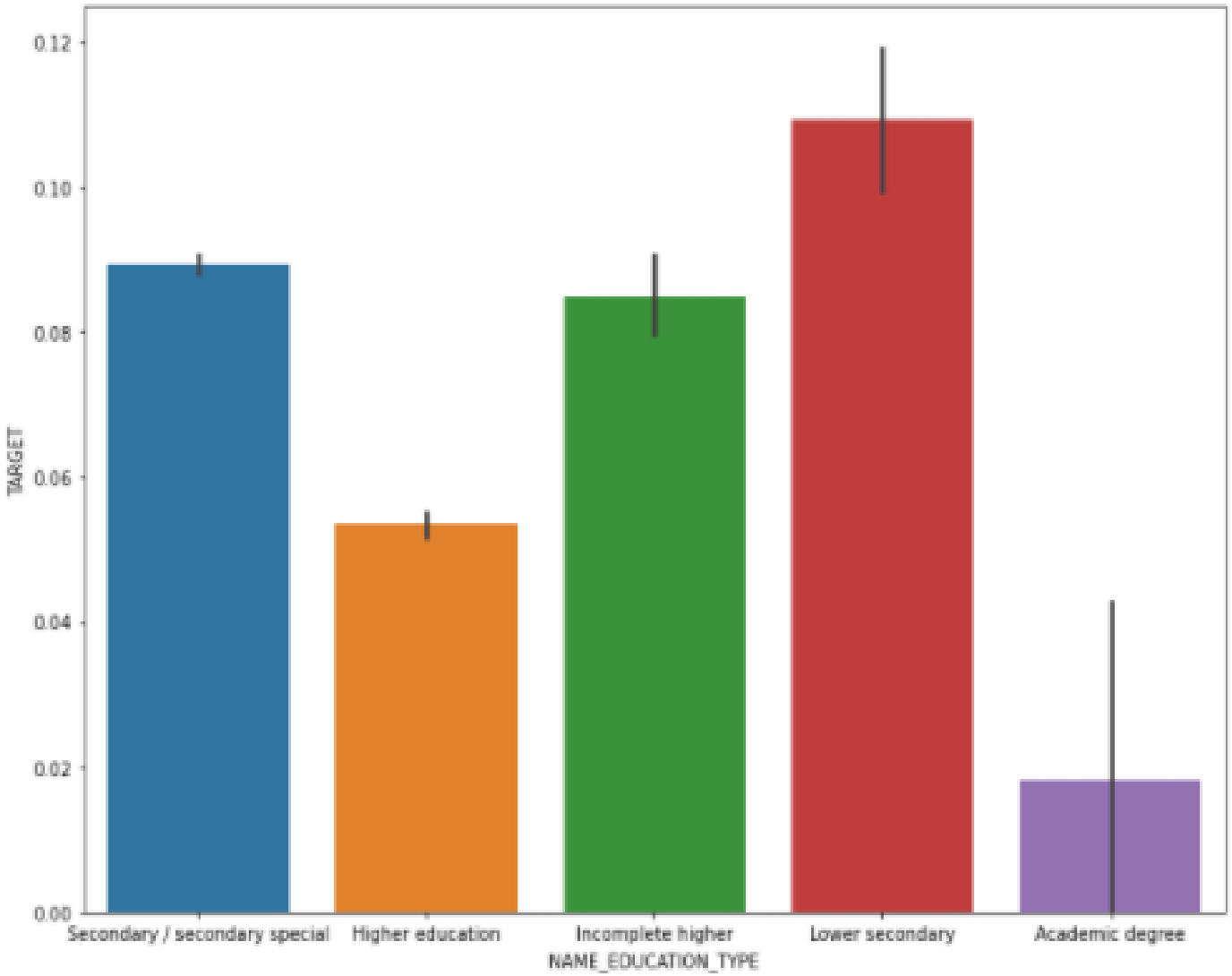
1. Housing Type

As we can see, those who are living in Rented Apartment and those who are living with their parents have a quite high rate of default.



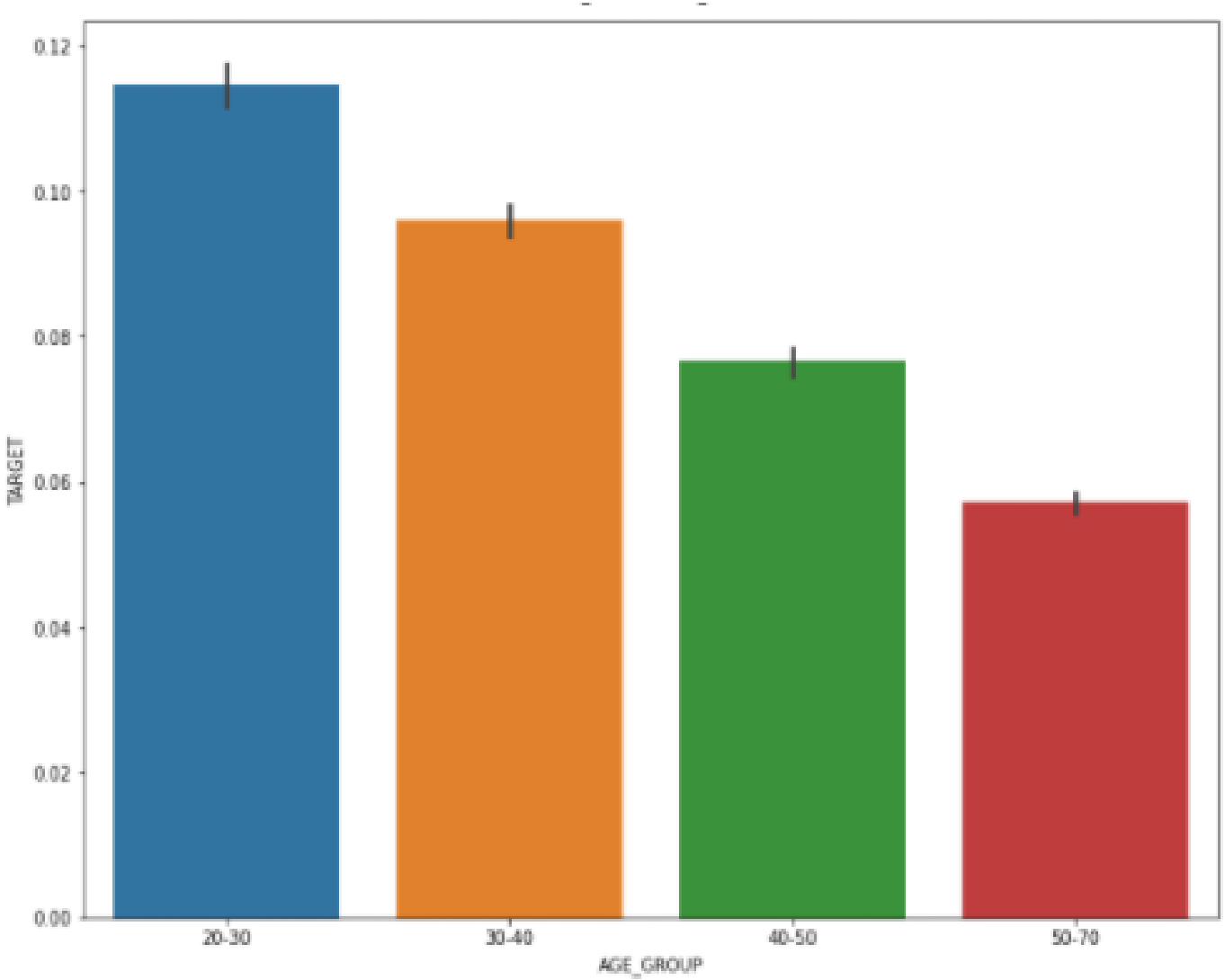
2. Education Type

People with Lower Secondary education are most likely to default among all the education types.



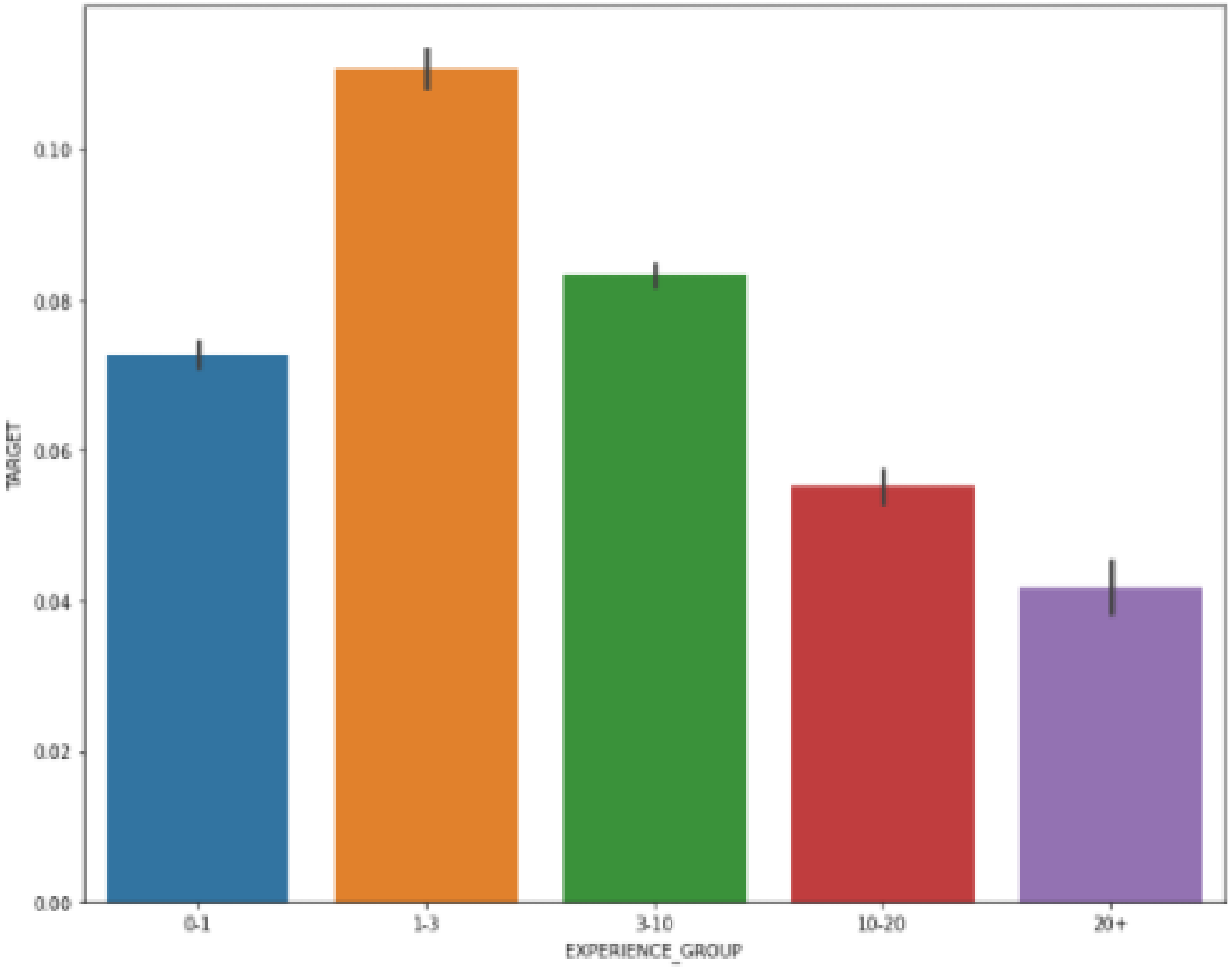
3. Age

It is clear from this plot that the rate of default decreases with increase in the age.



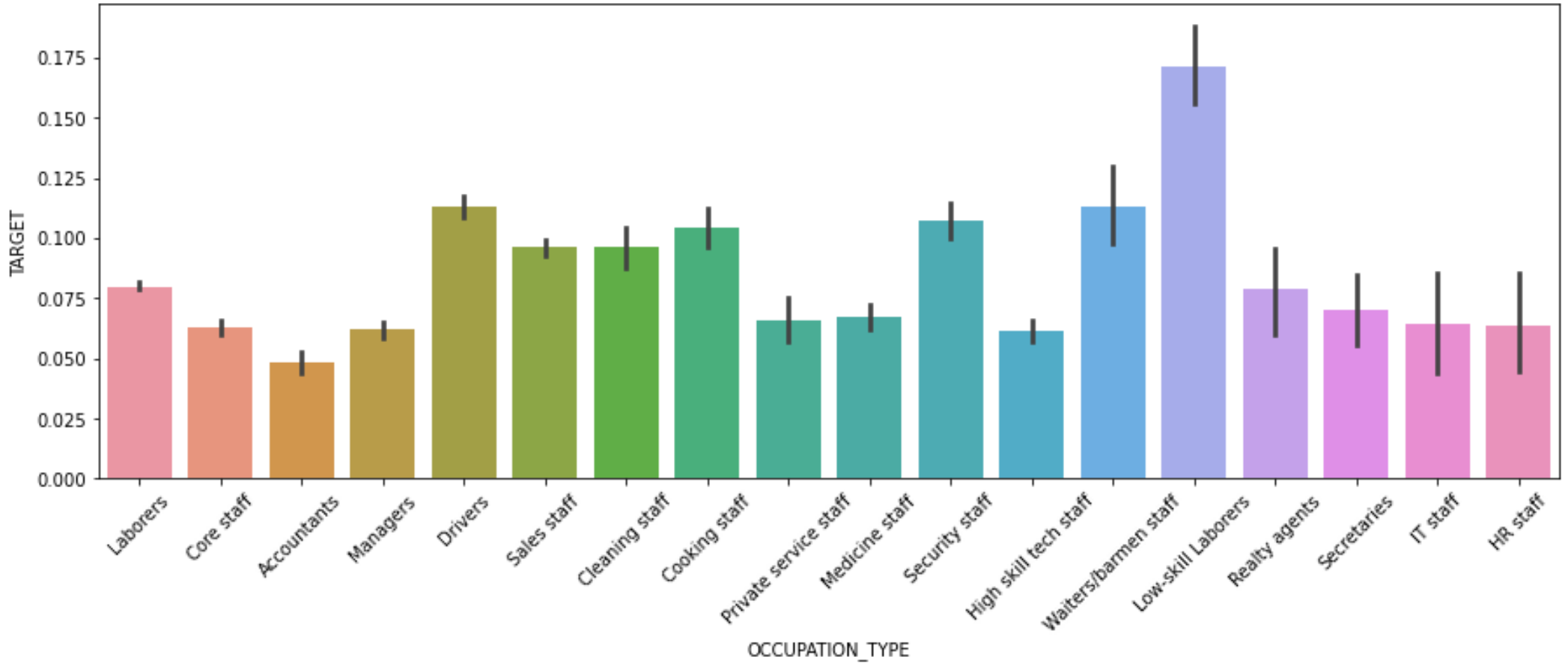
4. Experience

People with 1-3 years of job experience are most likely to default and their rate of default is quite high when compared to other experience groups.



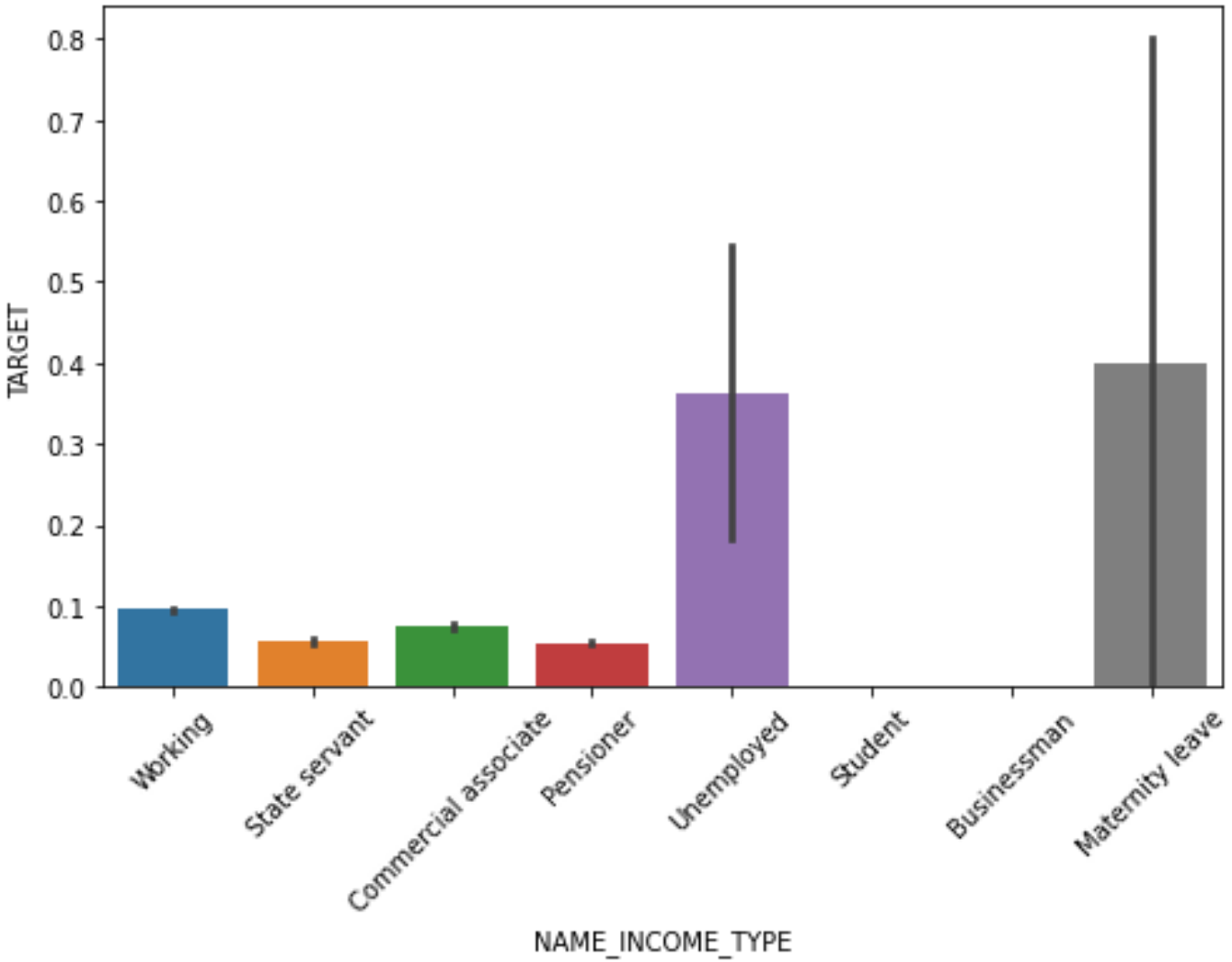
5. Occupation Type

Low skill Laborers have a very high rate of default (17%). So it would be good if the company avoids giving loans to this category.



6. Income Type People

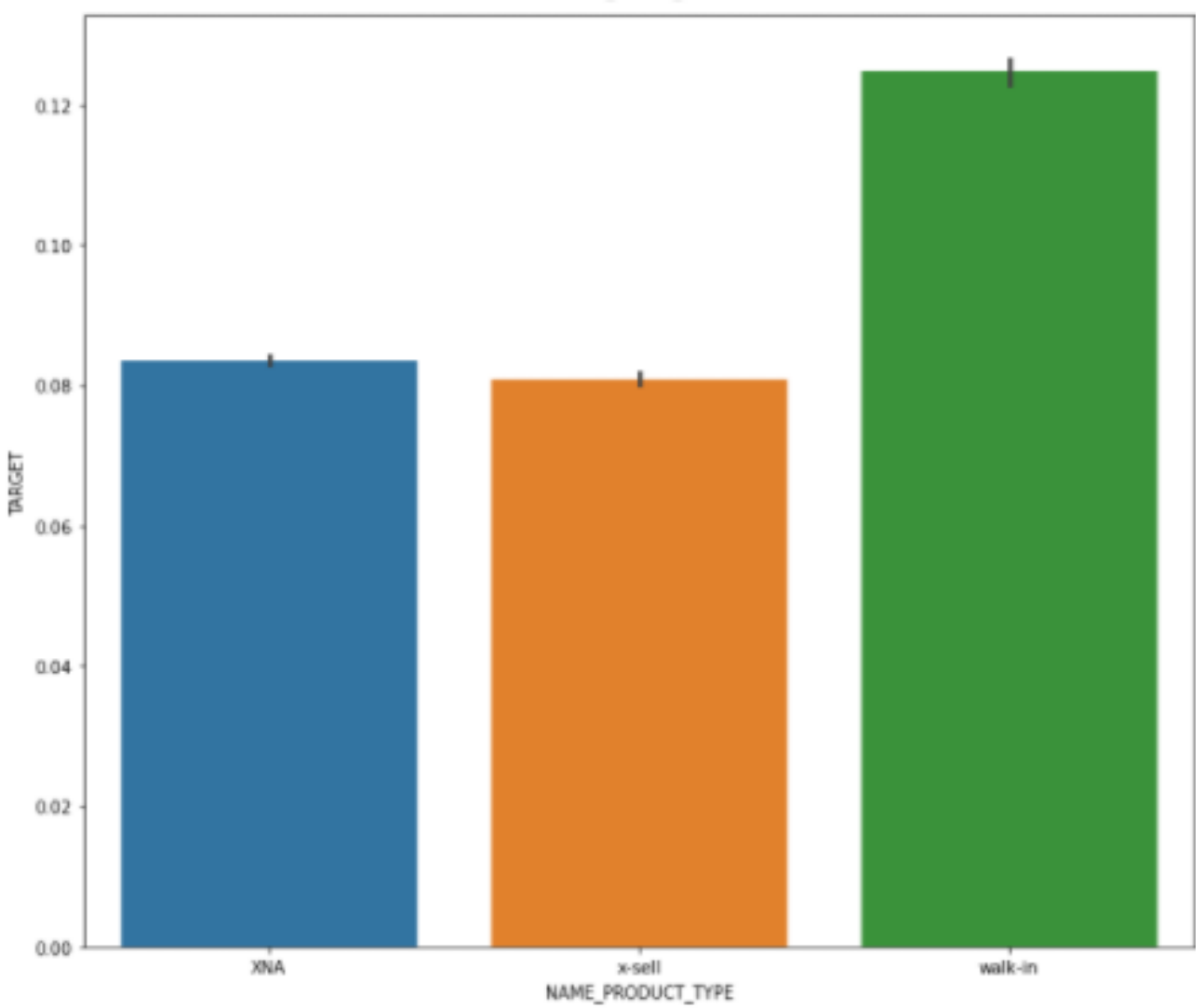
People who are Unemployed and who are on Maternity Leave have a very high rate of default (>40%). So it would be good if the company altogether avoid giving loans to this category.



PREVIOUS APPLICATION DATASET

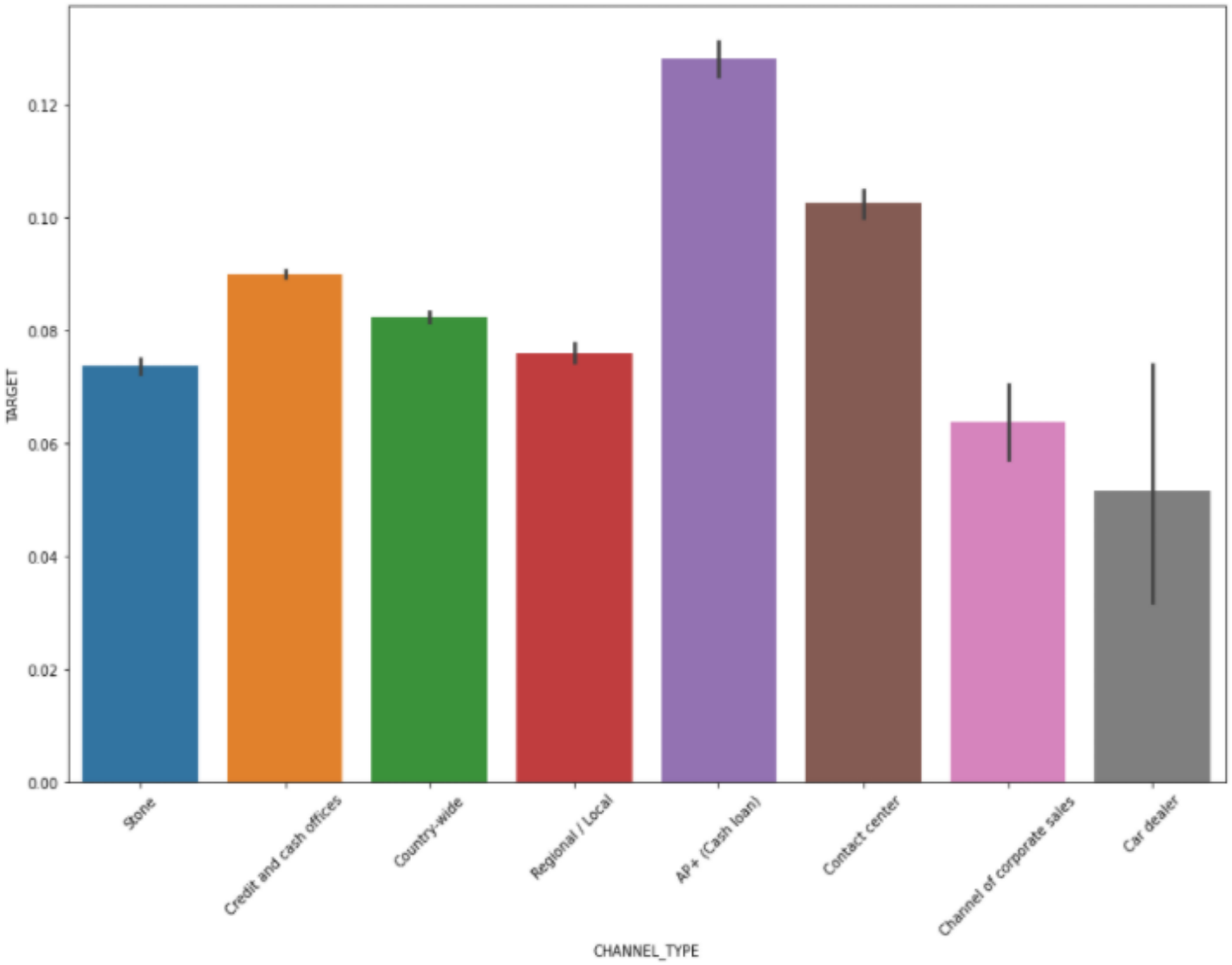
1. Product Type: -

People who had walk-in product in their previous application have the highest default rate.



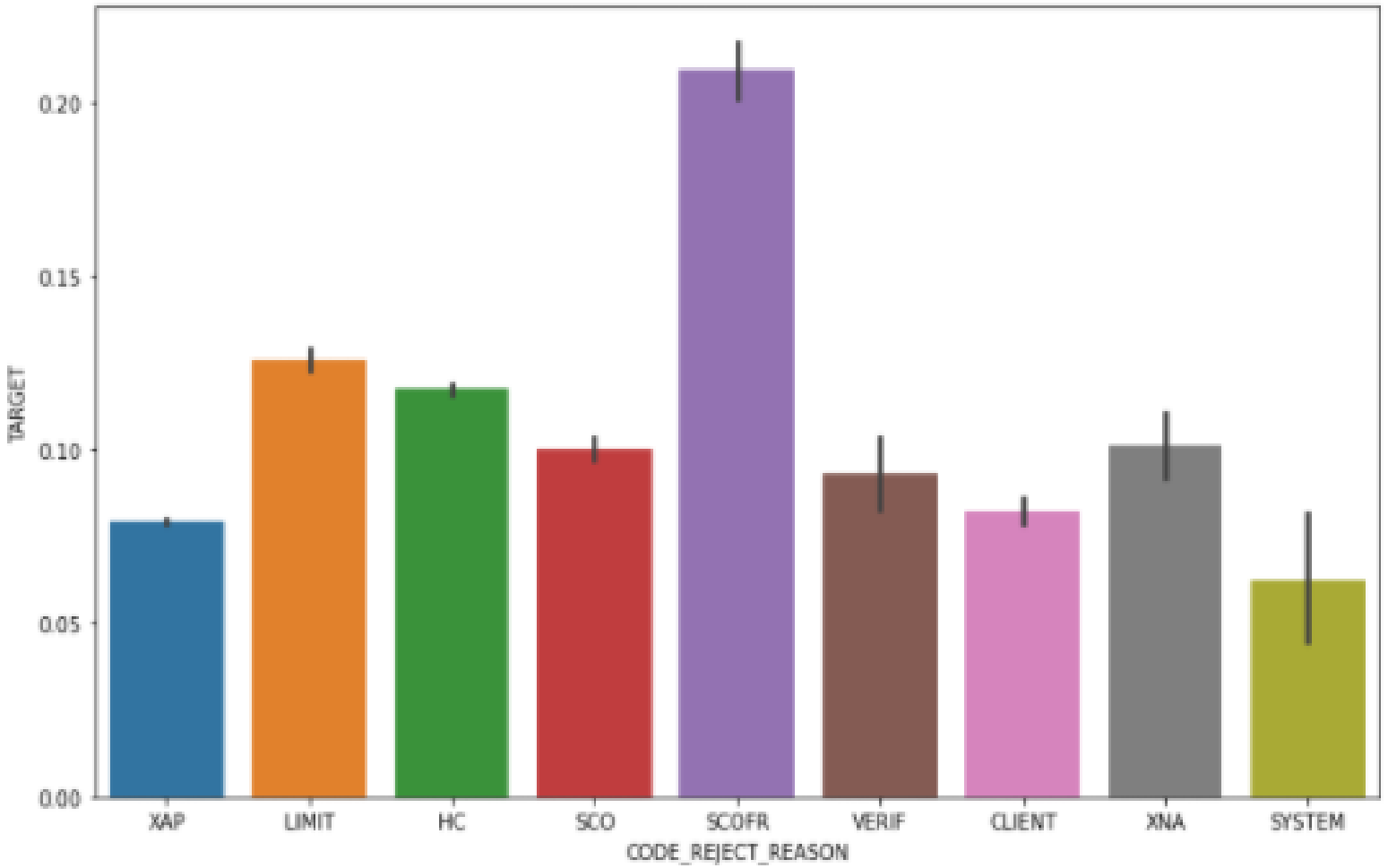
2. Channel Type

Customers acquired through the Channel AP+(Cash loan) in the previous application have the highest default rate.



3. Code Reject Reason

Applications with the Code Reject Reason SCOFR in the previous application have the highest default rate in the current application.



Safe Group

1. Those who have an Academic Degree.
2. Those who live in regions having high relative population (>0.05)
3. Those who have a work experience of greater than 20 years.
4. Those who are working as accountants.
5. Those whose income is from Trade Type 4.

Risky Group

1. Low skill Laborers
2. Those who work in Transport Type 3
3. Those who are Unemployed or who are on Maternity Leave.

CONCLUSION

Based on the above analysis the following steps should be taken by the bank to avoid bad debts : -

- Bank should focus on people who live in rented apartments and who live with their parents as there is a high rate of default in these groups.
- Bank should focus on people with lower secondary education because they are most likely to default among all other education types.
- Single and separated people need to be analyzed closely as they have a high rate of default compared to other family status types.
- Bank should give big loans to higher age people as the rate of default decreases with increase in age.

- Bank should not give big loan to people with 1-3 years of experience as they have a high rate of default.
- Bank should be wary of giving loans to people who have changed their phone number in the last one month.
- Bank should avoid giving loans to Low skill Laborers as the rate of default is quite high in this class.
- Bank should analyze people who are working in Industry type 8 & 13, and those who are working in Transport type 3.
- Bank should avoid giving loans to Unemployed people and those on Maternity Leave as the rate of default is very high for these groups.
- Bank should insist on getting Documents 6,7,8,9,11,13,14,15,16,17,18,19,21.

Thank You☺