Credit EDA Assignment

upGrad & IIITB | Data Science Program - September 2023

By Vishal Varma Akkala

Problem Statement 1

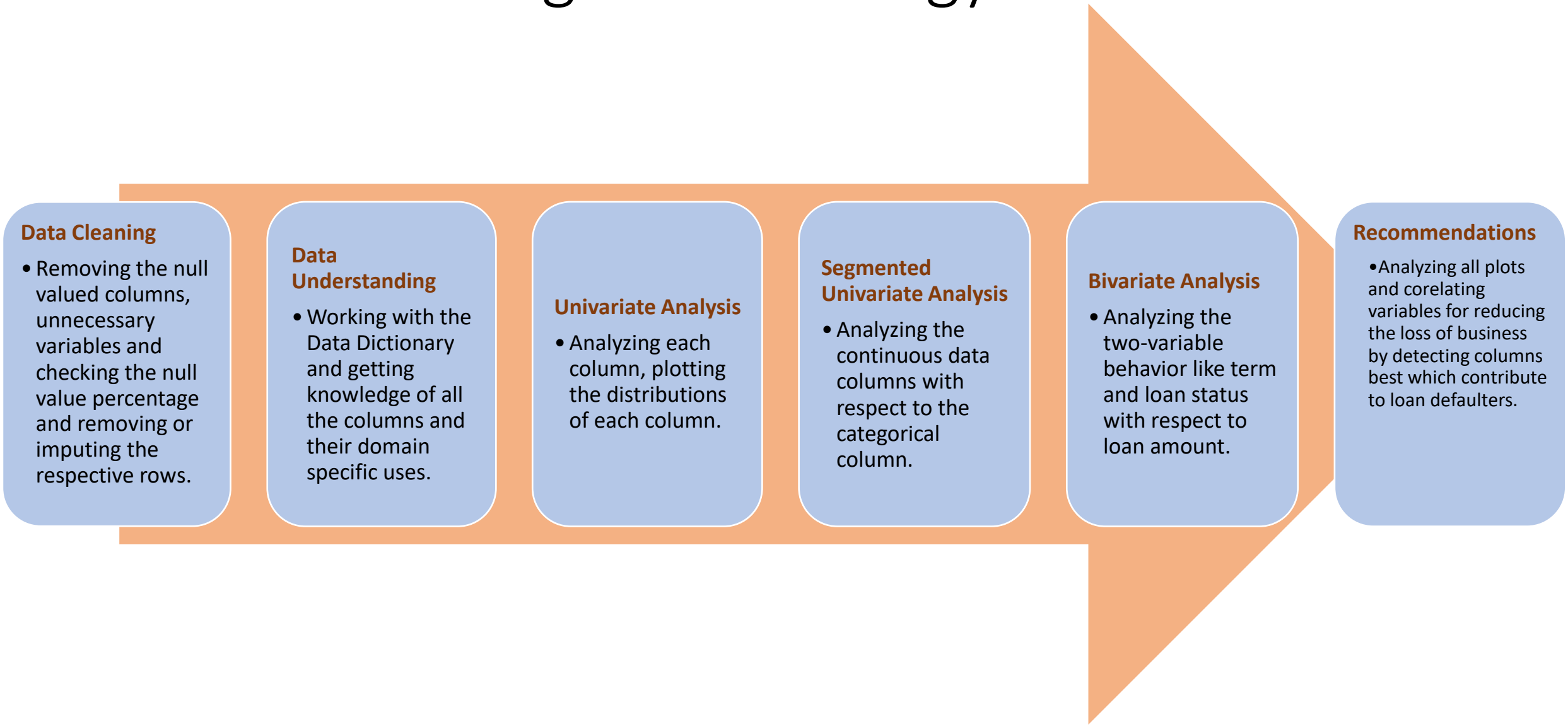
- A financial company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

Problem Statement 2

- Present the overall approach of the analysis in a presentation. Mention the problem statement and the analysis approach briefly.

- Identify the missing data and use appropriate method to deal with it. (Remove columns/or replace it with an appropriate value)

Problem Solving Methodology



Observations and Recommendations

- Bank is disbursing more cash loans than revolving loans, which is very high-risk factor to the loan book. 8% of the disbursed cash loans are defaulted.
- 92% of the loans disbursed are repaid and only 8% are defaulted. A healthy sign for the loan book.
- Customers who own a house or flat or live in an office apartment repay without default.
- Customers working as IT staff are more likely to repay loans while laborers are likely to default.
- Both business entities and student groups have repaid loans as against the usual risk of default with these segments.
- Customers from region rating 1 are more likely to repay.
- Customers who are employed and go on a maternity leave tend to default.
- Most of the customers applying for loan are females and are also good at repaying loan. Bank can consider to design their loan products targeting females.
- Customers who has a working mobile (reachable) and has submitted their work phone number are likely to repay loan without defaulting.
- More than 34% of the customers have reapplied for a fresh loan within a year and more than 23% have reapplied within two years. Assuming loan repayment period is for minimum of 5 years, it is advised to be extra cautious while underwriting recurring customer loan applications.
- Customers who have studied secondary education are likely to apply for a loan more and customers who have a higher education are likely to repay with defaulting.
- Customers who are married are likely to repay than to default.

Applications Data Set

```
In [4]: app_data = pd.read_csv("application_data.csv")
        app_data.head()
```

Out[4]:

NAME_TYPE_SUITE	NAME_INCOME_TYPE	NAME_EDUCATION_TYPE	NAME_FAMILY_STATUS	NAME_HOUSING_TYPE	REGION_POPULATION_RELATIVE	DAYS_BIR
Unaccompanied	Working	Secondary / secondary special	Single / not married	House / apartment	0.02	-94
Family	State servant	Higher education	Married	House / apartment	0.00	-167
Unaccompanied	Working	Secondary / secondary special	Single / not married	House / apartment	0.01	-190
Unaccompanied	Working	Secondary / secondary special	Civil marriage	House / apartment	0.01	-190
Unaccompanied	Working	Secondary / secondary special	Single / not married	House / apartment	0.03	-199

```
In [5]: #Shape of the applications df for rows and columns in the data set
        app_data.shape
```

```
Out[5]: (307511, 122)
```

- Dimensions of the raw file are 307511 X 122 [rows X columns]
- Data is having data types as object, int and float

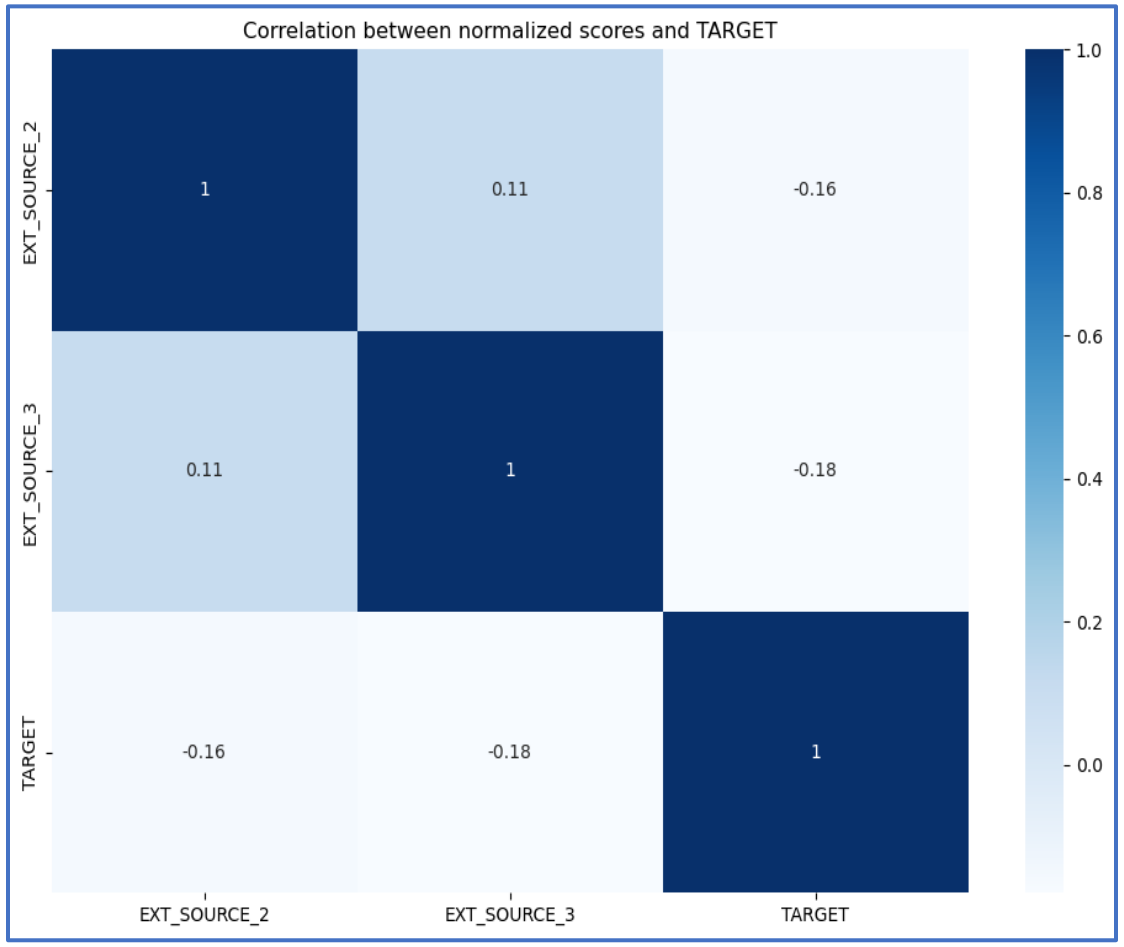
Null Values Handling

```
In [10]: app_null_50 = application_null_perc[application_null_perc>50]
         app_null_50

Out[10]: OWN_CAR_AGE                 65.99
         EXT_SOURCE_1                56.38
         APARTMENTS_AVG              50.75
         BASEMENTAREA_AVG            58.52
         YEARS_BUILD_AVG             66.50
         COMMONAREA_AVG              69.87
         ELEVATORS_AVG               53.30
         ENTRANCES_AVG               50.35
         FLOORSMIN_AVG               67.85
         LANDAREA_AVG                59.38
         LIVINGAPARTMENTS_AVG        68.35
         LIVINGAREA_AVG              50.19
         NONLIVINGAPARTMENTS_AVG     69.43
         NONLIVINGAREA_AVG           55.18
         APARTMENTS_MODE             50.75
         BASEMENTAREA_MODE           58.52
         YEARS_BUILD_MODE            66.50
         COMMONAREA_MODE             69.87
         ELEVATORS_MODE              53.30
         ENTRANCES_MODE              50.35
         FLOORSMIN_MODE              67.85
         LANDAREA_MODE               59.38
         LIVINGAPARTMENTS_MODE       68.35
         LIVINGAREA_MODE             50.19
         NONLIVINGAPARTMENTS_MODE    69.43
         NONLIVINGAREA_MODE          55.18
         APARTMENTS_MEDI             50.75
         BASEMENTAREA_MEDI           58.52
         YEARS_BUILD_MEDI            66.50
         COMMONAREA_MEDI             69.87
         ELEVATORS_MEDI              53.30
         ENTRANCES_MEDI              50.35
         FLOORSMIN_MEDI              67.85
         LANDAREA_MEDI               59.38
         LIVINGAPARTMENTS_MEDI       68.35
         LIVINGAREA_MEDI             50.19
         NONLIVINGAPARTMENTS_MEDI    69.43
         NONLIVINGAREA_MEDI          55.18
         FONDKAPREMONT_MODE          68.39
         HOUSETYPE_MODE              50.18
         WALLSMATERIAL_MODE          50.84
         dtype: float64
```

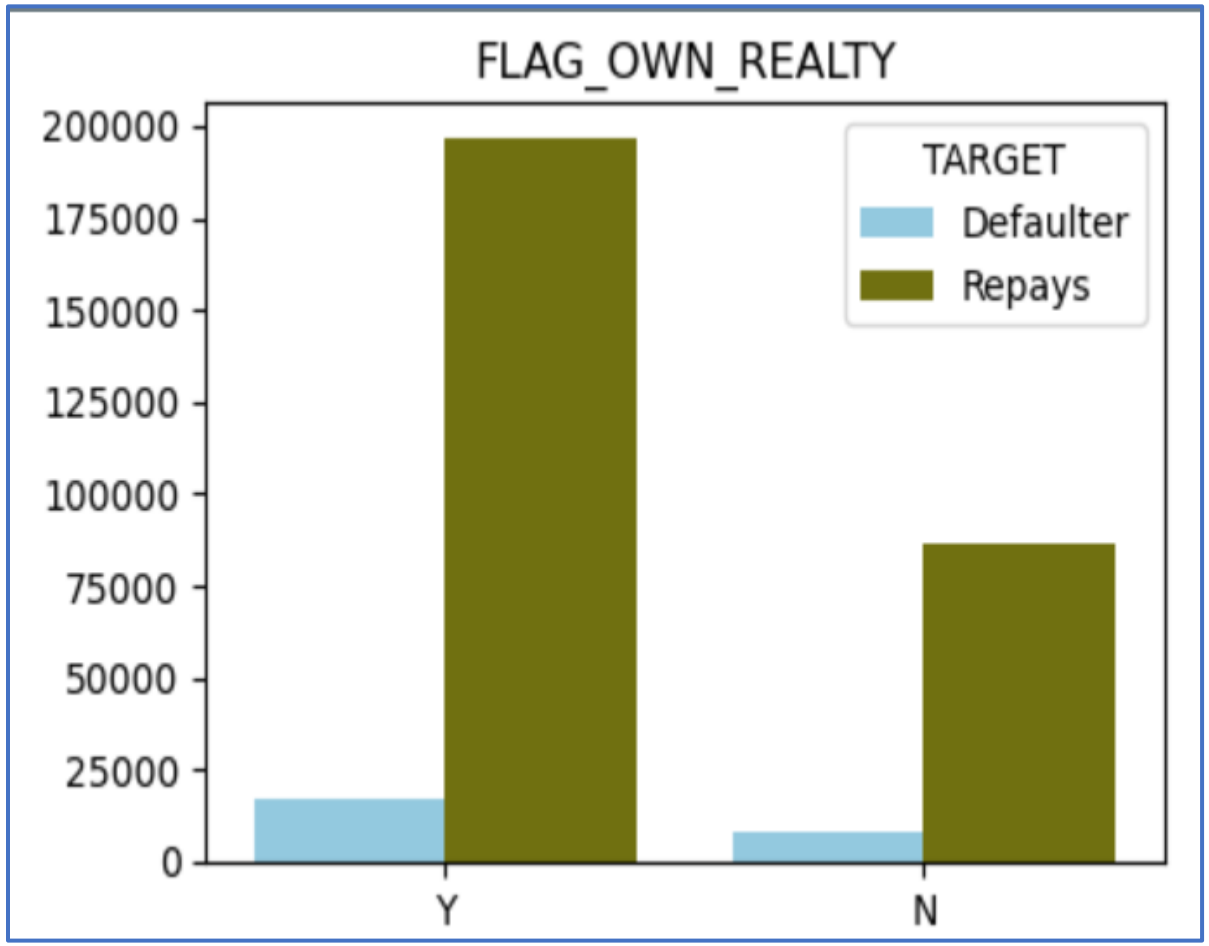
- There are 41 columns which are having null values more than 50%
- For our data analysis we will consider to delete all these columns such that we do not have these columns impacted on the target variable
- After dropping these 41 columns we are left with 81 columns

Null Values Handling Contd..



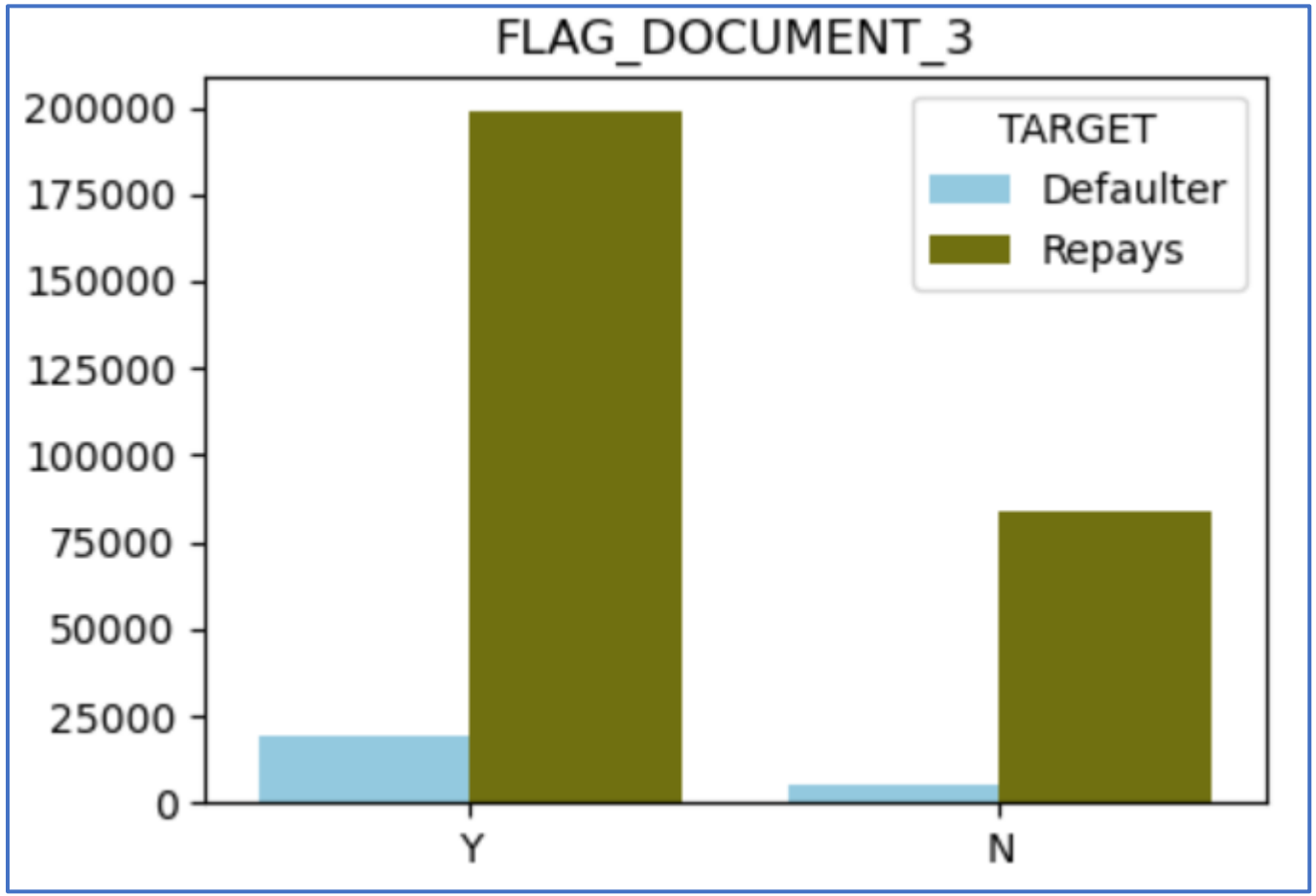
- Columns “EXT_SOURCE_2”,”EXT_SOURCE_3” have no linear correlation with TARGET variable
- We can consider to delete these columns as they are having null values
- Apart from “OCCUPATION_TYPE” column we may delete all other columns having null values greater than 15%
- After deleting these we are left with 71 columns

Observations on Indicators



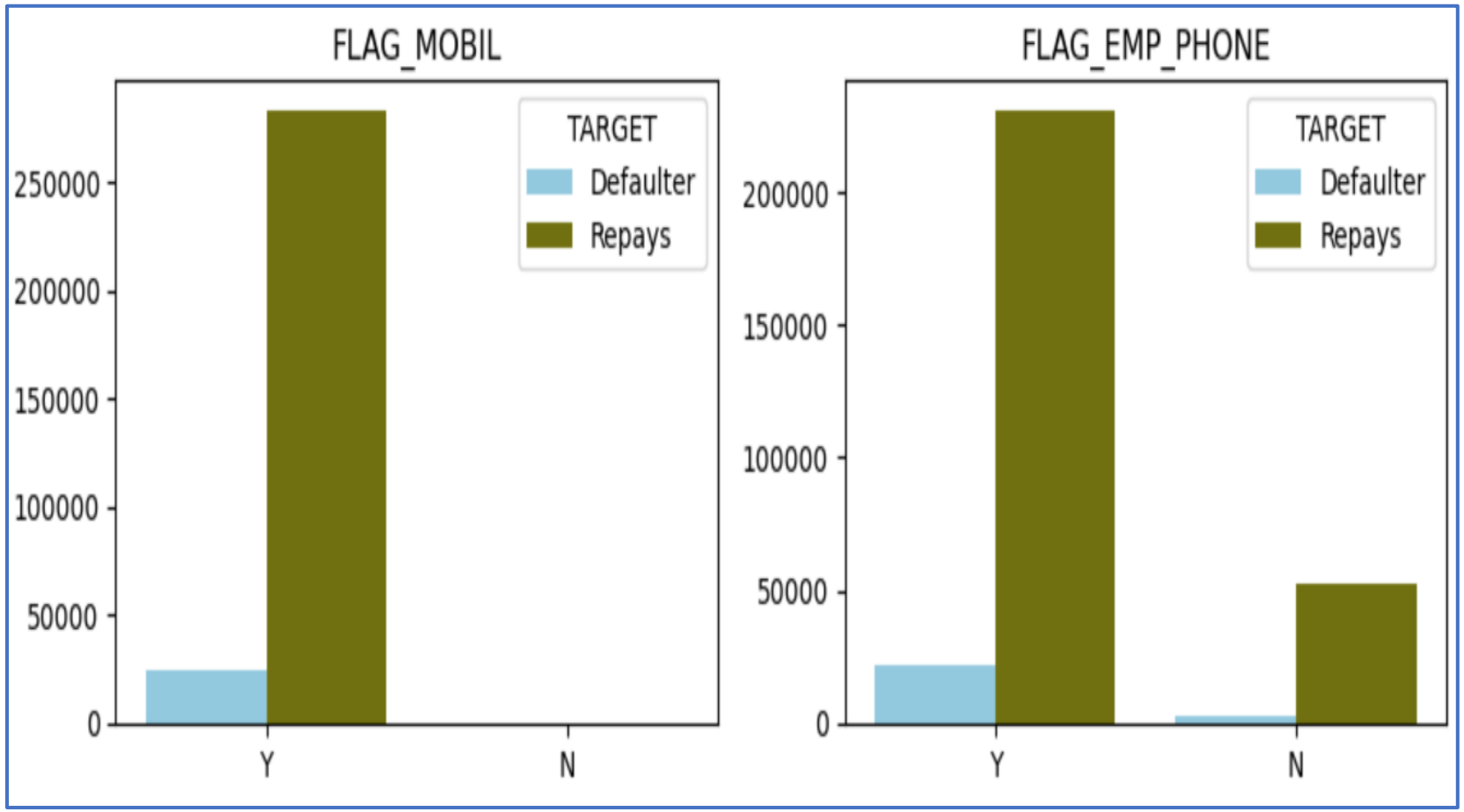
- Customers who own a house or flat repay their loans properly

Observations on Indicators contd..



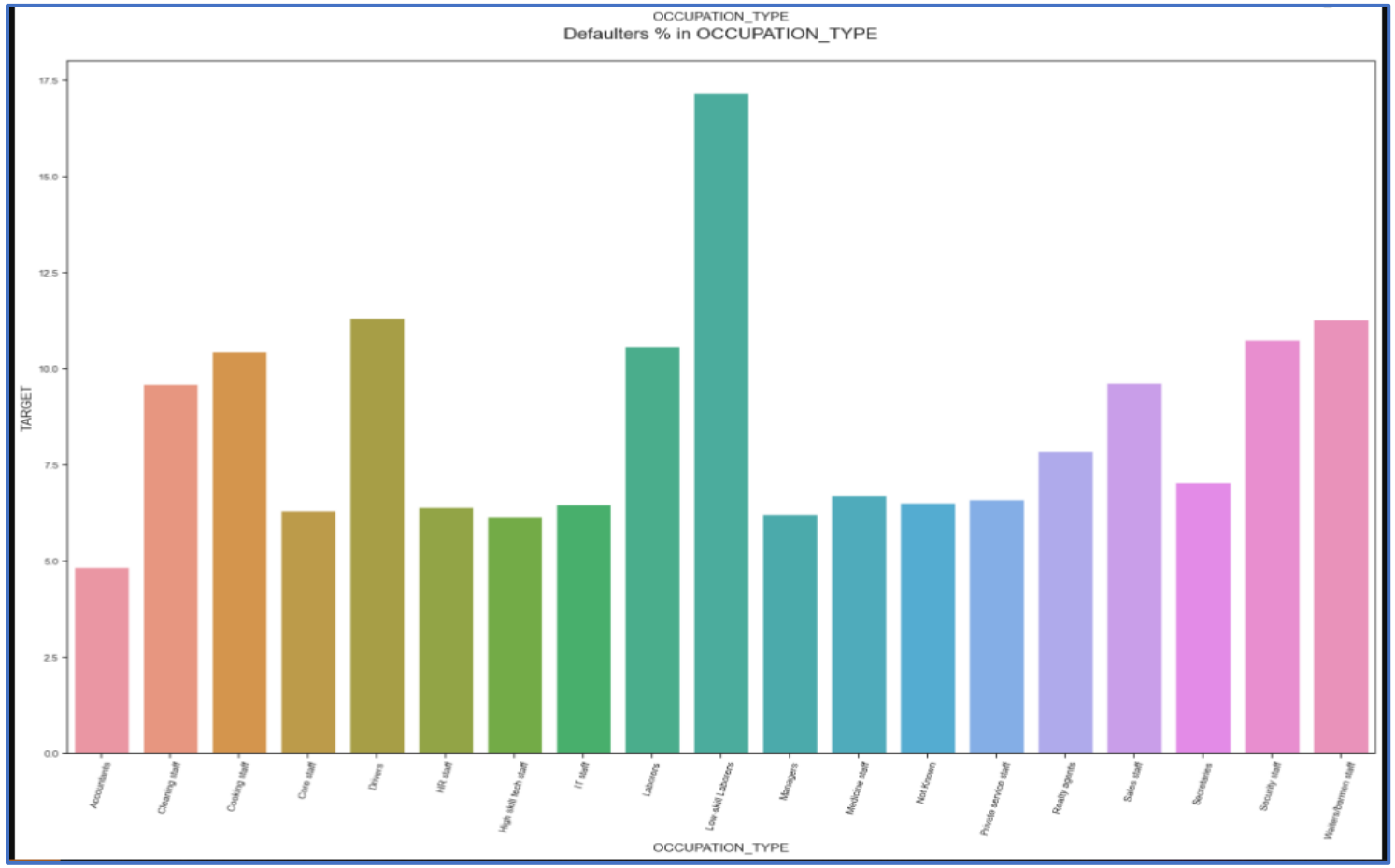
- Customers who have submitted a document as part of their loan application showed to repay loan more than default

Observations on Indicators contd..



- Customers who have submitted both personal mobile and work phone tend to repay loan

Observations on Indicators contd..



- Customers who work as IT Staff replay their loans properly
- Customers who work as laborers are more likely to default