Home Credit Default Risk

Sumeyra Bharuchi
Zahra Wieser
Rukayat Adeleke
Sana Shah
Feeza Sikandar

The day you sign a client is
the day you start losing them.

-Don Draper – Mad Men

PROBLEM RESEARCH

Goal

Minimize the number of clients whose credit loans are approved but in fact they are unable to pay the credit

Objective

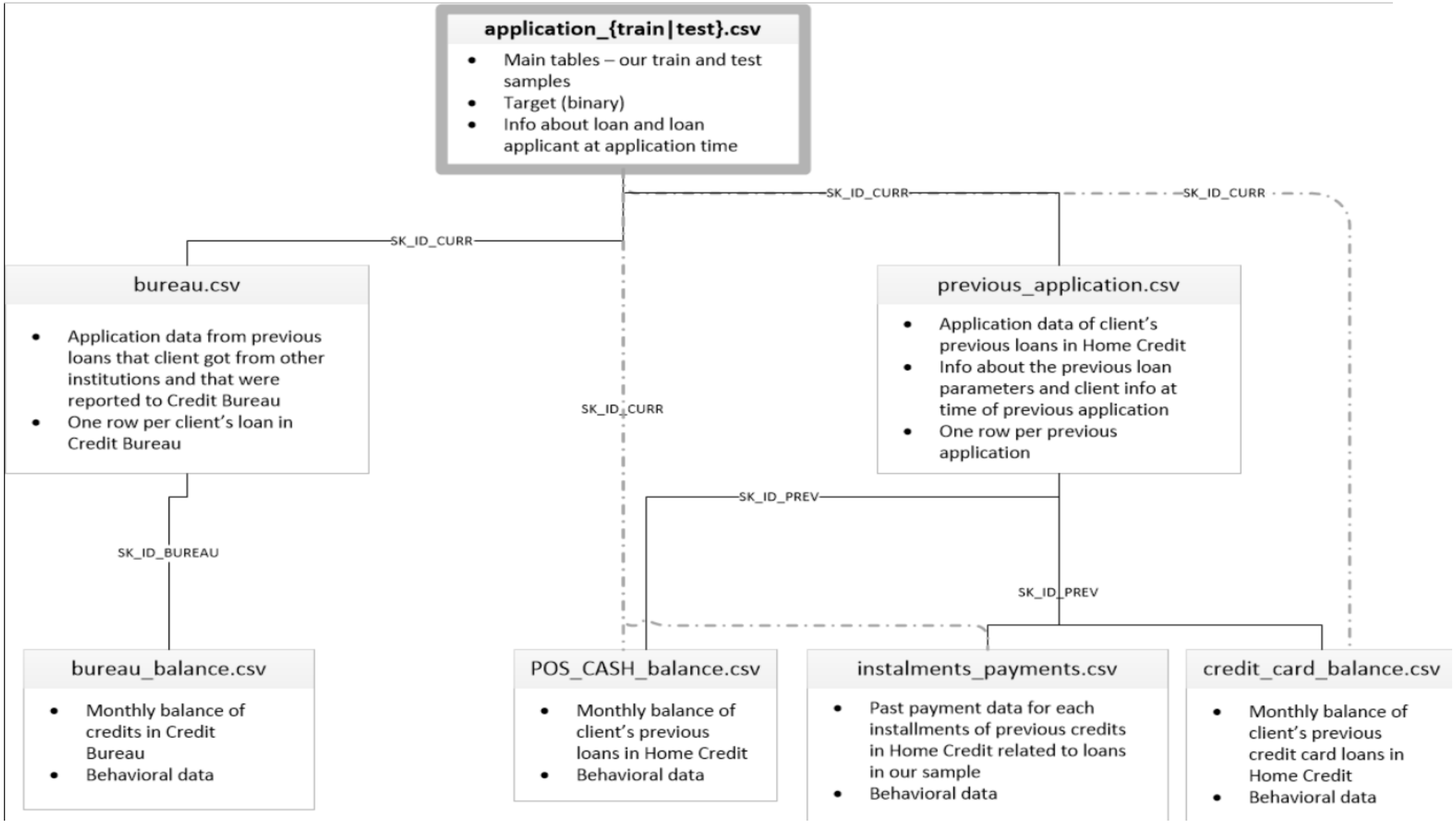
Build machine learning models to predict whether a loan applicant is capable of repaying the intended borrowed amount.

Bussines Metrics

LGD is an estimate of the amount of money that is lost from a borrowing institution when a borrower defaults on a loan.

DATASET

- The main dataset consists of 122 columns and 307511 rows. With 1 column target feature ["TARGET"]



EXPLORATORY DATA ANALYSIS (EDA), FEATURE ENGINEERING

EXPLORATORY EXPLORATORY DATA ANALYSIS (EDA), FEATURE ENGINEERING

1.Completing

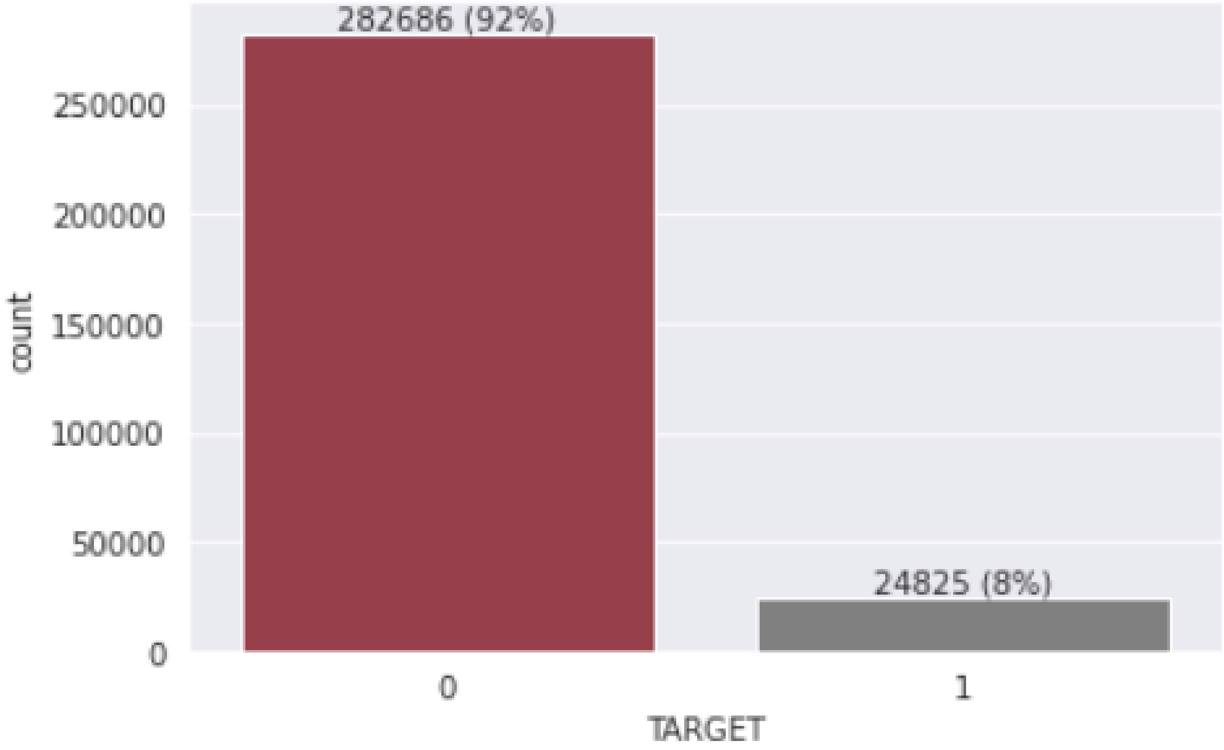
2.Correlation

3.Correcting

4.Conversion

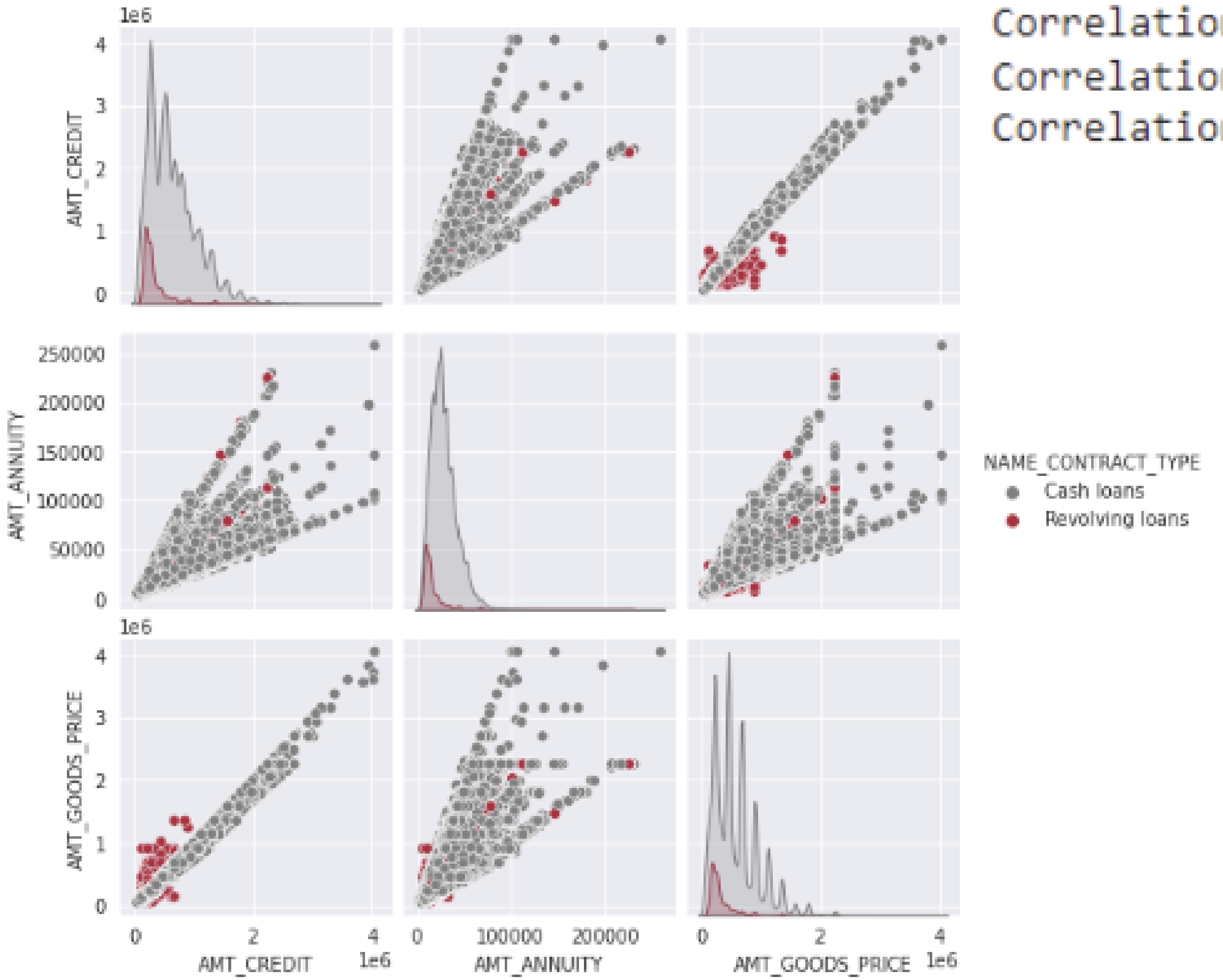
5.Creating

EXPLORATORY DATA ANALYSIS (EDA)



from total of customer there are **24825** **(8.0%)** individuals who **paid their loan** **(0)** and **282686 (92%)** individuals who **did NOT repay their loan (1)**

EXPLORATORY DATA ANALYSIS (EDA)

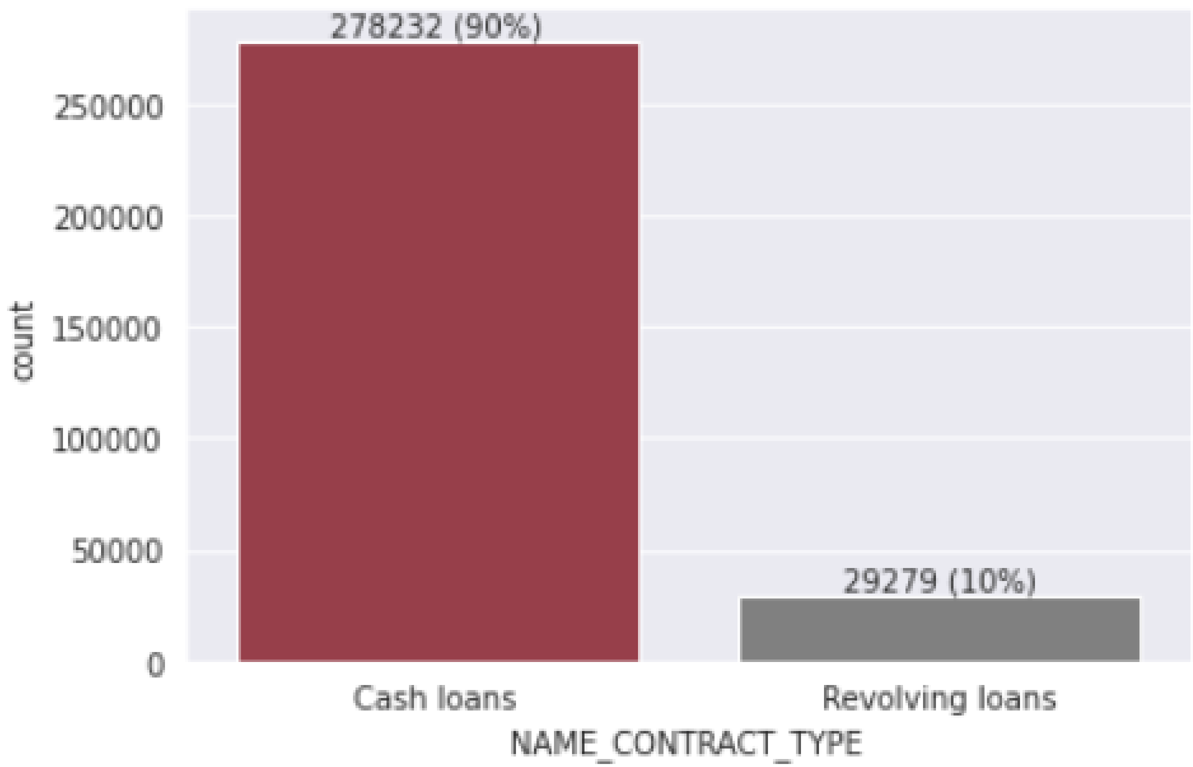



```
Correlation of Credit amount vs Price of goods: 0.99
Correlation of Annuity amount vs Credit amount: 0.77
Correlation of Annuity amount vs Price of goods: 0.78
```

Insight

- AMT_CREDIT and AMT_GOODS_PRICE are highly correlated (scoring 0.99), and has a positive linear slope - which makes sense because as the price of goods for which the loan is given gets higher, the credit amount of the loan gets higher too.
- AMT_ANNUITY is also highly correlated to AMT_CREDIT and AMT_GOODS_PRICE with a positive linearity. It's because the annuity is the monthly due amount

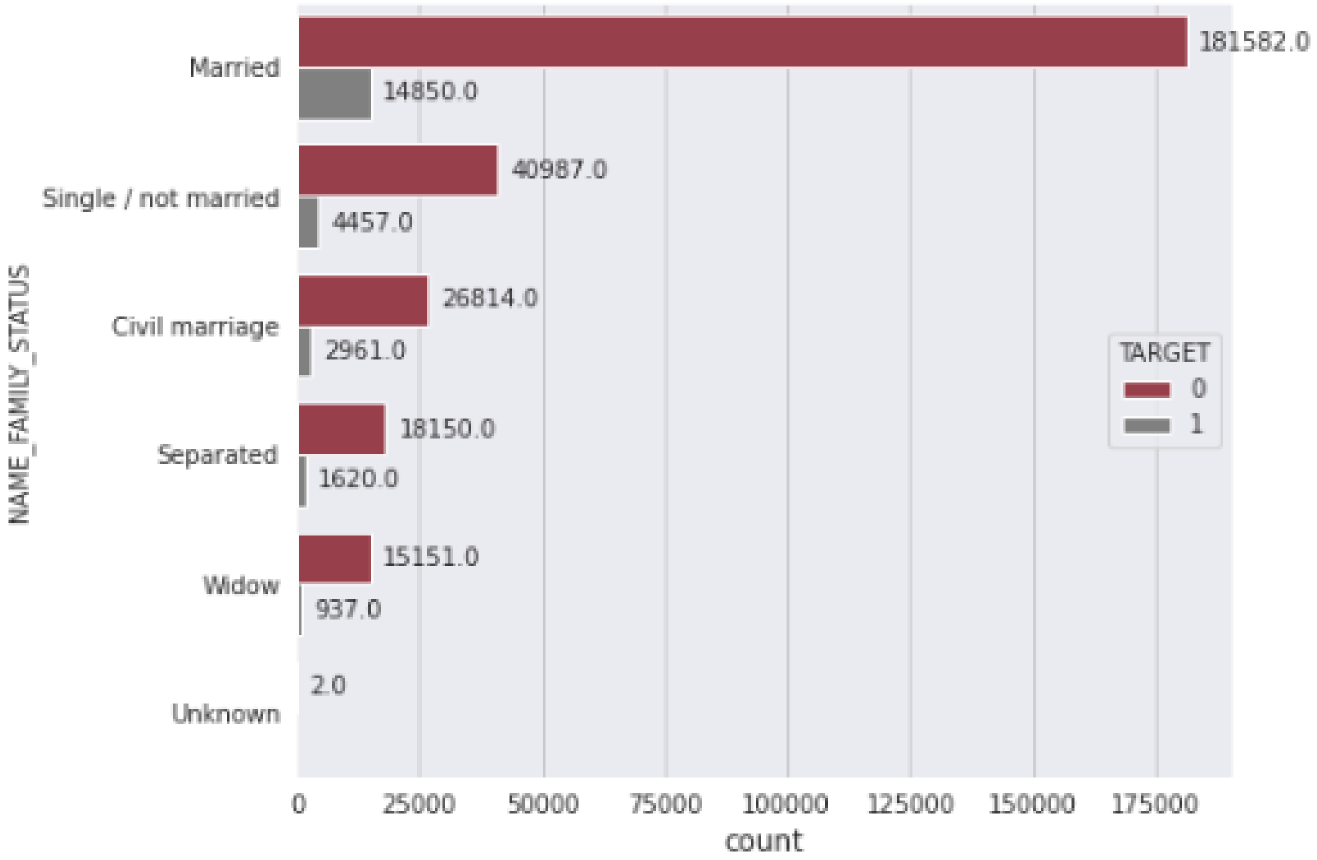
EXPLORATORY DATA ANALYSIS (EDA)



Insight

- Accounting for those who defaulted is much bigger in terms of cash loan than those with revolving loan, however, we must note that cash loan is significantly more popular to our sample consumers than the other.

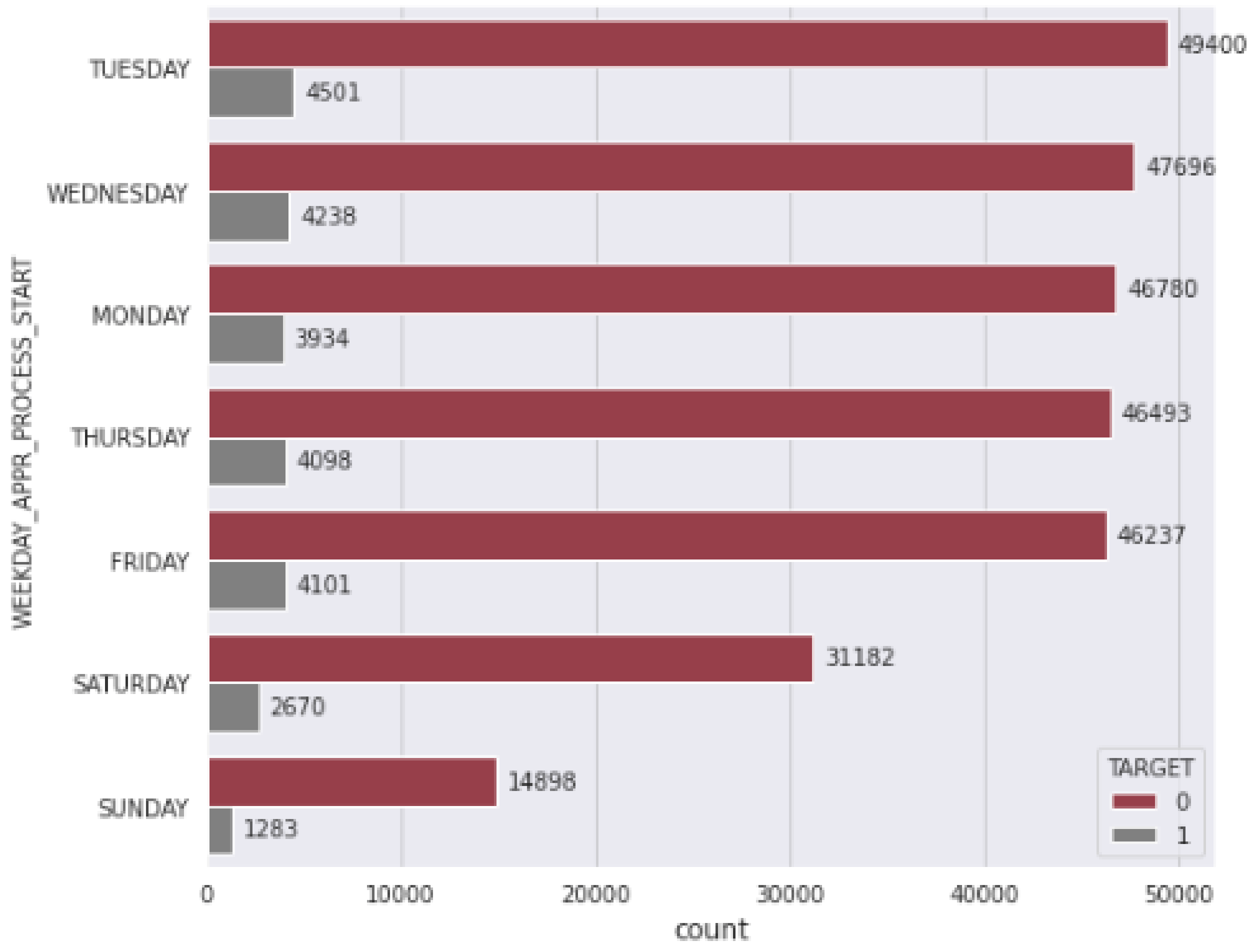
EXPLORATORY DATA ANALYSIS (EDA)



Insight

- We have a large number of married customers who are the most frequently defaulted individuals.

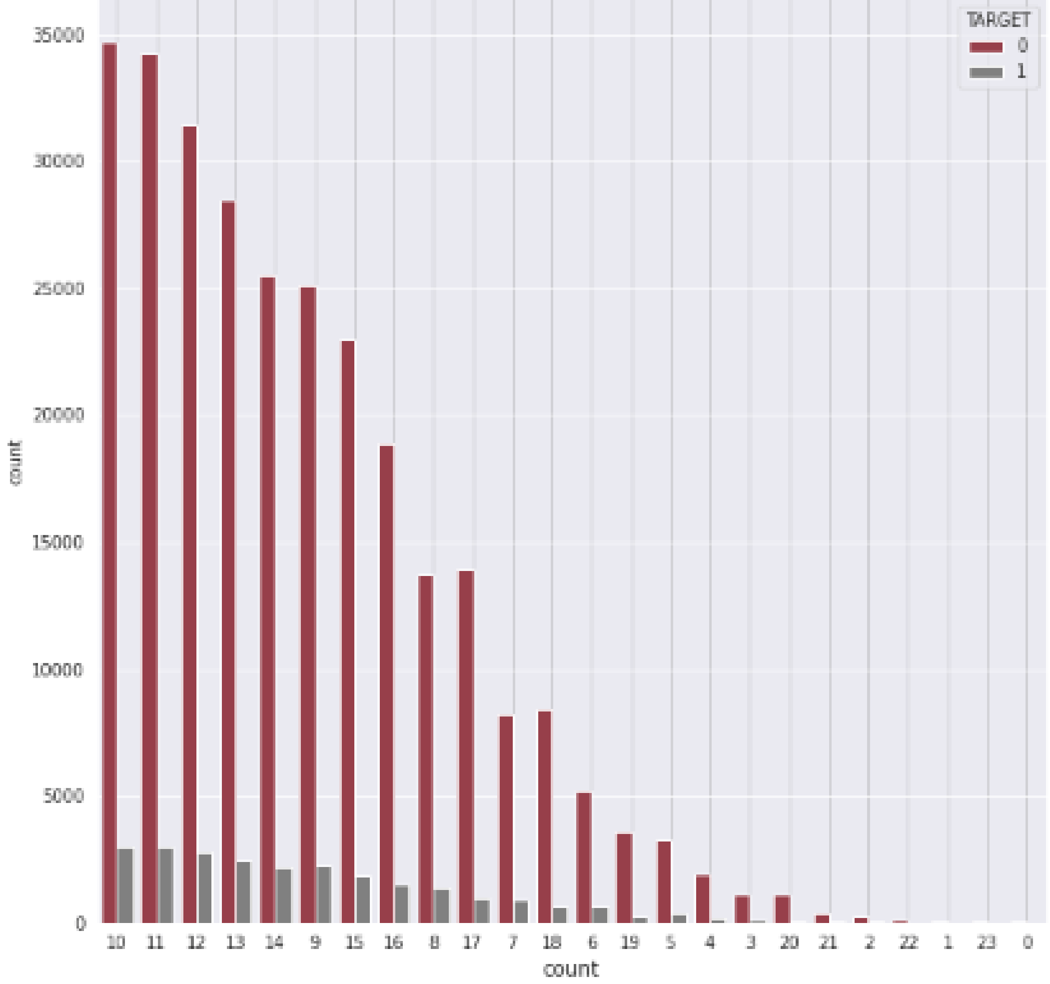
EXPLORATORY DATA ANALYSIS (EDA)



Insight

- Majority of the customers apply during weekdays, with a few on weekends. The trend on customers who weren't able to repay the loan is similar with that of those who did.

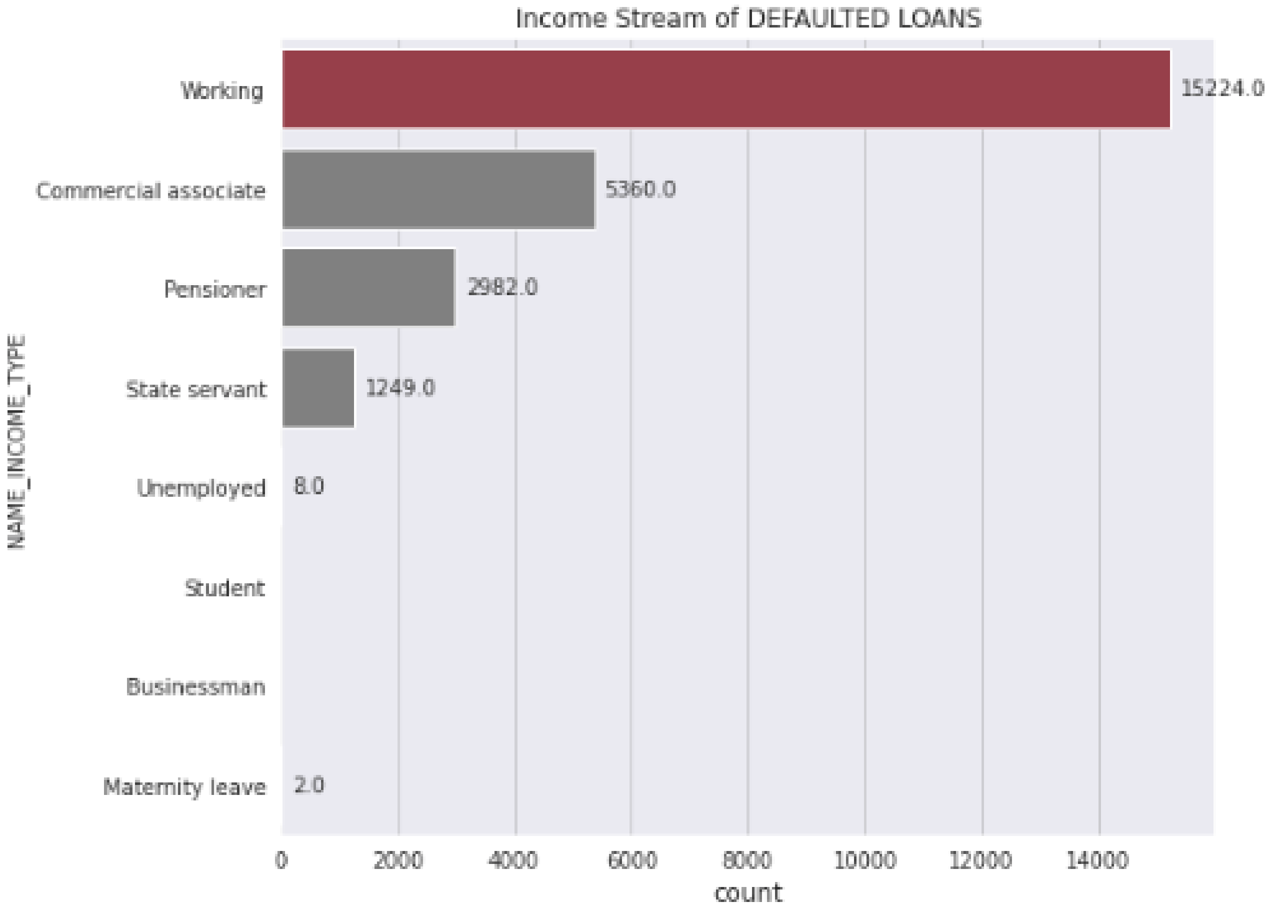
EXPLORATORY DATA ANALYSIS (EDA)



Insight

- Suspiciously, there are people applying for a loan account as early as 3am, and it gets denser throughout the day. Do note that those who defaulted on their loan has a similar pattern with those having good records.

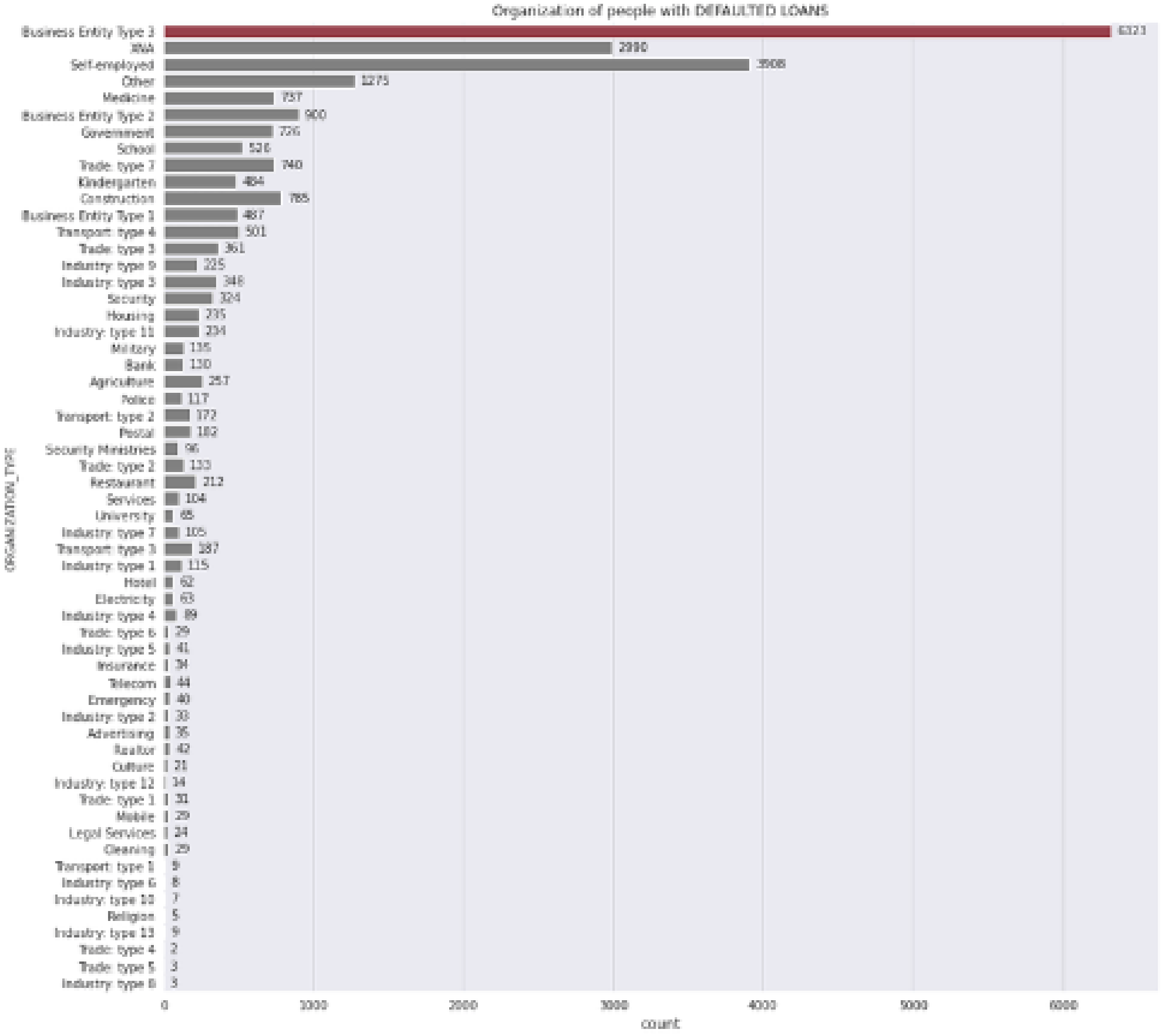
EXPLORATORY DATA ANALYSIS (EDA)



Insight

- The 'working' category is the most dense in terms of low wage high default customers. We also have very few samples on 'unemployed', 'student', 'maternity leave' and 'businessman'. Interesting to see that the 'businessman' category has an above average total income and has a high chance that they will maintain good credit scoring.

EXPLORATORY DATA ANALYSIS (EDA)



Insight

- The ORGANIZATION_TYPE is pretty diverse regarding where do these customers work. But base on the histogram, the category where the defaulting individuals are dominant are those in Business Entity Type 3, self-employed, and XNA.

DATA WRANGLING

After (finally) checking all our fields, it is time to proceed with data wrangling - also known as the data cleaning process.

Complete the null values of the following features:

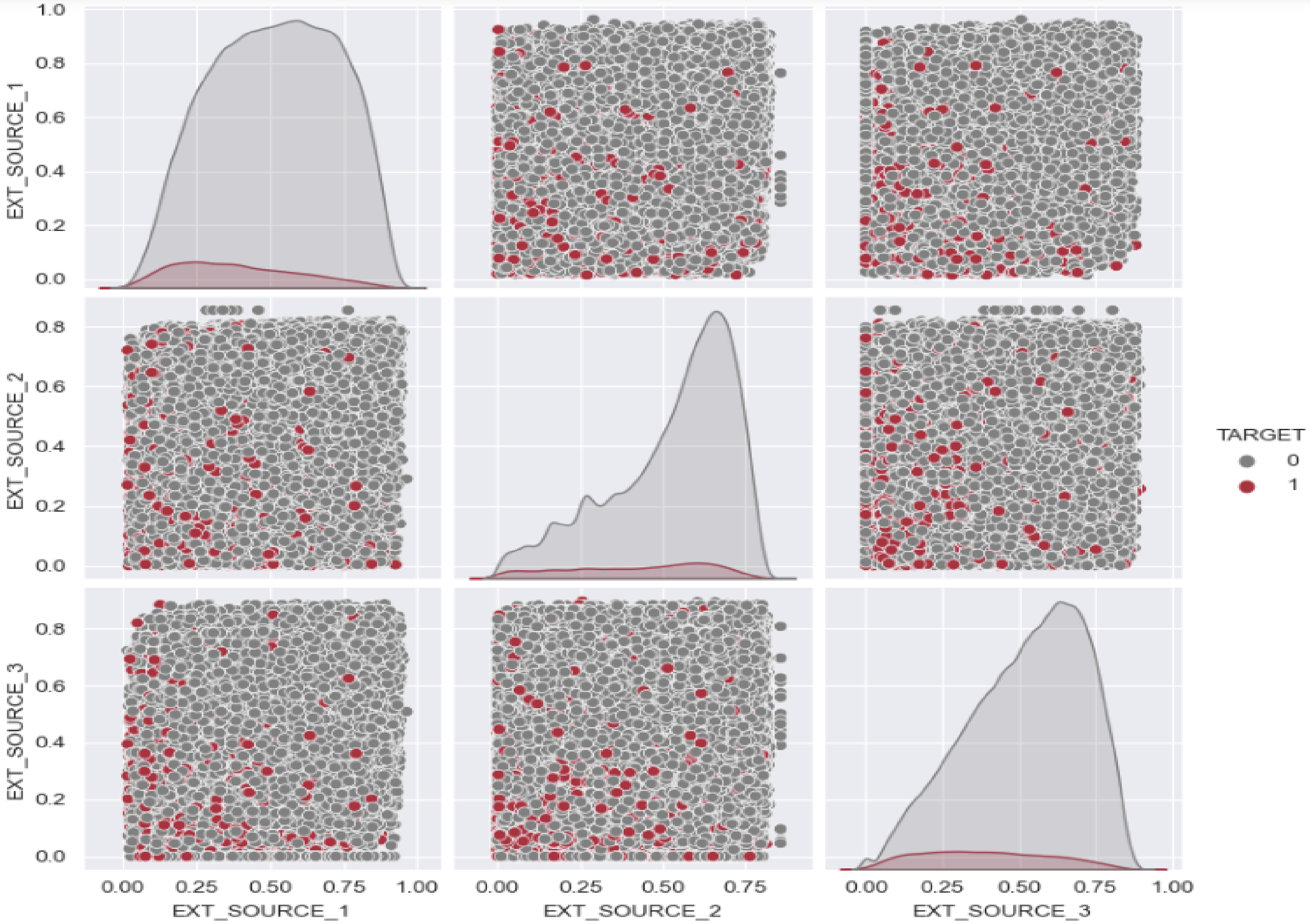
- 'EXT_SOURCE_1'
- 'EXT_SOURCE_2'
- 'EXT_SOURCE_3'
- 'CNT_FAM_MEMBERS'

Convert the anomaly data in 'DAYS_EMPLOYED'.

Convert the categorical text columns to numerical ones for:

- CODE_GENDER
- NAME_EDUCATION_TYPE
- ORGANIZATION_TYPE

DATA WRANGLING




```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 307511 entries, 0 to 307510
Data columns (total 4 columns):
 #   Column        Non-Null Count   Dtype
---  ------        --------------   -----
 0   TARGET        307511 non-null  int64
 1   EXT_SOURCE_1  134133 non-null  float64
 2   EXT_SOURCE_2  306851 non-null  float64
 3   EXT_SOURCE_3  246546 non-null  float64
dtypes: float64(3), int64(1)
memory usage: 9.4 MB
None
```

Observations:

These 3 fields are external data source score fields.

Base on the plot, those who were able to pay and did not pay can have scores fairly distributed on EXT_SOURCE fields, but it is quite evident that on the lower end of the normalized score mark (0.0-0.5), customers who paid (target=0, grey color) are much less prominent than those who didn't (target=1, red color)... and vice versa.

All 3 fields have missing values.

Logistic Regression

	Predicted 0	Predicted 1
Actual 0	70670	2
Actual 1	6205	1

	precision	recall	f1-score	support
payment difficulty	0.92	1.00	0.96	70672
other cases	0.33	0.00	0.00	6206
accuracy			0.92	76878
macro avg	0.63	0.50	0.48	76878
weighted avg	0.87	0.92	0.88	76878

Logistic Regression

Resampled training data

	Predicted 0	Predicted 1
Actual 0	47249	23423
Actual 1	2174	4032

	precision	recall	f1-score	support
payment difficulty	0.96	0.67	0.79	70672
other cases	0.15	0.65	0.24	6206
accuracy			0.67	76878
macro avg	0.55	0.66	0.51	76878
weighted avg	0.89	0.67	0.74	76878

Conclusion

Defaulting on your Home Credit payments can lead to serious risks and problems, including receiving terrorizing phone calls from debt collectors and even the risk of running away from your installments. However, if you’re having trouble paying your installments, don’t worry because there are solutions.

As a customer, make sure you understand all the terms and conditions in your loan agreement to avoid unnecessary fees.
Don’t be too hasty in choosing the loan amount you take and make sure you are able to pay the instalments on time and regularly. This way, you can minimize the risks and problems of taking out a loan with Home Credit.