EDA Credit Assignment

By: Sharanya Hegde

Index

- Problem statement
- Analysis approach
- Data imbalance – ratio (Target variable)
- Outcomes of univariate and sub-univariate analysis
- Correlation variables
- Conclusion

PROBLEM STATEMENT

Understanding driving factors behind loan default and Identify patterns which indicate if a client has difficulty paying their installments, so that the necessary steps can be taken.

ANALYSIS APPROACH

- Start off by importing, reading and understanding the data on python.
- Drop columns and rows with significant missing data, after ensuring the information being dropped are not important.
- Where necessary, impute missing data with the Mode, Median, Mean or with “Unknown” or “Others”.
- Ensure that the data is the right format ad units. Make the necessary changes.

ANALYSIS APPROACH (continued)

- Bin numerical data wherever necessary.
- Identify outliers. These have been retained for this assignent.
- Identified the imbalance in the data.
- Split the data based on targets – defaulters, and others (Non-defaulters).
- Perform univariate and sub-univariate analysis on different variables.

ANALYSIS APPROACH (continued)

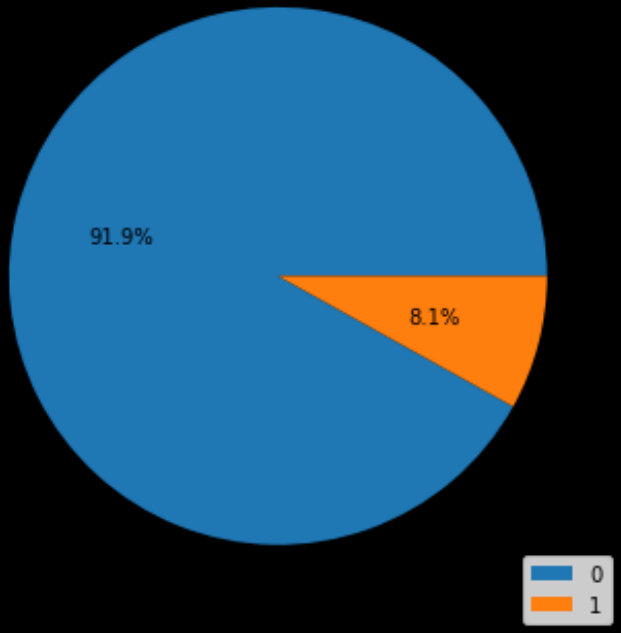
- Combine the Current ID and Target columns from the application data set to the previous application data set.
- Perform univariate and sub-univariate analysis on variables in the second data set.
- Identify the top 10 correlation between variables, for each target.

OUTCOMES

Data Imbalance:

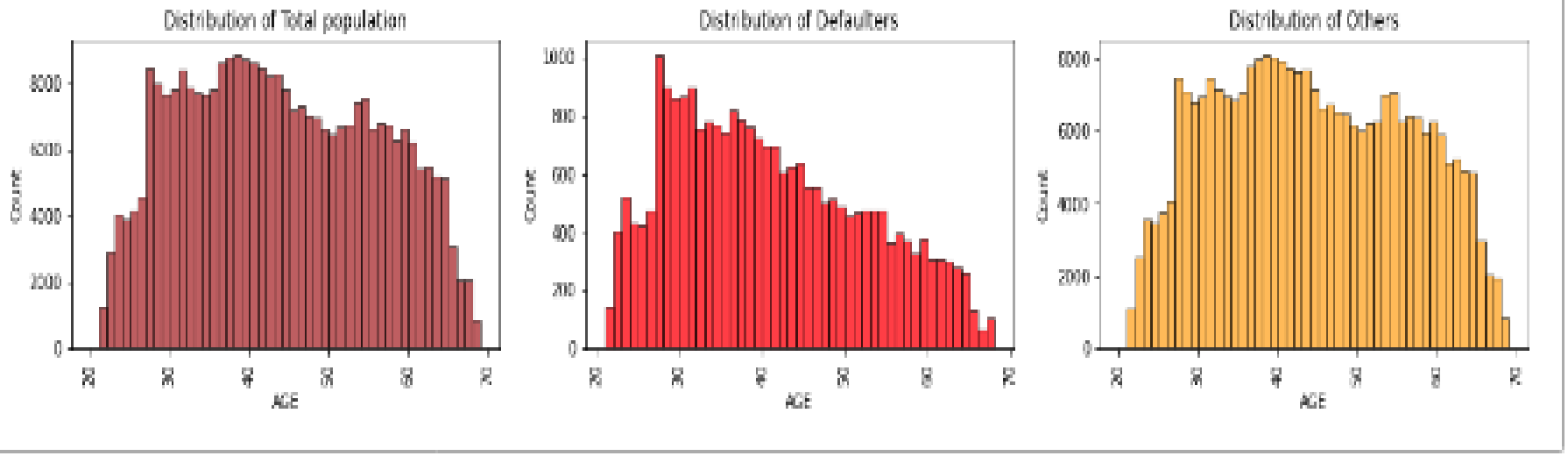
There was a huge imbalance in the Target Data, wherein just about 8% of the data pertained to Defaulters, and the rest 92% was with regards to Non-defaulters.

This makes the analysis less reliable since the population of Defaulters is vastly different from that of non-defaulters, there by not making the 2 comparable. Therefore, to identify the key factors leading to defaulters, we observed any difference in patterns in the 2 variables.



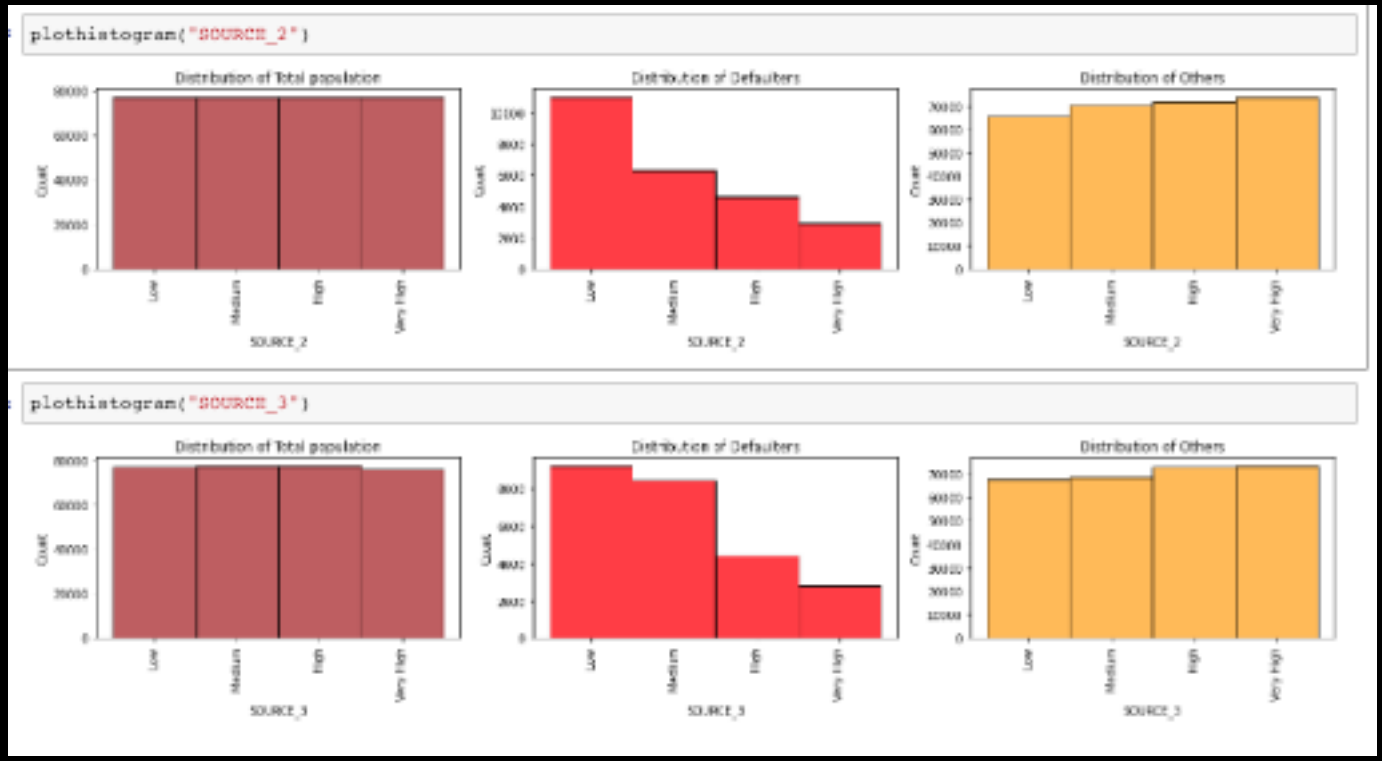
Univariate analysis - Age

```
plothistogram("AGE")
```



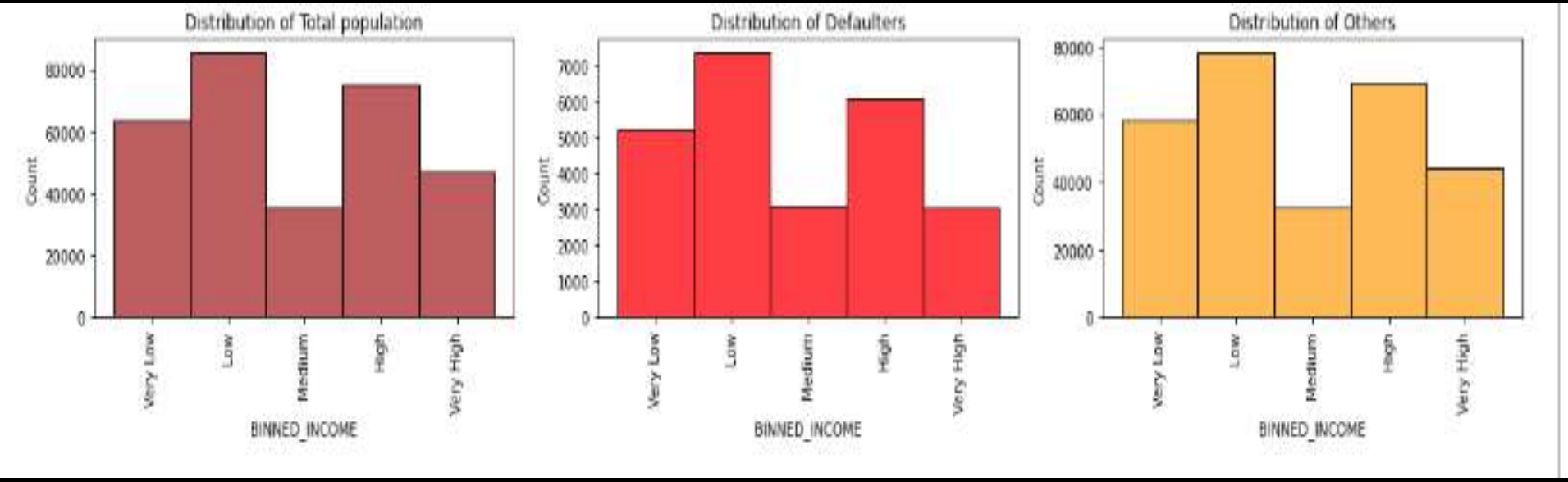
When one looks at the distribution as per age, you can see that most of the population tested is between the age group 35-45. But when it comes to defaulters, we see a higher number of people from the age of 25-35. This could mean that those from the age of 25-35 have higher chances of defaulting.

Univariate analysis – External sources



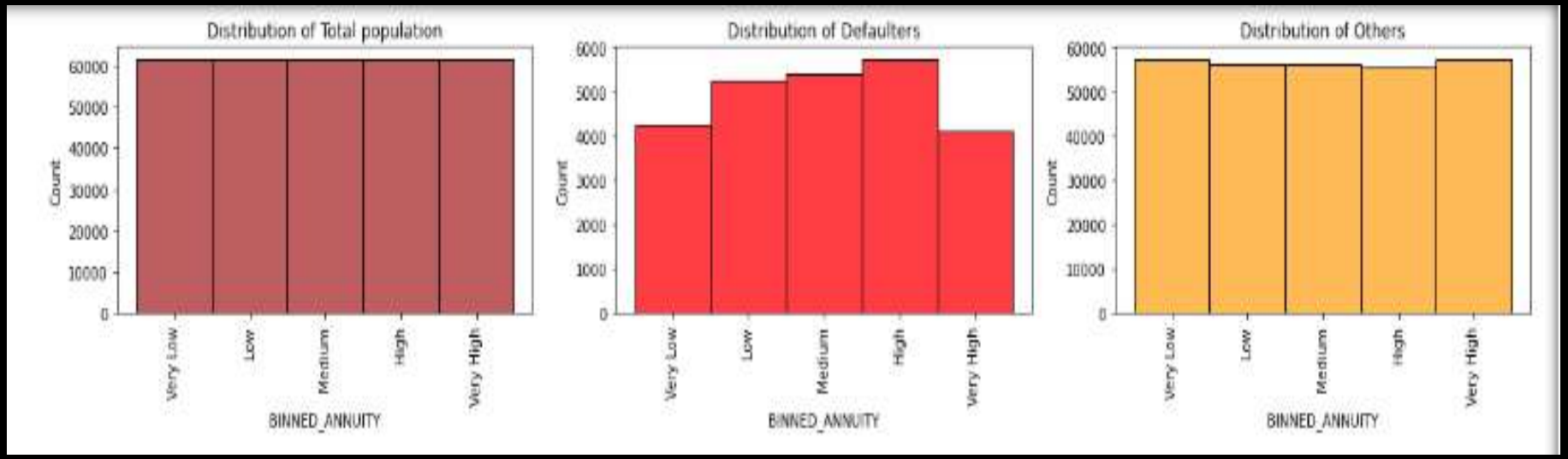
The distribution of outside sources are fairly equal across the total population. But we see a clear distinction in the defaulters population. Those with lower scores have a higher chance of defaulting.

Univariate analysis – Income



It has been observed that those of lower income groups have a higher chance of defaulting, since the percentage of low income candidates are the highest, in the defaulting data.

Univariate analysis – Annuity



It has been observed that with High and Medium annuity have a higher chance of defaulting. These are clients with annuity in the range of 22,000 to 38,000 (40-80% Quantile).

Correlation

The top 10 correlation variables for Defaulters are:

Variable 1	Variable 2	Coefficient
OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.99827
AMT_CREDIT	AMT_GOODS_PRICE	0.982566
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.956637
CNT_CHILDREN	CNT_FAM_MEMBERS	0.885484
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.869016
REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	0.847885
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.77854
AMT_CREDIT	AMT_ANNUITY	0.752195
AMT_ANNUITY	AMT_GOODS_PRICE	0.752022
REG_REGION_NOT_LIVE_REGION	REG_REGION_NOT_WORK_REGION	0.497937

Correlation

The top 10 correlation variables for Non-Defaulters are:

Variable 1	Variable 2	Coefficient
OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.99851
AMT_CREDIT	AMT_GOODS_PRICE	0.98688
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.950149
CNT_CHILDREN	CNT_FAM_MEMBERS	0.87857
REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	0.861861
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.859371
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.830381
AMT_ANNUITY	AMT_GOODS_PRICE	0.776251
AMT_CREDIT	AMT_ANNUITY	0.771297
REG_REGION_NOT_LIVE_REGION	REG_REGION_NOT_WORK_REGION	0.446101

Conclusion

Due to data imbalance the outcomes aren't conclusive. But some of the key factors observed are :

Age:

- Those between the age of 25-35 have higher chances of defaulting. Considering they may not have a steady a income or a job, it does make them likely to default.

External sources 1 & 3:

- Those with lower ratings from external sources have higher chances of defaulting.

Amount of Annuity:

- Clients with middle to higher annuity payments have a higher chance of defaulting. The range of annuity is from 22,000 to 38,000.

Income:

- Customers part of the lower income group, have a higher chance of defaulting. It is likely that lower income groups won't have the capacity to repay loans.