Credit EDA case study

BHAVINI D. GADHIWALA

SANKALP RUSIA

Data Understanding Techniques

- Division of categorical and numerical columns
- Box plots used to check for Outliers
- Distribution plots used for checking the skewness in data
- Pie plot used to check for data imbalance
- Pair plots used to see the emerging patterns
- Heatmaps used for drawing out the top 10 correlations in the data
- Count plots used to see the patten in the population

Data Cleaning Approach

CATEGORICAL COLUMNS

- A new value created and applied. For example ‘Unknown’ was applied where there was no value in Occupation type
- Created new values by binning continuous attributes for e.g. income, age and employment years
- Dropped columns where minimal data available and number of rows were low in numbers

NUMERICAL COLUMNS

- Mean applied
- Outliers treated with the formula
- Imputed with zeroes as applicable
- Imputed with a threshold value for example in the number of children

Categorical Columns observations

Married people applied for more loans 64%

Female consist of 66% of loan applicants and it is also observed that the percentage of mail defaulters are more

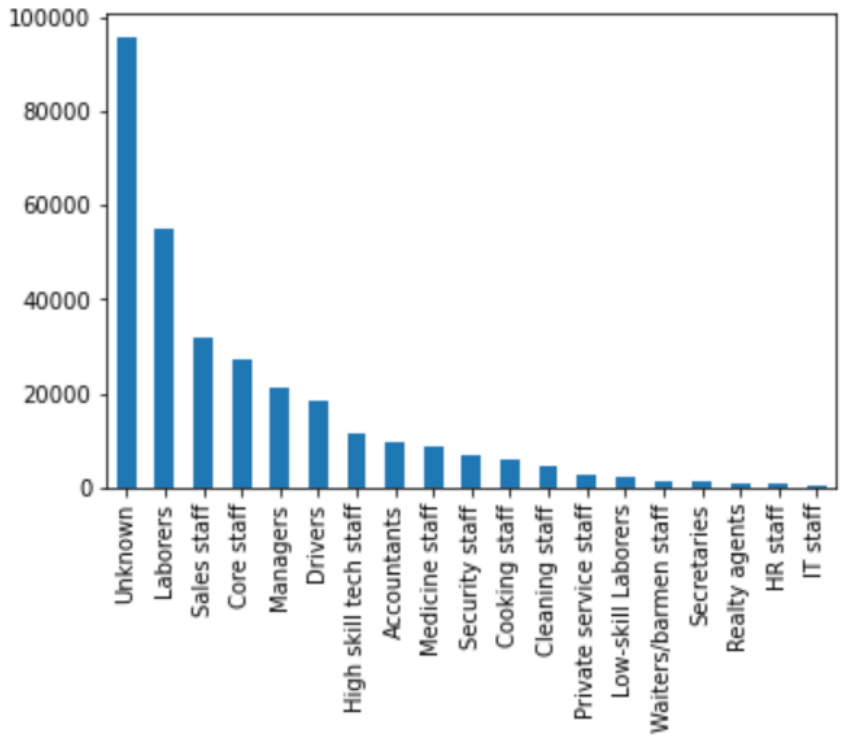
Cash loans more popular 91% compared to 9% revolving loans. Cash loans also has more defaulters

Day of the week is a completely independent variable

More than half of the loan applicants are Working

Educated people are observed to apply for loans

Occupation Types



- The occupation type attribute had a lot of missing values
- A new value of ‘Unknown’ was added for the missing values and that was observed to have the highest value count
- Another observations is the Laborers, which was the second highest value

Value Counts Distribution

Occupation type is a highly undisclosed attribute and among those known, Labourers were the highest type to apply for loans

Low and very-high income people tend to apply for more loans compared to High and Medium

This is also reflected in the age wise count plot where maximum loans are applied for by individuals between 30-40 years of age and with the years of employment of less than 5 years

Count of children is an independent variable

Lower loan amount of upto 5 lakh is most popular followed by 10 lakh then there is a huge drop observed.

Bi-variate Observations

Linear relation observed between credit amount and the annuity

Also found positive correlation between annuity and credit amount. The variation in the slope of it could be owing to changes in the rate of interest.

Previous Application Observations

Customers whose previous application was approved are mostly non defaulters

The defaulters are the applicants whose previous applications were refused maximum. So they should be avoided while approving the loan.