EXPLORATORY DATA ANALYSIS

By Koundilya Nitin
Gonda Priya Darshan

Case Study

- Inspecting null values
- Analyzing columns
- Check for irregular Values in application data
- Outliers in relevant columns
- Imbalance Percentage

Case Study

- Univariate Analysis for relevant categorical Columns
- Multivariate Analysis
- Correlation
- Conclusion

Case Study-Null values

As total number of rows and columns in application data is 307511 rows * 122 columns ,

It is not feasible to check each column for null values hence create a plot which shows column name where NA count is more than 50%

Case Study-Analyzing columns

Hence it is appropriate to impute missing values of columns by median of respective columns

Case Study-Outliers

- Looking at the box-plot it is pretty obvious that there is an outlier which is cut off the chart and is alone enough to disrupt true values of average AMT_INCOME_TOTAL.As compared to mean value it is 1000 times more than mean value.In such cases it is advisable to remove outliers and perform analysis
- AMT_CREDIT is the loan credited to applicants. Seeing the box plot it can be concluded that 3rd quartile is under 10,00,000 but there are outliers which are reaching to 40,05,000.Hence it can be concluded that an average of amt credited is 10,00,000.

Case Study-Outliers

- Same goes with AMT_ANNUITY, as it clear that max value is under 50,000,however little concentration of data points are available above 3rd quartile
- AMT_GOODS_PRICE can be seen that max number of loan credited for goods price are under 680000
- CNT_FAMILY_MEMBERS are generally around 2.3 family members but as per box plot there are 20 family members

Multivariate Analysis

- There are number of outliers in each of plot which are very much available in all the plots

Case Study-Imbalance Percentages

- It can be seen that data is quite unbalanced over applicant client who have none difficulty(91.93%) and who have difficulty(8.07%) in repaying any installment

Case Study-Univariate Analysis

- Cash loans are distributed almost 10 times more than revolving loans
- Female applicant is 31% more than male applicant for loan.
- It is interesting that around 70% of applicant have no child and 20% has one child

Case Study-Univariate analysis

- It is quite clear that applicant with secondary special degree has highest % age of Loan Application ,followed by higher education
- 64% are married,15% are single and there is no significant difference in default % age
- It is also visible that more than 85% live in house/apartments ,while 5% live with parents

Case Study-Univariate analysis

- Major Applicants Job type is Laborers(18%)
- 50% Applicant within income range of 1-2L
- 28% are within range of 2-5L
- Remaining are Lying in the range of 50k-1L
- Maximum density of Loan credit is between 2L-7.5L

Co-relation

- From the plots we can say that the correlation values is always within 1

Conclusion

- Understand
- Derive
- Interact

Thank You