Introduction to Risk Analytics in Lending: Exploratory Data Analysis

Understanding Loan Default Risk Through Data Insights

Introduction

- The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it to their advantage by becoming a defaulter.
- We can use EDA to analyze the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

Importance of EDA in Risk Analytics

- Helps in identifying factors influencing loan default.
- Guides decision-making to minimize risks and maximize profitability.
- Provides a foundation for predictive modeling and risk assessment strategies.

Data Understanding

- The Dataset given to me is having 307511 rows and 122 Columns
- 122 Columns have all the information about applicants like his Age, Gender, Income, Profession etc.
- The column “TARGET” has to unique values 1’s and 0’s.
- Target variable (1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample, 0 - all other cases)

Approach to Analysis

- Data loading and initial inspection.
- Handling missing data and outlier detection.
- Analysis of data imbalance.
- Types of visualizations used (histograms, scatter plots, etc.).
- Correlation analysis and insights derived.

Data loading and initial inspection

- I have gone through the data, There are more than 3 lakh Rows and 122 columns.
- I have checked for missing values and came to conclusion that there some columns which are having more than 50% missing values, I considered dropping those because analyzing those will not give proper insights.
- For remaining columns, I calculated the percentage of missing values and sorted them in descending order so it will be ease for me to handle them

Handling missing data and outlier detection.

- For “FLOORSMAX_AVG“ Column i.e : Normalized information about building where the client lives, it has highest number of missing values.
- Upon checking by plotting the Histogram and analyzing mean and mode of it, I can see it has no significant outliers, I can use its mean value to replace with misisng values in the column.
- Using similar approach and analyzing the graphs and checking the mean and mode values, I treated most of the numerical values.

- For “OCCUPATION_TYPE” column, Its categorical in nature.
- I can see “Occupation_Type” column has abut 31% missing values, We normally replace the missing values in categorical columns by its mode, But as we can see Mode is Laborers category, Its does not seem to be a good idea to replace it, I will keep it untreated
- This way by analyzing carefully I handled all the columns.

Analysis of data imbalance

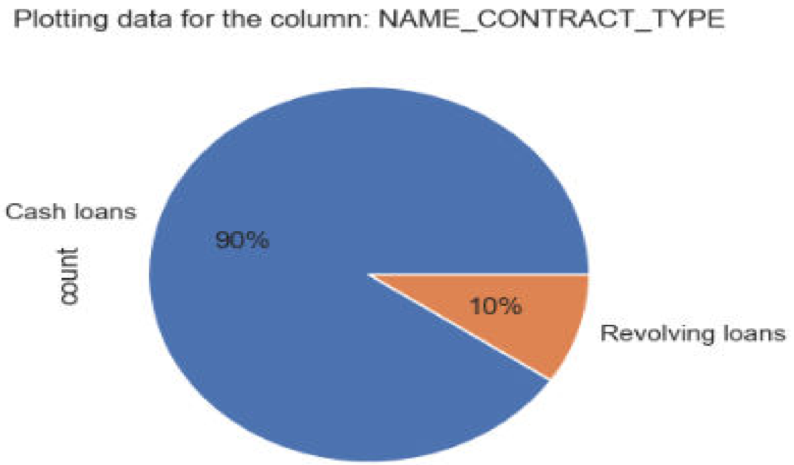
- Here we have are analyzing the "Target“ column.
- I have cheked the unique values in Target column.
- Calculated the No of 1's and 0’s in target column.
- Calculated Data imbalance Ratio in Target column
- Here the Data imbalance Ratio is around 11, We can conclude by saying for every 1 defaulter there 11 people who pays the loans
- This imbalance needs to be addressed when building predictive models, as it might cause the model to be biased towards the majority class (non-defaulters).

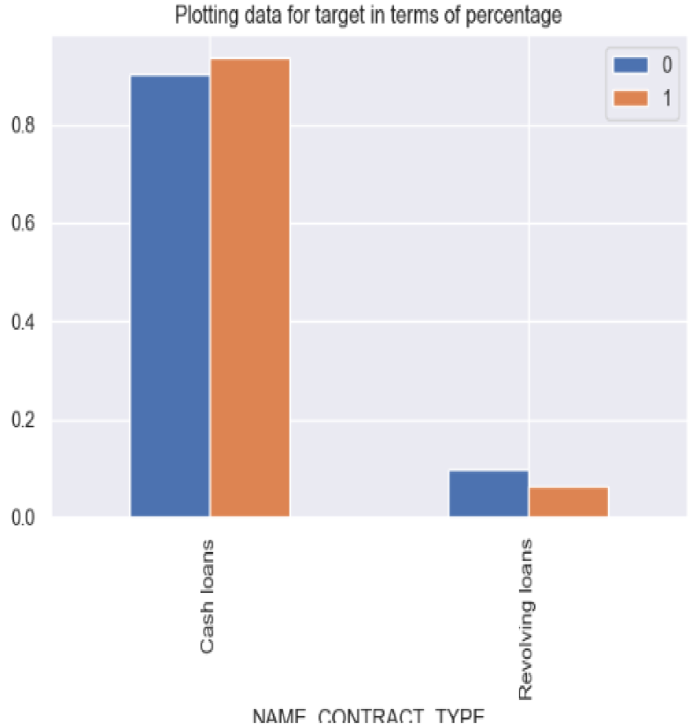
EDA Results

- A Plotting function is designed to create a set of three plots to analyze a particular column (column) in relation to the target variable (TARGET).
- We can draw some important insights by plotting the graphs

Univariate Analysis

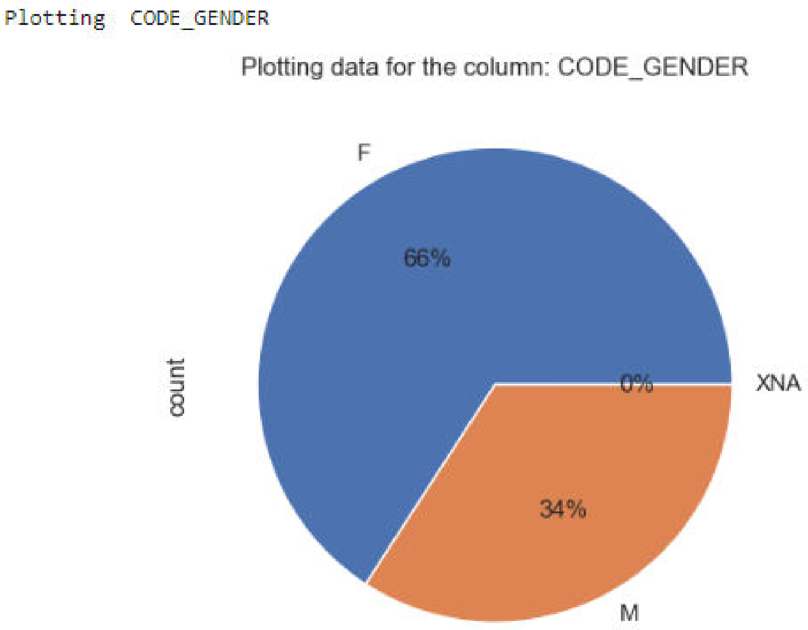
Plotting the pie diagram for NAME_CONTRACT_TYPE Column





- By analyzing the pie diagram we can say majority of the loans are of type Cash Loans(90%), and only 10% are of Revolving loans
- Percentage vice also We can see the same trend in the column

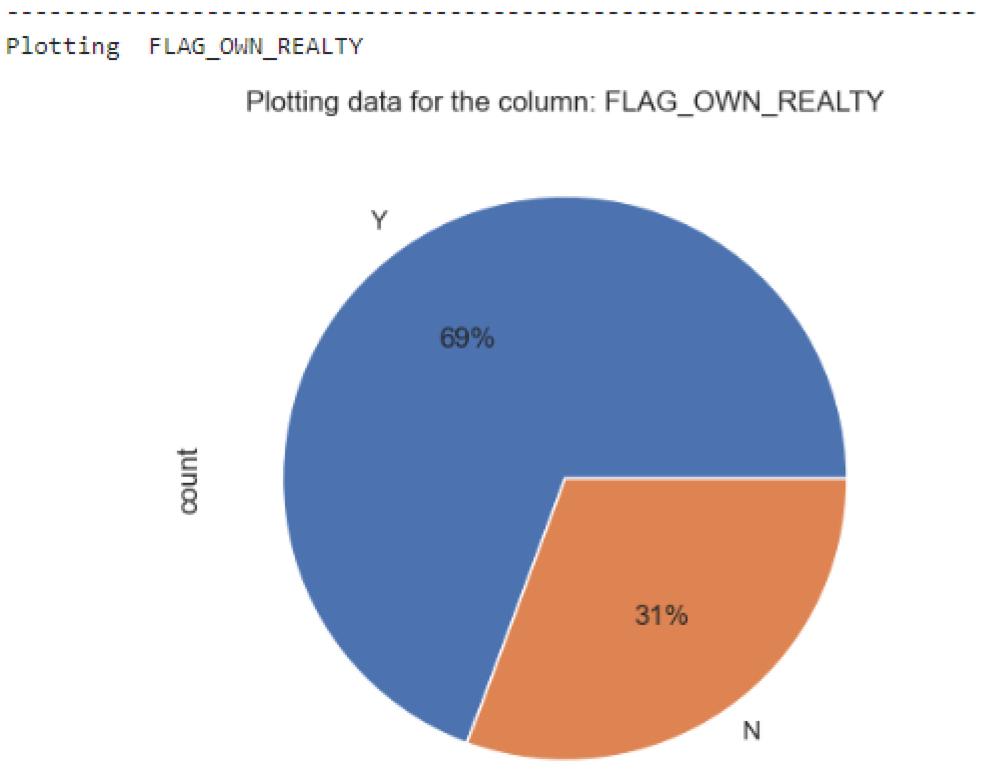
Plotting the pie diagram for Gender Column



```
Plotting  CODE_GENDER
```

- About 66% of applicants are Males and 34% are females.

Analyzing the Column which tells us about the property of Applicant



```
-------------------------------------------------------------------------
Plotting  FLAG_OWN_REALTY
```

Plotting data for the column: FLAG_OWN_REALTY

Correlation

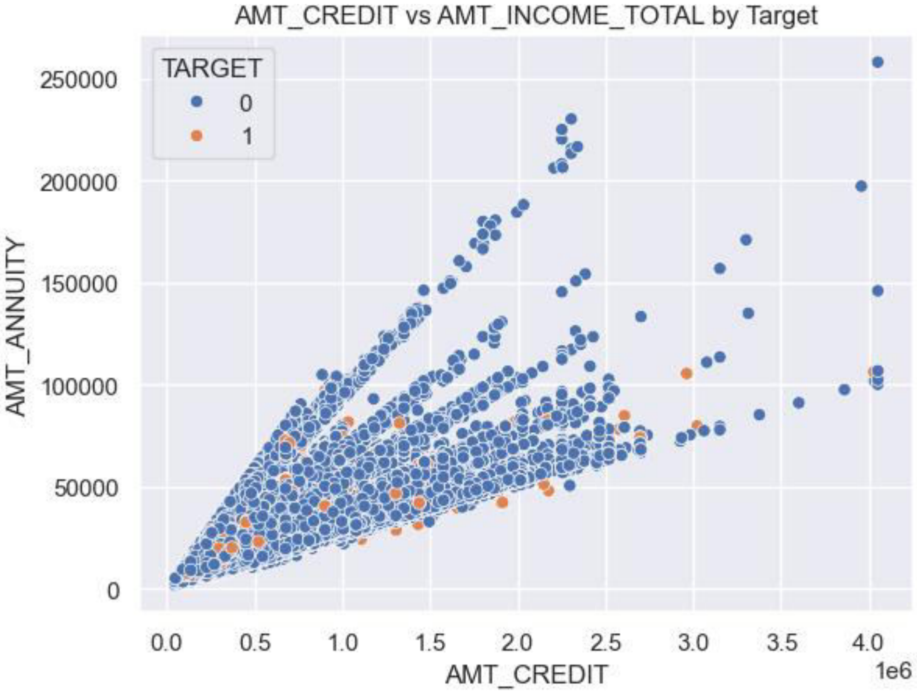
- we are calculating the absolute correlation values for the numeric columns when TARGET=1
  The output you've provided is a series of correlation pairs and their corresponding correlation coefficients. Each line shows:
- Two variables (e.g., REGION_RATING_CLIENT and FLAG_DOCUMENT_20).
- The correlation coefficient between these two variables (e.g., 0.000010).
- This value suggests an extremely weak linear relationship between these two variables. In practical terms, their values do not move together in any meaningful or predictable way.

- **BASEMENTAREA_MEDI - BASEMENTAREA_AVG (0.998250)**
- The correlation coefficient between BASEMENTAREA_MEDI and BASEMENTAREA_AVG is 0.998250.
- Here, the correlation is very strong, close to 1. This indicates that these two variables are highly correlated. In this case, it's likely that BASEMENTAREA_MEDI and BASEMENTAREA_AVG are different measurements or estimates of the same underlying feature (such as basement area), hence the high correlation.

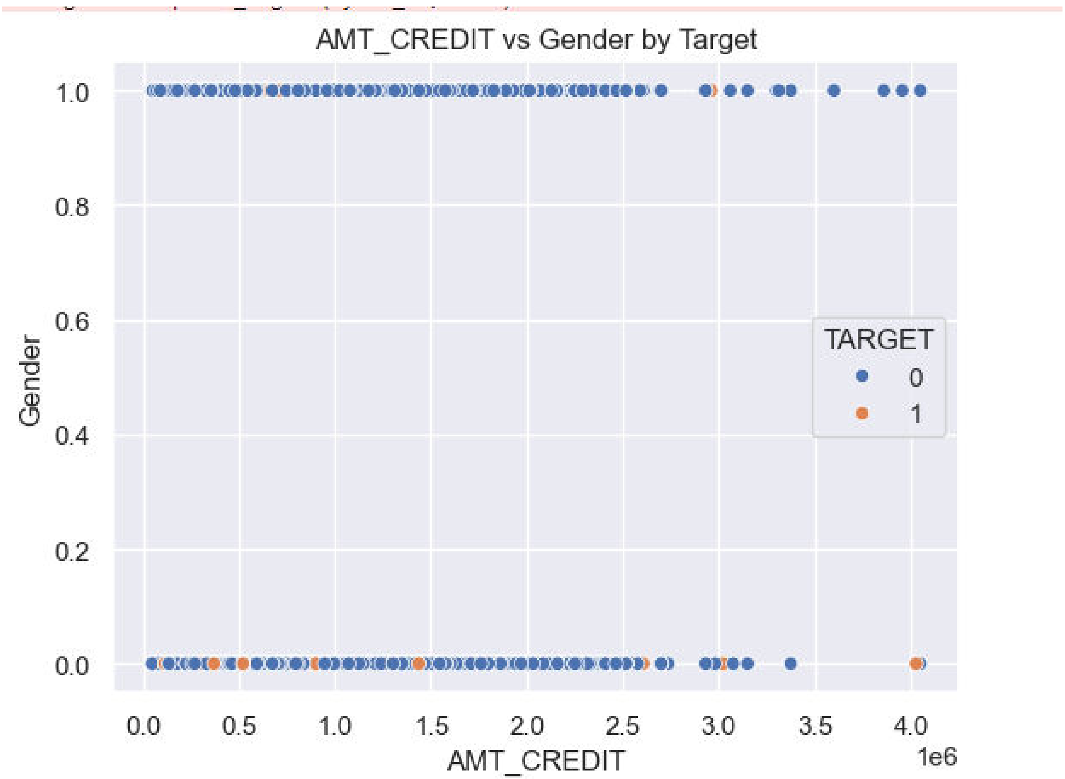
- After drawing Heatmap for Top 10 and least 10 correlated features with TARGET
- We can see that the column (REGION_RATING_CLIENT) finacial instiute’s rating of the region where client lives (1,2,3) and column(REGION_RATING_CLIENT_W_CITY) Our rating of the region where client lives with taking city into account (1,2,3) are two factors that affecting the defaulter rate, So we should consider this rating, according this rating we should make decisions whether to give loan or not

Bivariate Analysis

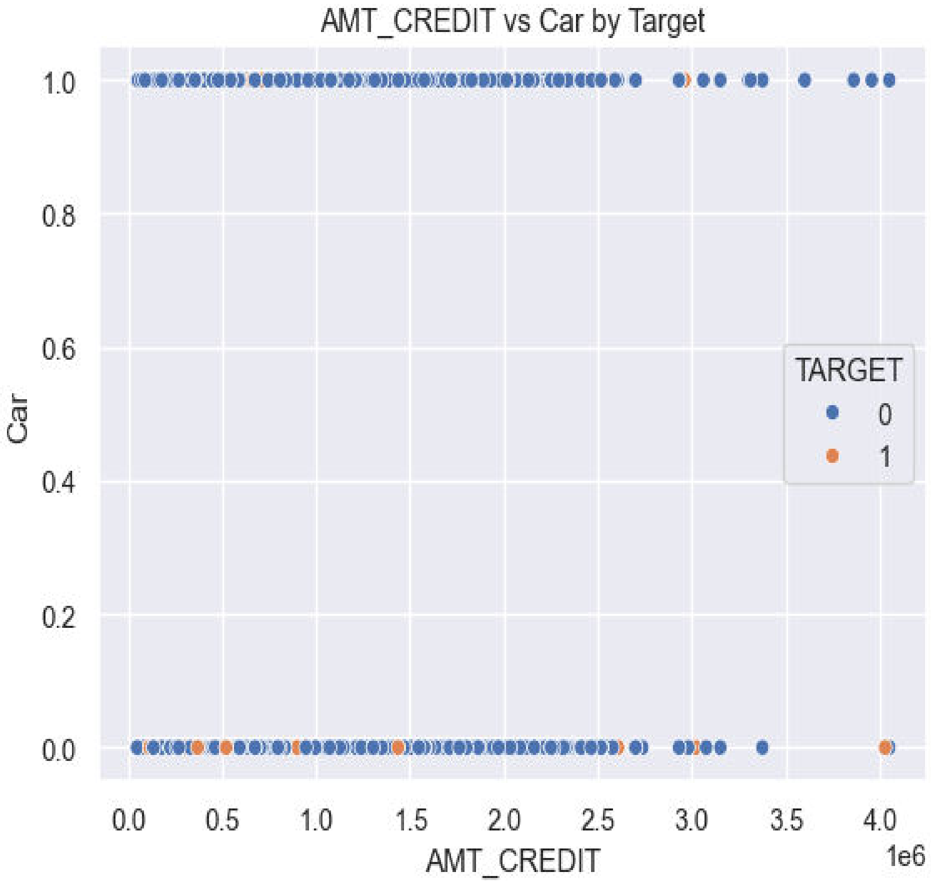
Upon examining amount of loan give and Amount annuity, We can see a trend that as the amount annuity increases the defaulter’s rate decreases



By analyzing amount of loan and gender with respect to target, We can see that for smaller loan amount and female clients, The defaulter rate is bit high



By looking at the graph we can say that people who don’t have car and who took smaller loan amount, The risk of defaulting the loan is bit high



Conclusion

- **Top Correlations with Default (TARGET=1):**Features such as DAYS_EMPLOYED, FLAG_EMP_PHONE, and OBS_60_CNT_SOCIAL_CIRCLE have high correlations with the default risk, indicating that longer employment gaps, owning a work phone, and high social circle observation counts are strong indicators of default risk.
- **Implication:** Understanding these correlations helps the company to refine the criteria for loan approval, focusing on applicants with higher employment stability and social network insights.

- **AMT_CREDIT vs. AMT_INCOME_TOTAL:**Scatter plots show that non-defaulters typically have a balanced relationship between credit amount and income.
- **Implication:** This reinforces the need for proportional lending, where the loan amount should be aligned with the applicant's income level.