LEADING CASE STUDY

Loan Risk Analysis

Followed Steps

- Data Cleaning
- Data Understanding
- Data Analysis
- Data Visualization

Data Cleaning

- Remove all Nan column and set a threshold of Nan value to 30% and then remaining nan rows is remove from the dataset.
- Drop duplicates values.
- In this dataset there are multiple columns have more than 80% nan values and some of columns has more than 40% nan values. So I dropped all the columns.
- Change the values of string column to integer.

Data Understanding

- Understand the stats of a dataset to figure out the distribution of multiple various.
- Check the value count of categorical values.
- Based on the understanding; select the multiple columns from dataset and drop some irrelevant columns from it.

Data Analysis

- Perform Univariate and Bi-variate analysis to understand the data in respect their distribution, counts and peaks.
- In Univariate analysis I figure out that there is major difference in terms of values in respective loan_status categories.
- Where in Bi-variate analysis there is no major difference in respective loan_status categories.
- Might be the data set is imbalance towards the Fully_paid loan status.

Data Visualization

- Create a graph of different categories. Distplot and histogram indicate the distribution and the peak of the attributes.
- Countplot indicates the values counts of each category.
- Boxplot indicates the stats of data as well as outliers too.
- Barplot indicates the count between open account and respective loan status values.
- Boxplot indicates the outliers in respect of different loan status and funded amnt.

Conclusion

- Data is imbalance in respect of loan status.
- Loan amnt is right skewed.
- Funded_amnt distribution is random.
- We can consider the grades and open account attributes.