Project 2 – Summary

**Observations based on Week 1 analysis:**

1. There are a total of 233154 customer observations. The data has a total of 40 features and 1 target value indicating if there is loan default or not.
2. 'Employee\_Type' parameter has 7661 null values. Remaining employess are classfied as "Self-Employed" and "Salaried". A good strategy to fill missing 'Emplyee\_Type' values is to check the 'Employee\_code\_ID' and fill missing values accordingly as the mode of 'Employee\_Type' for corresponding 'Employee\_code\_ID'. For example, 62 people with missing Employee\_Type have Employee\_Code\_ID as 908. Among remaining customers, 278 salaried individuals have code ID 908 and only 1 self-employed customer has code ID 908. This suggests that majority of Salaries employees have employee Code id 908. Therefore, it would be safe to mention 'Salaried' as the 'Employee\_Type' for missing values with 'Employee\_code\_ID' of 908.
3. Most customers have Disbursed amount and asset cost of 100,000 or less while these observations for others is much greater.
4. Most borrowers are young while as the age increases the number of borrowers decreases.
5. Most customers have very few primary or secondary accounts.
6. Most customers have average account age and credit history length of less than 5 years.
7. While 182543 are not defaulters, remaining 50611 are defaulters and need to be identified using appropriate model.
8. Customers who take survice from certain vehicle manufacturers and suppliers have higher tendency to default compared to other manfucturers and suppliers
9. Customers with higher loan to value of asset ratio have higher tendency to default.
10. Customers with higher disbursed amount have higher tendency to default.
11. Though age distribution of defaulters and non-defaulters is similar, the boxplots suggest that the age distrbution of defaulters tends to be marginally lower compared to no-defaulters.
12. While all the customers presented their mobile number, most of them presented their Aadhar card for proof (84.03%) and some others (14.49%) presented their Voter ID. However, very few customers present their Driving License, PAN card or passport.

**Observations based on Week 2 analysis:**

1. Boxplots show that non-defaulter have higher average bureau score compared to defaulters.
2. Further, those with bureau rating description indicating high-risk profile of customer showcase more defaulters while those indicating low risk profile showcase fewer defaulters.
3. If we ignore few extreme outliers, non-defaulters tend to have more number of primary accounts with higher balance and are sanctioned and disbursed with greater amount compared to defaulters.
4. Most customers (both defaulter and non-defaulter) do not have any secondary account. In case of few outliers, non-defaulters tend to have more number of secondary accounts with greater account balance
5. Though there is no difference between sanctioned and disbused amount for primary and secondary accounts of most customers, significant difference is observed for few customers.
6. Most customers do not make any inquiry. Among outliers, few defaulters and non-defaulter make inquires. However, only few non-defaulters in the data analyzed make more than 19 inquires. None of the defaulters makes more than 19 inquiries.
7. There is not significant correlation of new loans in last six month, loans defaulted in last 6 months and time since first loan and propability of defaulting`
8. However, this negligible correlation is observed because most customers have zero or very few new loan in 6 months and have zero loan defaults in 6 months.
9. If we only observe outliers separately, it is noticed that many customers who have new loans in last six months are non-defaulters. Similarly, many defaulters have loans defaulted in last 6 months.
10. The logistic Regression model developed predicts the correct result 78.28 % times
11. On observing the coefficients of the Logistic regression model, it is found that whether the customer is defaulter majorly depends upon disbursed\_amount, PRI\_DISBURSED\_AMOUNT, Bureau score description etc. among other features like branch id, state id, manufacturing id etc.
12. Customers with higher disbursed\_amount tend to have higher chances of being a defaulter.
13. Customers with higher PRI\_DISBURSED\_AMOUNT tend to have higher chances of being a defaulter.
14. Customers with Very Low risk bureau Score have lower chances of being defaulters while customers Higher risk bureau score have greater chances of defaulting.
15. Based on the confusion matrix, it is observed that the model is able to identify non-defaulters more accurately but is not able to find sufficient number of defaulters. This implies that if the model predicts a defaulter, it can be suggested to not give loan to such candidate. However, if the model does not predict a defaulter, it is difficult to indicate whether the candidate will default or not.

Observations based on Tableau analysis

Tableau Link:

<https://public.tableau.com/app/profile/raunaq.nayar/viz/loan_defaulter_analysis/BankLoanDefaulterDataAnalysis?publish=yes>

Chart, bar chart, funnel chart

Description automatically generated

Chart

Description automatically generated