

Questions & Answers:

1. **How can we make sure that our model can detect real defaulters and there are less false positives? This is important as we can lose out on an opportunity to finance more individuals and earn interest on it.**

Ans: To keep very less False Positives, oversampling techniques like SMOTE should be used in model creation. As we can see from the data, the percentage of defaulters is slightly higher than Banking industry

2. **Since NPA (non-performing asset) is a real problem in this industry, it's important we play safe and shouldn't disburse loans to anyone.**

Ans: Yes. LoanTap should not disburse loans to everyone. Company's internal policy and analysis should be in place to identify the correct persons. From data provided, 20% of people default on their loan, which in turn become NPAs for the company.

Low False positive means we should create the model with high Precision values. This can be achieved if we are keeping high threshold value in logistic Regression model.

But keeping too high values for threshold will increase False Negatives. This in turn may result in opportunity loss. In this case we will not give loans to persons which will not default but our model has predicted that they will default.

3. **What percentage of customers have fully paid their Loan Amount?**

Ans: Around 80% of customers have fully paid their Loan Amount

4. **Comment about the correlation between Loan Amount and Installment features**

Ans: There is very high co-relation between loan amount and installment. Pearson co-efficient is 0.95.

5. **The majority of people have home ownership as ____.**

Ans: Majority of the people (50%) have home ownership as mortgage followed by rent 40%.

6. **People with grades 'A' are more likely to fully pay their loan. (T/F)**

Ans: True. Out of all people with grade 'A', 93% got their loan approved

7. **Name the top 2 afforded job titles.**

Ans: Teacher and manager are two affordable job titles

8. Thinking from a bank's perspective, which metric should our primary focus be on...

- 1. ROC AUC**
- 2. Precision**
- 3. Recall**
- 4. F1 Score**

Ans: The best metric to consider is **F1 score** as we need to give importance to both precision and recall. We don't want to miss potential customers and at the same time we also don't want to give loan to defaulters.

9. How does the gap in precision and recall affect the bank.

Ans: Recall score: 0.82 and Precision score: 0.57. which tells us that there are more false positives than the false negatives. If Recall value is low, it means Bank is losing in opportunity cost. If Precision value is low, it means Bank's NPA (defaulters) may increase.

10. Which were the features that heavily affected the outcome?

Ans: Employee title, Pin code, Sub grade, open_acc and revol_util are the important features that heavily affected the outcome.

11. Will the results be affected by geographical location? (Yes/No)

Ans: Yes. Pin code (derived from address) has significant impact on the outcome

Insights:

- 76% data is of term 36 months
- Maximum 29% data is of loan grade B, followed by C with 27% data
- In the home ownership, 50% of data is of mortgage followed by rent 40%
- 35% data is of the verified income by LoanTap, 33% data of verified income source and rest 32% is not verified data
- 80% data is of fully paid loans and 20 % data of charged off loans
- 60% data is of F(full) as initial list status
- The dataset has almost all the data of individual application type
- 89% data with zero number of public record bankruptcies
- Loan Amount distribution / media is slightly higher for Charged off loan Status.
- Probability of Charged off status is higher in case of 60 months term.
- Interest Rate mean and media is higher for Charged off Loan Status.
- Probability of Charged off Loan Status is higher for Loan Grades are E ,F, G.
- G grade has the highest probability of having defaulter.
- Similar pattern is visible in sub grades probability plot.
- Employment Length has overall same probability of Loan status as fully paid and defaulter.
- That means Defaulters has no relation with their Employment length.
- For those borrowers who have rental home, has higher probability of defaulters.
- Borrowers having their home mortgage and owns have lower probability of defaulter.
- Annual income median is lightly higher for those who's loan status is as fully paid.
- Somehow, verified income borrowers probability of defaulter is higher than those who are not verified by loan tap.
- Most of the borrowers take loans for dept-consolidation and credit card payoffs.
- The probability of defaulters is higher in the small business owner borrowers.
- Debt-to-income ratio is higher for defaulters.
- Number of open credit lines in the borrowers' credit file is same as for loan status as fully paid and defaulters.
- Number of derogatory public records increases , the probability of borrowers declared as defaulters also increases. Specially for those who have higher than 12 public records.
- Total credit revolving balance is almost same for both borrowers who had fully paid loan and declared defaulter
- Revolving line utilization rate is higher for defaulter borrowers.
- Application type Direct-Pay has higher probability of defaulter borrowers than individual and joint.
- Number of public record bankruptcies increases, higher the probability of defaulters.
- Most important features/ data for prediction, as per Logistic Regression: Employee title, Pin code, Sub grade, open_acc and revol_util.

Recommendations:

- We should try to keep the precision higher as possible compare to recall, and keep the false positive low that will help not to miss out the opportunity to finance more individuals and earn interest on it. This we can achieve by setting up the higher threshold.
- Giving loans to those even having slightly higher probability of defaulter, we can maximise the earning , by this risk taking method
- Since NPA is a real problem in the industry, Company should more investigate and check for the proof of assets. Since it was observed in probability plot, verified borrowers had higher probability of defaulters than non-verified.
- Giving loans to those who have no mortgage house of any owned property have higher probability of defaulter, giving loan to this category borrowers can be a problem of NPA
- Around 80% of customers have fully paid their loan amount. The defaulters are 20%. From personal loan business perspective this ratio is high. These 20% will contribute in NPAs of LoanTap. To reduce the risk of NPAs,
 1. LoanTap should add slightly stringent rules to bring down this ratio to 5% to 6%.
 2. LoanTap should provide loans at slightly higher rate than other banks. This will offset the risks of defaulters and maintain the profitability of the business.