

<https://github.com/AnureetKaurTiwana/Logistic Regression Loan Tap/blob/main/Logistic Regression Model Loan Tap.pdf>

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.

=> To keep very less False Positives, oversampling techniques like SMOTE should be used in model creation. Also, we can use more complex algorithms like SVM, Decision-Trees, Random Forest and try various hyperparameter tuning.

=> As you 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.

=> Yes. Loan Tap 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?

=> Around 80.26% of customers have fully paid their Loan Amount.

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

=> The spearman correlation coefficient between `Loan Amount` and `Installment` is very high (i.e. 0.97).

=> This indicates high multicollinearity between these two features. This in turn resulted in high VIF value (i.e. 11.91).

=> Because of this HIGH VIF value, column `Loan Amount` was dropped in creating model.

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

=> Majority of people have `home_ownership = MORTGAGE`. This is about 50%.

=> Next highest is `RENT` (~40%)

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

=> True.

=> Out of all people with grade 'A', 93% got their loan approved.

7. Name the top 2 afforded job titles.

=> Teacher & Manager

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

=> 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?

=> Recall score: 0.99 and Precision score: 0.88. Which tells us that there are more false positives than the false negatives.

=> From Confusion Matrix it can be seen that FP = 10% of total cases & FN = 0.9% of Total Cases

=> If Recall value is low (i.e. FN are high), it means Bank is loosing in opportunity cost.

=> If Precision value is low (i.e. FP are high), it means Bank's NPA (defaulters) may increase.

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

=> address(pincodes), sub_grade, loan_amnt, home_ownership_rent

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

=> Yes.

=> pincodes (derived from address) has significant impact on the outcome.

Actionable insights and recommendations

- Around **80.26%** 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,
 - LoanTap should add slightly stringent rules to bring down this ratio to 5% to 6%.
 - 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.
- Overall Statistics of the Model:
 - **Accuracy = 89%**
 - **Precision = 88**
 - **Recall = 99**
 - **F1 -score = 93%**
- Model created has high values for accuracy, precision, recall & f1-score. This means, this model is a good classifier. Overall, it has good prediction capability in identifying right customers (which can be easily converted).
 - However this model has slightly low capability on correctly identifying defaulters. Overall data has 20% defaulters, model is able to predict 10% of them correctly.
 - Using this model, LoanTap can easily reduce the ration of defaulters in their portfolio.
- **Features which have significant impact on outcome are as follow:**
 1. int_rate: Interest Rate
 2. sub_grade: loan subgrade
 3. term : number of payments on the loan
 4. home_ownership
 5. purpose
 6. application_type
 7. pincode (from address)
 8. emp_title: job title supplied by the Borrower
- **sub_grade** and **grade** logic to classify persons by LoanTap is well created. From the model it is considered to be significant.
- **pincode** is significant feature.
 - Pincodes with Negative Coefficient: 93700, 11650, 86630, 48052, 30723, 70466, 22690
 - Pincodes with Positive Coefficient: 05113, 29597
 - LoanTap can increase their market presence in Pincodes with Positive Coefficient.
 - LoanTap should minimize their marketing/sales expenditure in Pincodes with Negative Coefficient.

- Pincode based market segmentation should be included at strategic levels.
- **emp_title** Owner & Driver has negative Coefficient. emp_title Project Manager has positive Coefficient. LoanTap can also decide their social media based marketing based on person's job-titles.
- application_type JOINT has positive Coefficient. Which means LoanTap can promote persons to apply for joint loan. Because of this, chances of default will reduce.
- Purpose Renewable energy has negative Coefficient. This means LoanTap should stick to giving loans to conventional purposes like Marriage, car etc.
- term 60 months has negative Coefficient. Which means LoanTap should focus more on Loans for shorter duration (i.e. 36 months). Their social media campaign and marketing strategy should be based on this consideration.