Questionnaire:

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

Ans : **80.38%** of customers have fully paid up their loan

2. Comment about the correlation between Loan Amount and Instalment features.

Ans : They are **close to perfect positive correlation** with value close 1 i.e 0.9

3. The majority of people have home ownership as **Mortgage**.

4. People with grades 'A' are more likely to fully pay their loan.

Ans : **True**

5. Name the top 2 afforded job titles.

Ans : **Teacher and Manager**

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

on.

- ROC AUC

- Precision

- Recall

- F1 Score

Ans : **Recall:** It measures the ability of the model to correctly identify all actual defaulters. High recall ensures that the bank catches as many risky borrowers as possible, thereby minimizing the number of approved loans that may default.

7. How does the gap in precision and recall affect the bank?

Ans :

Financial Losses:

False Positives (Low Precision): Approving loans to individuals who later default can result in financial losses for the bank. These non-performing assets (NPAs) not only reduce profitability but also tie up capital that could have been invested elsewhere.

False Negatives (Low Recall): Rejecting creditworthy applicants due to overly conservative risk assessment can lead to missed revenue opportunities. The bank loses out on potential interest income and customer relationships.

Reputation Damage:

False Positives: Approving loans to individuals who subsequently default can damage the bank's reputation. It may erode trust among customers and investors, affecting brand perception and market credibility.

False Negatives: Rejecting creditworthy applicants unfairly can lead to dissatisfaction among customers. Negative word-of-mouth, social media backlash, and complaints to regulatory authorities can tarnish the bank's reputation.

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

Ans : **Zip Code followed by Application Type and Verification status**

9. Will the results be affected by geographical location?

Ans : **Yes**, Zip code as part of geographical location highly affected the results

Insights:

- Approx. 80% of the loans are of 36 months duration
- Maximum Loans are from B grade followed by C,D,A
- Maximum Home Ownersip belong to MORTGAGE followed by RENT and OWN
- Fully Paid loans are almost 80% of the target variable loan_status
- Almost 90% of the applicants do not have derogatory Public Records
- Initial Listing Status of the loan is more in f category than w
- Almost 99% of the application types are individual
- Most of the applicants have got Mortage Account
- Almost 90% of the applicant have no Public Record Bankrupcies
- Almost 55% of the loans are taken against debt_consolidation followed by Credict card
- 2013 and 2014 were the years with maximum loans funding
- Percent share of default is much higher for long duration loans i.e 60 months
- Defaulters are highest for grade f and g and then decrease with grade. Sub-grade showing similar pattern
- Home Ownership: Charged-off % is high for None category followed by Rent, Own and Mortgage

- Surprisingly 100% defaulters observed for Zip codes 11650, 86630 and 93700. And zip codes with no defaulters at all are 00813, 05113, 29597
- pub_rec, pub_rec_bankruptcies, init_list_status and state have no impact
- In application type, Direct pay has maximum defaulters followed by individual and joint
- Applicants with mort_acc 0 have higher charged off % than ones with mort_acc category 1
- Applicants with small_business have high default rate followed by renewable energy and others
- No significant impact of employment length on loan repayments
- 2007 is the year with maximum percent of defaulters followed by 2015 and 2014
- Mean loan_amount,int_rate, dti, open_acc, revol_util is slightly higher for charged off
- Mean annual_inc is lower for charged off than fully paid
- The test score is slightly lower than both the training and validation scores but still close, indicating that the model generalizes reasonably well to unseen data.
- Precision for class 0: 0.92 means that out of all instances predicted as class 0, 92% of them were actually class 0.
- Precision for class 1: 0.58 means that out of all instances predicted as class 1, only 58% of them were actually class 1.
- Recall for class 0: 0.88 means that the model correctly identified 88% of all actual class 0 instances.
- Recall for class 1: 0.68 means that the model correctly identified 68% of all actual class 1 instances.
- F1-score for class 0: 0.90 is the harmonic mean of precision and recall for class 0.
- F1-score for class 1: 0.62 is the harmonic mean of precision and recall for class 1.
- Zip codes- 11650, 93700, 86630 signify strong positive relationship with the Loan Status
- Whereas zip codes 29597,00813,05113 show strong negative relatioship with target variable
- It shows that features such as emp_length, total_acc, revol_bal, annual_inc, int_rate, issue_d_year show no contribution at all.
- ROC Curve (AUC = 0.90) is observed
- Interpretation of PR AUC: A PR AUC value of 0.76 indicates the area under the precision-recall curve. It represents the integral of precision-recall pairs across all possible decision thresholds.

Recommendations:

Risk Mitigation:

- **Segment-Based Strategy**: Focus on higher grades (A, B, C) for initial rollouts while continuously monitoring performance. As the model proves effective, gradually extend to lower grades (D, E, F, G) with cautious parameters.
- **Loan Caps and Conditional Approvals**: Implement loan caps for high-risk segments and conditional approvals where additional guarantees or higher interest rates are applied.
- **Geographical Risk Assessment**: Given the strong relationship between certain zip codes and default rates, incorporate geographical risk factors into the model. Focus on high-risk zip codes with stricter criteria.

Enhancing Loan Approval Process:

- **Verification Process**: Strengthen the verification process for critical features like income, employment status, and home ownership to reduce misinformation.
- **Real-Time Monitoring**: Implement real-time credit monitoring for borrowers to identify early signs of financial distress and intervene before defaults occur.

Feedback Loop

1. **Continuous Monitoring**:

Performance Metrics: Continuously track key performance metrics such as precision, recall, F1-score, and AUC-ROC to evaluate model effectiveness.

Regular Audits: Conduct periodic audits of approved and denied loans to assess the model's decisions against actual outcomes.

2. **Iterative Improvements**:

Model Retraining: Regularly retrain the model with new data to capture changes in borrower behavior and economic conditions.

User Feedback: Incorporate feedback from loan officers and customers to identify areas of improvement in the model and process.

3. **Dynamic Risk Adjustments**:

Economic Indicators: Monitor macroeconomic indicators such as unemployment rates and economic growth to adjust lending criteria dynamically.

Anomaly Detection: Use anomaly detection techniques to identify and investigate unusual patterns in loan applications and repayments.