Q/A

1. **Why was the LendingClub dataset chosen for this analysis?**

The LendingClub dataset was selected because it provides a large, publicly available collection of real-world loan data containing diverse borrower characteristics. It offers sufficient depth for meaningful feature analysis and includes a clearly defined target variable—loan default—making it ideal for building and testing predictive credit risk models.

1. **How did you address the significant class imbalance in the dataset?**

The dataset exhibited an imbalance, with only about 11.6% of loans defaulting. To correct this, the Synthetic Minority Oversampling Technique (**SMOTE**) was applied to the training data. SMOTE generates synthetic samples of the minority (default) class, ensuring the model learns the underlying patterns of default cases rather than being biased toward non-default predictions.

1. **Why did you choose to compare Logistic Regression and Random Forest specifically?**

Logistic Regression was chosen for its interpretability and widespread use in credit risk modeling, as it provides clear insights into feature impact through coefficients. Random Forest was selected as a non-linear ensemble method that can capture complex relationships and interactions between features. Comparing the two allowed for a balance between transparency and predictive power.

1. **Why is Logistic Regression considered better despite its lower accuracy?**

Although Random Forest achieved higher accuracy, Logistic Regression produced much higher recall (70%), meaning it correctly identified most actual defaulters. In credit risk analysis, recall is more important than accuracy because missing a defaulter (false negative) poses a much higher financial cost than incorrectly flagging a non-defaulter (false positive).

1. **What do the feature importance results tell us about the primary drivers of loan default?**

Feature importance analysis revealed that **Age, Interest Rate, Months Employed,** and **Income** are the strongest predictors of default. Younger borrowers, higher interest rates, shorter employment durations, and lower income levels were all associated with increased default risk. These factors reflect financial stability and repayment capacity as critical elements of creditworthiness.

1. **How could this model be integrated into a real-world loan approval process?**

The model could be incorporated into a bank’s existing credit assessment workflow as a **decision-support tool**. It can flag high-risk applicants for further manual review, provide probability scores for each loan, and help adjust lending policies or interest rates based on predicted risk levels. Final decisions should still involve human analysts to ensure fairness and compliance.

1. **What are the business implications of the False Positives generated by the Logistic Regression model?**

False Positives represent customers incorrectly classified as potential defaulters. While this may temporarily limit credit access for some low-risk borrowers, it also reduces the likelihood of financial loss from undetected defaulters. From a business standpoint, this trade-off is acceptable because the cost of reviewing a few extra applicants is far lower than the cost of loan defaults.

1. **What ethical risks are associated with using models like this for lending decisions?**

Key ethical risks include potential bias against specific demographic or socioeconomic groups, lack of transparency in automated decision-making, and misuse of personal financial data. To mitigate these risks, models must be regularly audited for bias, use explainable algorithms, and operate under strict data privacy and regulatory compliance standards.

1. **How would you improve the model’s performance in future iterations?**

Future improvements could include tuning the probability threshold to balance precision and recall, testing advanced algorithms such as **XGBoost** or **LightGBM**, and incorporating macroeconomic indicators like interest rate trends or unemployment data. Feature engineering and dimensionality reduction techniques could also enhance model robustness and generalization.

1. **What steps are necessary to validate this model before deployment in a financial institution?**

Before deployment, the model should undergo rigorous **validation and compliance checks**, including:

* Out-of-sample testing on new, unseen data.
* Cross-validation using different time frames to ensure stability.
* Model bias and fairness audits.
* Review by credit risk experts and legal teams.
* Integration testing to ensure compatibility with existing systems and decision workflows.