1. Explain the pros and cons about your model, including limitation (can be both quantitative and qualitative).

Decision Tree

Pros:

- The decision tree provides a clear visualization of decision-making processes, making it highly interpretable and deployable even for non-experts.
- It shows the flow of decisions through branches and provides clear rules for classifying observations.
- They can capture non-linear relationships and complex interactions between features, which logistic regression may miss.

Cons:

- In the given context and dataset, the decision tree had a lower accuracy than the logistic regression model
- Decision trees are prone to overfitting, especially when the tree grows too deep. A fully grown tree can memorize the training data, leading to poor generalization on unseen data. This overfitting of the model could be the reason why the decision tree performed more poorly than the logistic regression model. To mitigate this, pruning is often required for decision trees. Pruning techniques are applied to reduce the size of the tree. Without pruning, decision trees tend to grow too deep, increasing complexity and the risk of overfitting. However, finding the right balance between pruning and accuracy can be tricky, further increasing the complexity of implementing a decision tree model.
- Decision trees are sensitive to small changes in the data. A small variation in the data can lead to an entirely different tree, making them less stable. This instability can affect reliability. However, an ensemble method like Random Forest can mitigate this issue.

Logistic Regression

Pros

- Had a higher accuracy than the Decision Tree model.
- Logistic regression models provide a clear and interpretable model where the relationship between the input variables and the output can be easily understood through the coefficients. This is especially useful in the field of finance, where interpretability is crucial. The coefficients tell you how much each feature influences the probability of an outcome.
- Logistic regression provides probabilistic outputs, allowing you to interpret the likelihood of a particular class rather than just a binary outcome like the decision tree.
- Logistic regression is also less prone to overfitting as compared to the decision tree model.

Cons

- Logistic regression assumes that there is a linear relationship between the independent variables and the log-odds of the outcome. If the data is highly non-linear, logistic regression might underperform. This may cause the model to fail to capture complex relationships in the data.
- Logistic regression is sensitive to multicollinearity (high correlation among independent variables), which can cause issues with the model's coefficients. This is often the case with financial data and may have been the case with the training data provided to train the model. (Eg. An unemployed individual may have a low annual salary and therefore as a result a low bank balance)

2. How to overcome the weakness of your model (future study).

Decision Tree

- **Overfitting:** Use pruning techniques like cost-complexity pruning to reduce the size of the tree by removing branches that have little importance. Another alternative is limiting the maximum depth of the tree to prevent it from becoming too complex. In this assignment, I used a max depth of 7 for the decision tree to improve the accuracy of the model and reduce the issue of overfitting.
- **Instability:** Use ensemble models such as Random Forest to reduce the variance in the decision tree and improve stability. Alternatively, bootstrap aggregating can be used to create multiple versions of the tree and average their predictions, reducing the effect of small data variations.

Logistic Regression

- **Addressing Linearity Assumption:** Introduce polynomial features or interaction terms to capture non-linear relationships.
- **Remove Highly Correlated Features**: Use techniques like Pearson Correlation Coefficient or correlation analysis to detect and drop highly correlated features.

3. Any descriptive analysis you could think of for this case. Example, confidence interval.

- **Correlation Analysis:** Compute the Pearson Correlation Coefficient matrix to identify relationships between the “Bank Balance” and “Annual Salary” variables. This matrix can then be visualized using heatmaps to easily spot strong correlations and potential multicollinearity issues which can be especially helpful for the Logistic Regression model.

4. What historical data variables are considered most influential in predicting loan defaults, and how are they weighted in your analysis?

Decision Tree

- For the Decision Tree model, “Bank Balance” was the most important feature, followed by “Annual Salary” and lastly, “Employed”. Their coefficients are listed in the table below. The importance value for each feature from the decision tree model is measured using the Gini impurity and it indicates how much that feature contributes to the decision-making process of the tree. Higher values mean that the feature is more important in making predictions.

Feature	Coefficient
Employed	0.832346
Bank Balance	0.151999
Annual Salary	0.015655

Logistic Regression

- For the Logistic Regression model, “Employed” was the most significant variable, followed by “Bank Balance and lastly, “Annual Salary”. Their coefficients are listed in the table below. The coefficient value for each feature from the logistic regression model indicates the strength and direction of the relationship between the feature and the target. The model estimates these weights during training to minimize the logistic loss function. A positive coefficient means that the likelihood of defaulting on a loan increases. A negative coefficient means that as the feature value increases, the likelihood of defaulting on a loan decreases.

Feature	Coefficient	Absolute Coefficient
Employed	5.198965e-01	5.198965e-01
Bank Balance	4.684461e-04	4.684461e-04
Annual Salary	3.885207e-07	3.885207e-07

5. How do economic indicators and market trends impact the accuracy of your loan default predictions, and what strategies are in place to adapt to changing conditions?

Economic indicators like unemployment rates, inflation, and interest rates significantly impact loan default predictions by affecting borrowers' financial stability. Market trends, such as economic downturns, can increase default risk, reducing the model’s accuracy. To adapt, the models should be regularly retrained on updated data, and should include macroeconomic variables to improve reliability.

6. How do you balance the need to mitigate default risk with the goal of providing access to credit for underserved or high-risk borrowers? Are there any ethical considerations in this decision-making process?

Balancing default risk with credit access for underserved or high-risk borrowers requires a careful approach. Mitigating risk involves using data-driven models to assess borrowers' ability to repay while minimizing discrimination. For underserved populations, alternative data (e.g., rent or utility payments) can improve credit assessments beyond traditional metrics, increasing access and financial equitability.

Ethical considerations include avoiding bias in models that might unfairly deny credit to certain groups. This can be done by ensuring the training data is unbiased, to prevent racism or discrimination from being “learned” and implemented in these models. Transparency in decision-making, fairness in model design, and ensuring that credit policies do not disproportionately impact vulnerable populations are critical. It’s essential to balance financial prudence with inclusivity, providing responsible lending options while managing risk.

7. The prescriptive analysis on loan default aims to enhance decision-making, reduce default risk, and optimize lending practices while maintaining a balance between profitability and risk mitigation. How do you think accurately predicting loan default can help in any decision making? (The importance of your model to the bank.)

Accurately predicting loan default is crucial for enhancing decision-making in lending. It enables banks to assess the creditworthiness of borrowers more effectively, minimizing losses from defaults. By identifying high-risk applicants early, banks can adjust lending terms, such as interest rates or collateral requirements, or deny credit where necessary. This risk-based approach optimizes lending practices, ensuring profitability while maintaining responsible lending.

Moreover, accurate predictions help allocate resources efficiently, improving customer segmentation and tailoring products to individual risk profiles. Ultimately, a reliable model reduces non-performing loans, boosts overall financial stability, and strengthens the bank’s long-term profitability.