Report on Loan Application Status

Approach Taken:

1. Data Preprocessing:

- Missing Data Handling: Missing values in the datasets were filled with '-1' to avoid issues with 'NaN' values during model training and predictions.
- Removed variables that do not affect our target variable as they may add noise and increase our computation time.
- Label Encoding: Categorical variables in both the training and test datasets were encoded using label encoding to convert them into numeric form.
- Feature Alignment: Ensured that the test data had the same features as the training data by aligning the columns before making predictions.

2. Modeling:

RandomForestClassifier: A decision-tree-based model was initially used due to its ability to handle a mix of categorical and numeric data and provide feature importance rankings.

XGBoost(XGBClassifier):To improve performance, the model was later upgraded to XGBoost, a gradient-boosting algorithm that often provides better accuracy due to its boosting mechanism.

3. Model Evaluation:

- Both models were trained using an 80-20 train-test split, and the accuracy was measured on the validation dataset.
- Feature importance rankings were extracted to understand which features impacted the predictions most.While the features in both cases got varied however the common features were considered.

Insights and Conclusions from Data:

Feature Importance: The most important features for predicting loan status were related to(common in both models):

- EMPLOY CONSTITUTION
- TOTAL ASSET COST

- CIBIL SCORE
- PAN NAME
- AGE
- MARITAL STATUS
- APPLIED AMOUNT

These features were consistently ranked as the most significant in both the RandomForest and XGBoost models, indicating their strong predictive power.

Model Performance Comparison:

- RandomForest Accuracy: 88.35%
- XGBoost Accuracy:88.45%

Performance on Train Data Set:

Accuracy:

The accuracy for both models on the train-test split was very close, with RandomForest achieving 88.35% and XGBoost slightly improving to 88.45%.

This indicates that both models performed well and generalized effectively on the validation set, with minimal overfitting.

Metrics Used:

Accuracy Score: The primary metric used to assess model performance. Both RandomForest and XGBoost delivered high accuracy, confirming that the models were effective in predicting loan status.

Conclusions:

- The small improvement in accuracy suggests that both RandomForest and XGBoost are well-suited for this dataset.
- The top features identified by both models align with real-world factors affecting loan approval decisions, such as credit score and income.
- Further performance improvements may require more advanced feature engineering or model tuning, but the current accuracy is strong at 88.45%.

Final Thoughts:

While XGBoost showed a slight improvement over RandomForest in terms of accuracy, the difference is negligible, and both models are equally effective for the given task. Therefore, the focus should shift from model selection to enhancing the data and refining the process.

For production use, either model can be selected based on practical considerations such as training time, interpretability, or ease of deployment, as performance is comparable.