### **Report: Loan Approval Model Training and Evaluation**

Kyle LaRavia and Bryan Ardon

#### **Preprocessing Steps and Rationale**

The preprocessing steps taken in this project include:

1. **Loading the Dataset:** The dataset was loaded using pandas for further exploration and analysis.
2. **Handling Missing Values:** Missing values were identified and addressed to ensure no null entries interfered with the model's performance.
3. **Column Standardization:** Whitespace was removed from column names to ensure consistent referencing during analysis.
4. **Label Encoding:**
   * The loan\_status column is processed into binary values of 1 for "Approved" and 0 for "Rejected."
   * Similarly, the education column was converted into binary values for categorical representation.
5. **Feature Scaling and Splitting:**
   * The data was then divided into training and testing sets to estimate model performance on unseen data.

#### **Main insights derived from the EDA / Data Analysis**

1. **Dataset Overview:**
   * The dataset contained relevant features, which included income, education level, credit score, and whether the individual was employed or not.
   * There were some missing values in some columns and were dealt with accordingly.
2. **Feature Engineering:**
   * Certain columns were transformed or engineered to have meaningful information.
   * Correlation Analysis indicated certain high predictors of loan status with the main 2 being credit score and income.
3. **Data Imbalance:**
   * The loan status class distribution seemed to have a data imbalance; it needs careful evaluation of model metrics apart from accuracy.

#### **Evaluation Results and Comparison of Models**

The project compared several models:

1. **Decision Tree Classifier:**
   * Pros: Interpretability and ease of understanding.
   * Performance Metrics: Moderate accuracy but prone to overfitting.
2. **Random Forest Classifier:**
   * Pros: Enhanced generalization through ensemble learning.
   * Performance Metrics: Higher accuracy and precision compared to the Decision Tree.
3. **K-Nearest Neighbors (KNN):**
   * Pros: Simple and intuitive.
   * Performance Metrics: Lower recall and longer computation time for large datasets.
4. **Logistic Regression:**
   * Pros: Suitable for binary classification problems.
   * Performance Metrics: The same accuracy as the Random Forest, but low recall.

The overall evaluation metrics are represented in terms of accuracy, precision, recall, F1 score, and ROC-AUC. The Random Forest model did the best out of the four models and gave a balance for accuracy versus recall, so it's the best to deploy.

#### **Business Insights and Recommendations**

1. **Insights:**
   * Credit score and income level are the two most influential predictors of loan approval.
   * Simple models like Logistic Regression may suffice for quick, albeit less nuanced, decision models.
   * Random Forest yields robust predictions with minimized false negatives, which is extremely critical in loan approval so as not to reject any eligible applicant.
2. **Recommendations:**
   * Use the Random Forest model in the loan approval process for higher accuracy.
   * Feature importance insights can be integrated into business decision-making processes for strategic customer targeting.
   * Retrain the model periodically with updated data to ensure continued accuracy and relevance.

#### **New Knowledge Gained**

1. **Practical Application of Machine Learning:** During this project, the importance of preprocessing, especially handling imbalanced data and feature engineering, was once more underlined.
2. **Model Evaluation:** We were further able to develop our understanding of the trade-offs between different models and metrics.
3. **Insights into Business Context:** More than the mere technical implementation, I learned how to translate the output of models into actionable business strategies.