### **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:** The dataset provided had no missing values but we made sure of this by calling the isNull() function on our DataFrame 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." This allows the models to distinguish between approved and rejected because models understand numbers not text.
   * Similarly, the education and self\_employed columns were converted into binary values for categorical representation.
5. **Preparing the Dataset:**
   * The data was then divided into 80% training and 20% testing datasets to estimate model performance overall.

#### **Key findings from the data exploration/analysis**

1. **Dataset Overview:**
   * The dataset contained relevant features for determining loan approval such as income, education level, credit score, and whether the individual was employed or not.
2. **Feature Engineering:**
   * Certain columns were transformed or engineered to have meaningful information for the model to understand.
   * Correlation Analysis indicated certain high predictors of loan status with the main two being credit score and income. We were able to find these impactful features by calling model.feature\_importances\_ which gave us the importance of each feature then we just sorted to get the top eight features.
     1. cibil\_score 0.827544
     2. loan\_term 0.079127
     3. loan\_amount 0.038479
     4. income\_annum 0.030363
     5. residential\_assets\_value 0.005846
     6. luxury\_assets\_value 0.005174
     7. loan\_id 0.004642
     8. commercial\_assets\_value 0.003687
3. **Data Imbalance:**
   * The loan status class distribution seemed to have a data imbalance issue. The percentage of loan approvals in dataset was about 62% which could lead the models to more likely predict approvals compared to rejections. Each model needs careful evaluation of model metrics apart from accuracy to ensure no overfitting occurs.

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

The project compared several models:

1. **Decision Tree Classifier:**
   * Pros: Interpretability and ease of understanding given that branches in the tree lead to direct answers.
   * Performance Metrics: Great accuracy but generally prone to overfitting.
   * Results:
     1. Accuracy: 0.9742
     2. Precision: 0.9813
     3. Recall: 0.9776
     4. F1-Score: 0.9794
     5. AUC: 0.9731

1. **Random Forest Classifier:**
   * Pros: Enhanced generalization through ensemble learning.
   * Performance Metrics: Higher across all metrics compared to the Decision Tree.
   * Results:
     1. Accuracy: 0.9778
     2. Precision: 0.9814
     3. Recall: 0.9832
     4. F1-Score: 0.9823
     5. AUC: 0.9986
2. **K-Nearest Neighbors (KNN):**
   * Pros: Simple implementation and can play around with hyper parameter n\_neighbors.
   * Performance Metrics: Lower metric performance overall and generally longer computation time for large datasets.
   * Results:
     1. Accuracy: 0.6019
     2. Precision: 0.6260
     3. Recall: 0.9086
     4. F1-Score: 0.7412
     5. AUC: 0.5146
3. **Logistic Regression:**
   * Pros: Suitable for binary classification problems.
   * Performance Metrics: Surprisingly not performing as well as we initially thought. Could be due the fact that credit score is the most impactful feature leading to the data being highly separated meaning a single feature almost independently predicts the outcome.
   * Results:
     1. Accuracy: 0.6370
     2. Precision: 0.6423
     3. Recall: 0.9515
     4. F1-Score: 0.7669
     5. AUC: 0.9986

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 so it’s the best for deployment.

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

1. **Insights:**
   * Credit score and loan term are the two most influential predictors of loan approval.
   * 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 highest performance overall.
   * Feature importance insights can be integrated into business decision-making processes for strategic customer targeting such as sending loan ads to people with high credit scores.
   * For optimal results, 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. Also learning how to export a model to be used on a real user interface was great!
2. **Model Evaluation:** We were further able to develop our understanding of the trade-offs between different models and metrics. We learned about the model.feature\_importance\_ attribute which directly gives you the weight of each feature according to the model.
3. **Insights into Business Context:** More than the mere technical implementation, I learned how to translate the output of models into actionable business strategies which is a key skill few people obtain.