Credit Scoring Model Explainability Report

1. Modeling Approach

The model development process consists of the following steps:

- **Train Multiple Shallow Decision Trees:** The process begins by training several shallow decision trees. Shallow trees are intentionally used to make the individual rules easier to understand. The report specifies a depth of 3 for these trees.
- **Extract Decision Rules and Convert to Binary Features:** The decision rules are extracted from the trained decision trees. These rules are then converted into binary features. For example, a rule might be "If Age > 30 and Income < 50,000, then ...". This rule would be transformed into a binary feature that is 1 if the condition is true, and 0 if it is false.
- **Fit L1-Penalized Logistic Regression:** A logistic regression model is then trained using the binary features created from the decision rules. L1 regularization is applied during the training process. L1 regularization helps to select the most influential rules by shrinking the coefficients of less important rules towards zero. This results in a more sparse and interpretable model.
- **Evaluate Predictions:** Finally, the model's performance is evaluated using appropriate metrics. The report mentions "metrics like accuracy and AUC," indicating that the model's ability to correctly classify credit risk and discriminate between different risk levels is assessed.

2. Explainability Techniques

The report uses the following explainability techniques to interpret the model:

- **SHAP (SHapley Additive exPlanations):** SHAP is a technique used to quantify the contribution of each feature to the model's prediction. It provides both global and local explanations:
  - **Global Explanations:** SHAP can be used to understand the overall importance of each feature in the model. This helps in identifying which rules are most influential in determining credit risk.
  - **Local Explanations:** SHAP can also explain individual predictions. For a specific borrower, SHAP values show how each rule contributed to the model's assessment of that borrower's credit risk.
- **LIME (Local Interpretable Model-agnostic Explanations):** LIME is another technique for explaining individual predictions. It works by approximating the model's behavior around a specific data point with a simpler, interpretable model (in this case a linear model). LIME helps to understand why the model made a particular prediction for a given borrower.

- **Partial Dependence Plots (PDP):** PDPs visualize the relationship between a feature and the model's predictions. Specifically, a PDP shows how the model's prediction changes as the value of a particular feature changes, while holding other features constant. This helps to understand how individual rules affect the predicted credit risk.