**Summary for Case\_study1\_model\_development**

**Objective:** The objective was to construct a classification model to determine customer load eligibility based on provided information.

**Feature Study:**

1. The majority of customers are male and have a higher loan disbursement rate.
2. The number of married customers is twice that of unmarried customers.
3. Most loan applicants do not have any dependents. However, customers with two dependents exhibit the highest disbursement rate. This could be because individuals starting families often require loans for homes or vehicles.
4. A large portion of customers are graduates, correlating with a higher disbursement rate.
5. Although self-employed customers are a minority, they demonstrate a higher disbursement rate.
6. The majority of customers have an available credit history.
7. Customer distribution across rural, urban, and semi-urban areas is nearly equal, yet semi-urban customers exhibit the highest loan disbursement rate.
8. Applicant income and co-applicant income were skewed. Log transformation was applied to rectify the distribution.

**Feature Engineering:**

Two new features were created:

1. Total Income: The combination of applicant and co-applicant income.
2. Loan Income Ratio: Loan amount divided by total income.

**Model Training:**

Three data samples were created:

1. Includes all variables along with log-transformed variables for applicant and co-applicant income, as well as total income.
2. Includes all variables except log-transformed ones.
3. Excludes original columns for applicant and co-applicant income and total income, using log-transformed variables instead.

**Model Evaluation:**

Confusion matrix, classification summary, learning curve, and ROC AUC curve were assessed to determine accuracy and generalization of each model.

**Insights and Findings:**

Although all three models perform similarly in terms of classification accuracy, the model with normal variables (without log transformation) exhibits a slightly higher AUC value of 84%. This suggests the model has good discriminatory power and effectively distinguishes between classes. Thus, we select the model trained on normal variables without log transformation.

While credit history has the highest importance in feature importance, other variables also hold uniform importance.

Logistic regression was also attempted. However, the model trained on log-transformed variables shows similar scores to the selected model above, but with a lower AUC value compared to the random forest model.