Two-Wheeler Loan Application Evaluation Report

1. Introduction

The objective of this project is to predict whether a two-wheeler loan application will be accepted or rejected based on the applicant's data. Using a labeled dataset containing historical loan application outcomes, I trained machine learning models to make predictions on new, unlabeled data.

---

2. Approach Taken

i. Data Preparation:

- **Data Loading:** The datasets Assignment_Train.csv and Assignment_Test.csv were loaded.

- **Data Cleaning:** Unnecessary columns were dropped, and missing values were handled using imputation strategies:

    - **Numerical Columns:** Imputed with median values.

    - **Categorical Columns:** Imputed with the most frequent value.

- **Feature Encoding:** Categorical features were one-hot encoded to prepare the data for machine learning algorithms.

- **Feature Scaling:** The features were standardized using StandardScaler.

ii. Exploratory Data Analysis (EDA):

- **Approval Rates by Branch State:** Analyzed the approval rates for each branch state, which revealed variability in approval rates across different states.

- **Distribution of Application Status:** Analyzed the distribution of loan approvals and declines, highlighting a higher number of approved applications.

- **Age Distribution:** Visualized the age distribution for approved versus declined loans.

- **Marital Status Analysis:** Examined the relationship between marital status and loan approval status.

iii. Model Training and Evaluation:

- **Model Selection:** Used RandomForestClassifier for its robustness and ability to handle complex data.

- **Model Training:** The model was trained using 80% of the data, and performance was evaluated on a validation set (20% of the data).

- **Model Evaluation:** Performance metrics were calculated using accuracy and classification report metrics.

iv. Prediction and Submission:

- **Prediction on Test Data:** The trained model was used to predict loan approvals on the test dataset.

- **Submission Preparation:** Created a submission file containing predictions for the test dataset.

---

3. Insights and Conclusions

- **Approval Rates by Branch State:** Approval rates vary significantly by branch state, with some states like Himachal Pradesh showing higher approval rates compared to states like West Bengal.

- **Application Status Distribution:** There were more approved applications (6677) compared to declined applications (3323), indicating a generally favorable approval rate.

- **Age Distribution:** The age distribution analysis showed variations in approval rates across different age groups.

- **Marital Status:** Marital status showed a strong correlation with loan approval, with married and single individuals both having high approval rates.

---

4. Performance on Train Data Set

- **Validation Accuracy:** 0.849 (84.9%)

- **Classification Report Metrics:**

  - **Precision:**

    - **Declined:** 0.96

    - **Approved:** 0.71

  - **Recall:**

    - **Declined:** 0.81

    - **Approved:** 0.93

  - **F1-Score:**

    - **Declined:** 0.88

    - **Approved:** 0.81

---

5. Use of Appropriate Metrics

- **Accuracy:** Provides a measure of overall correctness of the model.

- **Precision:** Indicates the correctness of positive predictions, important for understanding false positives.

- **Recall:** Reflects the model's ability to identify all relevant cases (approved/declined).

- **F1-Score:** Balances precision and recall, providing a single metric for model performance evaluation.

These metrics indicate that the model performs well in both predicting approved and declined applications.