Loan Risk Classification Model Report

1. Approach Taken

Objective:
The goal of this project was to develop a predictive model to classify loan applications as either APPROVED or DECLINED. By leveraging applicant-related features, the model aims to automate and improve the loan approval process.

Data Preparation:

- **Data Loading:**
  The training data (`Assignment_Train.csv`) and test data (`Assignment_Test.csv`) were loaded into the environment.
- **Feature Engineering:**
  - **Date Conversion:**
    The 'APPLICATION LOGIN DATE' column was converted into datetime format. I extracted the year and month (as `LOGIN_YEAR` and `LOGIN_MONTH`) to capture time-related trends.
  - **Cibil Score Handling:**
    The 'Cibil Score' was converted to numeric. Non-numeric values were coerced to NaN, and missing values were imputed.
- **Feature Selection:**
  Columns with excessive missing values or irrelevant information (like names and premium-related columns) were removed to focus on key predictors.
- **Handling Missing Values:**
  - **Numeric Features:**
    Median imputation was used to fill missing values, preserving central tendencies.
  - **Categorical Features:**
    Constant imputation with a placeholder value was applied to maintain consistency across the dataset.
- **Encoding Categorical Features:**
  One-Hot Encoding was applied to categorical variables, expanding the feature space to allow the model to effectively handle categorical data.

Model Selection:

I chose a **Random Forest Classifier** for its robustness and ability to handle complex feature interactions. This ensemble method aggregates predictions from multiple decision trees to improve accuracy.

Pipeline Creation:
A preprocessing pipeline was constructed to handle data transformations (imputation, encoding)

and integrate them with the Random Forest Classifier, ensuring seamless processing during both training and testing phases.

---

2. Insights and Conclusions from Data

Insights:

- **Feature Importance:**
  The Random Forest model revealed that `Phone Social Premium`, `Cibil Score`, and the time-based features (`LOGIN_YEAR`, `LOGIN_MONTH`) were strong predictors of loan approval status.
- **Data Quality and Impact:**
  Handling missing values and encoding categorical data were critical. Ensuring clean and consistent data improved the model’s overall performance.
- **Feature Engineering:**
  The engineered date-related features and normalized Cibil Score provided valuable predictive power, enhancing the model's interpretability and performance.

Conclusions:

The preprocessing steps, including feature extraction and careful handling of missing values, significantly contributed to the model's effectiveness. The Random Forest Classifier, combined with well-processed data, provided strong predictions, making it a useful tool for automating loan approval decisions.

---

3. Performance on Train Data Set

Training Metrics:

- **Accuracy:**
  86.2% — The model correctly classified 86.2% of instances in the training dataset, reflecting solid overall performance.

Classification Report:

- **Precision:**
  - **APPROVED:** 0.94 — Of the applications predicted as APPROVED, 94% were correct, showing the model’s strength in minimizing false positives.
  - **DECLINED:** 0.75 — 75% of predicted DECLINED applications were correct, indicating reasonable performance.
- **Recall:**

    - **APPROVED:** 0.85 — The model identified 85% of actual approved applications, showing good coverage in this class.
    - **DECLINED:** 0.89 — The model caught 89% of actual declined applications, demonstrating high recall for declined loans.
- **F1-Score:**
    - **APPROVED:** 0.89 — A balanced score reflecting strong performance in the APPROVED class.
    - **DECLINED:** 0.81 — A good performance measure for declined applications, although slightly lower than approved loans.
- **Macro Average:**
    - **Precision:** 0.84
    - **Recall:** 0.87
    - **F1-Score:** 0.85
      These scores reflect the average performance across both classes.
- **Weighted Average:**
    - **Precision:** 0.87
    - **Recall:** 0.86
    - **F1-Score:** 0.86
      These averages consider class imbalance, offering a comprehensive measure of overall model performance.

Interpretation:
The model performs well across both classes, particularly excelling in precision for APPROVED loans and recall for DECLINED loans. The metrics suggest a good balance between catching approved loans and identifying declined ones.

---

4. Use of Appropriate Metrics

Metrics Utilized:

- **Accuracy:**
  Provides a broad measure of performance, though it can be less informative with imbalanced datasets.
- **Precision:**
  Measures the accuracy of positive predictions. The high precision for APPROVED loans indicates that the model is very good at avoiding false approvals.
- **Recall:**
  Focuses on capturing all relevant instances. The high recall for DECLINED loans shows that the model is effective at identifying most rejected applications.
- **F1-Score:**
  Combines precision and recall, offering a balanced view of model performance across both classes.

Conclusion:

The chosen metrics provide a well-rounded evaluation of the model's ability to classify loans. High precision for approved loans ensures that the model is reliable in granting loans, while strong recall for declined loans indicates that it effectively filters risky applicants.