ML Project Report

**Credit Score Classification**

**Objective:**

* Build models that predict credit scores (categorized as "Good", "Standard", or "Poor") using historical credit data.
* Enhance credit risk assessment methodologies, allowing lenders to make more informed decisions.

**2. Dataset:**

* **Train Data**: 100,000 rows with various financial and personal features, such as:
  + Age, Occupation, Annual Income, Monthly Salary, Credit Card count, Loan details, and Payment Behavior.
* **Test Data**: 50,000 rows for evaluating the model performance.

**3. Data Preprocessing:**

* **Missing Values**: Addressed by imputing means for numerical columns and encoding categorical variables (e.g., month and occupation).
* **Feature Engineering**: Transforming categorical features (e.g., occupation, payment behavior) into numerical values and scaling numerical features (e.g., income, debt).
* **Handling Outliers**: Dealing with outliers like unrealistic ages (e.g., '-500') and cleaning unnecessary values.

**4. Modeling:**

* Two machine learning models were trained:
  + **Extreme Gradient Boosting (XGBClassifier)**: Achieved an accuracy of around **70.6%**.
  + **Light Gradient Boosting Machine (LGBMClassifier)**: Performed better, with an accuracy of **72.9%** and a lower log loss (0.59 vs. 0.64 for XGB).

**5. Evaluation:**

* **Confusion Matrix**: Analyzed model performance in predicting "Good", "Standard", and "Poor" scores.
* **ROC Curves**: Evaluated how well the models distinguish between different classes.
* **Accuracy & Log Loss**: LightGBM showed superior performance, making it the best-suited model for this task.

**6. Deployment:**

* The trained LGBM model is applied to the test dataset to predict credit scores.
* The predicted scores are added to the test data and saved for further analysis.

**7. Visualization:**

* Count plots and correlation heatmaps were used to visualize relationships between different features and credit scores, helping to better understand the dataset