**REPORT**

**1. Dataset Description and Preprocessing Steps**

* **Dataset:** Lending Club Loan Dataset (subset with relevant features)
* **Target Variable:**
  + not.fully.paid — indicates whether the loan was fully paid (0) or defaulted (1).
* **Features Used:**
  + Credit policy approval, purpose of the loan, interest rate, installment amount, log of annual income, debt-to-income ratio (dti), FICO credit score, days with credit line, revolving balance and utilization, number of inquiries in last 6 months, delinquencies, public records.
* **Preprocessing:**
  + Categorical feature purpose was one-hot encoded to convert it into numerical form.
  + Missing values were handled by filling with median values of respective features.
  + Features were scaled using StandardScaler to normalize the data.
  + Class imbalance was addressed using SMOTE (Synthetic Minority Over-sampling Technique) to balance the number of default and non-default cases.

**2. Model Implemented and Rationale**

* **Model:** Support Vector Machine (SVM) with RBF kernel
* SVM is effective for binary classification tasks and works well with high-dimensional data.
* The RBF kernel helps capture non-linear relationships in the data, useful for financial features with complex interactions.
* Probability estimates enabled to calculate ROC AUC and plot ROC curves.

**3. Key Insights and Visualizations**

* **Evaluation Metrics:**
  + Precision, Recall, and F1 Score were used to evaluate classification performance considering the imbalanced nature of the dataset.
  + ROC AUC score was calculated to evaluate the model's ability to discriminate between defaulters and non-defaulters.
* **Results:**
  + The SVM model showed balanced precision and recall, indicating good performance in identifying both defaulters and non-defaulters.
  + ROC curve visualized the trade-off between true positive rate and false positive rate, with an AUC score indicating strong model discrimination capability.

*(Insert ROC curve plot here if available)*

**4. Challenges Faced and Solutions**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | | **Challenge** | | --- | | Class imbalance in dataset | | Handling categorical data | | Feature scaling | | Missing values | | | **Solution** | | --- | | Applied SMOTE to synthetically balance classes | | Used one-hot encoding for the purpose column | | Applied StandardScaler for normalization | | Filled missing values using median imputation | | |

**5. Recommendations for Lenders**

* The model can assist lenders in identifying high-risk applicants before loan approval.
* Incorporating this model in the decision process can reduce loan defaults and financial losses.
* Continuous model retraining with fresh data is recommended to maintain accuracy over time.