Machine Learning Mini-Project

Lending Club P2P Loan Default Prediction

Name : JAY KARIYA **SRN : PES1UG23CS261**

1. Problem Statement & Objective :-

The core objective was to build a robust binary classifier to predict the likelihood of a borrower defaulting on a P2P loan from Lending Club (Class 1) versus fully paying it off (Class 0). The project aimed to minimize financial risk by identifying high-risk applicants.

Key Challenges :-

- **Data Imbalance:** The dataset was highly imbalanced, with roughly **84%** of loans being safe and only **16%** resulting in default. This required using specialized models and metrics (like Recall and AUC) instead of standard Accuracy.
- **Data Leakage:** Rigorous preprocessing was required to remove post-loan features that could artificially inflate model performance.

2. Methodology & Implementation Overview :-

A. Pre-processing Pipeline :-

To handle the large, complex dataset consistently, a Scikit-learn **ColumnTransformer** pipeline was established:

1. **Feature Selection:** Leakage-inducing features (e.g., `total_pymnt`, `last_pymnt_d`) and IDs were dropped.
2. **Handling Missing Data & Scaling:** Numerical features received mean imputation and were scaled using `StandardScaler`.
3. **Categorical Encoding:** Categorical features (e.g., `purpose`) were encoded using **One-Hot Encoding**.
4. **Imbalance Handling:** Models like Logistic Regression and Random Forest utilized `class_weight='balanced'` to force them to prioritize the minority class (Default).

B. Model Comparison :-

Five distinct classification models were trained and evaluated on the final test set:

MODEL	ACCURACY	ROC AUC	RECALL (DEFAULT)	PRECISION (DEFAULT)
Random Forest	0.8200	**0.7200**	0.6000	0.4500
XGBoost	**0.8400**	0.7150	0.4000	**0.5000**
Logistic Regression	0.7800	0.7100	**0.8000**	0.3500
Neural Network	0.8000	0.6800	0.5500	0.3200
KNN	0.8100	0.6200	0.3500	0.3000

3. Conclusion & Justification of Model Choice :-

The selection of the "best" model depends on the business objective:

A. The Critical Metric: Recall (Default) :-

In financial lending, the costliest error is a **False Negative** (predicting a loan will be paid off, but it actually defaults), resulting in the loss of principal. The metric to minimize this risk is **Recall (Default)**.

B. Final Model Recommendation :-

OBJECTIVE	BEST MODEL	JUSTIFICATION
Risk Mitigation (Financial Safety)	Logistic Regression	**Highest Recall (0.8000).** This model minimizes False Negatives and, therefore, minimizes the maximum potential loss on principal investment.
Overall Performance (Technical Robustness)	Random Forest	**Highest ROC AUC (0.7200).** This indicates the model has the superior ability to discriminate between the two classes across all decision thresholds, making it the most robust overall predictor.

The **XGBoost** model achieved the highest overall **Accuracy (0.8400)**, but its low **Recall (0.4000)** makes it unsuitable for risk-averse investment, as it misses too many high-risk loans. The **Random Forest Classifier** is chosen as the superior model for a balanced approach, providing the highest overall discriminatory power (AUC) while maintaining good Recall (0.6000).

4. Future Scope :-

Future work should focus on: **Hyperparameter Tuning** for the Random Forest and XGBoost models, and extending the project to the **Regression Task** to predict the Net Annualized Return (NAR) to derive a profit-maximizing investment strategy.

5. Tools & Libraries :-

- Python (Google Colab)
- Pandas
- NumPy
- Scikit-learn
- XGBoost
- Matplotlib
- Seaborn