Data Science Canvas

Project: Credit Default Risk

Team: Harshit, Sebin, Shreya, Tushar

Problem Statement

Business Case & Value Added

- Millions lack formal credit histories, leading to financial exclusion and unsafe lending.
- **Value:** Improved decision-making for financial institutions, identification of key risk factors.

Data Landscape

Required Data: Features like telco and transactional information.

Available Data: Application Data, Internal Credit History, External Credit History

Additional Data to be collected:

Model Selection

- **Logistic Regression** (baseline)
- **Random Forest** (robustness, non-linear)
- **LightGBM** (speed, efficiency, accuracy for large datasets)
- **XGBoost**
- **Ensemble Model**: LightGBM and XGBoost

Model Requirements

- High ROC AUC, Accuracy, Precision, Recall (F1-Score).
- Low Log Loss.
- Model Interpretability

Software & Libraries

- **Languages/Tools:** Python, Jupyter.
- **Libraries:** pandas,numpy, scikit-learn,lightgbm,matplotlib, seaborn, plotly,xgboo

Skills

- Data Acquisition & Cleaning
- Feature Engineering
- Machine Learning Modeling
- Model Tuning & Validation
- Data Visualization

Execution & Evaluation

Model Evaluation

- Measure success using ROC AUC, Accuracy, Precision, Recall, F1-Score, and Log Loss.
- Compare performance against baseline models.

Data Storytelling

The target group = **risk analysts + credit officers** who need:

- **Clear probability of default (PD)** for each applicant — not just a label.
- **Explainability** — why the model thinks someone may default (key features).
- **Reliability** — stable, validated predictions (AUC, calibration).
- **Actionability** — how the scores can support loan approval / rejection decisions.

Results can be communicated via **AUC metrics, key feature drivers , and simple visuals** that show how the model improves loan decisions.

Data Collection & Preparation

Data Selection & Cleansing

Relevant data: All applicant-level and history tables (application data, bureau data, previous loans, credit card balances, POS cash, installment payments).

Cleanup: the data must be cleaned for missing values, outliers, inconsistent categories, merging relational tables, and handling class imbalance before modeling.

Data Integration

Migrate all sources into a **centralized data warehouse or unified relational database (e.g., PostgreSQL/BigQuery)** so the data stays consistent, joinable, and ready for modeling.

Data Collection

Collect extra data through credit bureau checks, income verification, and transaction history, ensuring it is **accurate, up-to-date, consistent, and legally compliant** for reliable risk prediction.

Explorative Data Analysis

The EDA shows strong skewness and clear outliers in key financial fields (income, credit, annuity, days features), plus many missing-value structures that must be handled.

Descriptive statistics such as mean, median, percentiles, missing-value rates, and distribution plots reveal that defaulted clients typically have lower income, higher credit ratios, and more late payments.