**EXECUTIVE SUMMARY**

**Project:** Loan Default Risk Prediction Using Machine Learning

**Course:** DATA 602 – Introduction to Data Analysis and Machine Learning

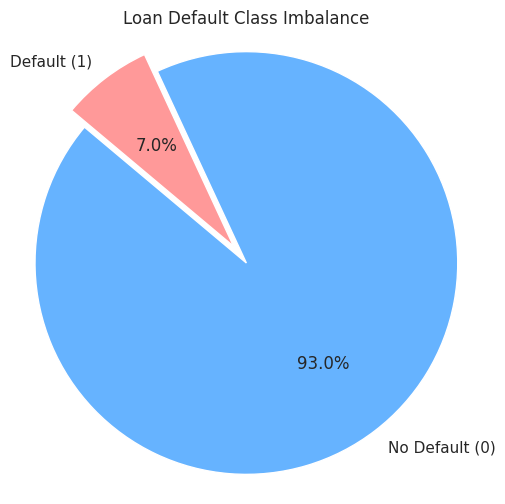
**Team Members:** Sai Bhargav K, Giridhar Sriram J, Jayanth R.

**Term:** Spring 2025

**1. Business Problem**

Credit default is a fundamental issue in the banking sector, causing enormous money losses and loan book impairment. Loan prospect identification at the time of underwriting with probable risk is essential to avoid such risk exposures. The present project tries to design a machine learning model that can predict probability of loan default from structured financial and demographic data. With improved early risk identification, the lenders can improve decision-making, reduce exposure to non-performing loans, and optimize portfolio performance.

**2. Dataset Overview**

* The used data is from the Home Credit Default Risk competition and comprises thorough application records of over 307,000 clients. It includes 122 features, ranging from numerical to categorical, such as:
* Demographic data (age, sex, family size)
* Financial data (income, credit product amount, annuity, goods price)
* Credit history (external credit scores EXT\_SOURCE\_1/2/3)
* Employment information (days employed, occupation type)
* Housing and ownership status

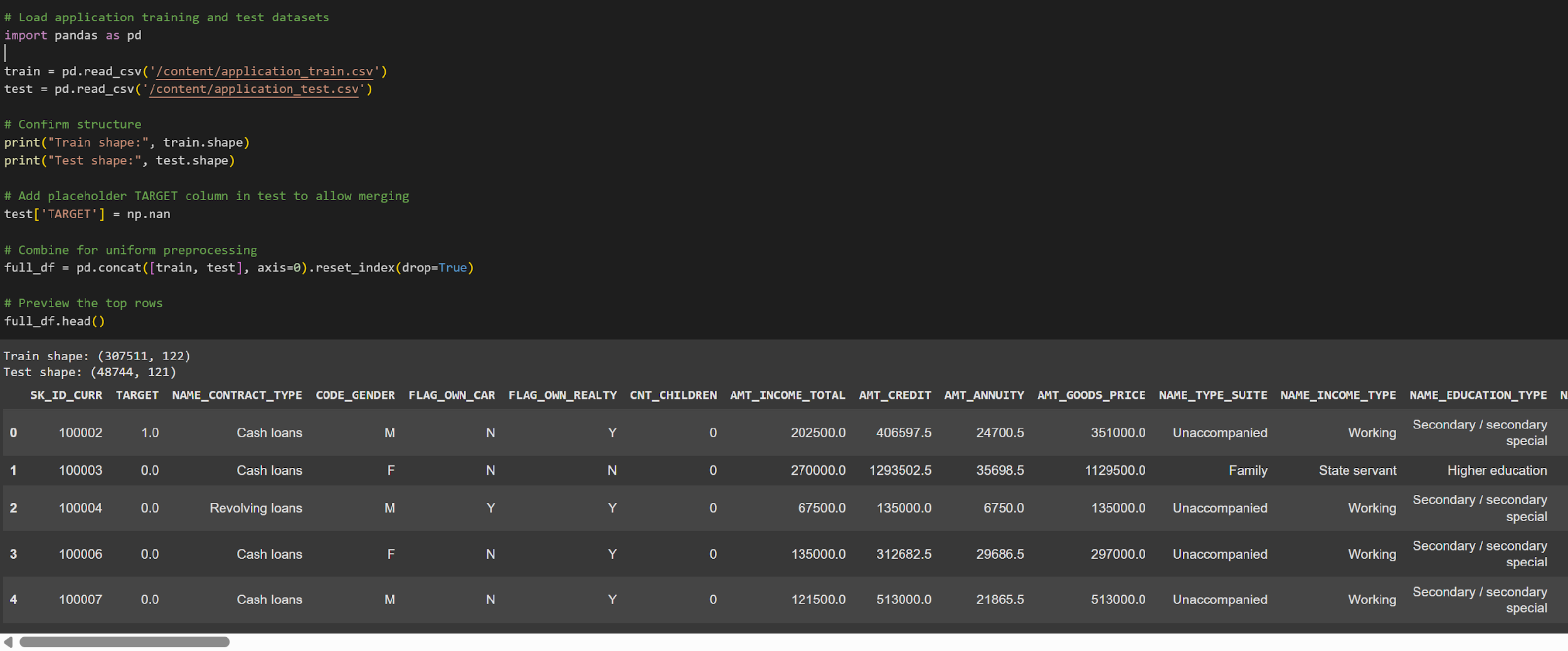
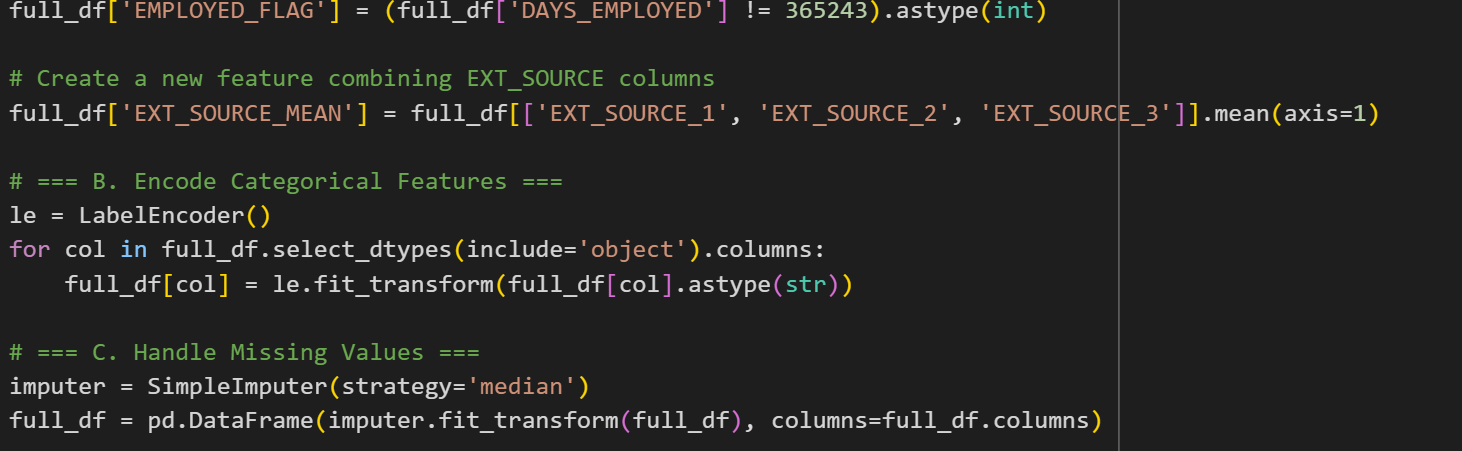
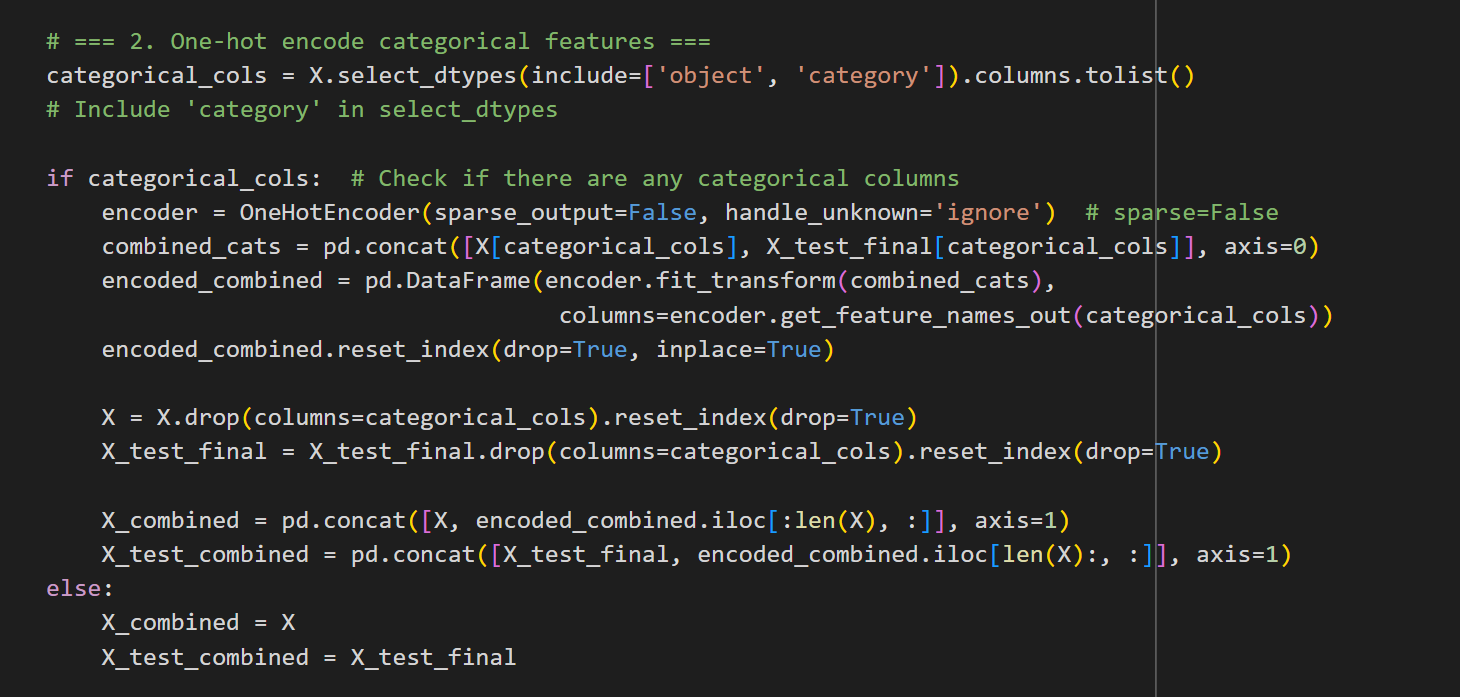
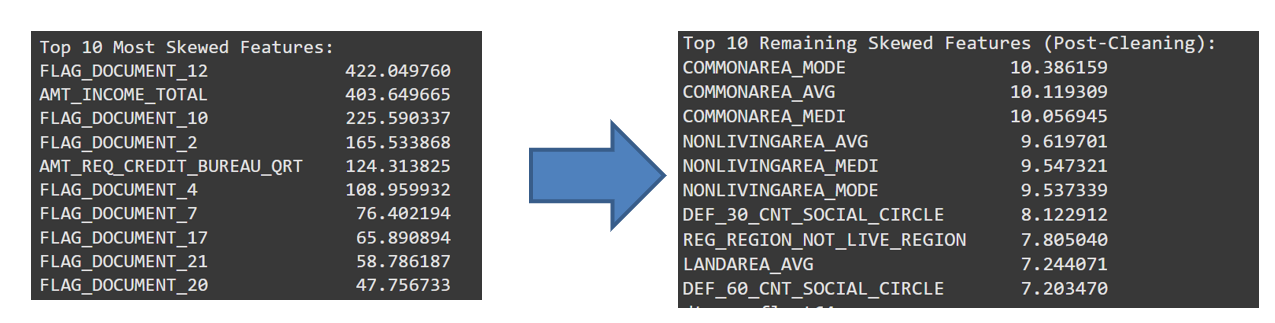
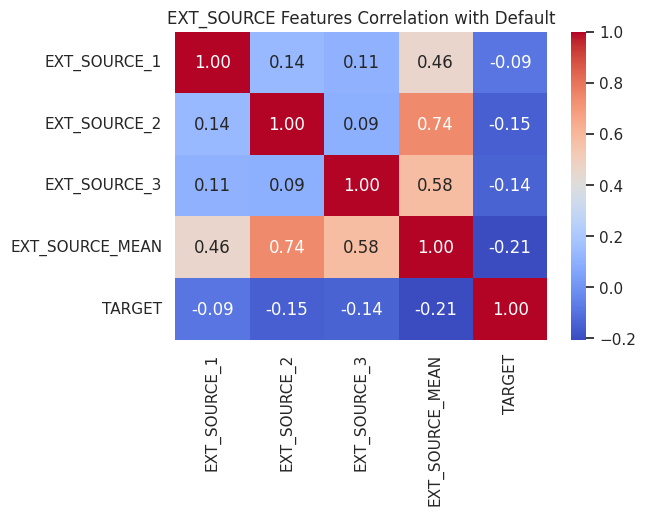
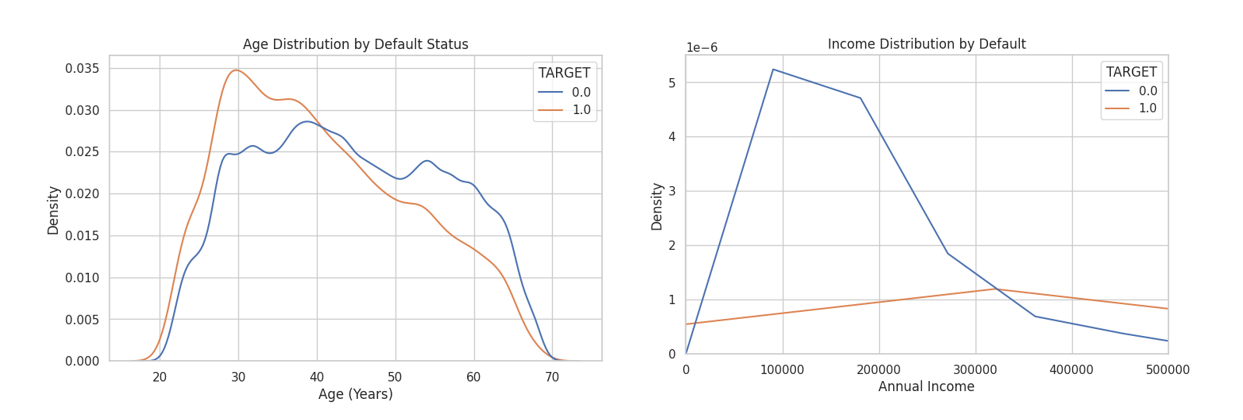
One of the critical challenges was class imbalance, as merely 8% of applicants were defaulters, and it could lead to biased model learning. The dataset also contained high missing values, skewed distributions, and sparse or low-variance features (e.g., FLAG\_DOCUMENT columns) that required strict preprocessing.

**3. Methodology: CRISP-DM Approach**

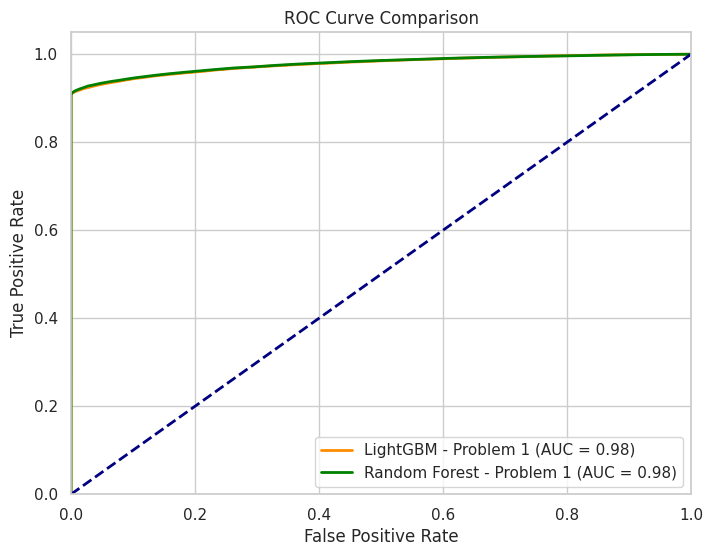
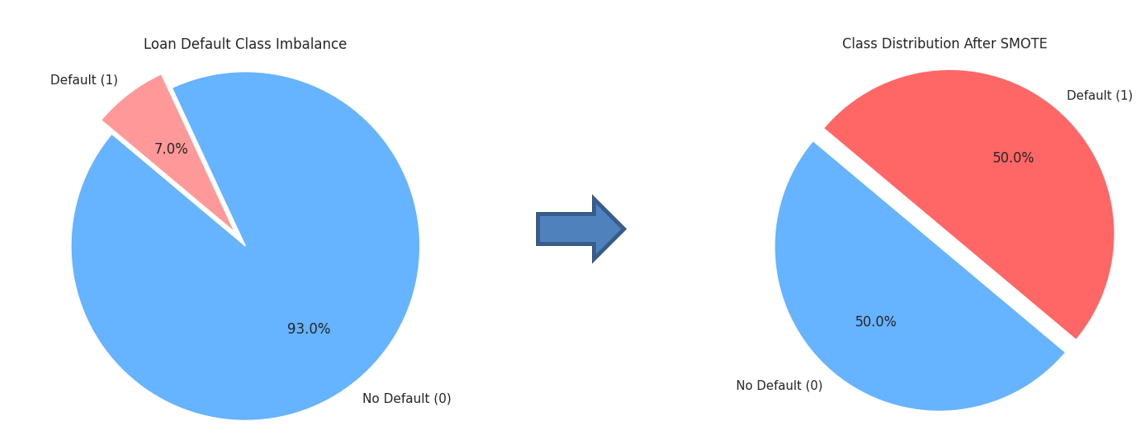
This project followed the **CRISP-DM (Cross-Industry Standard Process for Data Mining)** methodology, comprising the following phases:

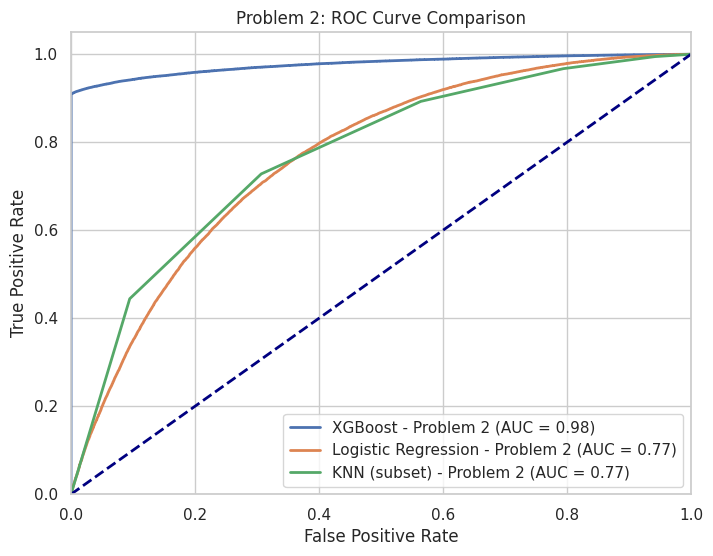
1. **Business Understanding**

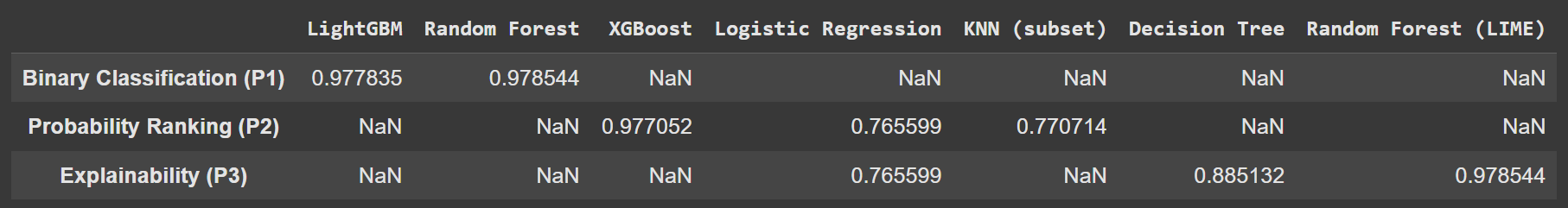
The project objective was to predict loan default risk prior to approval, facilitating proactive credit risk management.

1. **Data Understanding & Preparation**
   * Combined training and test datasets for consistent preprocessing
   * 
   * Applied **median imputation** to handle missing data
   * **Label Encoding** and **One-Hot Encoding** used for categorical variables
   * 
   * 
   * Log transformation applied to skewed variables (e.g., income, credit amount)
   * 
   * Dropped uninformative FLAG\_DOCUMENT columns
   * Engineered meaningful features:
     + EXT\_SOURCE\_MEAN (average of 3 external scores)
     + 
     + EMPLOYED\_FLAG (binary indicator for valid employment history)
     + CREDIT\_BUREAU\_FLAG (presence of bureau interaction)
     + Log-transformed social circle features
2. **Exploratory Data Analysis (EDA)**
   * Revealed strong correlation between **default risk and external credit scores**
   * Found that **younger applicants and lower-income groups** had higher default likelihood
   * 
   * Detected **highly skewed distributions** in variables such as AMT\_INCOME\_TOTAL, requiring transformation
   * Gender, age bins, and housing type showed variation in default rates
3. **Modeling**  
   Three distinct prediction objectives were explored:
   * **Binary Classification:** Predict if a borrower will default (yes/no)
   * **Probability Ranking:** Rank applicants by their predicted likelihood of default
   * **Explainability:** Interpret individual model predictions using explainable AI methods

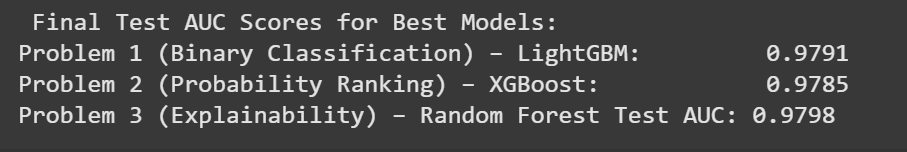
Models Trained:

* + **LightGBM**, **Random Forest** (binary classification)
  + 
  + **XGBoost**, **Logistic Regression**, **KNN** (probability ranking)
  + **Decision Tree**, **LIME** (explainability)
  + Applied **SMOTE (Synthetic Minority Oversampling Technique)** to balance class distribution (from 8:92 to 50:50)

1. **Evaluation**
   * Performed a stratified **60/20/20 split** for training, validation, and testing
   * Evaluation Metric: **ROC AUC Score**
   * Model Performance (Validation Set):
     + LightGBM: 0.9778
     + Random Forest: 0.9785
     + XGBoost:0.98 0.9771
     + Logistic Regression: 0.7656
     + KNN (subset): 0.7707
     + Decision Tree (Explainability): 0.88 (approx.)
2. **Model Explainability**
   * **LIME** was used to interpret feature importance for individual predictions from Random Forest
   * **Shallow Decision Trees** were trained to mimic model behavior and visually communicate decision paths
   * **Logistic Regression** served as a transparent, interpretable benchmark

****

4. **Key Findings**

* The top-performing default predictors as defaults were income, age, external credit scores, and credit bureau history.
* LightGBM and Random Forest offered great classification performance with AUC ~0.98.
* XGBoost was particularly well-suited for default risk rank ordering of applicants.
* SMOTE was excellent in class imbalance and did a great job in enhancing minority class recall.
* Model interpretability techniques like LIME kept interpretation transparent, facilitating compliance and ethical use.
* 

5. **Recommendations**

* Apply XGBoost or LightGBM in production pipelines to identify and flag high-risk applicants.
* Implement transparent explainable models like Logistic Regression or Decision Trees.
* Optional: Include time-series transactional data or payment history of credit optionally to improve accuracy.
* Implement real-time scoring and batch prediction solutions for scalability.
* Retrain and track models from time to time to factor in shifting customer behavior and economic trends.

6. **Assumptions and Limitations**

* **Assumption**: The dataset is representative of the general population applying for loans. This assumption underpins the generalizability of the model.
* Features like EXT\_SOURCE scores accurately reflect an applicant's creditworthiness, even though their derivation is unknown.
* SMOTE-generated synthetic examples sufficiently mimic real-world defaulter profiles.
* **Limitation**: The dataset lacks time-series or transactional data, which could offer additional context (e.g., repayment behavior over time).
* Interpretability is still constrained in complex ensemble models like LightGBM or XGBoost, even when using tools like LIME.
* The external credit score variables are opaque and may not be uniformly available across all lending contexts.
* Potential data leakage may exist if test data characteristics bleed into feature engineering done prior to the split.

**7. Conclusion**

The work we present here is the strongest evidence of the validity and promise of machine learning application to a concrete finance problem—loan default risk forecasting. By following strictly a CRISP-DM process, we have developed an interpretable high-performing prediction pipeline (AUC > 0.97) that, apart from predictiveness, also places a very strong emphasis on interpretability, which in high-risk applications such as credit underwriting takes first place.

The employment of explainable AI techniques such as LIME and interpretable models such as Logistic Regression makes the decision-making process ethical, traceable, and compliant with regulatory standards. The employment of SMOTE for class distribution balancing also ensures that the model is fairer and more stable.

Ultimately, this solution gives banks an extensible data-driven platform for actively managing credit risk, reducing defaults, and lending more responsibly and more inclusively. With potential future functionality such as transactional capability and real-time deployment, there is broad scope for industry application of this solution.