**LDRPM: Loan Default Risk Prediction Model Project**

|  |  |  |  |
| --- | --- | --- | --- |
| ***Member1*** | Chelsea Liu(ql2547) | ***Course Name*** | Big Data |
| ***Member2*** | Gesang Zeren(zg2442) | ***Section*** | D |
| ***Member3*** | Yuewei Shi(ys5795) | ***Semester Term*** | Spring 2024 |

## **Project Abstract**

The Loan Default Risk Prediction Model project addresses a critical challenge in consumer finance: accurately predicting loan default risks, particularly for clients lacking a substantial credit history. The primary objective is to develop predictive models that not only accurately assess the likelihood of default but also maintain their predictive power over time, ensuring long-term stability. This focus on stability is crucial, as traditional credit risk models, or scorecards, must be frequently updated to account for changing client behaviors, a process that is both time-consuming and resource-intensive. The absence of traditional data points such as credit history, often due to factors like youth or a cash preference, complicates this assessment, leading to a significant number of potential borrowers being unjustly denied. The project presents a solution to this problem by potentially offering more accurate predictions of a person's repayment capabilities, thus making loans more accessible to those who could benefit most.

## **Problem Statement**

In the consumer finance industry, accurately predicting the risk of loan default, particularly for individuals with little to no credit history, remains a significant challenge. Traditional credit scoring models, while effective to some extent, often fail to account for the dynamic nature of consumer behavior over time, leading to outdated risk assessments and potentially increasing the risk of default. The absence of a robust credit history further complicates the risk assessment process, resulting in a high rejection rate of potentially creditworthy clients. This scenario not only limits access to financial services for deserving individuals but also constrains the growth and inclusivity potential of consumer finance providers.

The Loan Default Risk Prediction Model project seeks to address these challenges by developing a predictive model that can accurately assess the default risk of loan applicants over time, with a particular focus on maintaining the model's stability and performance. This involves creating a solution that is resilient to changes in consumer behavior and economic conditions, ensuring that the model remains reliable and effective in predicting loan default risks in the long term. The project aims to leverage the latest advancements in data science, statistical, and machine learning methods to achieve a balance between model stability and performance, thereby reducing the need for frequent updates and adjustments.

## **Objectives**

1. Date preprocessing

Processing missing value, duplicate values and encoding categorical variables (if necessary).

1. Exploratory data analysis

We will conduct outlier detection, figure out the data trend, and conduct correlations analysis.

1. Feature engineering

Identify and incorporate pertinent features, such as deposit amount and incoming debit card transactions amount. Then we may need to employ aggregation functions that will condense the historical records associated with each “case\_id” into a single feature.

1. Feature Selection

We will use some methods, such as PCA(Principal Component Analysis) to examine the effect of feature importance and selecting only a certain number of the most important features on model performance.

## Model Selection

## We will use different models, such as CatBoost, XGBoost and LightGBM. Then, we will evaluate the models by using a gini stability metric. First, we get the weekly gini scores:

*gini= 2\* AUC -1 (1)*

Fitting the weekly gini scores, we obtain a linear regression a\*x+b, which aids in obtaining the stability metric:

*Stability metric = mean(gini) + 88.0\*min(0,a) - 0.5 \* std(residuals) (2)*

Then by performing cross validation or employing metrics such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), we will determine the best untuned model.

## Model Tuning and Prediction

## Lastly, we will tune a model and make predictions on the test dataset using this tuned model. Then, we predict a probability for the target score.

## **Datasource**

Name: Home Credit - Credit Risk Model Stability, Kaggle

Link: <https://www.kaggle.com/competitions/home-credit-credit-risk-model-stability/data>

Dataset File Size: 26.76GB

Approximate Number of Records: ~1.5millions

## **Proposed Technologies and Programming Languages**

Data Preparation and Processing: Python 3, NumPy, Pandas, PySpark

EDA: python packages(matplotlib, seaborn, mmlspark, statistics,etc.)

Machine Learning Models: XGBoost, LightGBM, CatBoost