VIETNAM NATIONAL UNIVERSITY, HO CHI MINH CITY
UNIVERSITY OF INFORMATION TECHNOLOGY

PROJECT REPORT

Courses: [CS313.P11] Data Mining and Applications

Instructor: Nguyễn Võ Lê Duy

Group Number:

	Student Name	Student ID	Task
1	Nguyễn Hà Anh Vũ	21520531	EDA, Demo (18%)
2	Nguyễn Đông Anh	21520569	Slide (16%)
3	Nguyễn Đình Minh Chí	21520648	Present (17%)
4	Trần Công Hải	21520811	Report (16%)
5	Phạm Đức Hiếu	21520856	Report, Slide (15%)
6	Huỳnh Nhật Hòa	21520860	Orientation, feedback, experiment (18%)

Project report for Data Mining and Applications - Group 3

CHAPTER I. INTRODUCTION

1. Introduction

- Personal credit plays an important role in the economy:

- Supports consumption, stimulates the economy.
- Promote personal investment.
- Increase financial inclusion.
- Allocates capital more efficiently.

-Risks of bad debt, which can have negative impacts on the economy:

- Loss of trust in the credit market.
- Significant financial system damage.
- Economic development in stunted.
- Creates risks that affect social development.

=> Therefore, predicting replayment ability is a typical and necessary task in the finacial sector today.

2. Target

- Building a model to predict whether a customer is likely to repay a loan, based on features such as income, credit history, and other related information.
- This model helps banks make more accurate lending decisions, minimizing the risk of bad debts by identifying potential loan profiles.

3. Motivation

- Minimize credit risk: Reducing the likelihood of defaults and bad debts by accurately predicting borrowers' repayment ability.
- Optimize loan portfolio profitability: Enhancing the performance of lending portfolio by making informed decisions base on data-driven insights.
- Apply machine learning to real-world problems with high business relevance.

Chapter II: PROBLEM SETUP

1. Problem definition

- Input: Personal information about customers such as: income, credit history, current loans, age,...

- output: The prediction results of model:

    - If the result is 1: it means the customer is likely to default or unable to repay the loan, in which case the bank will not approve the loan.
    - If the result is 0: it means the customer is unlikely to default or has the ability to repay the loan, in which case the bank will approve the loan.

2. Data features

- Source: Data train.csv and Test.csv from Kaggle

- Sample and Variable:

    - Sample: 58645
    - Number of features: 10 (including categorical and numeric variables), specifically 7 numeric variables and 4 categorical variables. We exclude 'id' feature as it doesn't affect).
    - Target variable: **'loan_status'** which has a value of 0 or 1.

- Distribution: Unbalanced Target variable.

- Main challenges:

    - Imbalanced dataset: which affects model performance.
    - Multi types of variable: Includes both categorical and numeric variables, requiring appropriate encoding.

Chapter III: METHOD.

1. Data Processing Procedures

- **Data Cleaning**:

    - Fill missing values using median (numeric variables) and mode (categorical variables). However, this is not necessary because there are no missing values.
    - Remove samples with extreme outliers: also not necessary because we will use meta-models such as XGBoost, CatBoost, and LightGBM, which are robust to outliers.

- **Feature Engineering**:

    - Create new variables, such as debt_to_income_ratio: We tried this, but it didn't improve the model's performance significantly.
    - Transform numeric variables into categorical variables (while keeping the original variables): This approach proved to be more effective.

- **Encoding**:

    - One-hot: Used for variables with few distinct values, such as 'loan_intent'.
    - Ordinal: Used for variables with many distinct values that have a natural order.

- **Scaling**: No need **- Handling Imbalanced Data**:

    - Using cross validation.
    - Using models that handle imbalanced data well.

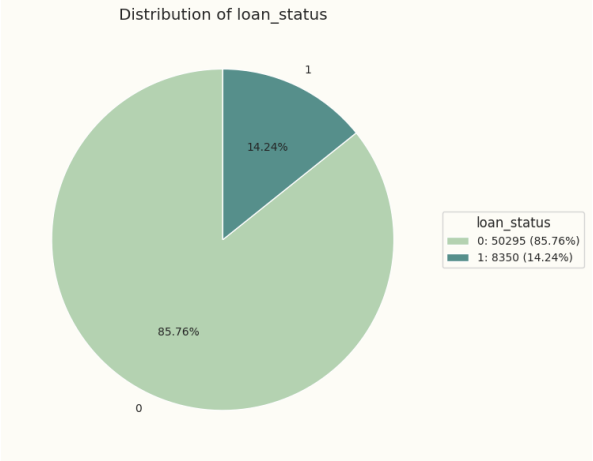
2. Model Used

- **XGBoost**: High performance with tabular data, and has mechanisms for handling class imbalance.
- **CatBoost**: Automatically supports categorical data with minimal preprocessing required.
- **LightGBM**: Fast and efficient with large datasets.
- **Voting**: Ensemble technique combines results from multiple models. In this case, soft voting is used: the final prediction is based on the average probabilities predicted by the sub-models (XGBoost, CatBoost, and LightGBM).

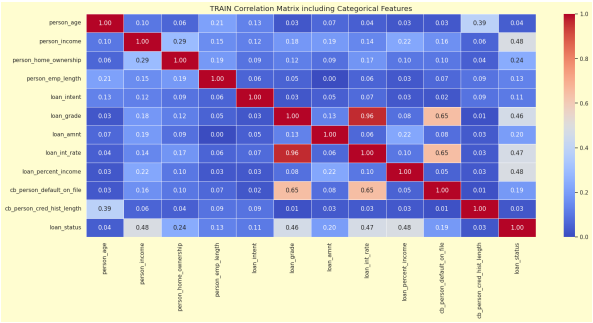
3. Evaluation Metrics

- **AUC- ROC**: Measures the ability of a model to distinguish between classed.

Chapter IV: EXPERIMENT



The distribution of **'loan_status'** shows an imbalance, which is why we use models like CatBoost, XGBoost,... and other Boosting-based algorithms for analysis, as they are effective in handling imbalanced datasets.
We use a correlation matrix to identify relationships between variables.



1. Features of the Correlation Matrix

- Correlation coefficient has colors represent the level of correlation:
  - Dark red: High positive correlation (close to 1).
  - Dark blue: Low or no correlation (close to 0).
  - White or light orange: Moderate correlation.
- Variable names: **'loan_status'**, **'person_age'**, **'loan_grade'**, ...

2. Analysis

- Highly Correlated Variables:

  - **'loan_grade'** and **'loan_int_rate'**: Correlation coefficient of 0.96, indicating a very strong relationship. This means loan grade and interest rate are closely related.
  - **'loan_status'** and **'person_income'**: Correlation coefficient of 0.48, suggesting individuals with higher income are more likely to repay loans on time.

- Moderately Correlated Variables:

  - **'cb_person_default_on_file'** and **'loan_grade'**: Correlation coefficient of 0.65, showing a moderate relationship between credit history and loan grade.
  - **'cb_person_cred_hist_length'** and **'person_age'**: Correlation coefficient of 0.39, indicating a moderate link between a person's age and the length of their credit history.

- Weakly Correlated Variables:

  - **'loan_amnt'** and other variables: Most correlations are close to 0, meaning loan amount has little significant relationship with other factors in this dataset.
  - **'loan_percent_income'** and **'cb_person_cred_hist_length'**: Low correlation (0.03), indicating a negligible relationship.

Additionally, other analyses are presented in the Notebook file.

3. Experimental results

A comparison table of results between models.

	Private test	Public test
LightGBM	0.95862	0.95768
CatBoost	0.95717	0.95728
XGBoost	0.95679	0.96096
Voting	0.95954	0.96028
Logistic	0.84918	0.84229

From the above table, we can see that Logistic Regression perform poorly on imbalanced data as they favor the majority class; instead, Boosting models like XGBoost, CatBoost, or LightGBM are better suited due to their handling of class weights and non-linear learning capability.

Chapter V: CONCLUSION

1. Result Summary

In the loan approval prediction task, based on features like financial status, loan purpose,... , models such as Boosting (XGBoost, LightGBM, CatBoost) and Voting Ensembles have demonstrated strong potential and effectiveness in predicting approval outcomes.

2. Further development

- Experiment with other methods to improve predictive outcomes.
- Find more optimal parameters for the model such as grid search, random search, optune,...