Subject: Machine Learning, CoE, PSU
Lecturer: Dr. Anant Choksuriwong
Name: Mr. Standa Na
ID: 6710130001

Credit Risk Prediction Model

1. Motivation

Credit risk is a critical concern for financial institutions, as it directly impacts their profitability and stability. Predicting whether a borrower will default on a loan is essential for minimizing financial losses and ensuring a stable lending environment. Developing an accurate prediction model allows institutions to better assess the risk of loan applicants and make more informed lending decisions. This helps reduce default rates, optimizes loan approval processes, and strengthens financial sustainability.

2. Project Workflow

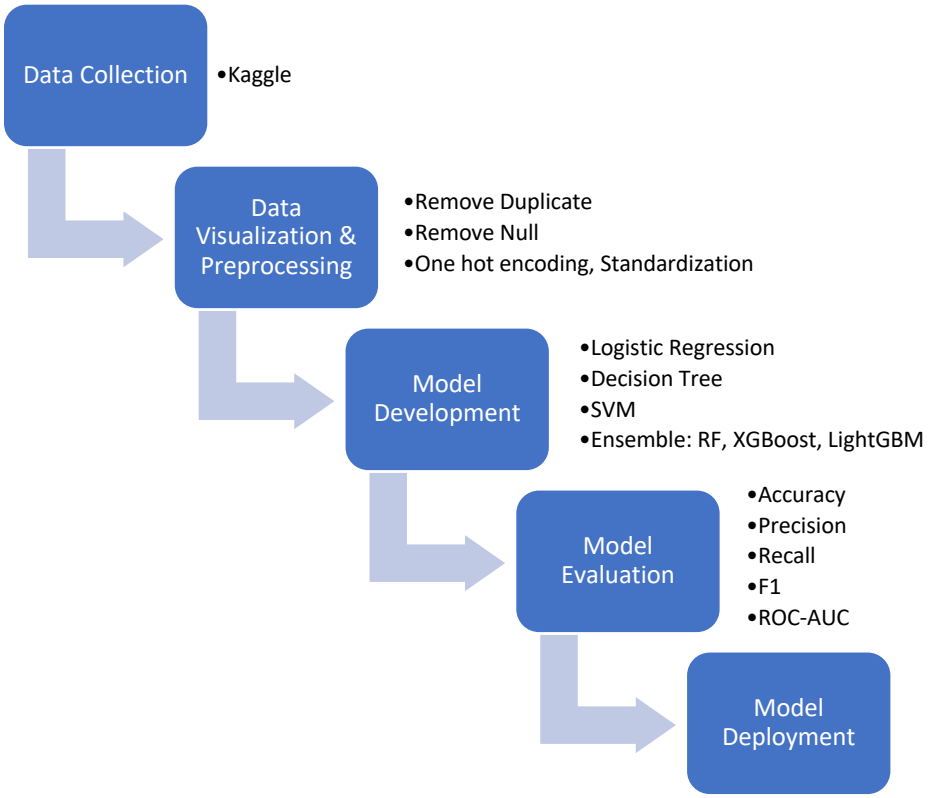

Figure 1 Project Workflow

3. Dataset

The dataset used in this study was sourced from Kaggle. It includes 11 features and a target variable, containing 32,581 records. The distribution of the target classes is as follows:

- **Non-default (label 0)**: 25,473 records
- **Default (label 1)**: 7,108 records

Table 1 Dataset Description

Column	Description	Data Type
person_age	Age	Int
person_income	Annual Income	Int
person_home_ownership	Home ownership	Object
person_emp_length	Employment length (in years)	Float
loan_intent	Loan intent	Object
loan_grade	Loan grade	Object
loan_amnt	Loan amount	Int
loan_int_rate	Interest rate	Float
loan_status	Loan status (0 is non default 1 is default)	Int
loan_percent_income	Percent income	Float
cb_person_default_on_file	Historical default	Int
cb_person_cred_hist_length	Credit history length	Int

4. Data Visualization

According to Figure 2, borrowers with loan grades A, B, and C tend to repay better than those with grades D and above. Figure 3 shows that features like the loan-to-income ratio and interest rate have a stronger correlation with loan repayment status than other features.

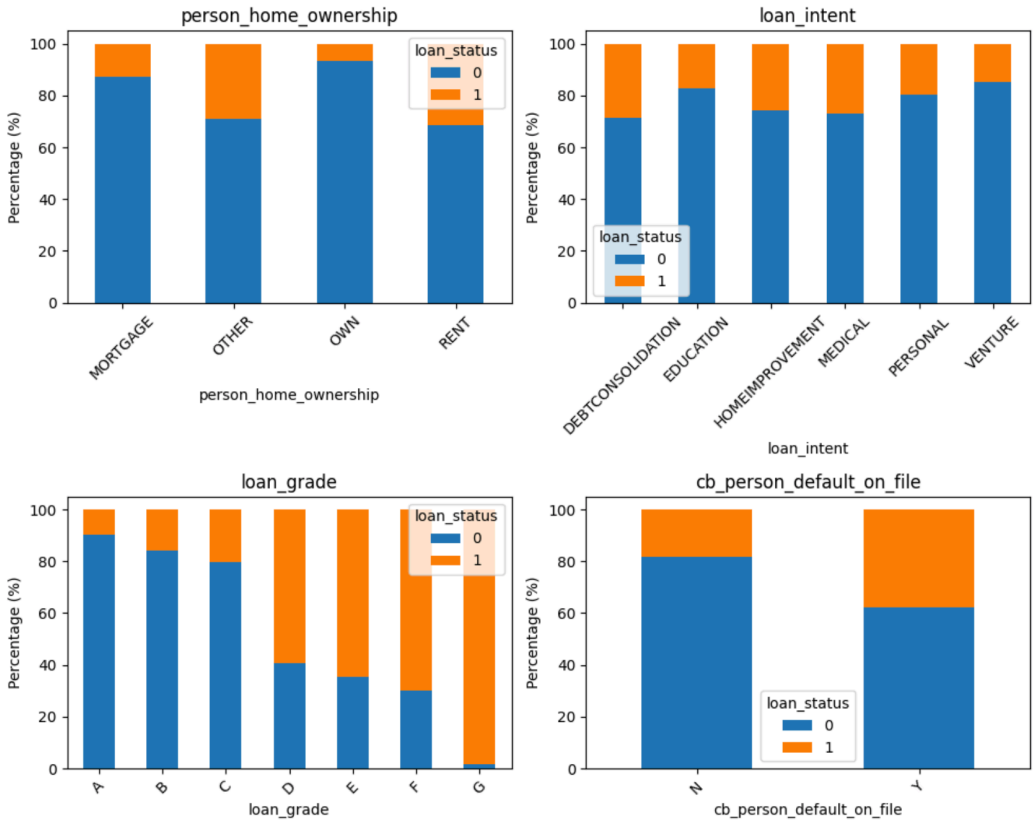

Figure 2 Loan Status Breakdown by Categorical Features (100% Stacked Bar)

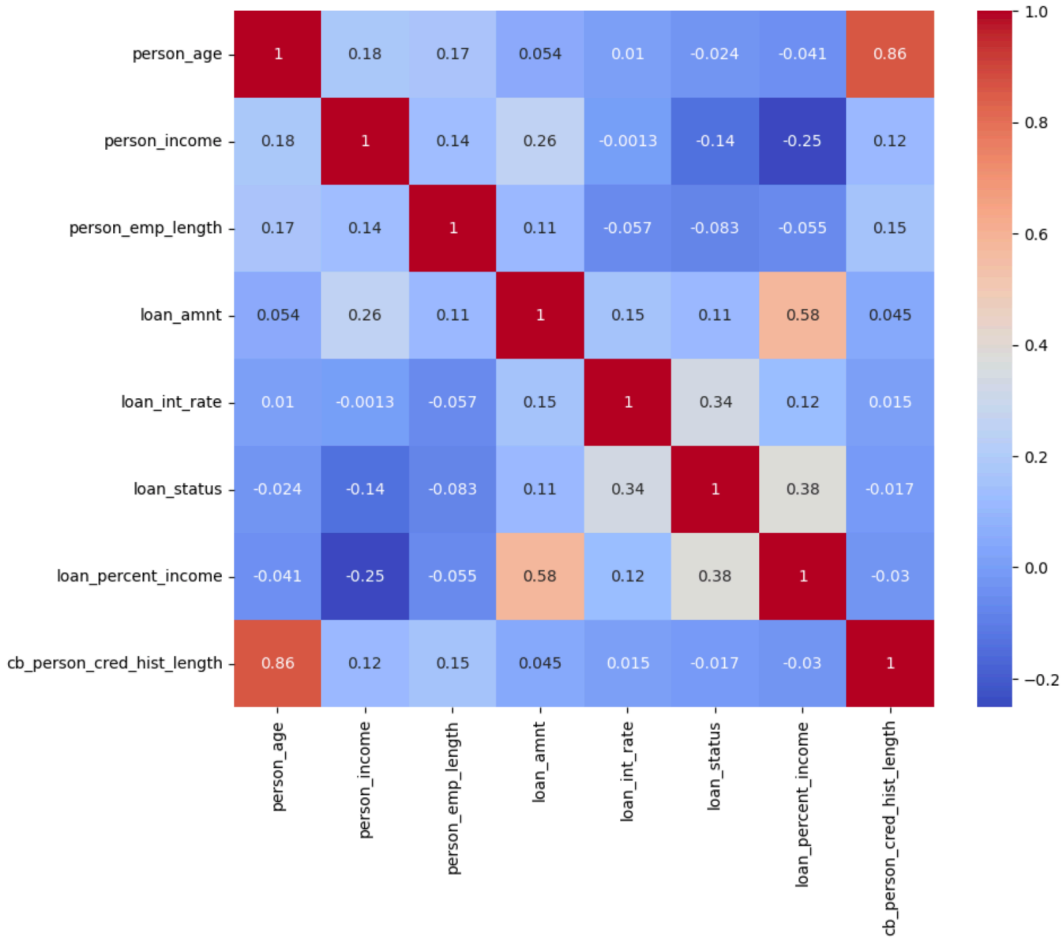

Figure 3 Pearson's Correlation between numerical features

5. Data Preprocessing

The data was cleaned by removing duplicate and missing records, leaving 28,501 records for analysis. We applied standardization to scale the numerical features and used one-hot encoding for categorical data.

6. Model Development

We split the dataset into 80% for training and 20% for testing. The machine learning models evaluated include:

- Logistic Regression (LR)
- Support Vector Machine (SVM)
- Decision Tree (DT)
- Random Forest (RF)
- Extreme Gradient Boosting (XGBoost)
- Light Gradient Boosting Machine (LightGBM)

Libraries used:

- **Pandas** (2.2.2)
- **Matplotlib** (3.7.1)
- **Seaborn** (0.13.2)
- **Scikit-learn** (1.5.2)
- **XGBoost** (2.1.1)
- **LightGBM** (4.5)

7. Model Evaluation

The models were evaluated using **Accuracy**, **Precision**, **Recall**, **F1 Score**, and **ROC-AUC**.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Where:

- TP = True Positives (correctly predicted positive cases)
- TN = True Negatives (correctly predicted negative cases)
- FP = False Positives (incorrectly predicted positive cases)
- FN = False Negatives (incorrectly predicted negative cases)

Table 2 The performance of each model on the Test set

Model	Parameters	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	ROC-AUC (%)
LR	Penalty:l2, C:1.0	86.70	79.25	53.96	64.21	87.01
SVM	C:1.0, kernel:rbf	91.01	93.18	64.04	75.91	89.11
DT	Criterion:gini	88.26	73.14	**74.12**	73.63	83.20
RF	Criterion:gini Estimator:100	93.00	95.84	71.42	81.85	92.51
XGB	Estimator:100, LR:0.3	93.07	94.35	73.01	82.32	**94.38**
LightGBM	Estimator:100, LR:0.1, num_leaves:31	**93.24**	**96.79**	71.82	**82.46**	94.30

8. Conclusion

According to the results, LightGBM achieved the best overall performance with an accuracy of 93.24% and an F1 score of 82.46%, while Decision Tree (DT) had the highest recall of 74.12%. XGBoost excelled in ROC-AUC, reaching 94.38%, showcasing its strong predictive power.

Reference

https://www.kaggle.com/datasets/laotse/credit-risk-dataset