A Classifier for Screening Out Loan Defaulters

1. Introduction

1.1 Background

According to a report by Trading Economics (2023), Hong Kong's private sector received loans worth 10.38 trillion Hong Kong dollars in August. However, there is a potential risk of repayment failures as loans mature. Evaluating the credit risk of loan applications is crucial for bankers to prevent defaults and minimize financial losses. Defaulting borrowers lead to additional costs for collections and disruptions in cash flow, making credit risk management vital for banks to reduce expenses.

1.2 Motivation

To mitigate the risk of loan default, financial institutions employ the "5C Principles" for assessing loan applications (Ni Kadek Ayu Rika et al., 2022). However, this process is time-consuming and subjective, requiring significant manpower. Adopting a machine learning model for loan defaulter classification becomes crucial to eliminating subjectivity and streamline the application process.

This study aims to identify loan defaulters and assess feature importance using multiple machine learning models. While TransUnion credit ratings (Furletti, 2006) have traditionally been used for credit card decisions, this analysis explores alternative features that may be more significant.

2. Dataset Preparation

2.1. Data Collection

The data for our analysis was obtained from Kaggle, an open-source community. The dataset used is called "Credit Risk Analysis" (Tse, 2020), provided in an Excel file format. It contains a total of 32,581 observations, with each observation having 12 variables. These variables capture demographic attributes of individuals who applied for a loan. As the data is from a third party, it is crucial to conduct thorough data cleansing to ensure data integrity and reliability before proceeding with further analysis.

2.2 Data Cleansing

During this phase, we conducted various checks on our dataset using Python. We examined missing data, incorrect formatting, and duplicate entries. We found 3,942 missing data points but no instances of duplicated or incorrectly formatted data. Additionally, we identified 4 records with abnormally high ages (144 and 123 years old), which we considered outliers based on domain knowledge and information from the Guinness World Records. To maintain data integrity, we removed these 4 outliers and 3,942 missing data points from the dataset. The exclusion of missing data and outliers is not expected to significantly impact on our analysis due to the ample number of remaining observations. Our new dataset size has been adjusted to 28,634.

2.3 Data Preprocessing

To ensure compatibility with different algorithm types, including non-tree-based and tree-based algorithms, we employed distinct transformation schemes. For categorical variables, we used label encoding to convert them into numeric form, effectively incorporating categorical information. To prepare the data for non-tree-based algorithms, we applied normalization techniques to scale features within a similar range, benefiting distance-based calculations or gradient descent optimization. Additionally, we addressed class imbalance by employing

stratified data splitting, dividing the data into 80% for training and 20% for testing. This approach-maintained class proportions and improved the model's ability to handle the imbalance. These transformation schemes aimed to optimize data compatibility, improve analysis performance, and effectively handle class imbalance.

3. Data Visualization

In this section, we will analyze two aspects of our dataset. Firstly, we will examine the proportion of loan defaulters and non-defaulters to understand the distribution of loan outcomes. Secondly, we will explore the relationship between categorical variables and loan status to identify any patterns or correlations.

3.1. Proportion of Loan Defaulters and Non-defaulters

In Figure 1, the data reveals that we have 21.7% loan defaulters and 78.3% non-defaulters, indicating that loan defaulters are a minority in the dataset. However, the imbalanced nature of the dataset can potentially result in less satisfactory outcomes when classifying loan defaulters. Therefore, in Section 5, we will evaluate our models by considering precision and recall.

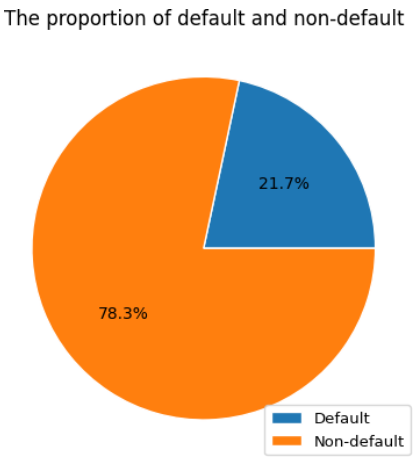

Figure 1: The proportion of loan defaulters and non-defaulters

3.2. Relationship between Categorical Variables and Loan Status

The bar plots in Figure 2 reveal important insights regarding the relationship between categorical variables and loan status. Specifically, two trends stand out: home ownership and loan grade. Home ownership is linked to individuals' wealth, with homeowners having a lower proportion of loan defaulters compared to mortgage or rent. Furthermore, as the loan grade decreases, the proportion of defaulters increases. Therefore, loan grade and home ownership are critical factors in determining loan defaulters. In Section 4, we will assess their significance using feature importance to strengthen our findings.

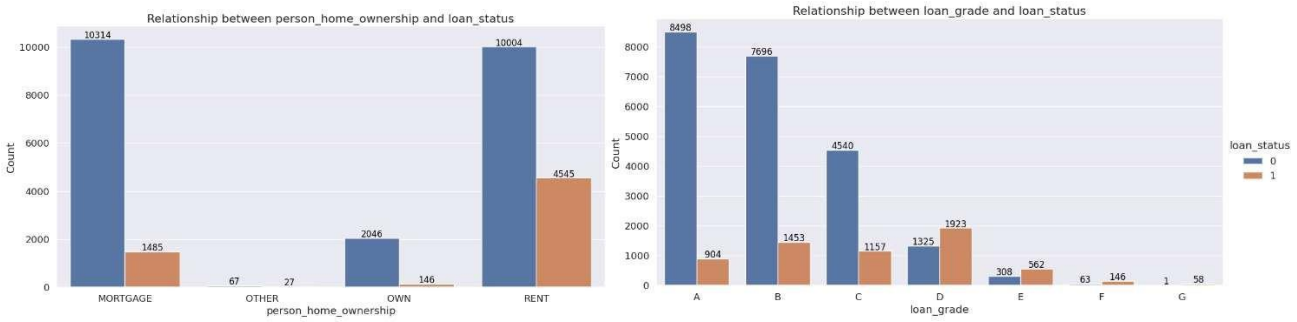

Figure 2: Relationship between home ownership, loan grade and loan status

4. Model Building

A tree-based classifier, comprising logistic regression, as well as non-tree-based classifiers such as decision tree, random forest, and light gradient boosting machine (LightGBM), was trained. Grid search with cross-validation was applied to tune the hyperparameters using the training data. To evaluate the performance of our models, we will calculate the accuracy, precision, recall, and F1-score using the testing data.

4.1. Logistic Regression

Logistic regression is used to predict loan default by providing probabilistic values. VIF score analysis removed highly correlated features which affect the model's performance. After hyperparameter tuning, the model achieved an accuracy of 84.32%, precision of 70.54%, recall of 47.46%, and F1-score of 56.74%. The relatively

low recall suggests that the model struggles to effectively identify loan defaulters, possibly due to the imbalanced nature of the data. Tree-based models are expected to perform better in this aspect.

4.2. Decision Tree

A decision tree can be employed by learning simple decision rules from data features associated with loan applications. Instead of using grid search, we evaluated the decision tree at various depths. Our findings indicate that max_depth = 8 yields the highest score for both the testing and training datasets. The performance of the model was assessed after tuning. With the tuned parameters, the model achieves an accuracy of 92.75%, precision of 96.83%, recall of 68.82%, and F1-score of 80.45%. This outperforms logistic regression, but it is anticipated that random forest, a more complex model, would yield better recall.

4.3. Random Forest

Random Forest relies on multiple decision trees and combines their predictions using majority voting to determine the final result. After the hyperparameter tuning process, the best parameters were found to be max_depth = 60 and n_estimators = 300. The model's performance was evaluated based on these tuned parameters, resulting in an accuracy of 93.00%, precision of 96.26%, recall of 70.43%, and F1-score of 81.34%. Although the model performs well, it is time-consuming to train. Therefore, considering the faster computation of the LightGBM model may be beneficial.

4.4. Light Gradient Boosting Machine

The LightGBM model, which combines multiple weak learners to create a more accurate model, was initially created without setting parameters, resulting in an accuracy of 93.54%, precision of 97.39%, recall of 72.12%,

and F1 score of 82.87%. After hyperparameter tuning, the best parameter values improved accuracy (+0.0002), recall (+0.0081), and F1-score (+0.0020), enhancing overall performance. However, precision decreased.

The feature importance analysis identified the individual's default history as the least important. Further analysis showed only 17.8% of individuals had defaulted, indicating minimal impact on the model's results. Thus, the 'cb_person_default_on_file' column was removed.

After removing the column and performing hyperparameter tuning (Learning_rate=0.1, Max_depth=8, Min_child_samples=30, N_estimators=300, Num_leaves=31), the model archived an accuracy 93.63%, precision 96.10%, recall 73.57%, and F1-score 83.34%. Accuracy, recall, and F1-score increased, while precision decreased, indicating a more cautious approach in predicting positive instances.

5. Model Evaluation

Model accuracy, precision, recall and F1 score are the main criteria for the model selection, and they are concluded from different machine learning model after tunning.

	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.843	0.705	0.475	0.567
Decision Tree	0.928	0.968	0.688	0.805
Random Forest	0.930	0.963	0.704	0.813
LightGBM	0.936	0.961	0.736	0.833

Table 1: 4 model classifiers' performance

Based on the provided table, it is evident that the LightGBM model surpasses the other three models in terms of accuracy, recall, and F1-score.

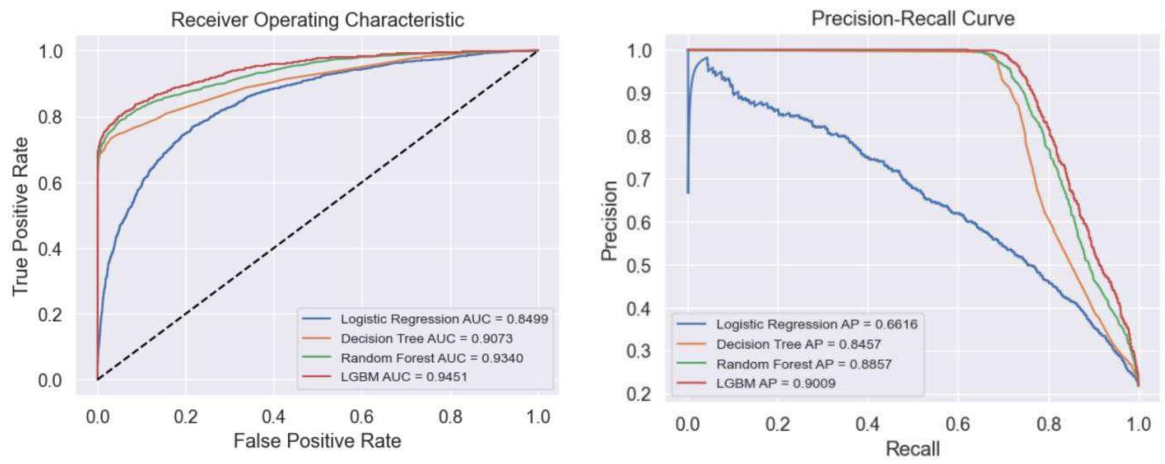

Figure 3: ROC curve and PR curve for the 4 classifiers

The LightGBM model demonstrates superior predictive performance compared to the evaluated classifiers, with an impressive AUC score of 0.9451 and average precision score of 0.9009, as observed from the ROC curve and Precision-Recall curve analysis.

6. Conclusion

In conclusion, the LightGBM model is the most effective classifier for identifying loan defaulters, achieving a precision of 96.1% and streamlining the loan screening process. However, attention should be given to non-defaulters, as the recall score is 73.6%, indicating potential missed loan default cases.

Our analysis emphasizes the significance of the debt-to-income (DTI) ratio, indicating higher repayment capacity for borrowers with lower DTI ratios. The credit grade of borrowers is also important, traditionally used in credit card application decisions. Therefore, prioritizing evaluation based on the DTI ratio is advised. However, it's important to note that the dataset considered only includes 11 variables, excluding crucial factors like family size, savings amount, and investment amount. Hence, there may be other important features in loan applications not accounted for in our analysis.

References

Furletti, M. J. (2006, September 14). *An overview and history of Credit Reporting*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=927487

Hong Kong Loans to Private Sector - September 2023 Data - 1978-2022 Historical. (n.d.). https://tradingeconomics.com/hong-kong/loans-to-private-sector

Huichen Z.(2023). *STAT 4001 statistics projects* https://goo.su/dI06cX

Ni Kadek Ayu Rika, R., I Made, S., & Luh Mei, W. (2022, September). The effect 5C principles assessment on non-performing loans at PT Bank. http://repository.pnb.ac.id/599/1/RAMA_62301_1815644109_artikel.pdf

Tse, L. (2020, June 2). *Credit risk dataset*. Kaggle. https://www.kaggle.com/datasets/laotse/credit-risk-dataset

Appendix

A. Further Results on Data Visualization

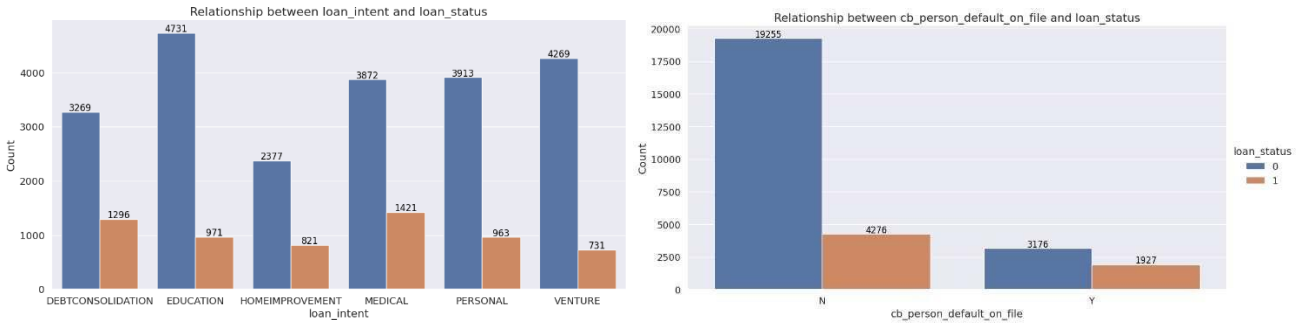

Figure A1: Relationship between loan intent, historical default, and loan status

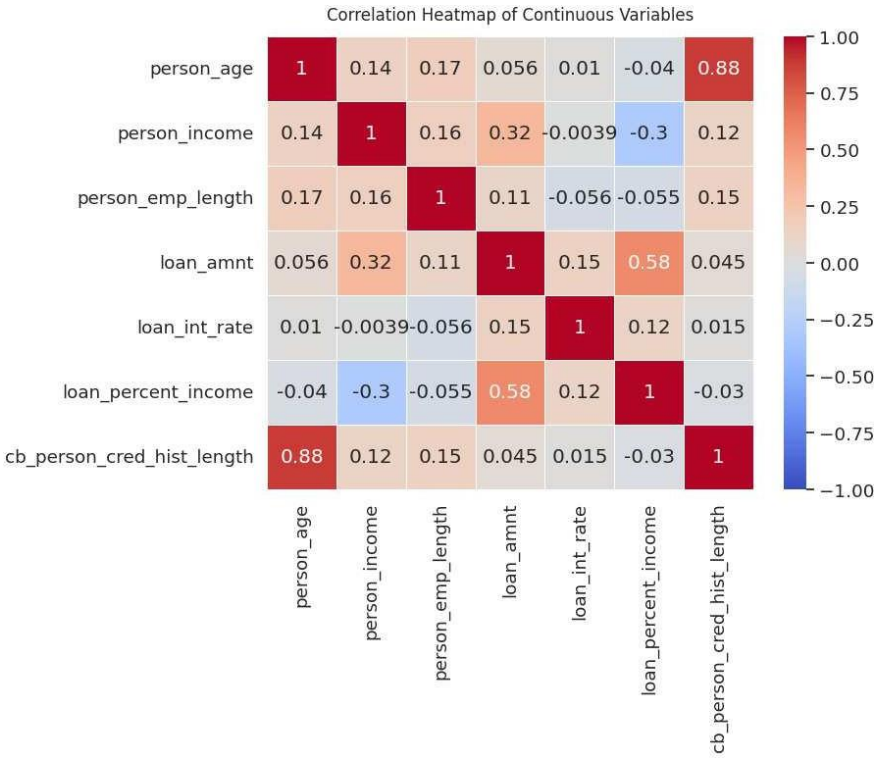

Figure A2: Correlation heatmap of continuous variables

B. Logistic Regression

Attribute	VIF Scores
person_income	3.124955
person_home_ownership	2.226121
person_emp_length	2.499063
loan_intent	2.782956
loan_grade	2.979433
loan_percent_income	2.948955
cb_person_default_on_file	1.710243
cb_person_cred_hist_length	1.902376

Table B1: VIF score for the remaining features

C	0.1	**1**	10	12	15
penalty	**l1**	l2			
solver	**liblinear**				

Table B2: Hyperparameter set for Logistic Regression hyperparameter tuning

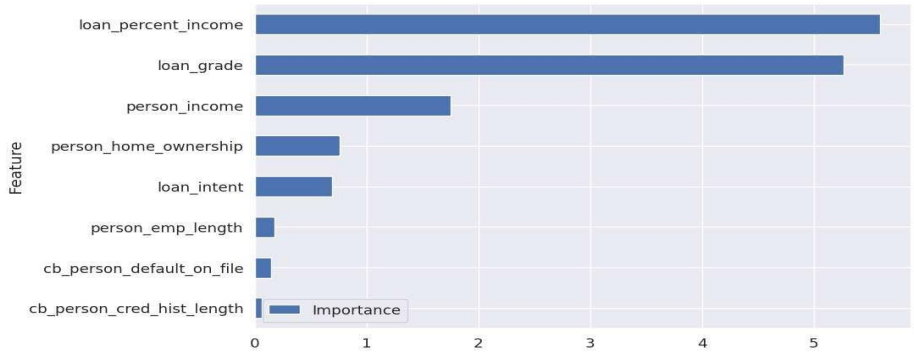

Figure B1: Feature importance for Logistic Regression

C. Decision Tree

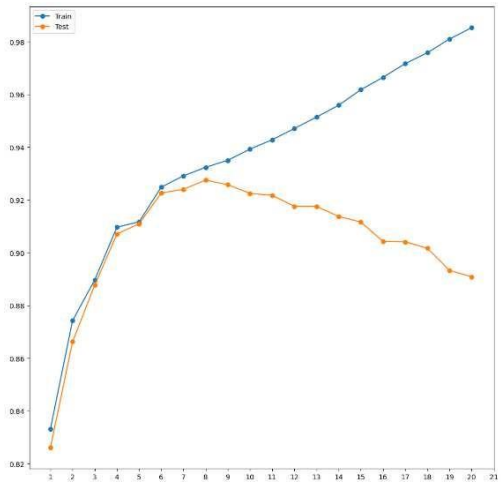

Figure C1: Hyperparameter tuning for Decision Tree

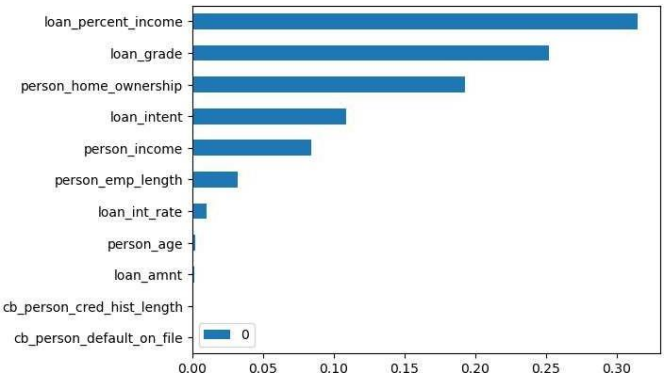

Figure C2: Feature importance for Decision Tree

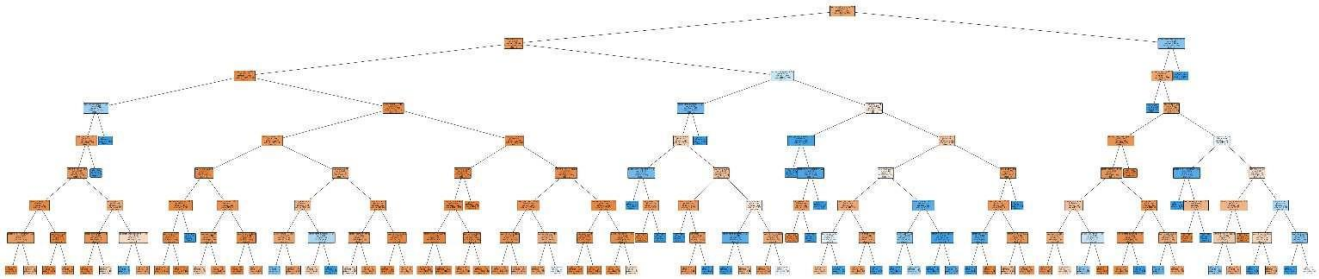

Figure C3: Decision tree visualization

D. Random Forest

n_estimators	100	200	**300**	
max_depth	20	40	**60**	80

Table D1: Hyperparameter set for Random Forest hyperparameter tuning

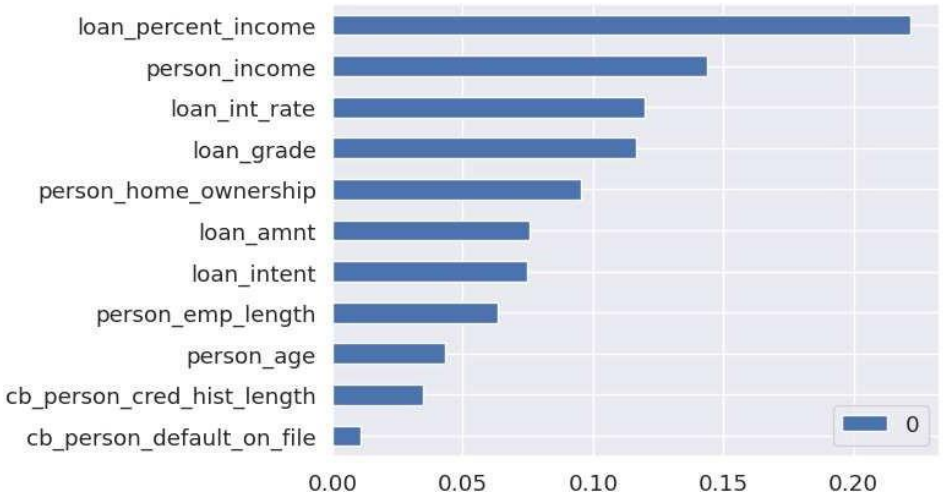

Figure D1: Feature importance for Random Forest

E. Light Gradient Boosting Machine

n_estimators	100	200	**300**
learning_rate	**0.05**	0.1	0.2
max_depth	**6**	8	10
num_leaves	**31**	63	127
min_child_sample	**20**	30	40

Table E1: Hyperparameter set for LGBM hyperparameter tuning

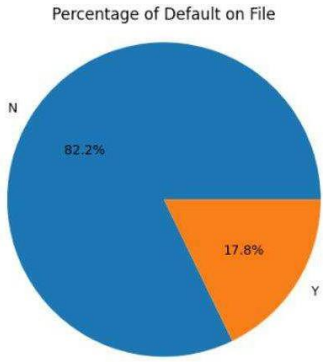

Figure E1: Percentage of Default on File

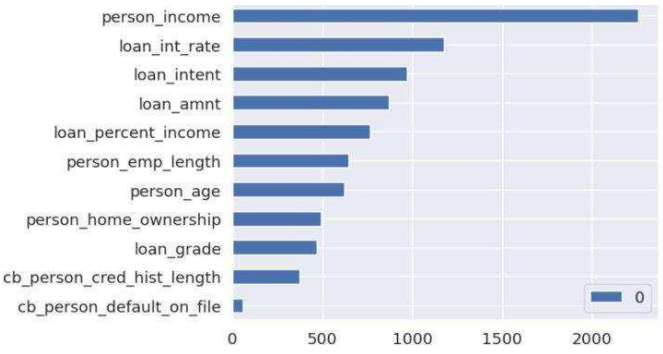

Figure E2: Feature importance of LGBM model (Parameter tuned)