Customer Analytics

in Large Organisations

Comparative Analysis of Credit Scoring Models Logistic Regression and XGBoost in Default Prevention

Name: Wangjun SHEN

Professor: Timofei Bogomolov

Institution: The University of South Australia

Course Name: Customer Analytics in Large Organizations

Course No: INFS 5096

[Abstraction]

This study compares Logistic Regression and XGBoost in predicting credit default on a Taiwanese dataset of 30,000 credit card users. It finds that Logistic Regression is easily interpretable, while XGBoost handles imbalanced data effectively, making it superior for identifying high-risk individuals. This analysis aids financial institutions in refining risk assessment strategies.

Content Table

Introduction

In the current financial market environment, the accuracy of credit scoring is crucial for effective risk management. Financial institutions require reliable methods to assess customers' credit risk, optimize loan approval processes, and mitigate bad debt losses. This study aims to enhance the identification of potential "subprime" customers by constructing and evaluating efficient credit scoring models, thereby supporting banks and other financial institutions in making precise and efficient credit decisions.

Subsequently, the study will employ the credit scorecard methodology to construct a baseline credit scoring model. Credit scorecards are widely utilized tools for credit risk assessment, assisting financial institutions in quantifying customers' credit risk by converting credit data into score formats. Additionally, the study will explore other predictive methods. By comparing and analyzing the applicability and performance of different models in real-world business scenarios, the study will provide scientific decision-making support for financial institutions.

The dataset utilized in this study originates from the Taiwan region and encompasses detailed credit information of credit card users. The dataset includes multiple dimensions of user information, such as credit limits (LIMIT_BAL), gender (SEX), education level (EDUCATION), marital status (MARRIAGE), age (AGE), repayment status from April to September 2005 (PAY_0 to PAY_6, but no PAY_1), and bill amounts (BILL_AMT1 to BILL_AMT6). Furthermore, the dataset records the actual payment amounts (PAY_AMT1 to PAY_AMT6) and the customer defaulted payment on the next month (default.payment).

In this study, data analysis and model construction techniques will be used to explore and identify the characteristics that affect customer default. Specifically, the study will build and evaluate the performance of credit scoring models and other predictive models in real-world applications, comparing their utility and feasibility across different scenarios. This report will provide valuable insights for bank loan officers, credit scoring analysts, and policymakers, assisting them in understanding and leveraging advanced credit assessment techniques to make more informed decisions in the highly competitive financial services market.

Dataset Description and Data Preprocessing

Variable	Description
ID	ID of each client
LIMIT_BAL	Amount of given credit in NT dollars (includes individual and family/supplementary credit)
SEX	Gender (1=male, 2=female)
EDUCATION	Education (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
MARRIAGE	Marital status (1=married, 2=single, 3=others)
AGE	Age in years
PAY_0	PAY_0: Repayment status in September 2005 (-1=paid in full, 1=1 month delay, 2=2 months delay, ..., 9=9+ months delay)
PAY_2 - PAY_6	Repayment status from April to August 2005 (scale same as PAY_0)
BILL_AMT1 - BILL_AMT6	Amount of bill statements from April to September 2005 (NT dollars)
PAY_AMT1 - PAY_AMT6	Number of previous payments from April to September 2005 (NT dollars)
default.payment	Customer defaulted payment on the next month (1=yes, 0=no)

Table 1: Description of Variables in Dataset

This data set contains 30,000 samples and 25 features including IDs and has no missing values. The meaning of these features and their corresponding data are referred to the description in Table 1. It can be found from the description that there is no distinction between 5 and 6 in EDUCATION because both numbers represent the same situation of "unknown", so 5 and 6 are merged into 5. There are 14 samples with EDUCATION values of 0, accounting for 0.05%; and there are 54 samples with MARRIAGE values of 0, accounting for 0.18%. Since the proportions of both are very small, they are be chosen to drop directly.

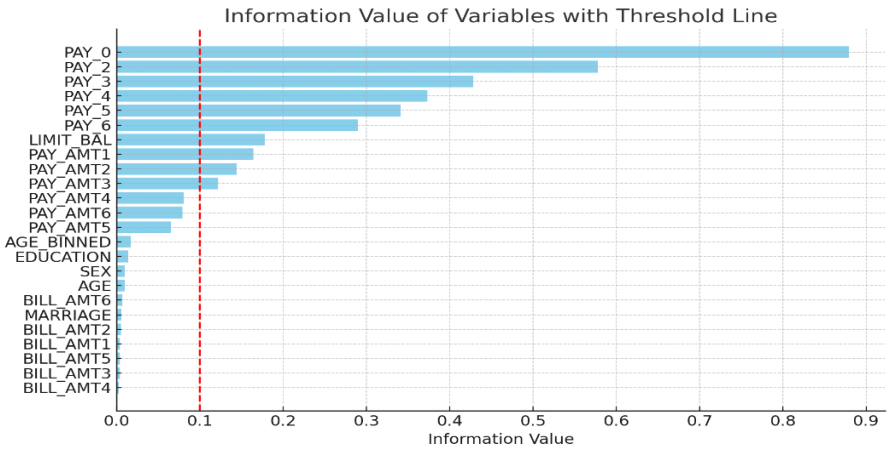

Figure 1: Information Value of Variables with 0.1 as Threshold Line

Information Value (IV) is a metric that measures a variable's predictive power for a

binary target variable like default/non-default. It quantifies how much predictive information an independent variable provides about the dependent variable, indicating its classification ability. As shown in Figure 1, the information value of different variables provides insights into their relative importance in predicting customer default risk.

PAY_0 (last month's repayment status) has an extremely high information value of 0.879, indicating it strongly predicts default risk. PAY_2 to PAY_6 (recent repayment history) has moderate to high information values above 0.1, highlighting the importance of recent repayment behavior. LIMIT_BAL (credit limit) has a moderate information value of 0.178, with higher limits linked to lower default probability. PAY_AMT1, PAY_AMT2, and PAY_AMT3 (recent repayment amounts) have higher information values, suggesting these recent repayment records crucially impact the prediction model.

Personal attributes like EDUCATION, SEX, MARRIAGE, AGE have low information values below 0.1, indicating weak direct default prediction ability but potential use for demographic analysis and market segmentation rather than primary risk assessment.

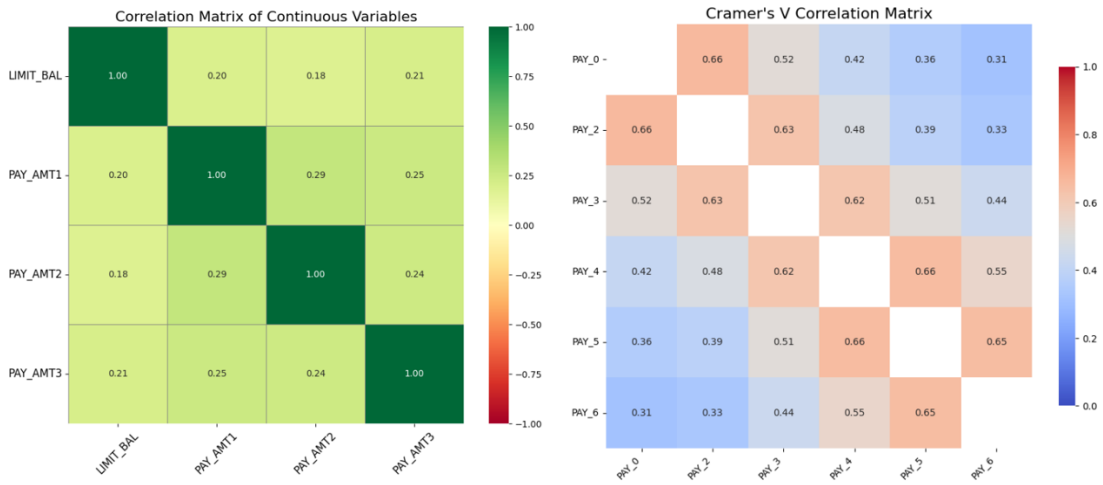

Figure 2: Correlation Matrix of Continuous Variables and Categorical Variables

After removing all variables with information values lower than 0.1, the correlation between the remaining variables needs to be analyzed to avoid multicollinearity problems that lead to poor performance of the logistic regression model and affect the creation of scorecard.

Figure 2 shows the correlation between the variables within the discrete variables and continuous variables in the selected features. It can be found that the correlations between continuous variables are very low, and no additional operations are required. However, the visualization results of discrete variables indicate that there is a high correlation between PAY_2, PAY_0 and PAY_3, and there is also a high correlation

between PAY_5, PAY_4 and PAY_6. Therefore, the two variables PAY_2 and PAY_5 are deleted to retain other variables as much as possible.

Credit Scoring Model and Credit Scorecard

Binning can convert continuous variables or categorical variables with many levels into a limited number of groupings to simplify the model and improve the interpretability of the results and create corresponding credit scorecards that non-technical people can understand and use.

Figure 3 shows the performance of the logistic regression model
The model's performance on the training set is evaluated. The K-S statistic (0.4042) indicates good discrimination between good (non-defaulted) and bad (defaulted) customers, while the ROC AUC (0.7654) suggests strong prediction accuracy. Notably, the model generalizes well, exhibiting a higher K-S value but similar ROC AUC on the test set.

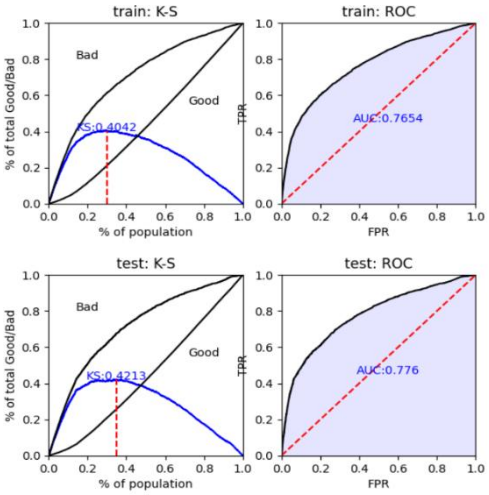

Figure 3: Model Performance Evaluation

The appendix section of this report includes a detailed credit scorecard form, a key part of the credit scoring model used to calculate a customer's credit score based on different credit behaviors and characteristics.

The credit scoring model evaluates creditworthiness using features like PAY_0, PAY_3, PAY_4, and PAY_6, which track payment statuses. Negative values indicate timely payments, scoring positively, while positive values show delinquency, resulting in severe penalties (e.g., PAY_0 above 2 scores -114 points). The LIMIT_BAL feature assesses credit limits, with lower limits linked to higher default risks. PAY_AMT1, PAY_AMT2, and PAY_AMT3 measure payment amounts, where very low payments increase default risk, and higher amounts decrease it.

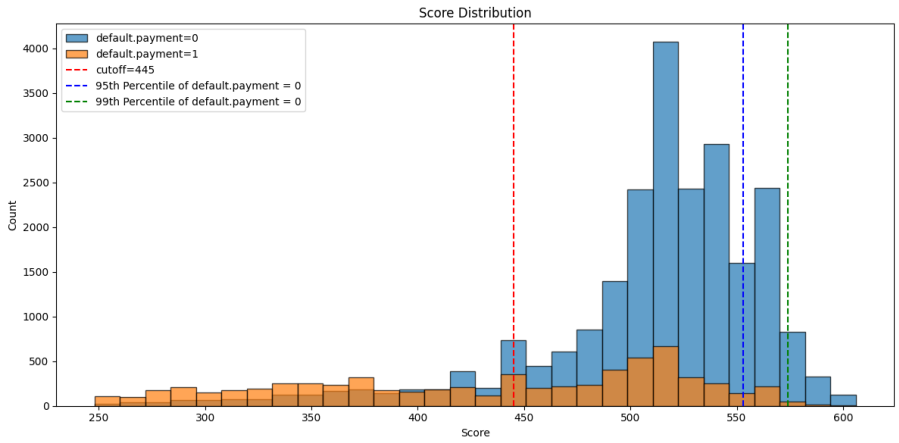

Figure 4: Credit Score Distribution with Default and Non-default Groups

Figure 4 illustrates the score distribution of customers categorized by their default status after applying the scorecard. Non-defaulting customers predominantly have scores ranging between 480 and 580, whereas the distribution of defaulting customers is more dispersed, with nearly half of them falling between 440 and 540. Utilizing the 95th or 99th percentile of scores from customers with default. payment of 1 as the cutoff would result in a significant misclassification of non-defaulting customers as defaulting. Therefore, the cutoff score is determined at the percentile corresponding to the maximum KS value. As depicted in the figure, most non-defaulting customers have scores above 450, while most defaulting customers have scores below 450. This cutoff score effectively identifies defaulting customers while retaining those with good credit, thus balancing business risks and returns.

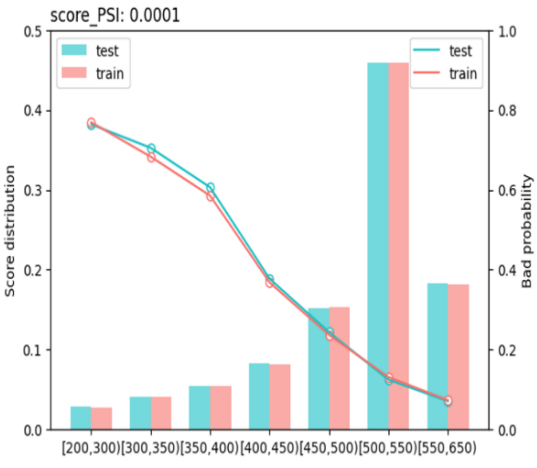

Figure 5: Score Distribution and Population Stability Index (PSI) Analysis

The Figure 5 illustrates the score distribution and bad probability for both the training and testing datasets, with an accompanying Population Stability Index (PSI) value. The bar chart depicts the score distribution across various score ranges, showing that the distributions for both the training set (in pink) and the testing set (in blue) are closely aligned. The line chart indicates the probability of bad outcomes, demonstrating a clear trend of decreasing bad probability with increasing scores for both datasets. The PSI value of 0.0001 signifies a minimal difference between the training and testing score distributions, confirming the model's stability and consistent performance across different datasets.

The XGBoost Model as Predict Model Method

When building the Credit Scoring model, a logistic regression algorithm was used initially. However, due to the imbalanced distribution of the target variable (as shown in Figure 6), XGBoost (Extreme Gradient Boosting) was employed for its ability to handle imbalanced data through sample weighting. Since decision trees, and by extension XGBoost, are less affected by multicollinearity, all features with an Information Value greater than 0.1 were included in the training process.

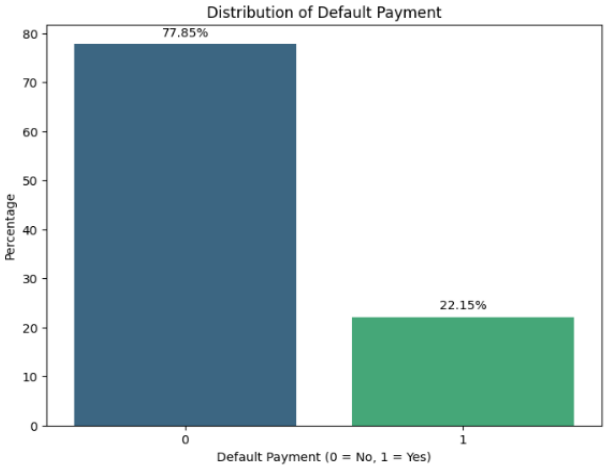

Figure 6: Distribution of Target Variables

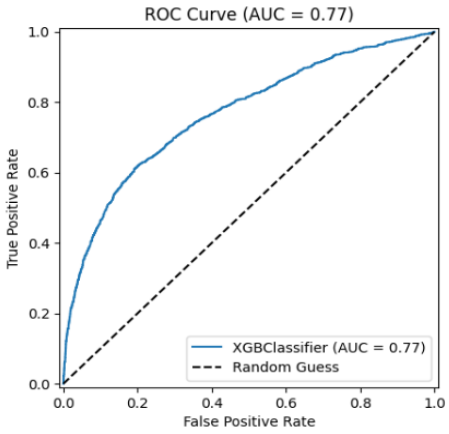

Figure 7： ROC-ACU Curve

The AUC value of 0.76 indicates that the model has good performance in distinguishing default and non-default samples. Specifically, the model is able to effectively identify most default samples while maintaining high accuracy on non-default samples.

The shape of the ROC curve also further verifies the robustness of the model. It is significantly higher than the baseline of random guessing (AUC=0.5), demonstrating the practical application potential of the model.

This model uses a total of 100 decision trees, but the depth of each tree only needs 3, and the positive and negative samples are balanced using weights. However, the complexity of the model leads to interpretability difficulties.

The model excels at predicting "good" customers (category 0), with a precision of 0.88, recall of 0.79, and F1 score of 0.83, indicating a high accuracy and reliability in identifying them. However, its performance in detecting "bad" customers (category 1) is moderate. The recall rate is 0.62, meaning it accurately identifies 62% of "bad" customers. Yet, the precision rate is only 0.46, implying that only 46% of predicted "bad" customers are truly bad, while 54% are misjudged "good" customers, potentially leading to higher costs due to misclassification.

Compare Credit Scoring model with XGBoost

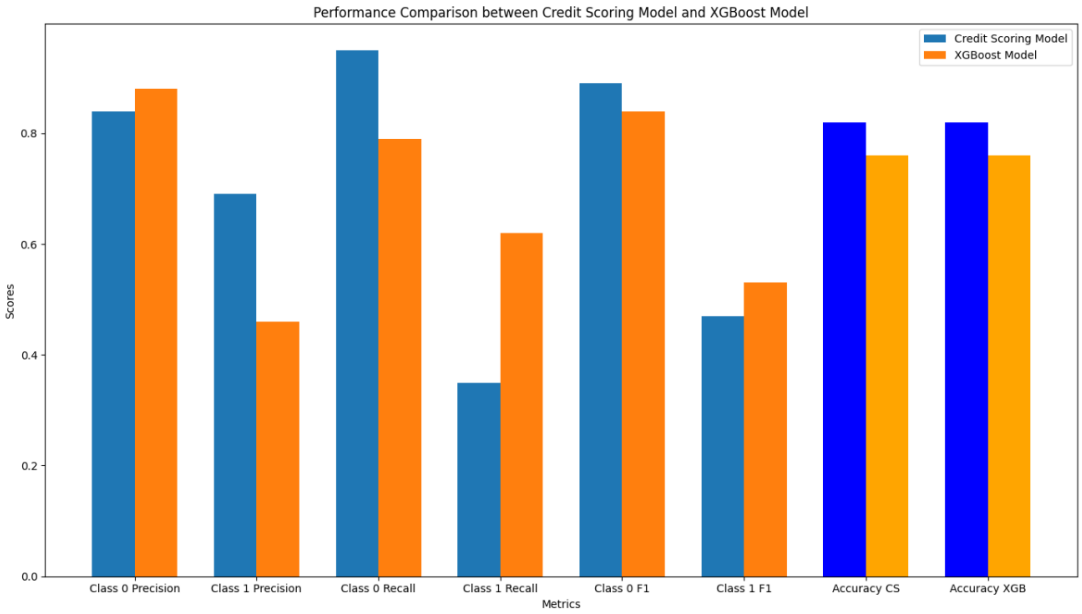

Figure 9: Compare Performance for tow Models

Figure 9 shows that Credit Scoring Model and the XGBoost Model exhibit different strengths and weaknesses in predicting loan defaults.

For class 1 (defaulters), the Credit Scoring Model has higher precision (0.69 vs 0.46), indicating better ability to accurately identify true defaulters and lower false positives. However, the XGBoost Model has higher recall (0.62 vs 0.35), suggesting it can identify more defaulters, albeit with more false positives. The XGBoost Model also has a higher F1 score (0.53 vs 0.47) for class 1, indicating better overall performance in balancing precision and recall for this class. For class 0 (non-defaulters), both models have relatively high precision and recall scores, with the Credit Scoring Model performing slightly better.

The logistic regression model is simple, easy to interpret, and fast to calculate, especially suitable for scenarios that require a high degree of interpretability. However, performance may not be as good as more complex models when dealing with complex data patterns and non-linear relationships. Logistic regression models are therefore suitable for general credit scoring, risk assessment and customer classification, especially where model decisions need to be explained to regulators.

XGBoost performs well when processing large-scale datasets and high-dimensional data. However, it is relatively complex, takes a long time to train, the model is not easy to interpret, and requires more computing resources. Therefore, the XGBoost model is suitable for scenarios that require high prediction accuracy, such as credit default

prediction, customer churn prediction, and fraud detection. Although its precision rate is slightly lower, its high recall rate makes it more suitable for identifying potentially risky customers for further manual review.

As shown in Figure 6 in previous, the number of samples in category 0 (non-default) is much larger than the number of samples in category 1 (default), which indicates that the overall default rate is lower. Such imbalanced data sets are common in the financial world because most customers are trustworthy.

In summary, the Credit Scoring Model is more suitable for business scenarios that require high accuracy and interpretability, such as bank credit scoring. XGBoost Model is more suitable for risk management scenarios that require high recall rates, such as credit default prediction and fraud detection. When selecting a model, you need to make trade-offs based on specific business needs and comprehensively consider performance indicators such as precision, recall, F1 score, and accuracy.

Conclusion

This study has comprehensively assessed the effectiveness of various credit scoring models, focusing primarily on logistic regression and XGBoost, using a dataset encompassing 30,000 credit card holders from Taiwan. We discovered significant insights about the predictive capabilities and practical applications of these models in real-world scenarios.

Our analysis highlights the paramount importance of payment history (e.g., PAY_0) in predicting default risk, with a high Information Value of 0.879. This is followed by factors such as credit limit and recent payment amounts, which also significantly impact creditworthiness assessments.

The logistic regression model demonstrated strong predictive capabilities with an ROC AUC of 0.7654, making it valuable for settings that require transparency due to its interpretability. On the other hand, the XGBoost model, with a comparable AUC of 0.76, excels in identifying potential defaulters due to its higher recall rate of 0.62 and a balanced F1-score of 0.53. This makes XGBoost suitable for applications where catching as many high-risk cases as possible is crucial, even at the expense of overall accuracy per case, as failing to identify a defaulter could result in significant financial loss.

Looking forward, the challenge of imbalanced datasets in credit scoring remains significant. We propose further exploration of undersampling techniques to balance the target variable, which could provide a more nuanced understanding of model performance across different segments and potentially lead to the development of

more robust credit scoring models.

The study's insights can guide financial institutions in enhancing credit scoring processes and risk assessment. A hybrid approach utilizing logistic regression for interpretability and XGBoost for handling complex patterns is recommended. This strategy leverages the strengths of both models, improving predictive accuracy and reliability, especially with large and diverse data sets. The integration of advanced techniques like XGBoost offers more nuanced risk assessments, allowing better credit risk management and reduction of losses due to defaults.

Appendix

Variables	bin	points
BasePoint	NaN	478
PAY_6	[-inf,-1.0)	3
	[-1.0,2.0)	6
	[2.0,inf)	-32
PAY_4	[-inf,-1.0)	2
	[-1.0,1.0)	3
	[1.0,inf)	-16
LIMIT_BAL	[-inf,40000.0)	-23
	[40000.0,140000.0)	-7
	[140000.0,380000.0)	12
	[380000.0,inf)	25
PAY_3	[-inf,-1.0)	3
	[-1.0,2.0)	5
	[2.0,inf)	-19
PAY_0	[-inf,0.0)	23
	[0.0,1.0)	36
	[1.0,2.0)	-33
	[2.0,inf)	-114
PAY_AMT1	[-inf,500.0)	-11
	[500.0,5000.0)	0
	[5000.0,17500.0)	10
	[17500.0,inf)	20
PAY_AMT3	[-inf,500.0)	-5
	[500.0,3000.0)	0
	[3000.0,5000.0)	2
	[5000.0,12500.0)	5
	[12500.0,inf)	10
PAY_AMT2	[-inf,500.0)	-10
	[500.0,2000.0)	-1
	[2000.0,5000.0)	1
	[5000.0,17500.0)	9
	[17500.0,inf)	23

Table 2: Credit Score Card Form

Log-odds formula for logistic regression models used in credit scoring models:

$$\log\_odds = -1.25 + (0.33) * PAY\_6\_woe + (0.16) * PAY\_4\_woe + (0.47) * LIMIT\_BAL\_woe + (0.19) * PAY\_3\_woe + (0.76) * PAY\_0\_woe + (0.28) * PAY\_AMT1\_woe + (0.18) * PAY\_AMT3\_woe + (0.27) * PAY\_AMT2\_woe$$