**Credit Scoring**

**Table of Contents**

[**Introduction**  . 2](#_Toc167659339)

[**Data Description and Preparation** 2](#_Toc167659340)

[**Cradit Scoring Model** 3](#_Toc167659341)

[**Predictive model** 6](#_Toc167659342)

[**Comparison of models** 7](#_Toc167659343)

[Potential application of these models 8](#_Toc167659344)

[About Population 8](#_Toc167659345)

[**Conclusion** 9](#_Toc167659346)

# 

# **Introduction** This report delves into the comparative analysis of two sophisticated credit scoring models designed to enhance decision-making in credit risk management. To build these models, we are going to use the data about the credit card information from China (Taiwan). Model 1, developed using SAS Enterprise Miner, and Model 2, crafted with the XGBoost algorithm, were both rigorously tested to ascertain their efficacy in identifying "bad" customers those likely to default within the next month. Employing a dataset split into 70% for training and 30% for validation, these models incorporate critical variables such as payment history and transaction amounts, aimed at generating a comprehensive predictive scorecard.

This analysis seeks to determine which model best meets the criteria for effective risk prediction through a meticulous examination of performance metrics including ROC AUC scores, accuracy, precision, recall, and F1 scores. Additionally, the report explores the practical applications of each model within different risk environments—ranging from conservative credit offerings to high-risk financial products—thus providing insights into their operational utility.

# **Data Description and Preparation**

|  |  |
| --- | --- |
| Variable | Description |
| ID | An identifier for each customer. |
| LIMIT\_BAL | The amount of credit given to the customer (in New Taiwan Dollar). |
| SEX | Gender of the customer (1 = male, 2 = female). |
| EDUCATION | Education level (1 = graduate school, 2 = university, 3 = high school, 4 = others). |
| MARRIAGE | Marital status (1 = married, 2 = single, 3 = others). |
| AGE | Age of the customer in years. |
| PAY\_0 to PAY\_6 | Repayment status from September 2005 to April 2005, where each month's status is coded from payment duly (-1) to a delay of nine months or more (9). |
| BILL\_AMT1 to BILL\_AMT6 | Bill statement amounts from September 2005 to April 2005. |
| PAY\_AMT1 to PAY\_AMT6 | Amounts paid by the customer from September 2005 to April 2005. |
| default.payment next month | Indicates whether the customer defaulted the following month (1 = yes, 0 = no). |

The dataset consists of credit card information from users in Taiwan, including both demographic and historical payment data. It features 25 variables spread across several categories, including…

**Data Preparation**

The data preparation process involved several steps to clean and transform the data to make it suitable for modelling. These steps included:

1. **Feature Engineering**: Created new features to better capture the credit behaviour of customers:

* AVG\_BILL\_AMT: The average amount billed over the six months.
* AVG\_PAY\_AMT: The average amount paid over the same period.
* UTILIZATION\_RATIO: The ratio of the average billed amount to the credit limit, providing an indication of how much of their available credit the customer is using.

1. **Data Transformation**:
   * Standardized features such as LIMIT\_BAL, AVG\_BILL\_AMT, and AVG\_PAY\_AMT using z-score normalization to ensure that all features contribute equally.

# **Cradit Scoring Model**

We developed a Credit Scoring model aimed at identifying "bad" customers who are likely to default in the subsequent month using SAS Enterprise Miner. We split data into two parts, 70% for training and 30% validation data. Model incorporates variables such as payment history, and transaction amounts to create a predictive scorecard as shown in figure below.

A diagram of a diagram

Description automatically generated with medium confidence

The Area ROC score is indeed a pivotal metric for evaluating the performance of credit scoring model. An AUC-ROC score falls between 0.5 and 1, where 0.5 represents a model with no discriminative ability (equivalent to random guessing), and 1 represents a perfect model that perfectly discriminates between the classes. Our model's **AUC ROC score of 0.7686** indicates that it has a good ability to differentiate between customers who will default and those who will not. This score is substantially better than random guessing but not close to perfect.

The cumulative lift chart below shows decreased probability as we increase the depth of the data. the first 20% of our population has higher (almost double) probability of being predicted correctly than the remaining population of data.

A graph of a graph

Description automatically generated

The graphs below show empirical odds trend of Training data and Validation data. The empirical odds of Training data start near a value of 1.0 when the average scorecard points are around 500. As the score increases, the odds of default steeply decrease, moving from 0 at approximately 550 points to about -3 at 650 points.

The empirical odds for validation data start near a value of 1.0 when the average scorecard points are around 500. As the score increases, the odds of default steeply decrease, moving from 0 at approximately 550 points to about -3 at 650 points.

A screenshot of a graph

Description automatically generated

The steeper slopes of the curves between approximately 580 to 620 points highlights a region where the average score is high with the less risk of defaulted payments by customer

A graph with a line

Description automatically generated

A screenshot of a table

Description automatically generated

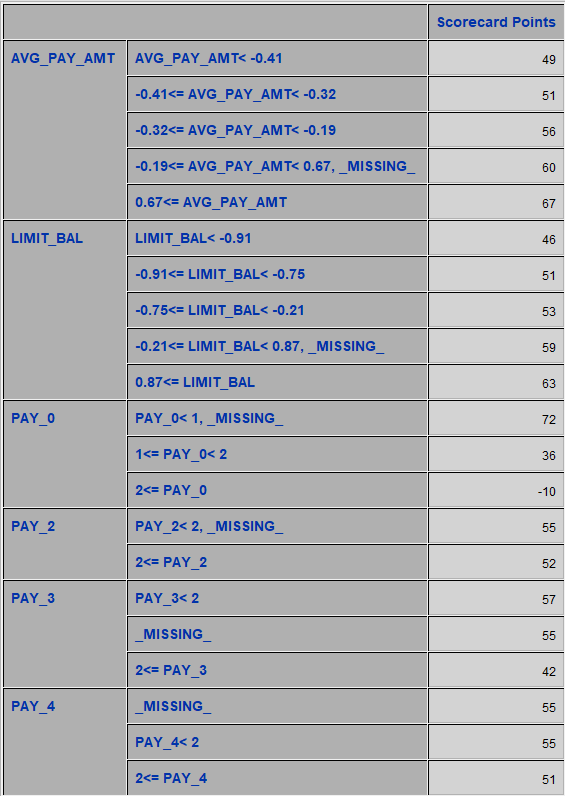
The Kolmogorov-Smirnov (KS) curve provided above reaches its maximum at around a score of 600, indicating this score as the optimal cutoff point where the difference between true good customers and bad customers is the largest.

The height of the peak, approximately 0.40, signifies a strong discriminatory power at this cutoff, suggesting that the model is effectively separating the two groups.

The odds of 3.52 at a scorecard point of 600 suggest that at this score, the probability of not defaulting is 3.52 times the probability of defaulting. Calculated using the formula 𝑝/(1−𝑝), where *p* is the probability of the positive class (non-defaulting).

**Final Scorecard**

The final scorecard given below is a comprehensive tool used in credit scoring to determine the creditworthiness of individuals.

****

A screenshot of a computer

Description automatically generated

* As the average payment amount increases to above 0.67, the scorecard awards 67 points, reflecting a stronger financial standing and reliability in making payments.
* Customers with a credit limit above 0.87 receive 63 points, indicating a perceived lower risk due to higher credit limits approved by lenders.

**Statistical Summary:**

* ROC AUC Score: 0.77
* Accuracy: Approximately 82.88% (Training), 81.70% (Validation)
* Precision: Approximately 67.85% (Training), 68.12% (Validation)
* Recall: Approximately 34.84% (Training), 35.12% (Validation)
* F1 Score: Approximately 45.92% (Training), 46.16% (Validation)

# **Predictive model**

Here we are going to make another predictive model to compare with our credit scoring model to predict bad customers. For this, we are going to use Extreme Gradient Boosting (XGBoost) method. We used the same pre-processed data that we used before in credit scoring. We kept the same 70 – 30 splits of the data (70% for training & 30% for testing).

The **AUC score of 0.7750** indicates a good ability to differentiate between customers who will default and those who will not. This score means that there is approximately a 77.5% chance that the model will rank a randomly chosen defaulting customer higher than a non-defaulting one in terms of risk. The **accuracy of 81.91%** implies that the model correctly predicts whether a customer will default or not about 82% of the time.

A line graph with orange and blue lines

Description automatically generated

* The ROC curve above shows that the true positive rate (TPR) starts increasing sharply from the origin, indicates that the model can identify a substantial proportion of actual defaulters correctly. On the other end, the false positive rate (FPR) initially remains very low as the TPR increases, which is ideal. This part of the curve indicates a high level of model specificity: non-defaulters are largely being correctly classified as such.

**Statistical Summary**

* ROC AUC Score: 0.7750 (77%)
* Accuracy: 0.8191 (81%)
* Precision: 0.6504 (65%)
* Recall: 0.3663 (36%)
* F1 Score: 0.4687 (46%)

# **Comparison of models**

|  |  |  |  |
| --- | --- | --- | --- |
| Metric | Model 1 (Training) | Model 1 (Validation) | Model 2 |
| ROC AUC Score | 0.7686 | 0.7686 | 0.775 |
| Accuracy | 82.88% | 81.70% | 81.91% |
| Precision | 67.85% | 68.12% | 65.04% |
| Recall | 34.84% | 35.12% | 36.63% |
| F1 Score | 45.92% | 46.16% | 46.87% |

* The table above shows the side-by-side comparison of both models using their statistical summary. Model 1 and Model 2 have almost identical ROC AUC scores (0.77 and 0.775, respectively). Model2 shows a slightly higher accuracy (81.91%) compared to Model 1's validation accuracy (81.70%). While the difference is minimal, it indicates that Model 2 may be slightly more reliable for correctly classifying both defaulters and non-defaulters.
* Model 1 has slightly better precision (approximately 68%) in its validation phase compared to Model 2's 65%. Higher precision indicates a lower false positive rate. Model 2 has a higher recall (36.63%) than Model 1 (approximately 35%). This suggests that Model 2 is slightly better at identifying all actual defaulting cases, though the difference is marginal. The F1 Score, which balances precision and recall, is slightly higher in Model 2 (46.87%) compared to Model 1's validation F1 Score (46.16%).

## Potential application of these models

**Model 1**

Higher precision in the validation phase (approximately 68%) suggests that Model 1 is particularly effective at minimizing false positives. Lower recall (approximately 35%) indicates that while it is conservative in predicting defaulters, it might miss identifying some actual defaulters.

Model 1 could be ideally suited for financial sectors or products where the cost of a false positive (e.g., a false default) is deemed higher than a false negative. For instance, in high-value loans or credit offers where the financial stability and trustworthiness of the customer are important, ensuring minimal false positives could be more critical than capturing every potential defaulter.

**Model 2**

With a slightly higher recall (36.63%), Model 2 demonstrates a better ability to identify actual defaulters. The slight trade-off in precision (65.04%) for higher recall means this model could generate more false positives.

Model 2's balance of high recall with adequate precision makes it suitable for environments where the business can manage or absorbing the implications of false positives. For instance, in a dynamic consumer lending environment where early detection of potential defaulters can lead to proactive account management strategies, such as restructuring of payments or offering financial counselling, this model could prevent larger losses.

## About Population

Based on the performance metrics and characteristics of the models we've analysed, the population targeted by these credit scoring models appears to be diverse in terms of creditworthiness and risk profile. The population includes individuals who range widely in their likelihood of defaulting on credit obligations, as evidenced by the necessity for models with different strengths in precision and recall.

The population includes both high-risk and low-risk borrowers. High-risk individuals are likely characterized by lower credit scores, higher utilization ratios, or poorer repayment histories. In contrast, low-risk individuals probably exhibit stable payment behaviours, lower utilization of credit lines, and higher overall creditworthiness.

Although specific demographic details aren't provided, typical credit scoring populations are often diverse in terms of age, income levels, education, and possibly other demographic factors like marital status and employment type. These factors can significantly influence credit behaviour and risk.

# **Conclusion**

In the comparative analysis of two credit scoring models, Model 1 (SAS Enterprise Miner) and Model 2 (XGBoost), each demonstrates distinct strengths tailored to different risk management needs in the financial sector. Model 1 excels in precision, achieving approximately 68% in validation, which minimizes false positives crucial in high-value lending where the implications of erroneous default predictions are significant. This model suits environments where preserving customer trust and relationship is paramount.

Conversely, Model 2 is characterized by its superior recall, approximately 36.63%, indicating its strength in identifying a larger portion of actual defaulters. This feature makes it ideal for high-risk credit products such as unsecured personal loans, where capturing as many potential defaulters as possible can significantly reduce financial losses, even at the expense of a slightly increased rate of false positives (65.04% precision).

Both models show similar ROC AUC scores, indicating good discriminative power to differentiate between defaulters and non-defaulters. However, **Model 2** may offer slightly better utility due to its higher recall and satisfactory precision, making it a versatile tool in both aggressive and conservative credit environments.