The study: Predicting Default of Credit Card Clients using Machine Learning

Ngo Gia Trang

National Economic University

E-mail: 11208187@st.neu.edu.vn

Abstract

The number of credit cards issued has increased rapidly in Taiwan and is considered a high-risk business compared to traditional bank loans. The purpose of this study is to build a prediction system. Effective risk prediction to detect possible defaults for credit card holders while also considering machine learning algorithm models. Data taken from UCI Machine Learning's Kaggle website titled "Default of Credit Card Clients Dataset" in cludes 25 variables and 29,899 observations (Any 101 observations were removed). The results show that the Neural Network and Random Forest models both give the same results and are the two most effective models in the models used in the article such as Lo gistic Regression, Decision Tree, Neive Bayes, k-NN models. Through the Logistic Regression model, we also see that the debt repayment status in the most recent month has a strong and positive impact on the credit default of Taiwanese peopl e. As a recommendation, it is crucial for the company to place a high priority on monitoring the most recent month's repayment status, billing statement amount, and the credit limit extended to borrowers.

1. INTRODUCTION

Credit card is type of payment payment card in which charges are made against a line of credit instead of the account holder's cash deposits. When someone uses a credit card to make a purchase, that person's account accrues a balance that must be paid off each month. Credit card default happens when you have become severely delinquent on your credit card payments. Missing credit card payments once or twice does not count as a default. A payment default occurs when fail to pay the Minimum Amount Due on the credit card for a few consecutive months. The global financial crisis and the increase in credit risk highlight the importance of this field (Hyeongjun Kim et al, 2020).

In Taiwan, where has suffered a credit crisis on unsecured lending products that began at the end of 2005 (Helen Lin, 2007). Growth in the issuance of credit cards in Taiwan over the past decade has been accompanied by a concerning surge in cardholder defaults. While financial institutions have shown great attention to the initial stages of credit card verification, the subsequent process of managing default risks has often been overshadowed. The escalation of delinquent payments and credit card insolvencies necessitates a fundamental shift in focus towards effective credit card risk management.

The effective management of credit card risk becomes a critical core competence for the long-term success of the financial and banking institutions (Tsung-nan Chou, 2007).

This study is aimed at predicting the case of customers default payments in Taiwan. To evaluate the correct prediction rate of the model for Taiwanese businesses by (...) through collecting data on defaulted payments, demographic factors, credit data, Payment history and billing statements of credit card customers in Taiwan in September 2005. The result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients.

2. THEORETICAL BACKGROUND

2.1 Theoretical Review

Researchers have sought to improve bankruptcy forecasting models using various quantitative approaches. For example, Ohlson (1980) was one of the first researchers to apply logistic regression analysis to default estimation. However, Begley, Ming, and Watts (1996) argued that the popular models based on Altman (1968) and Ohlson (1980) had become inaccurate and suggested the need for enhancements in the modeling of default risk.

Now, everyone are exploring artificial intelligence and machine learning tools to assess credit risk amid advances in computer technology (Barbora et al, 2017). Samuel (1959) introduced the concept of machine learning and defining it as "a discipline enabling computers to learn without explicit programming." The learning occurs when the program's performance, as measured by P, improves with experience. Regarding the application of machine learning in corporate default prediction, research is conducted in diverse ways, particularly within the field of computer science. Barboza et al (2017) contend that machine learning models demonstrate superior performance in predicting corporate bankruptcy. In most cases, the default prediction using machine learning applies a classification problem that classifies the status of the credit as being in one of two or more states, defined as normal (= 0) and default (= 1), and calculates the probability that the customer credit is in a particular state. Therefore, machine learning algorithms to solve classification problems are mainly used; Representative examples include SVMs, decision trees, and artificial neural network algorithms (Hyeongjun Kim et al, 2020).

Machine learning methods are considered to be among the most important of the recent advances in applied mathematics, with significant implications for classification problems (Tian, Shi, & Liu, 2012). Machine learning techniques assess patterns in observations of the same classification and identify features that differentiate the observations of different groups.

Machine learning studies are found across a wide range of research fields, such as Subasi and Ismail Gursoy (2010) and de Menezes, Liska, Cirillo, and Vivanco (2017) in medicine; Laha, Ren, and Suganthan (2015); Maione et al. (2016) and Cano et al. (2017) in chemistry; Bernard, Chang, Popescu, and Graf (2017) in education; and Cleofas-Sánchez, García, Marqués, and Sénchez (2016); Heo and Yang (2014); Kim, Kang, and

Kim (2015) and Gerlein, McGinnity, Belatreche, and Coleman (2016) in finance. My study focuses on the comparison of traditional statistical methods and machine learning techniques for predicting credit default. Some papers have studied credit default and machine learning (Danenas & Garsva, 2015; du Jardin, 2016; Tsai, Hsu, & Yen, 2014; Wang et al., 2014; Zhou et al., 2014), new studies, exploring different models, contexts and datasets. Shin et al (2005) apply an SVM to corporate default prediction and show that the SVM performs better than a back-propagation neural network model. Chen (2011) compares several default prediction models and claims that the SVM has high accuracy and performs well for short- and long-term default predictions. Liang et al (2016) also show that the SVM yields the best performance when predicting bankruptcy using financial ratios and corporate governance indicators. Jardin (2018) notably proposes a corporate default prediction model with ensembles of Kohonen maps and argues that this approach is highly efficient. Falavigna (2012) predicts the default risks of small Italian companies with insufficient account information using an artificial neural network algorithm. López Iturriaga and Sanz (2015) estimate and visualize banks' default risks by combining multilayer perceptrons and self-organizing maps. Azayite and Achchab (2016) improve a neural network's default prediction model by incorporating discriminant variables.

2.2 Methodology

This study uses default payments of credit card clients in Taiwan from 2005. This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

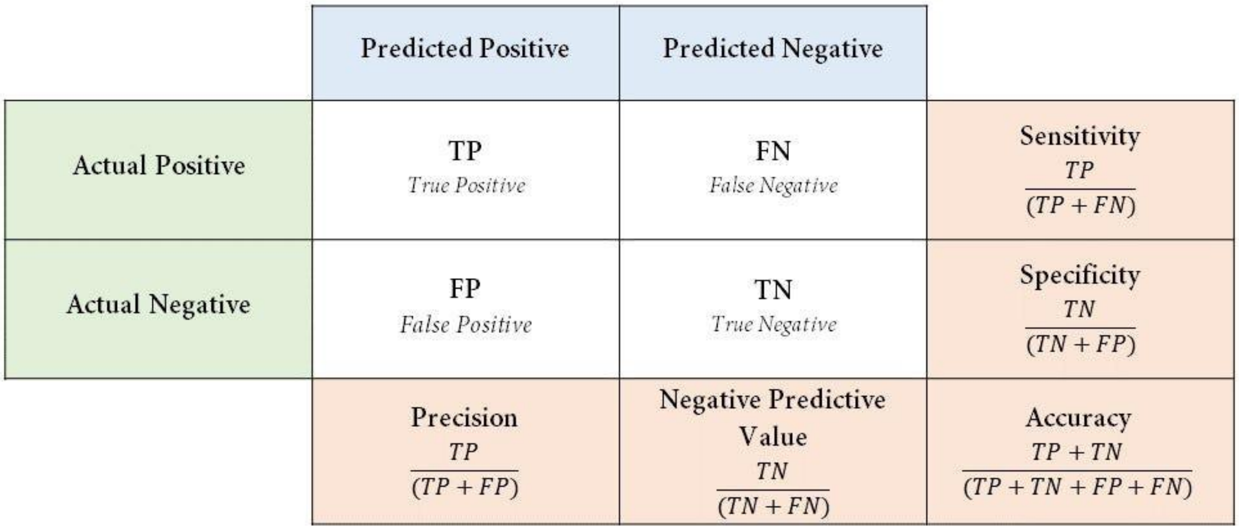
The study analyzes statistical models through the following steps:

- Step 1: Data preprocessing: descriptive statistics, handling missing values and outliers.
- Step 2: Check the correlation of independent variables.
- Step 3: Run the classification model on the training set.
- Step 4: Suitability of the model on the training set.
- Step 5: Evaluate the model on the test set.

The analysis will involve several stages, beginning with data preprocessing, followed by in-depth exploratory data analysis (EDA), supervised modeling, hyperparameter tuning, and ultimately, deriving recommendations based on the findings. Our primary objective is to predict the likelihood of default payments in the upcoming month, framing it as a classification challenge. The process will kick off with a comprehensive data analysis, followed by a detailed EDA to unearth insights and detect underlying patterns in the dataset. Subsequent steps encompass data preprocessing, encompassing cleaning and transformation to ensure data suitability for analysis. Moving forward, we'll apply various supervised modeling techniques, carefully selecting, training, and evaluating algorithms using relevant metrics. To enhance model performance, we'll conduct hyperparameter tuning, fine-tuning algorithm parameters to achieve the most optimal outcomes.

Because Machine Learning can be solved by many methods, resulting in different models. Faced with many choices, so that we can choose the most suitable model for the problem we are solving, I have used the following accuracy measures:

- Accuracy: Accuracy simply evaluates how often the model correctly predicts. Accuracy is the ratio between the number of correctly predicted data points and the total number of data points.
- Confusion Matrix: The drawback of relying solely on accuracy is that it only provides an overview of the model's prediction correctness, lacking insights into the nature and extent of prediction errors. Therefore, we require an additional evaluation method - the Confusion Matrix. The Confusion Matrix is a powerful tool for assessing model performance in classification problems. It tabulates the true positives, true negatives, false positives, and false negatives, offering a detailed breakdown of how many data points are correctly or incorrectly classified within each class.



	Predicted Positive	Predicted Negative	
Actual Positive	TP *True Positive*	FN *False Negative*	Sensitivity $\frac{TP}{(TP+FN)}$
Actual Negative	FP *False Positive*	TN *True Negative*	Specificity $\frac{TN}{(TN+FP)}$
	Precision $\frac{TP}{(TP+FP)}$	Negative Predictive Value $\frac{TN}{(TN+FN)}$	Accuracy $\frac{TP+TN}{(TP+TN+FP+FN)}$

Source: Medium

- Precision and Recall: Precision answers the question: among the data points classified by the model into the Positive class, how many data points actually belong to the Positive class. On the other hand, Recall helps us know how many real data points in the Positive class are correctly classified by the model among all real data points in the Positive class.
- F1-score: A good model is when both Precision and Recall are high, indicating that the model has little misclassification between classes as well as a low rate of missing objects belonging to the class of interest. However, the two values of Precision and Recall are often not balanced with each other

(when one value increases, the other value often tends to decrease). To evaluate both Precision and Recall at the same time, we use the F-Score measure

- Area Under the ROC curve (AUC): AUC is the area under the ROC curve, used to evaluate the classification performance of models against each other. The larger the AUC model (the closer the ROC curve is to the high left corner), the more accurate the results will be. On the contrary, the closer the ROC curve approaches the 45-degree diagonal (blue dashed line in the image above), meaning the lower the AUC, the worse the results. The higher the AUC, the easier it is for the model to correctly classify both positive and negative classes.

2.3 Models

a) Logistic Regresssion

Logistic Regression is a supervised statistical technique to find the probability of dependent variable. The Logistic Regression instead for fitting the best fit line, condenses the output of the linear function between 0 and 1. Logistic Regression is one of the simplest algorithms which estimates the relationship between one dependent binary variable and independent variables, computing the probability of occurrence of an event.

b) K-Nearest Neighbors (k-NN)

The k-nearest neighbors' algorithm is a non-parametric supervised learning method, which used for classification and regression. In k-NN regression, the output is the property value for the object. This value is the average of the values of k nearest neighbors. If k = 1, then the output is simply assigned to the value of that single nearest neighbor. There are three steps:

- Step 1: First, the distance between the new point and each training point is calculated.
- Step 2: The closest k data points are selected (based on the distance).
- Step 3: The average of these data points is the final prediction for the new point.

c) Native Bayes

The Naïve Bayes classifier is a supervised machine learning algorithm, which is used for classification tasks, like text classification. It is also part of a family of generative learning algorithms, meaning that it seeks to model the distribution of inputs of a given class or category. Unlike discriminative classifiers, like logistic regression, it does not learn which features are most important to differentiate between classes. Naïve Bayes is also known as a probabilistic classifier since it is based on Bayes' Theorem, which is represented with the following formula:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Bayes' Theorem stands out for its utilization of sequential events, where new information obtained subsequently modifies the initial probability. These probabilities are referred to as the prior probability and the posterior probability. The prior probability represents the initial probability of an event prior to being considered in a specific context or condition, often known as the marginal probability. The posterior probability, on the other hand, is the probability of an event after considering newly observed data.

d) Decision Tree

The Decision Tree algorithm is a member of the supervised learning family. What sets it apart is its versatility—it can effectively tackle both regression and classification problems, distinguishing it from other algorithms within supervised learning. The main objective behind employing a Decision Tree is to build a training model capable of predicting the class or value of the target variable. It achieves this by learning simple decision rules derived from historical data (training data). The algorithm segments a feature space, formed by a combination of explanatory variables X_1, X_2, . . ., X_p, into J non-overlapping regions R_1, R_2, . . ., R_j, and it makes the same prediction for observations belonging to the same domain.

One notable advantage of the Decision Tree algorithm lies in its model interpretability and intuitive nature. However, a significant limitation is the propensity for overfitting during the process of segmenting the feature space and generating branches. Overfitting can ultimately decrease prediction accuracy, undermining the effectiveness of the model.

e) RandomForest

The Random Forest is a popular machine learning algorithm, coined by Leo Breiman and Adele Cutler, that amalgamates the outcomes of numerous decision trees to derive a final result. Known for its ease of implementation and flexibility, this algorithm adeptly handles both classification and regression problems. Although decision trees are widely used in supervised learning, they can be susceptible to issues like bias and overfitting. However, in the Random Forest algorithm, when multiple decision trees collaborate as an ensemble, they yield more accurate predictions, particularly when the trees are uncorrelated. In the Random Forest algorithm, a specified number of explanatory variables are selected through randomization when creating a new decision tree. Typically, this selection involves choosing the square root of p, representing the total number of explanatory variables. Denoting this number as k, the Random Forest algorithm generates multiple decision trees, each utilizing k randomly chosen explanatory variables. For classification problems, the model's prediction is determined based on the most frequently predicted outcome across the decision trees.

f) Neural Network

A Neural Network is a set of algorithms designed to identify fundamental relationships within a given dataset by emulating the functioning of the human brain. Essentially, an artificial neural network is a system composed of artificial neurons. Neural Networks possess the capability to adapt to changes in input, allowing them to produce optimal results without requiring a redesign of the output criteria. This concept finds its roots in artificial intelligence and is rapidly gaining traction in the development of electronic trading systems.

In the realm of finance, artificial neural networks play a crucial role in supporting various processes including algorithmic trading, time series forecasting, stock classification, credit risk modeling, and indicator development. They mimic the behavior of human neural networks, where each neuron in an artificial neural network is a mathematical function responsible for collecting and categorizing information based on a specific structure. Neural Networks seamlessly integrate with standard statistical methods such as curve plotting and regression analysis.

A Neural Network comprises layers that house interconnected nodes, with each node resembling a perceptron structured similarly to a multiple linear regression function. In a multilayer perceptron, these nodes are organized into interconnected layers. The input layer gathers input samples, while the output layer collects classifications or output signals corresponding to the reflected input samples.

3. DATA

3.1 Data Analysis

The original dataset "Default of Credit Card Clients Dataset" was presented on Kaggle by the UCI Machine Learning Repository. There are 25 variables and 29899 observations:

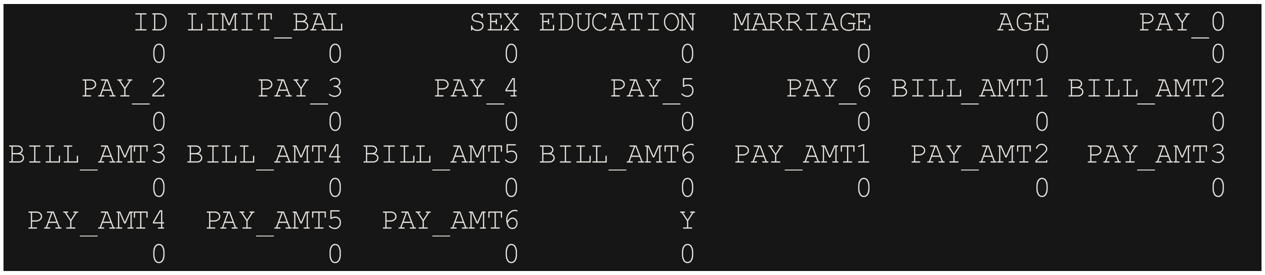
Table 1. Default of Credit Card Clients Dataset

Variable name	Description
ID	ID of each client
LIMIT_BAL	Amount of given credit in NT dollars
SEX	Gender (1=male, 2=female)
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 to PAY_6	(-2 = No consumption, -1 = paid in full, 0 = use of revolving credit (paid minimum only), 1 = payment delay for one month, 2 = payment delay for two

	months, ... 8 = payment delay for eight months, 9 = payment delay for nine months and above)
BILL_AMT6, BILL_AMT5, BILL_AMT4, BILL_AMT3, BILL_AMT2, BILL_AMT1	Amount of bill statement from April to September (Dollar)
PAY_AMT6, PAY_AMT5, PAY_AMT4, PAY_AMT3, PAY_AMT2, PAY_AMT1	Amount of previous payment from April to September (Dollar)
default.payment.next.month	Default payment (1=yes, 0=no)

MARRIAGE and EDUCATION have undefined data. Looking at the description table of the variables above, we can see that the variable 'MARRIAGE' takes values 1, 2 and 3, and the variable 'EDUCATION' takes values from 1 to 6. However, the dataset contains undefined values denoted by 0 for these variables. Upon examination, these undefined data seem accurate and should be included in the model. We will group the undefined and unknown values into the category 'Others'.

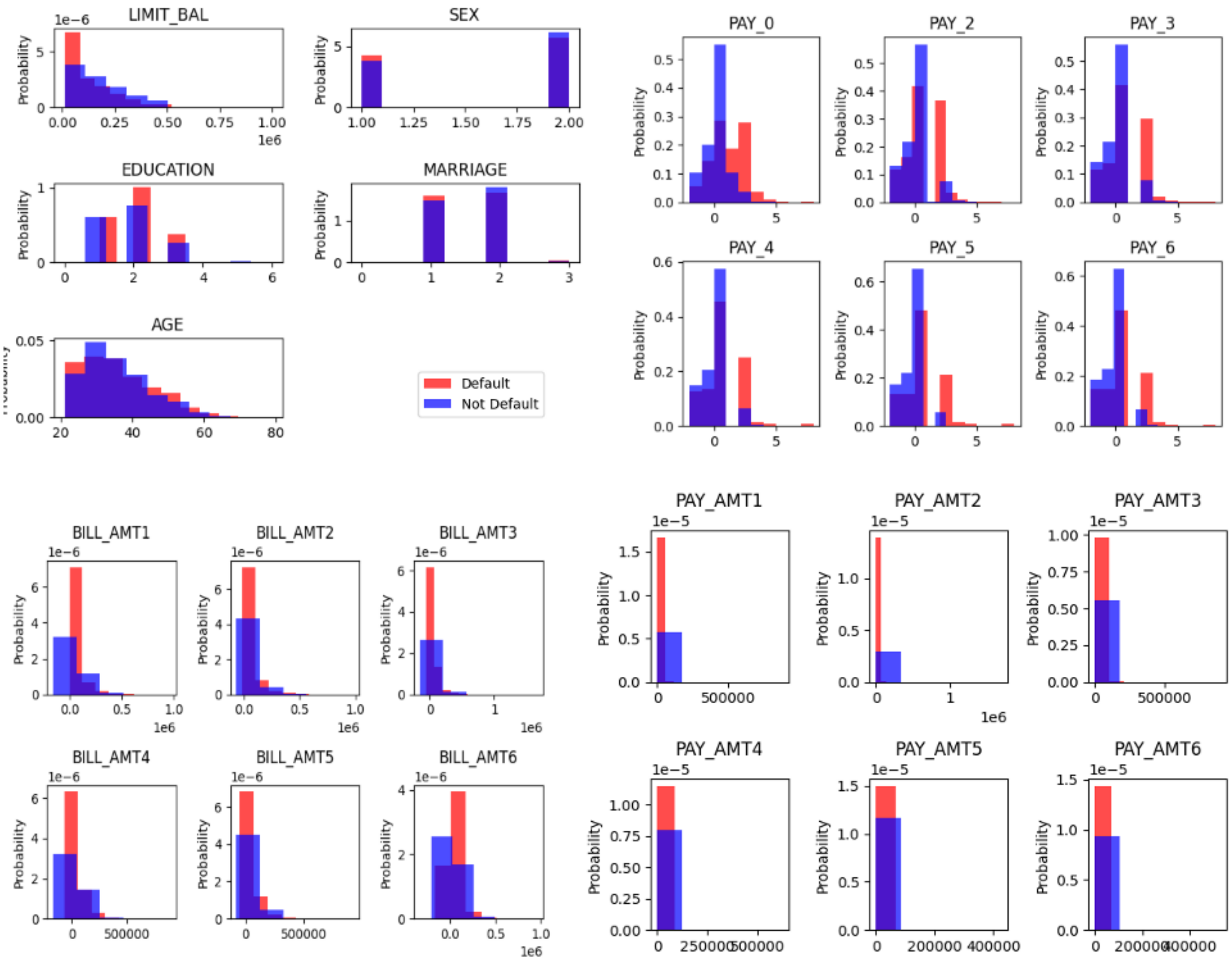
Checking missing data gives the result that there are no variables with missing data.



```
        ID LIMIT_BAL       SEX EDUCATION  MARRIAGE       AGE     PAY_0
         0         0         0         0         0         0         0
     PAY_2     PAY_3     PAY_4     PAY_5     PAY_6 BILL_AMT1 BILL_AMT2
         0         0         0         0         0         0         0
BILL_AMT3 BILL_AMT4 BILL_AMT5 BILL_AMT6  PAY_AMT1  PAY_AMT2  PAY_AMT3
         0         0         0         0         0         0         0
  PAY_AMT4  PAY_AMT5  PAY_AMT6         Y
         0         0         0         0
```

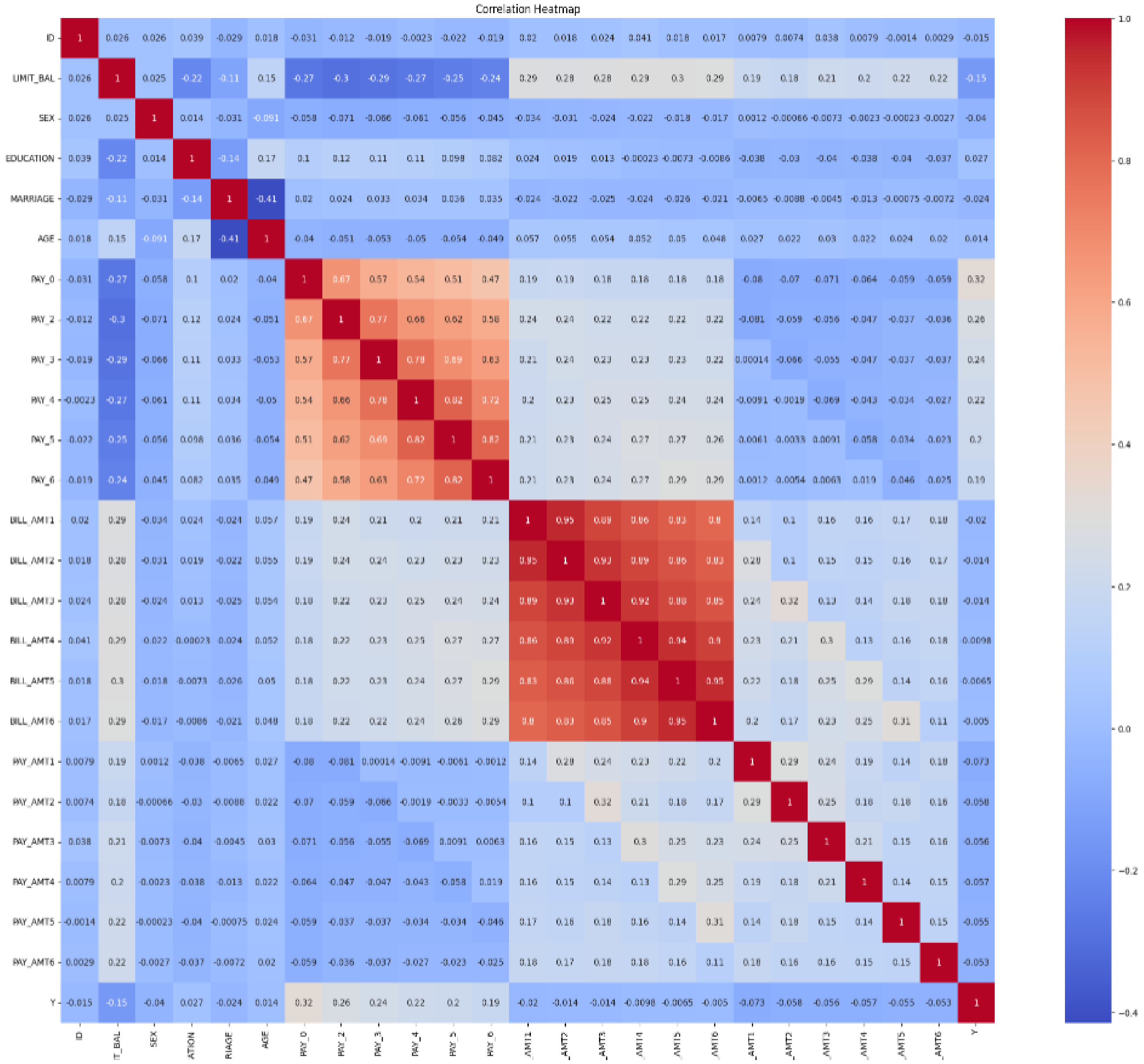
High cardinality for category variables may impact the model performance. Hence, I need to check for category variable. The result is no cardinality issue. Now I can confirm the dataset is clean.

Graph 1. Default rate is based on client background



The greater the amount of credit extended to an individual, the higher the likelihood of default. Moreover, higher levels of education are associated with an increased susceptibility to credit default. Individuals opting for revolving credit, paying only the minimum amount, exhibit a higher incidence of both default and non-payment compared to their counterparts. Clients who have delay payment in the latest month tends to have default payment next month.

Graph 2. Correlation matrix



The variables within the "Amount of Bill Statement" (BILL_AMT) group exhibit high positive correlations, as do the variables within the "Repayment Statistics" (PAY) group. This may cause multicollinearity when we run the model.

3.2 Data Processing

Information value (IV) is a measure of the predictive power of a qualitative variable for predicting the outcome (or label) of a classified variable. WoE is the intermediate to calculate IV (cumulative IV to be precise). According to Siddiqi (2005), by convention the values of the IV statistic in credit scoring can be interpreted as follows. If the IV statistic is:

- Less than 0.02, then the predictor is not useful for modeling (separating the Goods from the Bads)
- 0.02 to 0.1, then the predictor has only a weak relationship to the Goods/Bads odds ratio
- To 0.3, then the predictor has a medium strength relationship to the Goods/Bads odds ratio
- To 0.5, then the predictor has a strong relationship to the Goods/Bads odds ratio
- 0.5, suspicious relationship

I noted that there could be multicollinearity in Pay and Bill_Amt. So, I will only keep the highest variable in modelling.

Table 2. IV of dataset's variable

Variable	IV
ID	0,0165
LIMIT_BAL	0,1791
AGE	0,0191
SEX	0,0068
EDUCATION	0,0179
MARRIAGE	0,0034
PAY_0	0,8479
BILL_AMT1	0,0238
PAY_AMT1	0,1896
PAY_AMT2	0,1715
PAY_AMT3	0,1359
PAY_AMT4	0,1153
PAY_AMT5	0,1041
PAY_AMT6	0,1078

Based on the information above, we see that there are 5 variables IV < 0.02 (ID, AGE SEX, EDUCATION, MARRIAGE). From there, eliminate the above 5 variables so that the predictor is useful for modeling.

4. FINDINGS

a) Model

After cleaning the data, I built a logistic regression model using the python tool and got the results below:

$$\log\left(\frac{p_i}{1-p_i}\right) = -0.1817 - 0.1019 \times LIMIT_{BAL_i} + 0.6936 \times PAY_{0_i} \\ - 0.0972 \times BILL_{AMT1_i} - 0.1706 \times PAY_{AMT1_i} - 0.1892 \times PAY_{AMT2_i} \\ - 0.0657 \times PAY_{AMT3_i} - 0.0414 \times PAY_{AMT4_i} - 0.0312 \times PAY_{AMT5_i} \\ - 0.0622 \times PAY_{AMT6_i}$$

Looking at this model, we observe that only the variable 'repayment status in September' has a positive effect, while all other variables have a negative effect on the probability of defaulting credit card.

b) Confusion matrix

Using the python tool, we obtain the following confusion matrix values:

- **Logistic Regression (LR)**

		Predict	
		Good	Bad
Actual	Good	4683	2304
	Bad	669	1314

- **K-NN**

		Predict	
		Good	Bad
Actual	Good	4681	2306
	Bad	801	1182

- **Naïve Bayes (NB)**

		Predict	
		Good	Bad
Actual	Good	1189	5798
	Bad	125	1858

- **Decision Tree (DT)**

		Predict	
		Good	Bad
Actual	Good	5619	1368
	Bad	1151	832

- **Random Forest (RF)**

		Predict	
		Good	Bad
Actual	Good	6298	689
	Bad	1116	867

- **Neural Network (NN)**

		Predict	
		Good	Bad
Actual	Good	6298	689
	Bad	1116	867

Naive Bayes and Logistic Regression models have the highest number of accurate forecasts but also have a large number of type I errors. Besides, the Random Forest and Neural Network models show that they both give the same results and have a superior level of effectiveness compared to other models. From the above confusion matrices, applying the formulas we obtain table 3.

Table 3. Accuracy measures

	Precision	Recall	F1-score	Accuracy	Specificity	AUC
LR	0.36	0.66	0.23	0.67	0.67	**0.67**
k-NN	0.34	0.60	0.22	0.65	0.67	0.63
NB	0.24	**0.94**	0.19	0.34	0.17	0.55
DT	0.38	0.42	0.20	0.72	0.80	0.61
RF	**0.56**	0.44	**0.24**	**0.80**	**0.90**	**0.67**
NN	**0.56**	0.44	**0.24**	**0.80**	**0.90**	**0.67**

Logistic Regression, Random Forest and Neural Network,all have the same and highest AUC index (0.67). But Random Forest and Neural are having very high accuracy up to 80%. Other indicators of the above two models all produce the highest v alues, such as precision (56%), F1 (24%) and Specificity up to 90%. From there, we can see that the above two methods are very effective for predicting credit default rates in Taiwan. Logistic regression produces relatively good results, not falling into the very good or very bad range. Therefore, the Logit Model is a popular model for banks and other financial institutions in the near future due to its simple understandability.

5. CONCLUSIONS & DISCUSSION

In this study, we conducted an analysis of credit card user data in Taiwan to address the challenge of predicting credit default risks. Additionally, we reviewed prevalent research methodologies, categorizing them into statistical models and machine learning algorithms. Notably, while reviewing the literature, I observed a prevalent approach of treating corporate default prediction as a classification problem in machine learning research. Conversely, in financial engineering, structural and risk models analyze corporate defaults in a sequential manner. It's worth noting that within the domain of corporate bankruptcy prediction using machine learning methods, macroeconomic factors are often rare or overlooked, despite their significant influence. While the financial condition of a company is a primary determinant of corporate default, the impact of macroeconomic conditions should not be underestimated.

We found that the Neural Network and Random Forest models both yielded better results than the models applied in this analysis in terms of prediction accuracy. Through the Logistic Regression model, we also see that the debt repayment status in the most re cent month has a strong and positive impact on the credit default of Taiwanese people.

However, the collection of credit card holder data is limited to the card issuer. Therefore, only a portion of the cardholder data collected from a financial institution was applied in this study to evaluate the performance of the integrated model.

References

Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance*, *23*(4), 589–609. https://doi.org/10.2307/2978933

Azayite, F. Z., & Achchab, S. (2016). Hybrid Discriminant Neural Networks for Bankruptcy Prediction and Risk Scoring. *Procedia Computer Science*, *83*, 670–674. https://doi.org/10.1016/j.procs.2016.04.149

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, *83*, 405–417. https://doi.org/10.1016/j.eswa.2017.04.006

Begley, J., Ming, J., & Watts, S. (1996). Bankruptcy classification errors in the 1980s: An empirical analysis of Altman’s and Ohlson’s models. *Review of Accounting Studies*, *1*(4), 267–284. https://doi.org/10.1007/bf00570833

Bernard, J., Chang, T.-W., Popescu, E., & Graf, S. (2017). Learning style Identifier: Improving the precision of learning style identification through computational intelligence algorithms. *Expert Systems with Applications*, *75*, 94–108. https://doi.org/10.1016/j.eswa.2017.01.021

Cano, G., Garcia-Rodriguez, J., Garcia-Garcia, A., Perez-Sanchez, H., Benediktsson, J. A., Thapa, A., & Barr, A. (2017). Automatic selection of molecular descriptors using random forest: Application to drug discovery. *Expert Systems with Applications*, *72*, 151–159. https://doi.org/10.1016/j.eswa.2016.12.008

Chen, M.-Y. (2011). Bankruptcy prediction in firms with statistical and intelligent techniques and a comparison of evolutionary computation approaches. *Computers & Mathematics with Applications*, *62*(12), 4514–4524. https://doi.org/10.1016/j.camwa.2011.10.030

Cleofas-Sánchez, L., García, V., Marqués, A. I., & Sánchez, J. S. (2016). Financial distress prediction using the hybrid associative memory with translation. *Applied Soft Computing*, *44*, 144–152. https://doi.org/10.1016/j.asoc.2016.04.005

Danenas, P., & Garsva, G. (2015). Selection of Support Vector Machines based classifiers for credit risk domain. *Expert Systems with Applications*, *42*(6), 3194–3204. https://doi.org/10.1016/j.eswa.2014.12.001

de Menezes, F. S., Liska, G. R., Cirillo, M. A., & Vivanco, M. J. F. (2017). Data classification with binary response through the Boosting algorithm and logistic regression. *Expert Systems with Applications*, *69*, 62–73. https://doi.org/10.1016/j.eswa.2016.08.014

du Jardin, P. (2018). Failure pattern-based ensembles applied to bankruptcy forecasting. *Decision Support Systems*, *107*, 64–77. https://doi.org/10.1016/j.dss.2018.01.003

Falavigna, G. (2012). Financial ratings with scarce information: A neural network approach. *Expert Systems with Applications*, *39*(2), 1784–1792. https://doi.org/10.1016/j.eswa.2011.08.074

Gerlein, E. A., McGinnity, M., Belatreche, A., & Coleman, S. (2016). Evaluating machine learning classification for financial trading: An empirical approach.

Expert Systems with Applications, *54*, 193–207. https://doi.org/10.1016/j.eswa.2016.01.018

Heo, J., & Yang, J. Y. (2014). AdaBoost based bankruptcy forecasting of Korean construction companies. *Applied Soft Computing*, *24*, 494–499. https://doi.org/10.1016/j.asoc.2014.08.009

Kim, H., Cho, H., & Ryu, D. (2020). Corporate Default Predictions Using Machine Learning: Literature Review. *Sustainability*, *12*(16), 6325. https://doi.org/10.3390/su12166325

Kim, M.-J., Kang, D.-K., & Kim, H. B. (2015). Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction. *Expert Systems with Applications*, *42*(3), 1074–1082. https://doi.org/10.1016/j.eswa.2014.08.025

Laha, D., Ren, Y., & Suganthan, P. N. (2015). Modeling of steelmaking process with effective machine learning techniques. *Expert Systems with Applications*, *42*(10), 4687–4696. https://doi.org/10.1016/j.eswa.2015.01.030

Liang, D., Lu, C.-C., Tsai, C.-F., & Shih, G.-A. (2016). Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. *European Journal of Operational Research*, *252*(2), 561–572. https://doi.org/10.1016/j.ejor.2016.01.012

林靜慧. (2007, January 1). *台灣消費性信用緊縮之研究─以雙卡風暴為例*. Airiti. https://www.airitilibrary.com/Publication/alDetailedMesh1?DocID=U0004-2910200810315710

López Iturriaga, F. J., & Sanz, I. P. (2015). Bankruptcy visualization and prediction using neural networks: A study of U.S. commercial banks. *Expert Systems with Applications*, *42*(6), 2857–2869. https://doi.org/10.1016/j.eswa.2014.11.025

Machine Learning, U. (2016). *Default of Credit Card Clients Dataset*. Www.kaggle.com. https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset/code

Maione, C., de Paula, E. S., Gallimberti, M., Batista, B. L., Campiglia, A. D., Jr, F. B., & Barbosa, R. M. (2016). Comparative study of data mining techniques for the authentication of organic grape juice based on ICP-MS analysis. *Expert Systems with Applications*, *49*, 60–73. https://doi.org/10.1016/j.eswa.2015.11.024

Ohlson, J. A. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research*, *18*(1), 109–131. https://doi.org/10.2307/2490395

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, *3*(3), 210–229. https://doi.org/10.1147/rd.33.0210

Siddiqi, N. (2005). Credit risk scorecards: developing and implementing intelligent credit scororing. Wiley.

Shin, K.-S., Lee, T. S., & Kim, H. (2005). An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications*, *28*(1), 127–135. https://doi.org/10.1016/j.eswa.2004.08.009

Subasi, A., & Ismail Gursoy, M. (2010). EEG signal classification using PCA, ICA, LDA and support vector machines. *Expert Systems with Applications*, *37*(12), 8659–8666. https://doi.org/10.1016/j.eswa.2010.06.065

Tian, Y., Shi, Y., & Liu, X. (2012). RECENT ADVANCES ON SUPPORT VECTOR MACHINES RESEARCH. *Technological and Economic Development of Economy*, *18*(1), 5–33. https://doi.org/10.3846/20294913.2012.661205

Tsai, C.-F., Hsu, Y.-F., & Yen, D. C. (2014). A comparative study of classifier ensembles for bankruptcy prediction. *Applied Soft Computing*, *24*, 977–984. https://doi.org/10.1016/j.asoc.2014.08.047

Tsung-nan, C. (2007). *Sci-Hub | A Novel Prediction Model for Credit Card Risk Management. Second International Conference on Innovative Computing, Informatio and Control (ICICIC 2007) | 10.1109/ICICIC.2007.68*. Sci-Hub.se. https://sci-hub.se/https://ieeexplore.ieee.org/abstract/document/4427856

Wang, G., Ma, J., & Yang, S. (2014). An improved boosting based on feature selection for corporate bankruptcy prediction. *Expert Systems with Applications*, *41*(5), 2353–2361. https://doi.org/10.1016/j.eswa.2013.09.033

Zhao, Z., Xu, S., Kang, B. H., Kabir, M. M. J., Liu, Y., & Wasinger, R. (2015). Investigation and improvement of multi-layer perceptron neural networks for credit scoring. *Expert Systems with Applications*, *42*(7), 3508–3516. https://doi.org/10.1016/j.eswa.2014.12.006