(IJISE) 2017, Vol No. 6, Jul-Dec **e-ISSN: 2454-6402, p-ISSN: 2454-812X**

DEVELOPING A MACHINE LEARNING BASED SMART MODEL TO EFFECTIVELY ANALYZE AND GRADE CREDIT RISK

Somya Panchal

Gargi College, University of Delhi

ABSTRACT

Paying for goods and services with a credit card is quick and easy. However, the likelihood of late payments increases as debt grows over time, especially in light of the pandemic and rising unemployment. The global financial crisis of 2007-2008 illustrates the need for commercial institutions to anticipate their customers' credit risk. As the quantity of Mastercard clients has expanded, banks have been confronting a heightening charge card default rate. This paper proposes a clever viable procedure to section clients by their anticipated likelihood of defaulting on instalments to assist monetary establishments with surveying risk before giving charge cards. We used the Taiwan credit card default dataset to present our method and findings. However, the proposed method can be applied to credit card default datasets from other nations because it has been generalized.

INTRODUCTION

Recent technological advancements, e-commerce developments, and socioeconomic shifts, such as India's demonetization [1], have brought the world closer to a cashless economy. The COVID-19 pandemic and the ease of online ordering have made credit cards and other forms of payment increasingly popular. Local pharmacies, online retailers, and grocery stores prefer card-based or digital payments to cash to avoid the hassle of maintaining cash.

Although credit cards are convenient and simple, hundreds of thousands have suffered mental and financial losses due to misuse [2]. Because you can get new credit cards in minutes, many consider credit cards equivalent to "free" money. According to several studies, credit card users spend up to 100 per cent more than they normally do [3]. Even worse, credit card debt accrues a lot of interest over time, making it harder to pay it back as the amount owed rises and the interest keeps rising. Credit card debt can quickly become unmanageable as a result of these factors.

Unemployment, inflation, and the global economy are just a few of the other factors that indirectly impact a person's credit health. People now have to choose between paying for their day-to-day needs and making credit card repayments, which reduces their ability to repay credit card debt as inflation and living costs rise [4].

The rise in unemployment, particularly since the pandemic, also impacts this. As a result, credit card defaults and late payments eventually occur. Risk management comes into play in this situation [5].

(IJISE) 2017, Vol No. 6, Jul-Dec **e-ISSN: 2454-6402, p-ISSN: 2454-812X**

Banks and other financial institutions often use risk management to evaluate a customer's ability to pay on time. Banks can predict a customer's credit risk using financial data like financial statements, customer transactions, and repayment records. Machine Learning is the obvious answer to this issue because of this industry's large amount of data. Customer credit card risk has been accurately predicted using Machine Learning [6]. Financial institutions can then use these predictions to reduce credit lines and prevent defaults, resulting in annual savings of millions of dollars [7]. Nonetheless, the misfortune of losing a potential record that may not default can likewise be high. This necessitates a solution that reduces false positives and accurately predicts risks.

The common focus of popular research is predicting the probability of credit card default versus no default. Still, despite its academic interest, this is not an outcome that financial institutions can use. For instance, predicting a customer's likelihood of defaulting on a credit card may not help financial institutions determine a customer's final interest rate. This paper proposes a viable procedure to portion likely clients by their anticipated likelihood of defaulting on instalments. The Taiwan credit card default dataset [8] is used to evaluate various Machine Learning methods for predicting default probabilities. After that, we suggest using an iterative Decision Tree method to divide these probabilities into groups of customers with very low to very high risk. This would provide institutions with a straightforward and concrete method for determining the default risk of potential customers.

DATA

A credit card default dataset for Taiwan clients [8] from the UCI Machine Learning Repository with payment data for October 2005 was used in this study. There are 30000 instances and 24 variables in the dataset, 23 of which are independent. The dependent variable is a binary variable labelled "default payment for next month." The default status of the customer is recorded as either Yes (shown as 1) or No (shown as 0). The following information can be found in Table I. for the remaining 23 variables Algorithm

Table 1

Columns	Description	Values
X1	Amount of the given credit	NT dollar
X2	Gender	1 (Male), 2 (Female)
X3	Eduction	1 (graduate school), 2 (university), 3 (high school), 4 (others)
X4	Marital status	1 (married), 2 (single), 3 (others)
X5	Age	Year
X6 - X11 (X6 - September 2005, X7 - August 2005, and so on till X11 - April 2005)	History of past payment	-1 (paid duly), 1 (payment delay for one month), 2 (payment delay for two months) and so on.
X12 - X17 (X12 - September 2005, X13 - August 2005, and so on till X17 - April 2005)	Amount of bill statement	NT dollar
X18 - X23 (X18 - September 2005, X19 - August 2005, and so on till X23 - April 2005)	Amount of previous payment	NT dollar

A. Decision Tree Classifier the Decision Tree (DT) classifier, which is a tree-structured classifier with "if" statement-like branches that divide the dataset based on how well the condition classifies the data was developed by Algorithm A. [16].

The depth of the tree is frequently limited to prevent overfitting. Internal nodes, branches, and leaf nodes represent the outcome or class, decision rules, and dataset features. Like the "if" statement, the classifier selects the feature that best divides the data into classes at each node.

B. K-Nearest Neighbor Classifier The K-Nearest Neighbor (KNN) classifier, as described in [17], determines how similar the brand-new data is to its neighbouring k data points. The KNN classifier looks for the pattern closest to the given unknown data point based on how close it is in terms of distance. The larger part of the class among the adjoining k information focuses on choosing the new or obscure information class. Various metrics, such as the Manhattan and Euclidean distances, calculate the distance between the new and k neighbouring points.

(IJISE) 2017, Vol No. 6, Jul-Dec **e-ISSN: 2454-6402, p-ISSN: 2454-812X**

C. XGBoost Classifier

XGBoost [19] utilizes angle drop to get to the next level feeble tree-based students gradually. It is the most effective model because it combines numerous decision tree models. It is a good choice for building accurate models on large datasets without access to computing resources because it can perform parallel computation on a single machine. Due to its distributed implementation of the Gradient Boosting architecture, XGBoost outperforms Gradient Boosting in speed and performance.

PROPOSED APPROACH

Our most memorable perception was that the dataset is imbalanced the reliant variable contained around 6636 clients(22.12%) with a default instalment ('yes') and the excess 23 thousand 300 64 clients (77.88%) with no default instalment ('no').

Therefore, to correct the imbalance, we divided the dataset into a 70:30 train: We used Decision tree, K.NN, Random forest, and XGBoost classifiers with various parameters and 10-fold cross-validation to evaluate the classifier once against the dataset U and once against the dataset D. We analysed the results to determine the classifier and the optimal set of parameters that achieved the best AUC. Test split and then resampled the training data to build an upsampled dataset U and a downsampled dataset D. On the dataset U. Fig., we discovered that XGBoost produced the best results. The diagrammatic representation of our suggested method is shown in l.

To generate the probability of payment (probability of class 0, or probability of "no" default) on dataset U, we used the XGBoost classifier with the parameters that produced the best results on the training set for dataset U. We then created dataset B with this probability and target label. We then created ten bins by multiplying the probability by 1000 (0 to 100, 101 to 200, and so on). The number of I's (the default), and O's (not the default) within a specific bin interval was then determined. As shown in Fig., we utilized these data and plotted a histogram. 2, from which we concluded that the bins at the end of the histogram contained the greatest number of O's and l's. The 0 to 100 bins, which represents a chance of up to 10% that the customer will make a payment and contains the greatest number of clients who defaulted on their loans, was also found to contain the greatest number of O's and l's. Similarly, the 900 to 1000 bin contains the maximum number of customers who did not default on their loans and represents a chance of 90-100% that the customer will make a payment. This brought back the XGBoost classifier's ability to predict the payment likelihood.

RESULTS OF EXPERIMENTS

To evaluate the Machine Learning Algorithms, we conducted a series of experiments described in this section. Our study used decision trees, K-nearest neighbours, random forests, and XGBoost classifiers. To advance the exhibition of these calculations, we tuned the boundaries of these calculations. On the U and D datasets, we used 10-fold cross-validation to evaluate the Machine Learning algorithms.

(IJISE) 2017, Vol No. 6, Jul-Dec e-ISSN: 2454-6402, p-ISSN: 2454-812X

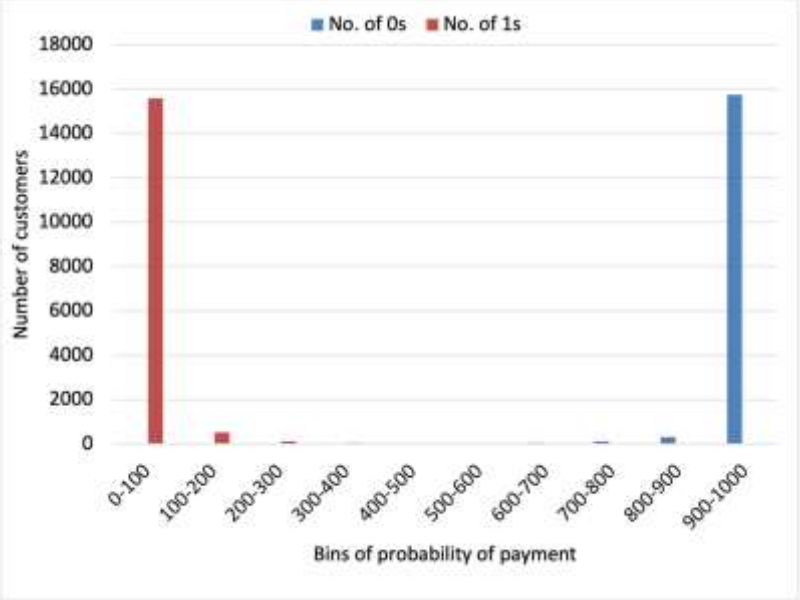

Fig 1: Histogram

Fig. The performance of the chosen algorithms on dataset U is shown in Figure 4. We can see that the AUC for XGBoost is the highest, at 0.97, while the AUC for Decision Tree is the lowest, at 0.88.

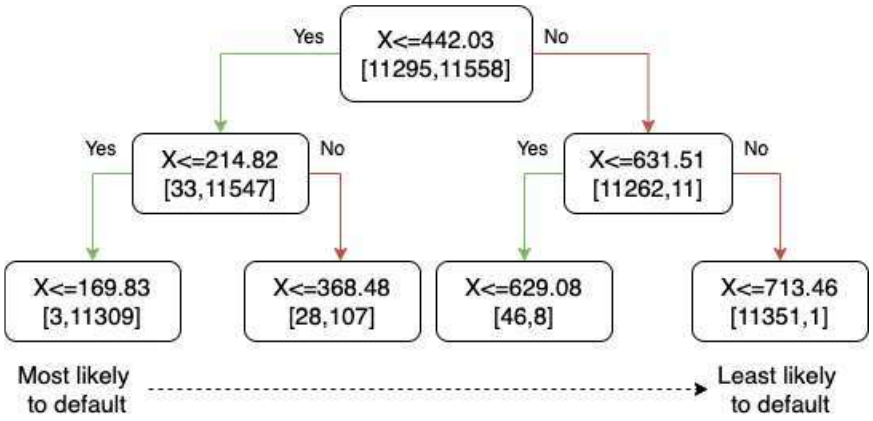

Fig. 2. Decision Tree of training dataset

Fig. The performance of the selected algorithms on dataset D is shown in Figure. We can see that the performance of dataset D was lower than that of dataset U. The AUC for XGBoost is the highest, at 0.79, while the AUC for Random The Forest is the lowest, at 0.78.

(IJISE) 2017, Vol No. 6, Jul-Dec e-ISSN: 2454-6402, p-ISSN: 2454-812X

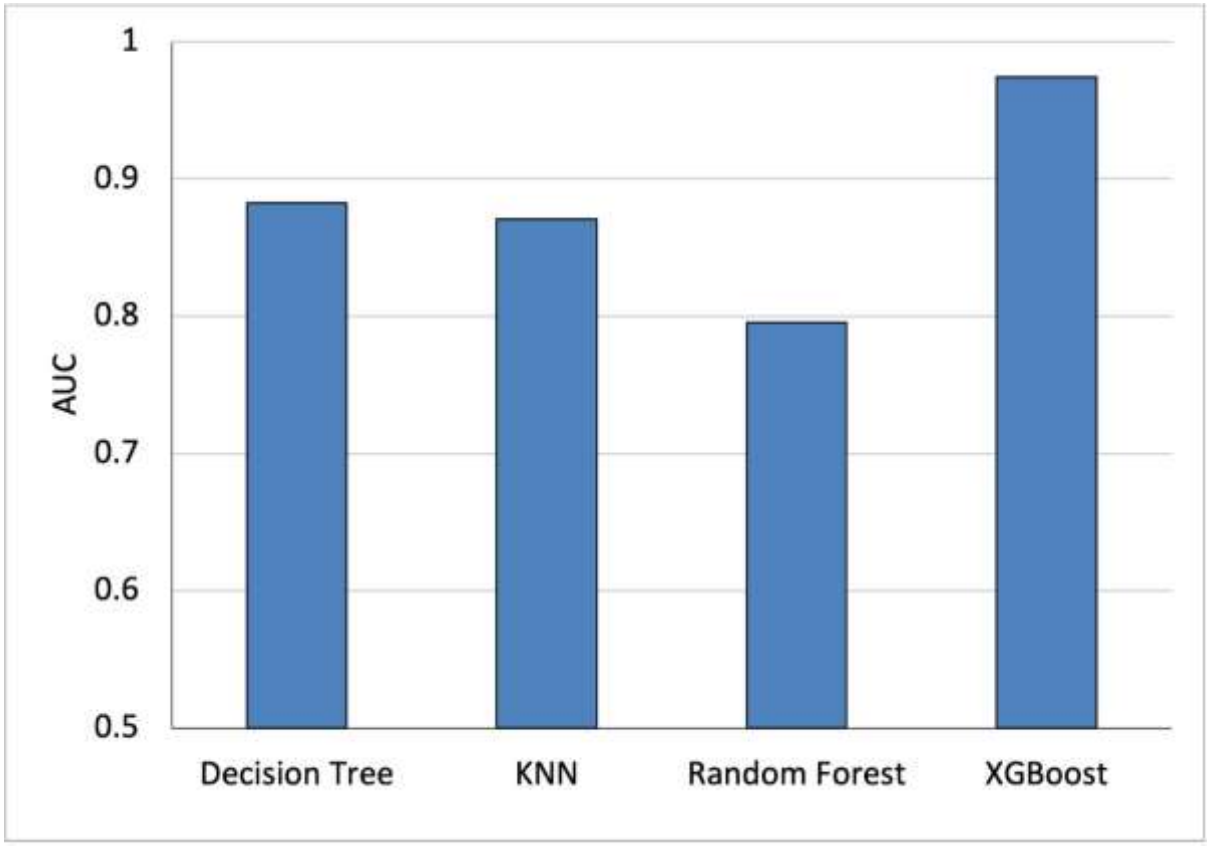

Fig. 3. AUC of Upsampled dataset

We used the XGBoost algorithm and the parameters with the highest AUC, as well as the upsampled dataset, for the remainder of the study and analysis because the observations demonstrate that XGBoost performed best with dataset U. We prepared the model on the dataset U utilizing the XGBoost calculation and tried utilizing 10-crease cross-approval which provided us with a typical AUC of 0.97. Utilizing the dataset U and XGBoost calculation, we made a dataset B that contained the anticipated likelihood of default and the objective variable.

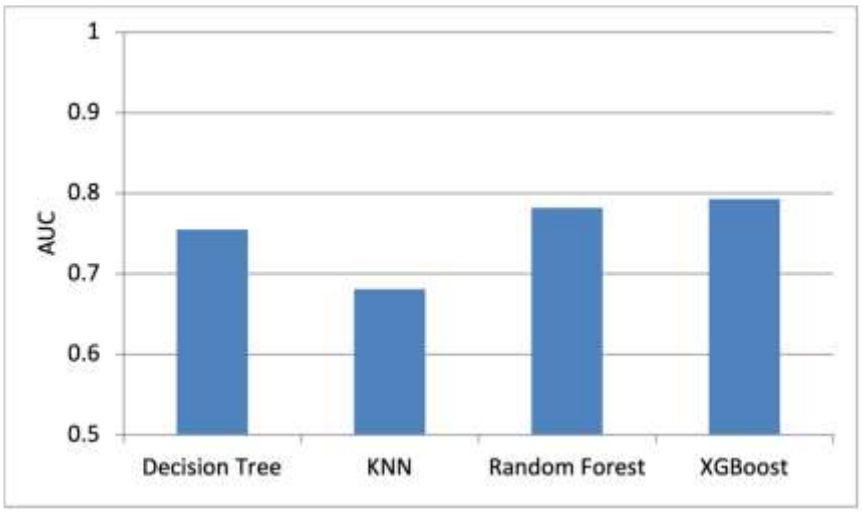

Fig. 4. AUC of Downsampled dataset

(IJISE) 2017, Vol No. 6, Jul-Dec e-ISSN: 2454-6402, p-ISSN: 2454-812X

The split points from dataset B were then applied to the testing dataset S, which the Decision Tree depicts in Fig., using the iterative Decision Tree technique described in the preceding section. 6. The leftmost subset (lowest probability of payment) has 60% of records with an actual 1 (default) value. In comparison, the rightmost subset (highest probability of payment) has 85% of records with an actual value of 0 (non-default). This can be seen by comparing the tree's leaf nodes on the left and right. Using this method, we can divide the customers in the testing dataset and place them on a scale from the lowest to the highest risk of default. This makes it possible for financial institutions to evaluate risk in a way that is more objective than simply looking at the probability of default, which can be interpreted subjectively.

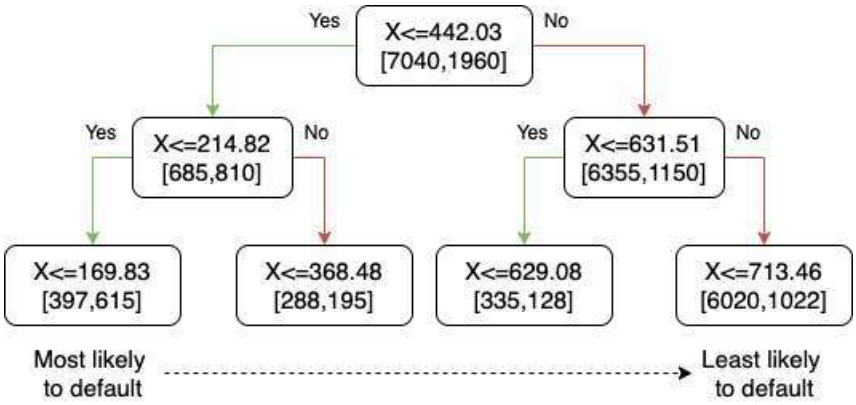

Fig. 5. Decision Tree of testing dataset

CONCLUSION

As we saw, there is a lot of research that uses machine learning to predict whether or not a person will pay their credit card balance, home mortgage, or car loan on time.

However, no machine learning method can accurately predict such events. Even though financial institutions like to avoid taking risks, they can't just say "yes" or "no" to every potential customer. Instead, they can select a lending strategy based on the degree of risk by employing the practical solution presented in this paper to determine how much risk each client represents. For instance, regardless of whether a monetary foundation decides to dismiss giving a charge card or a home credit to people in the fragment addressed by the furthest left leaf hub of the choice tree, they can decide to give a credit line to people in different sections with a pretty much the loan fee determined in the agreement to counterbalance the gamble presented.

REFERENCES

1. M. A. J. Shirley, (2017); "Impact of demonetization in India," *International journal of Trend in Research and Development,* vol. 17, pp.20-23.
2. Frech, J. Houle, D. Tumin, "Trajectories of unsecured debt and health at midlife," *SSM-Population Health*, p.l 00846.
3. D. Prelec, D. Simester, (2001); "Always leave home without it: A further investigation of the credit-card effect on willingness to pay," *Marketing letters*, 12 (I), pp. 5-12.

4. J. Geanakoplos, P. Dubey, (2010); "Credit cards and inflation," *Games and Economic Behavior,* vol. 70, no. 2, 2010, pp. 325-353.
5. K. Brown, P. Moles, (2016); "Credit Risk Management," *Great Britain*, 2nd edition edition, 2016.
6. F. Butaru, Q. Chen, B. Clark, S. Das, A. W. Lo, A. Siddique, (2016); "Risk and risk management in the credit card industry," *Journal of Banking & Finance* 72, 2016, pp. 218-239.
7. Default of credit card clients Data Sethttps://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients
8. Charleonnan, (2016); "Credit card fraud detection using RUS and MRN algorithms," 2016 *Management and Innovation Technology International Conference (MITicon),* pp. MIT-73-MIT-76, doi:10.11 09/MITICON.2016.8025244.
9. 1-Cheng Yeh and Che-hui Lien, (2009); "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients," *Expert Systems with Applications,* vol. 36, no. 2,pp.2473-2480.
10. Y. Y. Song, Y. Lu, (2015); "Decision tree methods: applications for classification and prediction," *Shanghai Arch Psychiatry*, vol. 27(2), 2015, pp.130-135.
11. T. Denoeux, (1995); "A k-nearest neighbor classification rule based on Dempster-Sbafer theory," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 25(5), pp. 804-813.
12. L. Breiman, (2001); "Random Forests," *Machine Learning*, vol. 45, pp. 5-32.
13. T. Chen, C, Guestrin, (August 2016) "XGBoost: A scalable tree boosting system," *Proceedings of the 22nd ACM sigkdd international conference on knowledge discovery and data mining*, pp. 785-794.