© September 2024 | IJIRT | Volume 11 Issue 4 | ISSN: 2349-6002

Analyse effectiveness in predicting credit card default using Artificial Neural Networks (ANNs) & other Machine Learning algorithms in Microsoft Azure Studio

HA QUANG SON
Golden Gate University, San Francisco, USA

Abstract— This study aims to analyse the effectiveness of various machine learning algorithms, including Artificial Neural Networks (ANNs), in predicting credit card defaults using Microsoft Azure Machine Learning Studio. This research will compare the performance of different models, such as decision forest, support vector machine, boosted decision trees, logistic regression, in accurately forecasting if a credit card holder will default on the payment. Amongst these algorithms, ANNs is found to be of the highest accuracy. By leveraging on the open-sourced & large dataset of a popular consumer bank in Taiwan offering credit card facilities, the conclusion of this study will contribute to the research of predictive accuracy of different classification machine learning models in Microsoft Azure Studio. The interest of this study comes from the rising credit card delinquencies in the recent years, posing risks & increasing costs for the banks in managing consumer credit lines.

Index Terms- Credit Card, Default, Neural Network, Machine Learning, Microsoft Azure

I. INTRODUCTION

Credit card usage has become increasingly popular in modern financial landscape, with millions of transactions occurring daily. However, not all credit card owners have the same credit history or behaviours, hence posing a risk of default in repayment to the credit-issuing institution. Accurately predicting credit card defaults is crucial for managing credit risk and minimising losses, allowing financial institutions to proactively identify customers with high risk of defaults and take necessary actions.

Machine learning algorithms have emerged as powerful tools for classification problems, offering the potential to analyse complex relationships between borrower characteristics, credit history, and economic factors and predict whether default payment will occur. Artificial Neural Networks (ANNs), in particular, have shown promising results in capturing non-linear patterns and making accurate predictions.

This study aims to analyse the effectiveness of various machine learning algorithms, including ANNs, in predicting credit card defaults using Microsoft Azure Machine Learning Studio. By comparing the performance of different models, such as ANNs, decision forests, support vector machines, boosted decision trees, and logistic regression, we aim to identify the most accurate and reliable approach for forecasting credit card default probabilities.

The research will utilise a large dataset consisting of 30,000 customer profiles with varying educational backgrounds, ages, and payment histories. The data will be pre-processed, normalised, and divided into training and testing sets in 80/20 ratio to ensure the validity of the models. Various performance metrics, such as Accuracy, Recall, Precision and F1 Score, will be used to evaluate the models' predictive capabilities.

II. MACHINE LEARNING & CLASSIFICATION ALGORITHMS

Machine learning (ML) is a branch of artificial intelligence that enables computers to learn from data. [1] By analysing large datasets, ML algorithms are used to identify patterns and make predictions or classification, and functioning autonomously with minimal human intervention. The rise of big data and advanced computational capabilities has made machine learning essential across various domains, including communication, healthcare, and finance.

Classification is a supervised machine learning method where the model tries to predict the correct label of a given input data. [2] The data is typically split between training data where the model is trained on, and testing data where the model is used to make prediction. In classification, prior to training the model, there are several algorithms to choose from in order to compare and select the best performing one.

- Logistic Regression (LR): This algorithm is used for binary classification problems. It predicts the probability of a binary outcome based on one or more variables. However, it cannot properly predict with the problem of non-linear variables.
- Decision Trees: This algorithm uses a tree-like structure to make decisions based on feature values. Decision trees are intuitive and easy to interpret but can be prone to overfitting.
- Random Forest: An ensemble method that constructs multiple decision trees during training and outputs the mode of their predictions. Random forests improve accuracy and control overfitting compared to individual decision trees.
- Support Vector Machines (SVM): SVMs are powerful classifiers that find the hyperplane that best separates data points of different classes. They are effective in high-dimensional spaces and are particularly useful for complex datasets.
- k-Nearest Neighbour's (k-NN): This algorithm classifies data points based on the classes of their nearest neighbours in the feature space. It is simple and effective but can be very slow because it requires distance calculations for all training samples.
- Naive Bayes: This algorithm assumes independence between features. It is particularly effective for text classification tasks and works well with large datasets.
- Artificial Neural Networks (ANNs): These algorithms are inspired by the human brain and consist of interconnected nodes (neurons) that process input data. This algorithm can properly predict with the problems of non-linear variables.

Each of these algorithms has its strengths and weaknesses, making them suitable for different types of classification problems. The choice of algorithm is usually based on specific evaluation metrics, such as accuracy, precision, recall, and F1-score, in order to conclude its performance.

III. MICROSOFT AZURE STUDIO

Microsoft Azure Studio is a cloud service for accelerating and managing the machine learning (ML) project lifecycle. [2] It allows user to train and deploy ML models in a no-code environment using drag-and-drop interface for both dataset and other components in the architecture. It supports collaboration, allowing users to share and find assets, resources, and metrics for projects through the Machine Learning Studio UI. [3]

IV. LITERATURE REVIEW

The prediction of credit card defaults has attracted significant interests in recent years due to its implications for financial institutions and its risk management frameworks. In general, the number of customers with overdue repayments is significantly lower than those who have paid their credit card loans on time. In previous studies, most examples dealt with imbalanced data, therefore giving rise to the need of using various different techniques such as resampling methods to avoid the potential issues. Various machine learning algorithms, particularly Artificial Neural Networks (ANNs), have been employed to enhance predictive accuracy in this domain due to their capability to model complex and nonlinear correlations in data. This literature survey reviews several studies that have contributed to the understanding and application of machine learning techniques for predicting credit card defaults.

Several papers have explored different methodologies in using machine learning models to predict credit card defaults. For instance, an undersampling technique combined with clustering methods is used to improve model accuracy on imbalanced datasets, demonstrating that innovative sampling approaches can enhance predictive performance. In another example, an enhanced version of both oversampling and undersampling methods is used, thereby addressing the challenges posed by imbalanced data distributions.

The application of ANNs has been particularly notable in credit risk prediction. Studies have shown that ANNs can effectively capture complex non-linear relationships within the data. However, several other studies were related to credit card fraud instead of credit card defaults. For example, a deep learning model is designed to specifically targets imbalance datasets, achieving improved prediction accuracy compared to traditional methods.

Comparative analyses of various machine learning algorithms have also been conducted. For instance, some experiment results indicated the deep neural network performed better in most evaluation metrics and achieved an impressively high accuracy of 0.93 if compare to the machine learning models. This highlights the importance of model selection and evaluation in achieving the best results.

Despite the progress made, challenges remain in the field of credit card default prediction. The inherent imbalance in credit default datasets continues to pose difficulties for model training and evaluation. It will bring tremendous value to study a real world dataset that has reasonable distribution in default vs non-default target variables. In addition, there were little mention regarding the provider of such machine learning tools, hence limiting the effective comparison of popular and accessible tools.

In summary, the literature indicates a growing body of work focused on leveraging machine learning, particularly ANNs, for predicting credit card defaults. While significant advancements have been made, ongoing research is necessary to refine these models, address data imbalance, and introduce the right tools to use. The findings from this study will help future research of credit card defaults using Microsoft Azure Machine Learning Studio, which is one of the popular tools available for researchers as well as the general public.

V. METHODOLOGY

To conduct a performance analysis on credit card default prediction using different machine learning algorithms in Microsoft Azure Machine Learning Studio, the following end-to-end ML lifecycle is followed:

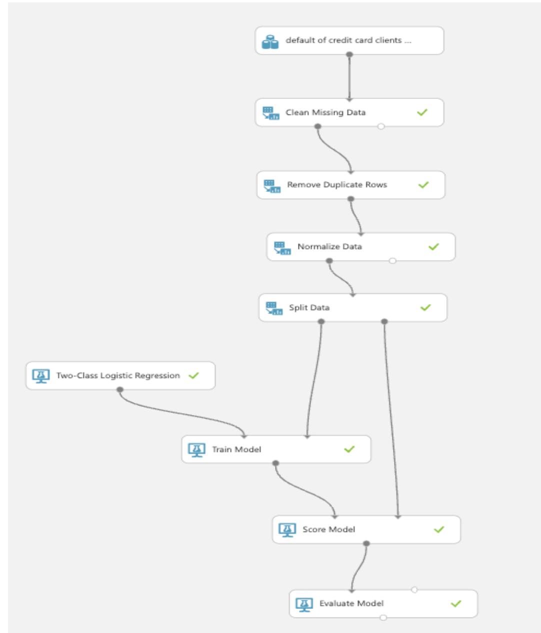

Fig.1 – Illustration of end-to-end ML lifecycle using Microsoft Azure Studio

5.1. Data Collection

Gather the credit card dataset from UCI Machine Learning Repository. The dataset consists of 30,000 records with target variable of defaulting payment accounting for 22.12%, hence avoiding the case of imbalance data. Other input variables are as follows:

- LIMIT_BAL: this indicates the credit limit given in Taiwanese currency (NT Dollar)
- SEX: this indicates gender, where 1 = male, 2 = female
- EDUCATION: this indicates the education background of the credit card owner, where 1 = graduate school; 2 = university; 3 = high school; 4 = others
- MARRIAGE: this indicates marital status, where 1 = married; 2 = single; 3 = others
- AGE: this indicate age in year
- PAY 0 – PAY 6: this indicates the delay period of past payment, where -1 = no delay, 1 = delay for 1 month, 2 = delay for 2 months, etc, 9 = delay for 9 months and longer
- BILL_AMT1 – BILL_AMT6: this indicates amount in the bill statement
- PAY_AMT1 – PAY_AMT6: this indicates amount of payment

5.2. Data Pre-Processing

Pre-process the data by cleaning missing values, removing duplicate data. In addition, normalising data is also required due to differences in the values of the feature variables. The most common normalisation technique is the Min-Max normalisation. The dataset is also split into training and testing sets in 80/20 ratio.

5.3. Model Training

The pre-processed dataset is used to train various machine learning models, namely:

- Logistic Regression
- Decision Forest
- Support Vector Machine
- Boosted Decision Tree
- Bayes Point Machine
- Artificial Neural Network

5.4. Model Scoring

Allow the model to predict the outcomes using the 20% testing data and compare with the true values of the target variable.

5.5. Model Evaluation

Evaluate the performance of all the trained model using several metrics, such as:

- Accuracy: indicates how many percent of the predictions is correct
- Recall: indicates how good the model at giving real correct predictions
- Precision: indicates how many percent of the positive predictions is correct
- F1 Score: indicates the harmonic means of Precision and Recall

In this research, accuracy is chosen to be used as the right metric, to determine which algorithm has the highest percentage of correct predictions amongst all the algorithms.

5.6. Model Deployment and Maintenance

Although it is not shown in the aforementioned illustration, the best-performing model will be selected for deployment as an integrated function of the financial institution. In order to improve its accuracy over time, more data is required and hyper-parameter finetuning is necessary for a long-term predictive capabilities of the model.

VI. RESULTS

The evaluation results of the different classification algorithms are shown below (Figs. 2-7).

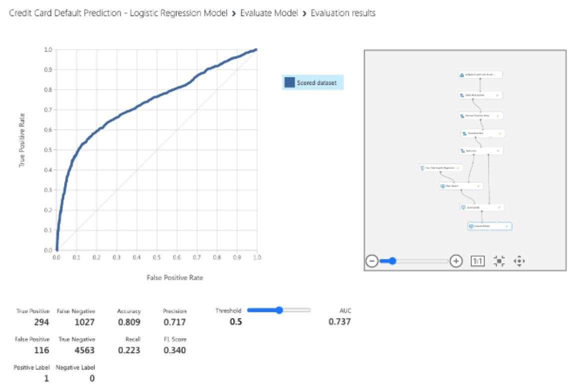

Fig. 2 – Evaluation results of Logistic Regression model

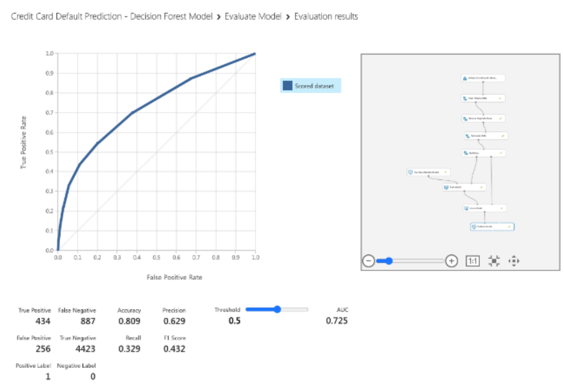

Fig. 3 – Evaluation results of Decision Forest model

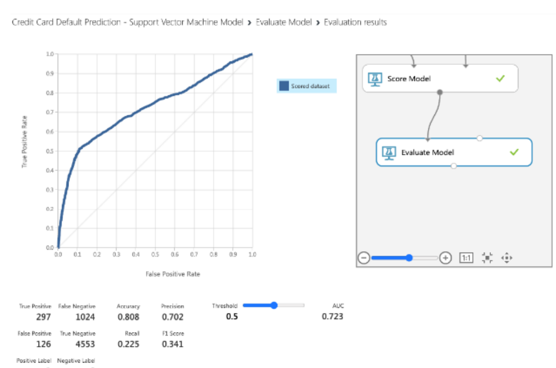

Fig. 4 – Evaluation results of Support Vector Machine model

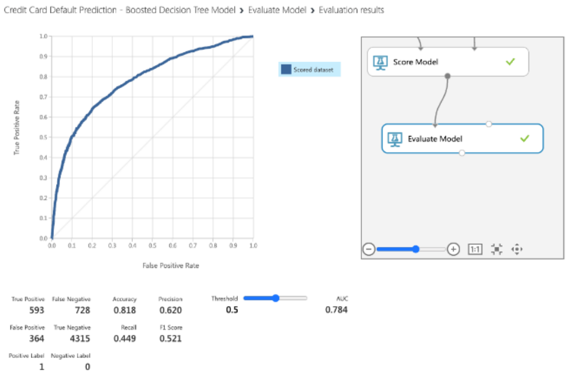

Fig. 5 – Evaluation results of Boosted Decision Tree model

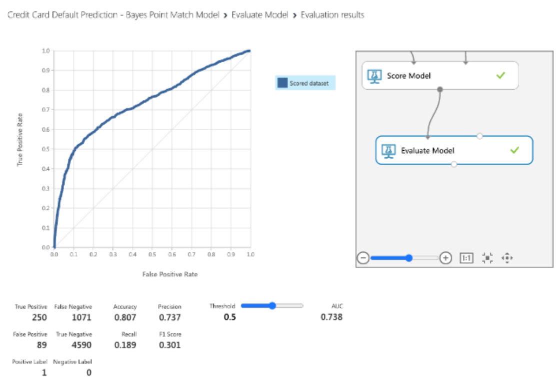

Fig. 6 – Evaluation results of Bayes Point Machine model

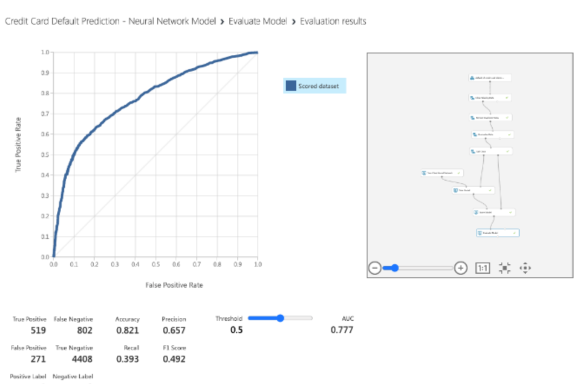

Fig. 7 – Evaluation results of Artificial Neural Networks (ANNs) model

When using accuracy as the performance metrics, the classification results show the performance of the 6 machine learning algorithms are ranked (from lowest accuracy to highest accuracy) as follows:

	Classification Algorithms					
Evaluation Metrics	Bayes Point Machine	Support Vector Machine	Logistic Regression	Decision Forest	Boosted Decision Tree	Artificial Neural Network
Accuracy	0.807	0.808	0.809	0.809	0.818	0.821
Recall	0.189	0.225	0.223	0.329	0.449	0.393
Precision	0.737	0.702	0.717	0.629	0.620	0.657
F1 Score	0.301	0.341	0.340	0.432	0.521	0.492

Table 1 – Classification Algorithm results

CONCLUSION

In this study, we first implemented the end-to-end ML lifecycle to the large data set of credit card transactions. After applying the aforementioned 6 classification algorithms in Microsoft Azure Studio, the best performing model in terms of accuracy is Artificial Neural Networks (ANNs), meaning it has the highest percentage of correct predictions amongst all the algorithms. In order to improve its performance over time, it is necessary to collect more data to retrain the model in case there are changes in business environments or overall customer behaviours. It is also noteworthy to mention that Recall value can be improved by a better data processing and good hyper-parameter tuning.

REFERENCES

[1] I-Cheng Yeh, 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, 36 (2009) 2473–2480*

[2] Tsungnan Chou and Mingmin Lo (2018). Predicting Credit Card Defaults with Deep Learning and Other Machine Learning Models. *International Journal of Computer Theory and Engineering, Vol. 10, No. 4, August 2018*

[3] TALHA MAHBOOB ALAM, KAMRAN SHAUKAT, IBRAHIM A. HAMEED, SUHUAI LUO, MUHAMMAD UMER SARWAR, SHAKIR SHABBIR1, JIAMING LI, AND MATLOOB KHUSHI *(2020).* An Investigation of Credit Card Default Prediction in the Imbalanced Datasets. *IEEE Access · October 2020*

[4] Yuge Han (2024). An Investigation of Machine Learning Applications in the Financial Fraud Detection. *Proceedings of Finance in the Age of Environmental Risks and Sustainability - ICFTBA 2024 (ICFTBA 2024)*

[5] Tianpei Xu, Min Qu (2024). Novel embedding model predicting the credit card's default using neural network optimized by harmony search algorithm and vortex search algorithm. *Heliyon, Volume 10, Issue 9, e30134, May 15, 2024*