**CREDIT CARD DEFAULT PREDICTION**

**RAHUL SINGH WALDIA**

**Data science trainees,**

**AlmaBetter, Bangalore**

**Abstract**—Since the overdue amount of credit cards has been increasing year by year, the rising credit card delinquencies might prevent the commercial banks to allocate more funds in profitable investments. At the same time, the high processing costs of credit card delinquencies through manual verification also affect the competitiveness of credit card issuers. Because the market competition becomes strict in the era of financial technology, to predict correctly whether cardholders will be unable to pay off credit card debt and establish an effective risk prediction model is the major purpose of this study. We first implemented four machine-learning approaches to predict the default cases, however, most models encountered challenges to resolve imbalance problem of delinquency cases in data sets and reported lower predictive accuracy. Two inference strategies including grey incidence analysis and fuzzy decision tree were proposed to improve the predictive performance. The average accuracy for both strategies were increased from 0.82 to 0.86 and 0.89 respectively. In addition, the deep learning approach integrated with various network structures was also incorporated to evaluate model performance. The 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. Finally, three feature selection methods were employed with the deep learning model, and the results showed similar predictive accuracy as the original deep learning models with slightly better performance being reported by filtering variables with the grey incidence analysis. This research work could be extended to apply more complicated deep learning algorithms to learn and trace the behaviors of the credit card holders and reduce the default risks for banking industries. Index Terms—Default prediction, machine learning, deep neural network, deep learning.

**I. INTRODUCTION** As more and more consumers rely on the credit card to pay their everyday purchases in online and physical retail store, the amount of issued credit cards and the overwhelming amount of credit card debt by the cardholders have rapidly increased. Therefore, most financial institutions have to deal with the issues of credit card default in addition to the credit card fraud such as credit card dump. Both the credit card verification applied to the cardholders and the default risk management after card issued are crucial to the future success of most financial institutions. Under those circumstances, to reduce the delinquent payment of credit card holders, and perform the effective management of credit card risk will help financial and banking institutions to achieve competence. A default analysis system empowered with data mining techniques will allow the credit card issuers to monitor and eliminate the ongoing transactions of potential risky account prior to credit fraud and default. Many researchers have developed various risk prediction systems to detect the credit card defaults for banking industry, some research works such as [1], [2] created the default prediction systems based on statistic techniques. More studies that are recent have reported that Artificial Intelligence (AI) methods might perform better than the statistical methods for credit risk assessment. Data mining techniques and machine learning algorithms can assist data managers to improve their problem-solving techniques by discovering meaningful patterns and trends from large amounts of data. Meanwhile, the fast development of artificial intelligence and deep learning techniques also bring great impacts to the financial industries. This study aims to compare the prediction accuracy for the newer developed deep learning methods against with the conventional machine learning based on the financial information of the credit card holders. The implementation of the framework comprised three stages. The first stage applied conventional machine learning techniques including decision tree (DT), naive Bayes (NB), K nearest neighbor (KNN), and logistic regression (LR) to implement default inference systems for predictive tasks of financial institutions. A modified fuzzy decision tree (FDT) was also implemented with different number of membership functions to evaluate whether the fuzzy partition and K-Means clustering could improve the predictive performance. The experiment result from our previous study [3] applying grey incidence analysis (GIN) was included in this paper for comparison with the KNN models. The second stage used a deep learning system based on the customized modification of layers, neurons and activation functions. The last stage combined three feature selection methods to reduce the dimensionality of input variables for the deep neural models. The overall objective of this framework intended to extract useful features from cardholder data, and applied the artificial intelligence techniques to assist financial institutions to reduce their risk of credit card default. Both the personal information and repayment details of the cardholder were used to trace the consumer behavior for the predictive models. The rest of this paper was organized as follows. Section II documented the proposed models and related methodologies. Section III discussed the data processing and evaluated experimental results, and finally the Section IV provided a conclusion of the work.

As the original ML models used all 24 variables as inputs, high-dimensional data might increase the analytical difficulty for machine-learning models. Therefore, three feature selection methods were applied in the last experiment to reduce the number of input variables for models. Especially, the grey incidence analysis was also incorporated as one of the feature selection due to its ability of ranking the relational order of variables. The methods and strategies used in this study were briefly described as follows.

**A. Machine-Learning Models** Four most common used machine-learning techniques including logistic regression, naive Bayes, decision tree, and K-nearest neighbor methods addressed in [4] were employed to predict the delinquency cases of credit card based on the original 24 variables. The logistic regression was a statistics methodology used to account for the relationship between variables; especially the dependent variable was categorical or binary. On the other hand, naive Bayes method was a supervised learning algorithm based on Bayes theorem and worked very well in many research applications. The approach only required a small amount of training data to estimate the required parameters. The C4.5 decision tree [5], improved from the original ID3 algorithm could handle training data with missing values and performed classification tasks with improved computational efficiency. Moreover, K-nearest neighbor was another supervised learning technique mostly used for classification tasks such as the default and non-default cases. As the KNN searched the entire training dataset for the K most similar cases from the neighboring instances and provided a summarizing outcomes for those K neighbors, this method was vulnerable to erroneous and outlier data. In comparison to the KNN method, the grey incidence analysis measured the closeness of a relationship between sequences by evaluating the similarity of the geometrical curves formed by the sequences, which were considered as an analogous to the corresponding K neighbors in KNN model using different distance measures. The prediction of a testing sample by KNN model was fulfilled with the five nearest neighbors based on Euclidean distance in this study. Meanwhile, the GIN model calculated the metric of incidence degree for each variable in sequences, and then used the summarized metric to represents the geometrical similarity between all variables. As a result, the more similar for the geometrical shape of curves formed by all variables, the higher grey incidence degree between them. For the purpose of providing comparison with different searching algorithm and feature selection, the predictive result of the GIN model in our previous experiment [3] was involved as a counterpart for performance comparison with KNN model. The GIA-A model used the original 24 variables. In addition, the GIA-B model comprised 10 out of 24 variables selected by four different incidence measurements fusing with evidential reasoning approach. Moreover, GIA-C model derived from the GIA-B model but worked with a dynamic weighting. In addition, for the purpose of improving the performance of decision tree in the first experiment, another fuzzy decision tree model based on fuzzy partition and K-Means clustering [6], [7] was also employed as a counterpart for comparison to the classical decision trees. Both the input and output variables were partitioned by various fuzzy membership functions. The maximum partition size and the number of membership was defined by K-Means clustering method for each active variable. The K-Means method was an unsupervised learning technique used for clustering, and K was the number of clusters the method tried to learn from the dataset. The fuzzy decision tree was also pruned to reduce its complexity, and the paths from the root node to the leaf nodes were converted as decision rule [8]-[10]. The FDT model was examined with three partition methods, which were denoted as FDT-M3, FDT-M5, and FDT-M7 models respectively. B. Deep Learning Models The conventional artificial neural network built with multiple hidden layers has been successfully solving various image classification and recognition problems under complex scenarios. Each hidden layer could learn the abstraction of previous layer output. Especially, the deep learning algorithms and models have been well developed for many years, several important networks, such as deep neural International Journal of Computer Theory and Engineering, Vol. 10, No. 4, August 2018 106 networks (DNN), convolutional neuron networks (CNN), recurrent neural networks (RNN), autoencoder (AE), were pervasively applied all over the science and technology [11], [12]. An autoencoder neural networks, including compression and decompression functions, could be stacked to learn the representation of variables without assigning labels for supervised learning. A recurrent neural network was normally employed with the application based on time sequence data. On the other hand, the convolutional neuron networks normally was used to learn patterns and components of an image. However, training neural networks with multiple hidden layers could be time-consuming and difficult if the structure of network became more complex. The deep neural networks applied in this study included several steps. The first step was to initialize a sequential DNN model consisted of different stack of layers, and followed by appending customized input layer, hidden layers and an output layer to the model according to what problems the model tried to solve in application. As the hidden and output layers controlled the DNN inner workings, the number of hidden layers and their corresponding processing units might affect the performance of predictive model. The hidden layers within the sequential model were experimented with 2 to 5 layers, and the processing units were tested for 64, 32, 16 units respectively. Furthermore, we also used dropout layers and set up cutoff threshold value to control over-fitting problem in this study. The dropout rate was set to 0.2 to remove the processing units with weights below 20%. Both the Sigmoid and ReLU activation functions were used in different layers, and the max-pooling layer, where a small 2\*2 pooling region acted as a filter was applied to filter the output of the previous layer and reduced the dimensionality of input pattern by eliminating non-maximal values. Finally, the last layer was a fully connected layer and similar to a traditional neural network. The Softmax function employed in this layer classified the outcomes into the final default categories. The implementation of deep neural network was fulfilled by the TensorFlow open-source software library and Keras API based on the following processes. Step 1: Select the top N variables as inputs based on feature selection. Step 2: Construct a sequential deep neural network and initialize the sequential model. Step 3: Define the number of hidden layers and corresponding network nodes. Step 4: Add dropout layers and specify the dropout rate. Step 5: Assign the ReLU and Sigmoid activation functions for layers. Step 6: Reshape the output variable with one-hot encoding. Step 7: Apply a fully connected output layer. C. Feature Selection Methods To evaluate the influence of feature selection on accuracy of deep learning models, the grey incidence analysis, Pearson correlation analysis, and information gain were applied in the final experiment. By organizing the 24 input variables as data series, the grey incidence analysis measured the similarity between the testing and training data, and identified the most similar pattern matched within the training data series by calculating incidence coefficient for each variable in the two observed data series [13]. The individual coefficients were aggregated as the degree of grey incidence to represent the overall measurement of two data series according to different methodologies described in our previous study. On the contrary, a data series could also be formed by all instances collected by a specific variable, and the degree of incidence was then used to rank the relationship between variables. If we used grey incidence analysis to compare series of instances rather than that of variables, then we could acquire a ranking list of variables based on the calculated degree of incidence. The Pearson correlation analysis explored the strength and direction of linear relationships between the variables and was another alternative for feature selection. Despite this method required the assumption that the variables were normally distributed and encountered with the problem of multicollinearity to select relevant variables, the result of correlation test could be used in our experiment to choose the top ten variables as inputs, which were highly relevant to the dependent variable for models. Additionally, information gain was another simple option to get the important attributes for feature selection. The entropy of each input variable was calculated for the dependent variable, and the input variables that contribute more information were selected.

**III. EXPERIMENT RESULTS** The experiment results were divided into three parts in this section. The first part comprised three experiments based on conventional machine-learning models, but incorporated different modifications to improve predictive performance. Moreover, the second part implemented the deep learning technique with various network structure and contained another two experiments. The third part employed three different feature selection methods with the deep neural network models. Totally four performance metrics derived from the confusion matrix were used to evaluate the model performance including accuracy, sensitivity, specificity, precision, and F-score. A. Performance of the ML Models As shown in Table I, four conventional machine-learning models were evaluated for the prediction performance, and among all models, the DT model was inferior to other three models in accuracy metric although all model achieved the accuracy over 0.76. Because 76% of the collected default cases in dataset were positive, therefore, we dealt with the imbalanced dataset in this study. As the result, the DT model obtained the lowest accuracy and was not regarded as better than the accuracy of random guessing.

**IV. CONCLUSION** In this study, we first implemented four machine-learning approaches to predict the default cases, however, most models encountered challenges to resolve imbalance problem of the delinquency cases in our datasets. To improve the predictive performance, we proposed another two strategies including grey incidence analysis and fuzzy decision tree to improve the predictive accuracy. The average accuracy were increased from 0.82 to 0.86 and 0.89 for each strategy respectively. The performance of DT model in machine learning approaches was improved by integrating the fuzzy partition with K-Means clustering method, and the predictive accuracy of the proposed FDT model was risen to higher value at 0.893. Moreover, the input variables were rearranged for the grey incidence analysis to predict the default cases based on searching the geometrical similarity of data series. The predictive accuracy was increased from 0.806 to 0.891 if compared to its counterpart of KNN model using similar searching algorithm. Although the proposed two strategies integrated with machine learning approaches moderately increased the average accuracy to 4.5% and 7.2% respectively, we also implemented the newly developed deep learning approaches to examine whether the new techniques could improve the predictive accuracy. The DNN models were evaluated with a different number of hidden layers and neurons, which also combined with the dropout layers and various activation functions to improve the training process. The experiment result indicated the DNN model with different number of layers increased the average accuracy to 0.930. On the other hand, the DNN models evaluated with different number of neurons suggested that the DNN-64N model achieved the highest accuracy of 0.931 if compared with other two models within the same group. Finally, the DNN models were evaluated with three different feature selection, and the experiment result reached the highest average accuracy of 0.934. The result also suggested that grey incidence analysis employed to rank variables achieved the highest accuracy of 0.935 and was considered as the best model if compared to the results of all experiments. In conclusion, the experiment results showed the deep neural network performed better in most evaluation metrics and achieved an impressively high average accuracy of 0.930 if compare to the machine learning models which was 0.851. This research work could be extended to implement other deep learning algorithms to learn and trace the behaviors of the credit card holders and reduce the default risks for banking industries in the future.

**REFERENCES**

[1] J. Haslem and W. Longbrake, “A credit scoring model for commercial loans: Comment,” Journal of Money, Credit, and Banking, vol. 4, no. 3, pp. 733-734, July 2015.

[2] A. Steenackers and M. J. Goovaerts, “A credit scoring model for personal loans,” Insurance Mathematics Economic, pp. 31-34, 1989.

[3] T. Chou, “Apply two-stage grey incidence analysis with dynamic weights to improve the quality of default prediction,” Journal of Quality, vol. 22, pp. 405-426, 2015.

[4] I. Witten, E. Frank, M. Hall, and C. Pal, Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed. Morgan Kaufmann, 2016.

[5] A. Galathiya, A. Ganatra, and C. Bhensdadia, “Improved decision tree induction algorithm with feature selection, cross validation, model complexity and reduced error pruning,” International Journal of Computer Science and Information Technologies, vol. 3, no. 2, pp. 3427-3431, 2012.

[6] J. Alcala-Fdez and J. Alonso, “A survey of fuzzy systems software: Taxonomy, current research trends and prospects,” IEEE Transactions on Fuzzy Systems, vol. 24, no. 1, pp. 40-56, 2016.