Journal of Strategic and International Studies

Volume XVIII Number 1 2024
ISSN 2326-3636

EDITOR-IN-CHIEF

PROF. DR. DIETER FLAMIG
Professor, Technical University of Berlin
Berlin, GERMANY

Academic Organization for Advancement of Strategic and International Studies™

Volume XVIII Number 1 2024

Journal of Strategic and International Studies

Academic Organization for Advancement of Strategic and International Studies™

Journal of Strategic and International Studies™

Print ISSN *2326-3636* Issued by the Library of the Congress of The United State of America, Washington, D.C., USA

Typeset and Graphic Design by Kalina@AcademicOASIS©, Florida, The United States of America

Published by the Academic OASIS™ Publishing Miami Beach, Florida, The United States of America

Printed by the Academic OASIS™ Publishing, Miami Beach, Florida, The United States of America

Journal of Strategic and International Studies

Volume XVIII Number 1 2024

ISSN 2326-3636 Issued by the Library of the Congress of The United State of America, Washington, DC, USA

TABLE OF CONTENTS

Journal of Strategic and International Studies

Volume XVIII Number 1 2024

ISSN 2326-3636 Issued by the Library of the Congress of The United State of America, Washington, DC, USA

COMPARISON OF SEVERAL MACHINE LEARNING METHODS IN CREDI DEFAULT CLASSIFICATION

Maysam Khodayari Gharanchei, Carnegie Mellon University, Pittsburgh Pennsylvania, USA
Prabhu Prasad Panda, Carnegie Mellon University, Pittsburgh Pennsylvania, USA

ABSTRACT

The default prediction of credit card holders in finance industry will help the financial firms, dealing with credits and monthly cash flows, to have not only a develop a better marketing plan to target the best clients but also to form a precise risk management system. In this research, the performance of several machine learning methods, taught in the machine learning I and II courses of the MSCF program are evaluated to predict the default possibility of 30,000 clients in Taiwan. To do so, at first the most suitable criteria to predict the default risk are proposed according to the literature review and past works done by the noble machine learning experts. Then, after some data analysis and dataset review, two approached are used to answer the research question. Since about 10% of the card holders didn't default during past six months but they defaulted in the seventh payment, in this research it's tried to see if we can assume them as non-defaulters (or not) to come with a better classifier.

Keywords: Machine Learning, Default risk, Risk Management, Supervised and Unsupervised Learning

1. INTRODUCTION

For having an up-to-date risk management measures in financial firms providing credit (Lee, 2023), it's necessary to perform statistical analysis (Gaganis, 2023) (Patidar and Sharma, 2011) and machine learning models on the cumulative big data from their clients (Liu, 2018). This project presents and discusses several supervised and one unsupervised machine learning models for predicting the defaulters among the credit card users. Data include details like limit balance (Alam, 2020), age, sex, amount of bill statement, repayment status and amount of previous payment. The research discusses which variables are the strongest predictors (Patidar and Sharma, 2011) of default risk, and to make predictions on which customers are likely to default (Stolfo, 1997). Also we posed a key question to see is it a good idea to suppose that the first-time defaulters (Wu, 2023) (with a clean record over past six months) be interpreted to non-defaulters? For this purpose, we implemented the logistic regression (Sayjadah, 2018) model with cross validation on the both datasets to compare the results. Then we progress the research with the better dataset to compare 11 machine learning models performance including Logistic Regression (via cross-validation) (Yang and Zhang, 2018), KNN Method, Decision Tree, Decision Tree (via cross-validation) (Yang and Zhang, 2018), Random Forest, Random Forest (via cross-validation) (Sahin and Duman, 2011), Naive Bayes, Random Forest (via RFE feature selection) (Yang and Zhang, 2018), Random Forest (via Ada-boosting) and Neural Networks. We also tried ensembling the models (Maes, 2002) that performed the best.

The original dataset is available from kaggle.com. The covariates are selected after a thorough review of the prior works (Patidar and Sharma, 2011) done in different countries. (Patidar and Sharma, 2011) Used logistic regression, decision tree, and random forest to find the best model to test the variables in predicting credit default by measuring higher accuracy and area under the curve.

In their research, (Liu, 2018) compared traditional machine learning models, i.e. Support vector Machine, k-Nearest Neighbors, Decision Tree and Random Forest, with Feedforward Neural Network and Long Short-Term Memory. They observed that the two neural networks achieve higher accuracies than traditional models. By emphasizing on gradient boosting method, (Zaslavsky and Strizhak, 2006) compared five data mining methods, Logistic regression, SVM, neural network, Xgboost and LightGBM on the same dataset. Also, (Joshi, 2015) developed a model to predict credit default risk by employing various credit-related datasets for resolving the issue of class imbalance (Abdel-Zaher and Eldeib, 2016), as in this research is done by using Synthetic Minority Oversampling Technique, or SMOTE (Shukla and Yadav, 2015).

2. DATA PREPARATION AND METHODS

According to the mentioned studies and measures mentioned in literature review, parameters are selected and listed in Table (1).

TABLE 1. COVARIATES FOR CLASSIFYING DEFAULTERS

Covariate	Description
ID	Each credit card holder ID
LIMIT_BAL	Monetary term of each client available credit
SEX	0 = Female 1 = Male
AGE	Age in term of years
MARRIAGE	1 = married 2 = single 3 = others
EDUCATION	0 = University 1 = Graduate School 2 = High school 3 = Unknown 4 = Others
PAY_0, PAY_1, …., PAY_6 (Last six months Payment status)	Pay duly i = -2, 1, 0 Payment delay for i month i = 1, 2, …, 9
BILL_AMT1, BILL_AMT2, …., BILL_AMT6	Amount of bill statement in ith month i = 1, 2, …, 6
PAY_AMT1, PAY _AMT2, …., PAY _AMT6	Amount of previous payment in ith month i = 1, 2, …, 6

To get a good perception upon the dataset, the following information in form of tables and plots show the bivariate and multivariate distributions of different covariates with respect to the classes, 1 for defaults and 0 for non-defaults. Table (2) briefly shows demographic distribution of our data.

TABLE 2. QUANTITY DISTRIBUTION OF DEMOGRAPHIC VARIABLES

Status	Sex	Education	Marriage	Default
0	18112	14030	15964	23364
1	11888	10585	13659	6636
2	-	4917	323	-
3	-	345	54	-
4	-	123	-	-

FIGURE 1. BIVARIATE DISTRIBUTION OF SEX AND DEFAULT STATUS

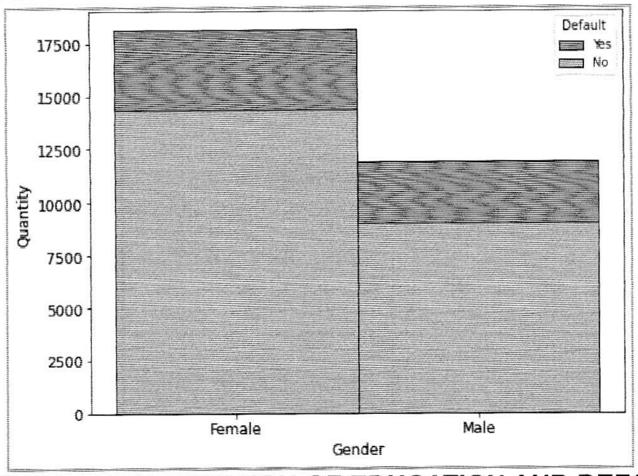


FIGURE 2. BIVARIATE DISTRIBUTION OF EDUCATION AND DEFAULT STATUS

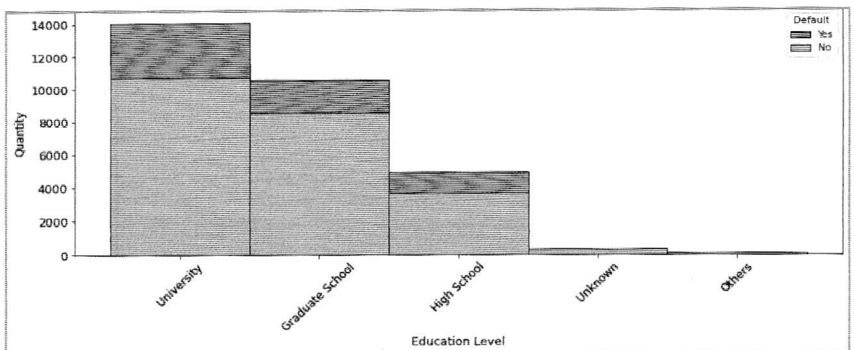
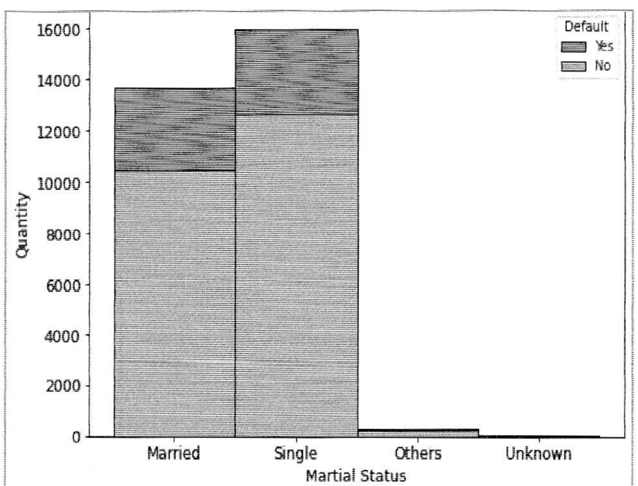


FIGURE 3. BIVARIATE DISTRIBUTION OF MARRIAGE AND DEFAULT STATUS



For a visionary presentation of the dataset, figures (1) – (7) depicts different bivariate distributions of the covariates. As we can see, females outnumber the quantity of men and number of defaulted women is slightly greater than men which may shows that women adhere more to their financial obligations in Taiwan. Also, number of single and married defaulters are almost equal and they have almost a balanced distribution. In figure (3), we see that people with higher graduation level includes more sample. We are interested to see the bivariate distribution of balance limit and level of education which is shown in figure (4).

It's interesting to find that the majority of defaults are occurring in a close range of balance limits among different educational level with various balance limits. Figure (5) and Figure (6) clearly depict the bivariate distribution of balance limit and age with the classes, respectively.

Now that we saw the bivariate distribution of age and balance limit with the classes, it should be interesting to check the multivariate distribution of age, limit balance, and the classes in figure (7).

FIGURE 4. BIVARIATE DISTRIBUTION OF EDUCATION AND BALANCE LIMIT

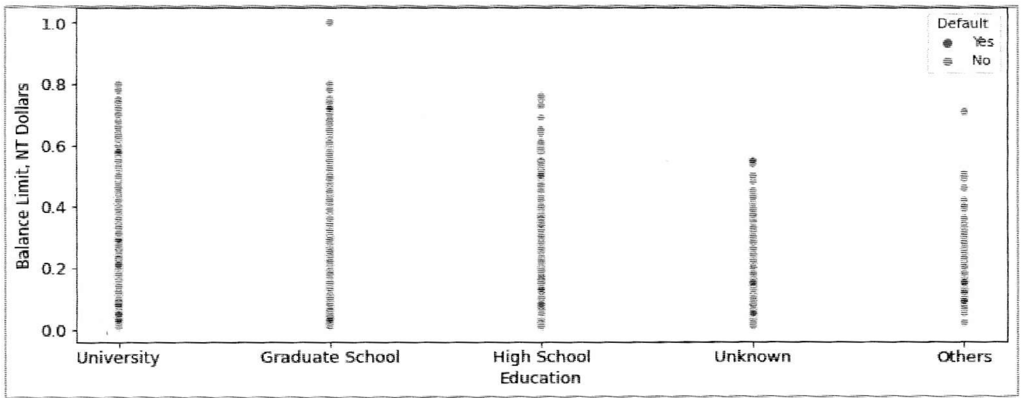


FIGURE 5. BIVARIATE DISTRIBUTION OF AGE AND DEFAULT STATUS

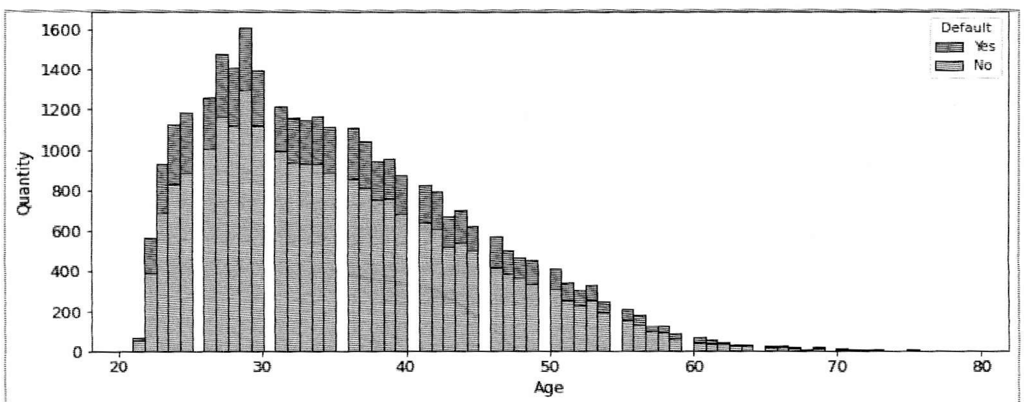
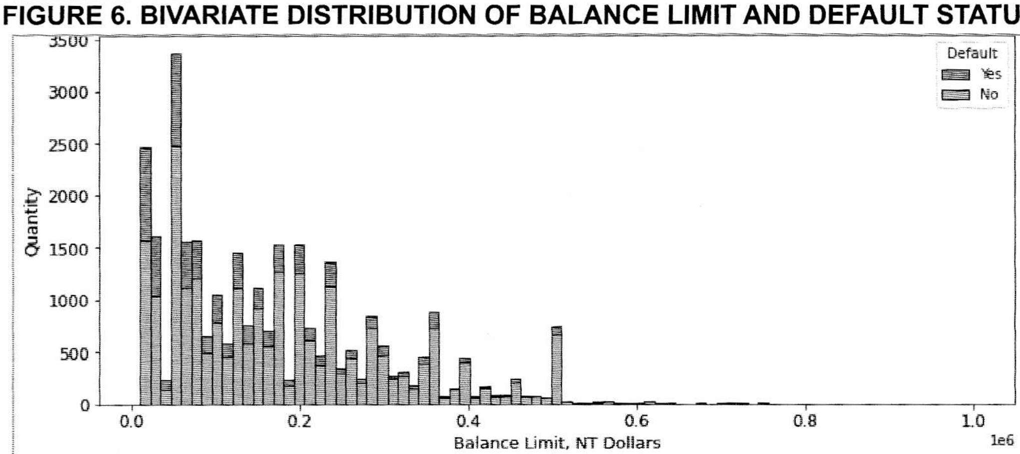
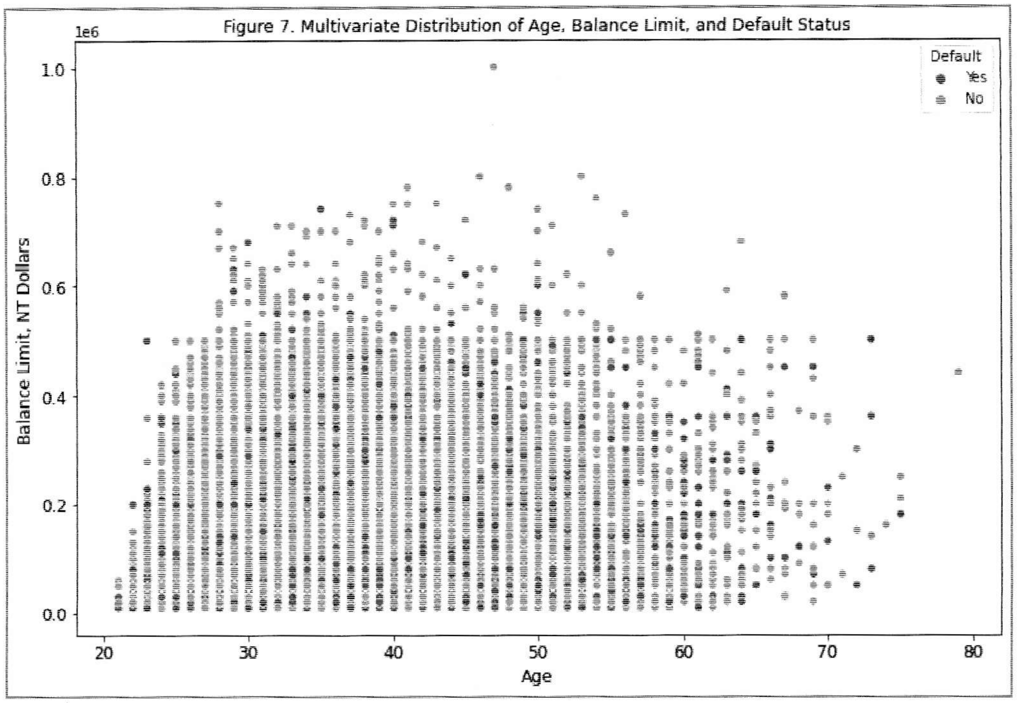


FIGURE 6. BIVARIATE DISTRIBUTION OF BALANCE LIMIT AND DEFAULT STATUS



Also, we are curious about knowing the distribution of last six months' payments status and payment amounts. While the histogram bivariate distribution plots are available in the appendix file, the correlation between amounts paid are shown in figures (8). This heatplot can help us to see if any multicollinearity there exists between numerical covariates especially between the ones which are in a same group like PAY_ j, BILL_AMT i, and PAY _AMT i; i = 1, 2, ..., 6, j = -2, -1, ..., 9.

FIGURE 6. BIVARIATE DISTRIBUTION OF BALANCE LIMIT AND DEFAULT STATUS



We care about multicollinearity because the multicollinear covariates are not actually independent, and it is difficult to test how much the combination of the independent variables affects the default risk within the regression models and impose some degree of biasness. We used the variance inflation factor, VIF for measuring the multicollinearity among the numerical covariates.

VIF provides a measure of multicollinearity among the independent variables in a multi-dimension regression model. A large VIF on an independent variable indicates a highly collinear relationship to the other variables that should be considered or adjusted for in the structure of the model and selection of independent variables.

As we can see in Table (3), the VIF index for covariates BILL_AMT i are so high. The generally accepted cut-off for VIF is 5, regarding the prior works, with higher values denoting levels of multicollinearity that could negatively impact the regression model. Hence, all of the six BILL_AMTvariables are replaced with their average value as a new variable titled as "Average_Bill_AMT". The old and new VIF values are presented in the following table.

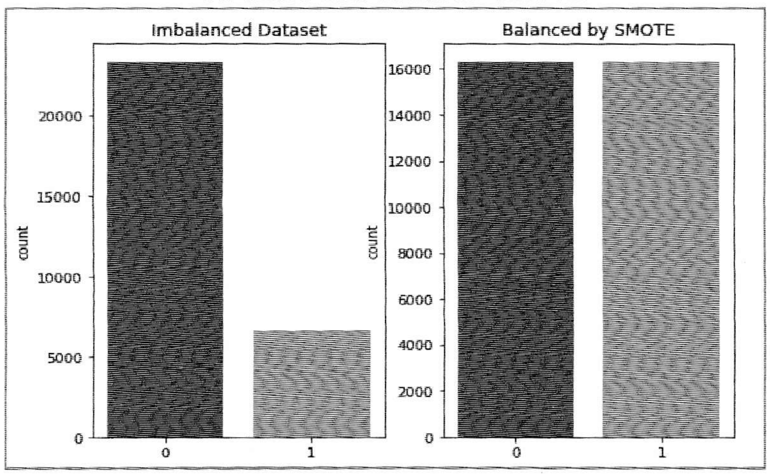
TABLE 3. VIF BEFORE AND AFTER ADJUSTMENT

Covariate Name	VIF before adjustment	VIF after adjustment
LIMIT_BAL	3.21	3.19
AGE	2.74	2.73
BILL_AMT1	20.77	-
BILL_AMT2	38.15	-
BILL_AMT3	31.88	-
BILL_AMT4	29.53	-

BILL_AMT5	36.25	-
BILL_AMT6	21.72	-
Average_Bill_AMT	-	1.78
PAY_AMT1	1.87	1.33
PAY_AMT2	2.36	1.25
PAY_AMT3	1.88	1.27
PAY_AMT4	1.76	1.23
PAY_AMT5	1.84	1.21
PAY_AMT6	1.26	.120

We discovered that are various number of outliers in the dataset which are shown by box plots in the appendix. We let them to stay in the dataset because they are the natural part of population and we don't have some justifiable reasons to remove or replace them with different samples. Also, since the classes distribution is very imbalanced, Synthetic Minority Oversampling Technique – SMOTE is employed to redistribute the data regarding the original distribution features. Following figures show the imbalance distribution and balanced one after implementing SMOTE. As we can see the difference, the distribution of the date is pretty balanced. However, we will compare results for both balanced and imbalanced datasets over each approach to acquire a comparable conclusion.

FIGURE 6. COMPARING IMBALANCED DISTRIBUTION OF BALANCE DISTRIBUTION AFTER PERFORMING SMOTE



At this point we have 4 different datasets. The first for balanced original dataset, the second for imbalanced original dataset, the third for balanced altered dataset (assuming defaulters with a clean payment record as non-defaulters), and the last dataset for imbalanced dataset. The following plots are confusion matrixes and ROC AUC scores for logistic regression that was implemented separately on each dataset.

3. CONCLUSION

Since Roc curves for all datasets are almost surely equal (see the appendix file), For getting a better resolution over the accuracy of the logistic regression over either of the datasets, it's better to check the confusion matrixes above. In credit default risk management, the risk of predicting defaulters as non-defaulters (FN) is so important as is true prediction of true defaulters (TP), we like to choose a model that has the maximum number of TPs and possibly the minimum number of FNs. Regarding this criteria, we continue our project with the altered balanced data. Then, we employed 13 different classifying model to compare several model outcomes and their scores on the selected dataset. Among, Logistic Regression (HPT), KNN Method, KNN Method (HPT), Decision Tree, Decision Tree (HPT), Random Forest, Random Forest (HPT), Naive Bayes, and Neural Networks, Neural Networks and random forest (HPT) showed a better MSE and ROC-AUC score. Next, we implemented RFE and Adaboosting methods on the selected random forest model to compare the new models result with the original one. The covariates being used in the random forest after RFE implementation on the dataset are { LIMIT_BAL, AGE, PAY_i, i = 1, 2, ..., 5, and Average_Bill_AMT }. The tables show that the RF after HPT is the best. Finally, we ensembled the results of RF after HPT with Neural Networks in two ways – 1) AGGR: Predict default if either of the models predict default else predict no default; 2) AVG: Average the scores of both the models and predict based on that. Table (4) and figure (11) show the models results and the ROC curves, respectively.

TABLE 4. RESULTS OF DIFFERENT CLASSIFIERS ON THE BALANCED ALTERED DATASET

Classifier	MSE	Roc Auc Score	False Negative (# - rate)
Logistic Regression (HPT)	0.1214	0.9479	234 - 0.1684
KNN Method	0.1381	0.9143	227 - 0.1634
KNN Method (HPT)	0.1320	0.8028	405 - 0.2915
Decision Tree	0.1321	0.7935	437 - 0.3146
Decision Tree (HPT)	0.1436	0.9364	179 - 0.1289
Random Forest	0.1190	0.9478	298 - 0.2145
Random Forest (HPT)	0.1512	0.9474	79 - 0.0569
Naive Bayes	0.1487	0.9137	161 - 0.1159
Random Forest (RFE)	0.1856	0.7541	1431 - 0.7301
Random Forest (Adaboosting)	0.1136	0.9488	287 - 0.2066
Neural Networks	0.1596	0.9490	80 – 0.0576
Ensemble - 1 (AGGR)	0.1681	0.9502	37 – 0.0266
Ensemble - 2 (AVG)	0.1567	0.9499	67 – 0.0482

REFERENCES

abdel-Zaher, A. M. & Eldeib, A. M. 2016. Breast cancer classification using deep belief networks. *Expert Systems with Applications*, 46, 139-144.

Alam, T. M., Shaukat, K., Hameed, I. A., Luo, S., Sarwar, M. U., Shabbir, S., Li, J. & Khushi, M. 2020. An investigation of credit card default prediction in the imbalanced datasets. IEEE Access, 8, 201173-201198.

Gaganis, C., Papadimitri, P., Pasiouras, F. & Tasiou, M. 2023. Social traits and credit card default: a two-stage prediction framework. *Annals of Operations Research*, 325, 1231-1253.

Joshi, S., Pandey, B. & Joshi, N. 2015. Comparative analysis of Naive Bayes and J48 classification algorithms. *International Journal of Advanced Research in Computer Science and Software Engineering*, 5, 813-817.

Lee, W., Lee, S. & Seok, J. Credit card default prediction by using Heterogeneous Ensemble. *2023 Fourteenth International Conference on Ubiquitous and Future Networks* (ICUFN), 2023. IEEE, 907-910.

Liu, R. 2018. Machine learning approaches to predict default of credit card clients. *Modern Economy*, 9, 1828-1838.

Maes, S., Tuyls, K., Vanschoenwinkel, B. & Manderick, B. Credit card fraud detection using Bayesian and neural networks. *Proceedings of the 1st international NAISO congress on NEURO fuzzy technologies*, 2002. 207

Patidar, R. & Sharma, L. 2011. Credit card fraud detection using neural network. *International Journal of Soft Computing and Engineering* (IJSCE), 1.

Sahin, Y. & Duman, E. Detecting credit card fraud by decision trees and support vector machines. *Proceedings of the International Multi Conference of Engineers and Computer Scientists*, 2011. 1-6.

Sayjadah, Y., Hashem, I. A. T., Alotaibi, F. & Kasmiran, K. A. Credit card default prediction using machine learning techniques. 2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA), 2018. IEEE, 1-4.

Shukla, S. & Yadav, R. 2015. Voting based extreme learning machine with accuracy based ensemble pruning. International Journal of Computer Applications, 115, 14-18.

Stolfo, S., Fan, D. W., Lee, W., Prodromidis, A. & Chan, P. Credit card fraud detection using meta-learning: Issues and initial results. AAAI-97 Workshop on Fraud Detection and Risk Management, 1997. 83-90.

Wu, J., Zhao, X., Yuan, H. & Si, Y.-W. 2023. CDGAT: a graph attention network method for credit card defaulters prediction. *Applied Intelligence*, 53, 11538-11552.

Yang, S. & Zhang, H. 2018. Comparison of several data mining methods in credit card default prediction. Intelligent Information Management, 1 122-115,0

Zaslavsky, V. & Strizhak, A. 2006. Credit card fraud detection using self-organizing maps. Information and Security, 18, 48.

AUTHOR PROFILES

Maysam Khodayari Gharanchei is M.S. Computational Finance Alumnus, December 2023, and affiliated with Tepper School of Business, Carnegie Mellon University, Pittsburgh Pennsylvania, USA

Prabhu Prasad Panda is affiliated with Carnegie Mellon University, Pittsburgh Pennsylvania, USA