Predictive Analytics for Default of Credit Card Clients

Alžbeta Bačová
Deparment of Cybernetics and Artificial Intelligence
Faculty of Electrical Engineering and Informatics, Technical University of Košice
Letná 9, 040 01 Košice, Slovakia
alzbeta.bacova@student.tuke.sk

František Babič
Deparment of Cybernetics and Artificial Intelligence
Faculty of Electrical Engineering and Informatics, Technical University of Košice
Letná 9, 040 01 Košice, Slovakia
frantisek.babic@tuke.sk

***Abstract*— Predictive analytics has a significant potential to support different decision processes. We aimed to compare various machine learning algorithms for the selected task, which predicts credit card clients' default based on the free available data. We chose Random Forest, AdaBoost, XGBoost, and Gradient Boosting algorithm and applied them to a prepared data sample. We experimentally evaluated the classification models within metrics like accuracy, precision, recall, ROC, and AUC. The results show a very similar performance of the selected algorithms on this dataset. The Gradient boosting (0.7828) achieved the best performance within AUC, but the best precision for target class 1 reached the Bagging algorithm (0.72). The simple data processing brought only minimal improvements in individual metrics. Our results are comparable to the mentioned studies instead of MCC metrics that resulted in better value (0.4111) achieved by the Gradient Boosting model.**

Keywords—prediction, ROC, AUC, precision, bagging, boosting

I. Introduction

The banking industry produces a significant amount of data every day that holds valuable information. The analysts can apply predictive methods to bank transaction data, clients 'data, credit card history, customer experience, and stock market data. Based on the results, the financial institutions know how to prevent any unwanted consequences, provide services with minimal loss, or offer services to customers. Many banking companies see the importance of machine learning (ML) and artificial intelligence and their usefulness to predict future occurrences in various areas. Such a trend of applying technologies in business is crucial for competitiveness, growth, and profit. According to the survey The Rise of AI in Financial Services from 2016[1], 32% of financial executives said they uses new technologies, such as predictive analytics or voice recognition systems. Many financial institutions are still afraid of applying new procedures or there are many legal obstacles.

SPD Group presents many benefits of using Artificial Intelligence (AI) in the banking sector[2], like in 2030, by implementing AI technologies, the banks will save up to 1 trillion dollars. Chinese banks used facial recognition to determine which client is lying about their financial situation that decreased losses by 60% from unpaid loans. Out of 33 000 customers, 54% said they want to receive suggestions produced by ML technology.

The Accenture study proposes a scenario that AI will transform financial service providers into data- and AI-based businesses [1] .

The survey on AI in Financial Services conducted by the World Economic Forum in collaboration with the Cambridge Centre for Alternative Finance at the University of Cambridge Judge Business School and supported by EY and Invesco aimed to understand the opportunities and challenges resulting from the mass adoption of AI in financial services [2]. Some of the main highlights from this survey are 77% expect that AI will become essential to their business within two years; 52% are currently implementing AI-enabled products and processes.

The paper is organized as follows: a short introduction with related work and selected machine learning methods. The second chapter describes the experiments focusing on comparing various machine learning methods applied to credit cardholders' data from Taiwan. The last chapter concludes the paper and evaluates the achieved results.

A. Related Work

Yeh and Lien applied six classification methods on the same dataset: logistic regression, discriminant analysis, decision trees, neural networks, k-nearest neighbour, and Naïve Bayesian classifier [3]. The first task dealt with the credit cardholders' classification who default when the monthly repayment is due based on a metric called area ratio. The experiments showed that neural networks, decision trees, and Naïve Bayesian generated the best classification models. The second task focused on selecting a model with the best predictive ability when identifying defaulter using the Sorting Smoothing method. In this case, the neural networks achieved the highest predictive power according to the coefficient of determination R2 = 0.9647, comparable to Naïve Bayesian classifiers' coefficient R2 = 0.8994 or logistic regressions' coefficient R2 = 0.794.

Neema and Soibam compared seven machine learning algorithms through the best predictive performance of defaulters within the same data: decision tree, logistic regression, linear discriminant analysis, k-nearest neighbor, neural networks, Naïve Bayes classifier, and Random forest [4]. They used confusion matrix, cost function, and Matthews' correlation coefficient (MCC) for evaluation. Due to imbalanced data, they applied several resampling techniques on train data, such as under/oversampling, SMOTE, and ROSE techniques. Results showed that

[1] https://narrativescience.com/resources/blog

[2] https://spd.group/machine-learning/machine-learning-in-banking/

978-1-7281-8053-3/21/$31.00 ©2021 IEEE

Random forest and neural networks obtained the best performance among all algorithms, especially on original and undersampled data. Results for Random forest were 0.37 (MCC) and 9478 (cost) and neural networks 0.38 (MCC) and 9817 (cost).

Sariannidis et al. aimed at a similar task, i.e., to compare selected seven machine learning algorithms on the Taiwan cardholders dataset: Support Vector Classifier, linear Support Vector Classifier, logistic regression, decision tree, random forest, Naïve Bayes, and k-nearest neighbour [5]. They started with several descriptive statistics, like the age structure. The first part of the experiments dealt with the original dataset and 5-fold cross-validation; the second part used the same validation combined with the forward feature selection method. Most models showed very similar results, but the Support Vector Classifier resulted in a model with the highest accuracy score (82.21%) tested on data with six selected attributes and linear SVC (81.65%) tested on original data. Finally, the analysis of features importance finished with the two most important attributes: PAY 0 – repayment status in September 2005 and EDUCATION attribute.

Chishti and Awan analyzed this dataset in 2019 using neural networks (NN) as part of deep learning combined with non-linear data transformation [6]. The NN architecture contained a set of different parameters and regularization methods to prevent overfitting. The final model consisted of three hidden layers with 28 neurons. The input layer included 28 attributes. The output layer consisted of 1 output (classification to either 0 or 1) using sigmoid function; each hidden layer used activation function. This model achieved the best accuracy score: 81.83%, 84% precision score for value 0, and 67% for 1.

Çiğşar and Ünal compared six machine learning techniques, including Naïve Bayes, Bayesian network, logistic regression, decision tree (C4.5), Random forest, and artificial neural network, offered by the WEKA 3.9 software [7]. Also, the experiments investigated the importance of risk factors. The authors used the Turkish dataset containing more than sixty thousand observations, twelve descriptive attributes, and one target: 1 = repaid, 2 = non-repaid. The evaluation of the generated models used the following six metrics: accuracy, precision, sensitivity, F-measure, Receiver operating characteristic (ROC), and Root-mean-square error (RMSE). The logistic regression generated the best model: accuracy score (83.11%), RMSE (0.342), ROC score (0.843), F-measure score (0.824), precision score (82.2%) and sensitivity score (0.831). From this model, the authors extracted the most important factors based on the odds ratio, such as women are more likely to repay a debt (precisely the odd is 1.1174 times higher) than men are.

B. Methods

Our study applied CRISP-DM (Cross-industry process for data mining) as a process methodology for successful data analysis, modeling, and knowledge discovery [8]. We identified our business problem in *business understanding (*the default risk of bank customers and the decision-making process*)* and transformed it into the data mining problem (a classification task). *Data understanding* included exploratory data analysis and data visualization. *Data preparation* involved various techniques to clean and transform data into a shape that will be most efficient for the next phase – *modeling*.

In the modeling phase, we applied three machine learning algorithms from a group called ensemble classifiers: Random forest [9], bagging (decision tree as base model) [10], and boosting algorithms (AdaBoost [11], XGBoost[3], and Gradient Boosting [12]). *The Random forest* creates an ensemble of decision trees producing on a subset of data. Each decision tree is evaluated, and the best accuracy (based on the voting process) is calculated for overall accuracy for the whole model. *Bagging* algorithm, which stands for *bootstrapped aggregating*, applies a classifier on the subsets of data and is used for minimizing model variance. Each classification model calculates the accuracy score, and voting ensures the best accuracy score. *Boosting* minimizes model bias by weighing an observation that was incorrectly classified. New classifiers assign the target class to the observation correctly, and the accuracy score is improved.

The evaluation phase consists of selected performance metrics and meeting business goals. We decided on the following metrics: accuracy, precision, recall, ROC, and AUC (Area under curve). Also, we investigated the importance of the features within the generated models.

II. Experiments

A. Data

We chose public data of credit card holders from Taiwan from April to September 2005 [13]. This dataset includes thirty thousand records described by twenty-five attributes: *ID, default.payment.next.month* – target attribute and 23 independent features. The binary target attribute stands for each individual's default state; class null represents *non-defaulter*, one as a *defaulter*.

The meaning of 23 independent attributes is as follows:

- LIMIT_BAL: amount of given credit in NT dollars (New Taiwan dollars);
- SEX: gender (1 = male, 2 = female);
- EDUCATION: (0 = unknown, 1 = graduate school, 2 = university, 3 = high school, 4 = others, 5 = unknown, 6 = unknown);
- MARRIAGE: marital status of individual (0 = unknown, 1 = married, 2 = single, 3 = others);
- AGE: in years;
- PAY 0, PAY 2,…, PAY 6: repayment status from September to April 2005 (-2,-1,0 = pay duly, 1 = payment delay for one month, 2 = payment delay for two months,…, 8 = payment delay for eight months);
- BILL AMT1, …, BILL AMT6: amount of bill statement from September to April 2005 in NT dollars;
- PAY AMT1, …, PAY AMT6: amount of previous payment from September to April 2005 in NT dollars.

[3] https://github.com/dmlc/xgboost

We found a few inconsistencies in attributes PAY and EDUCATION attributes, e.g., EDUCATION values 0, 5, and 6 that refer to the unknown. The dataset contained no missing values. We calculated several essential descriptive characteristics to better understand the data, like:

- the average age of cardholders was 35 years (Fig. 1),
- the maximum amount of given credit was 1 000 000 NT dollars (nearly $35 000),
- 46.77% of credit card holders were university educated, and 23.73% of them were defaulters
- 60.37% of clients were women, and 20.78% of them defaulted.

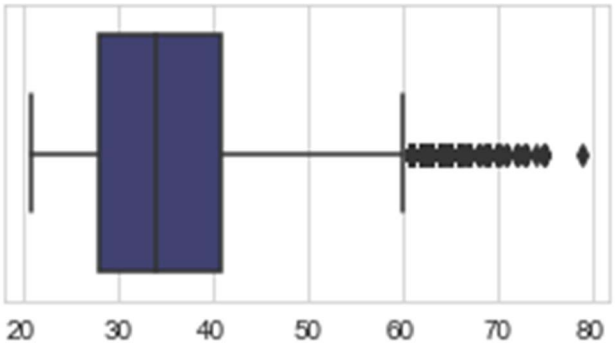

Fig. 1. The age structure of the cardholders in the dataset.

The correlation matrix allowed us to see a level of independence among numerical attributes. In Fig. 2, we plotted correlations of LIMIT BAL, AGE, BILL AMT1 to BILL AMT6, PAY AMT1 to PAY AMT6 variables. The strongest relationships were between BILL AMT attributes, e.g., Pearsons' standard correlation coefficient between BILL AMT1 and BILL AMT2 is 0.9515.

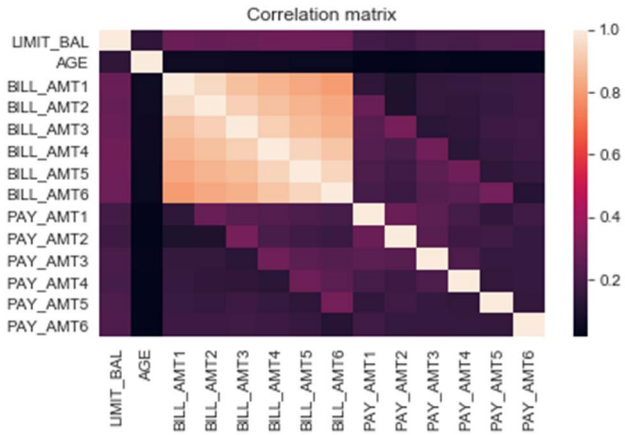

Fig. 2. Correlation matrix.

From the point of binary classification, it is essential to know the distribution of the input attributes' values. We analyzed three of them in histograms; all are skewed to the right; see Fig. 3 to 5.

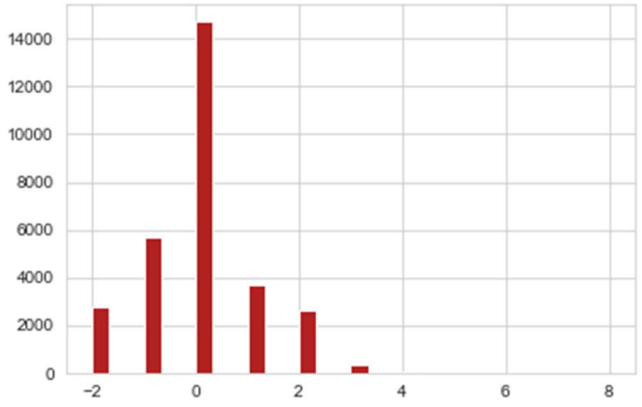

Fig. 3. Values distribution of the PAY_0 attribute.

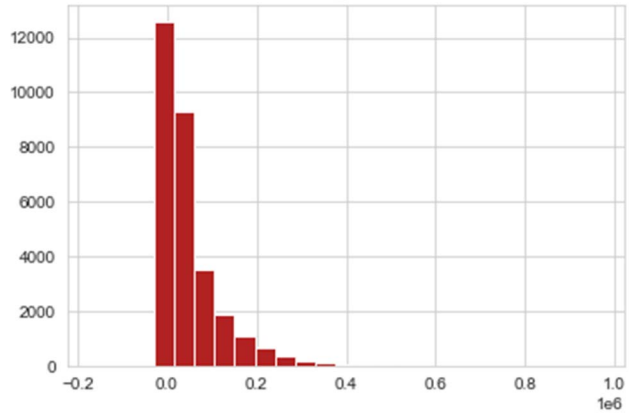

Fig. 4. Values distribution of the BILL_AMT1 attribute.

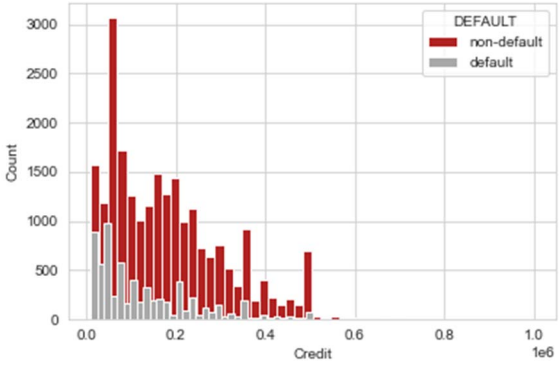

Fig. 5. Values distribution of the LIMIT_BAL attribute.

Target attribute *default.payment.next.month* was highly imbalanced; it contained only 22.12% records of target class 1 (Fig. 6).

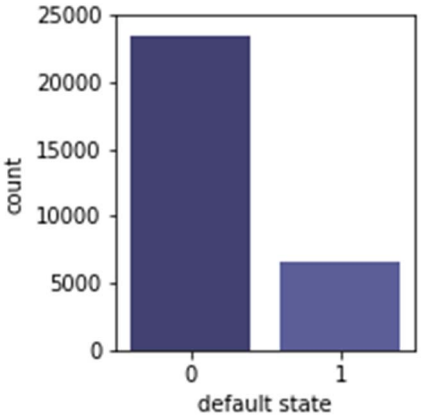

Fig. 6. Target attribute histogram.

B. Results

We divided *data* into two samples: original data without any preprocessing to create a baseline model and preprocessed dataset to improve the initial results (inconsistencies removing, data standardization). Next, we used the stratified splitting into training (70%) and testing data (30%). Table 1 visualizes the results: 1 – original dataset, 2 – processed data.

TABLE I. PERFORMANCE OF THE GENERATED CLASSIFICATION MODELS

Algorithm	Metrics				
	Accuracy	*Precision*		*Recall*	
		Class 0	Class 1	Class 0	Class 1
Random forest[1]	0.82	0.84	0.65	0.95	0.37
Random forest[2]	0.82	0.84	**0.66**	0.95	**0.36**
Bagging[1]	0.82	0.84	0.72	0.96	0.33
Bagging[2]	0.82	0.84	0.72	0.96	0.33
AdaBoost[1]	0.82	0.83	0.69	0.96	0.32
AdaBoost[2]	0.82	0.83	0.69	0.96	0.32
XGBoost[1]	0.81	0.84	0.64	0.94	0.37
XGBoost[2]	**0.82**	0.84	**0.66**	0.94	0.37
Gradient boosting[1]	0.82	0.84	0.69	0.95	0.37
Gradient boosting[2]	0.82	0.84	0.69	0.95	0.37

We also experimentally evaluated the data ratios 60/40 and 80/20, but the metrics did not improve significantly.

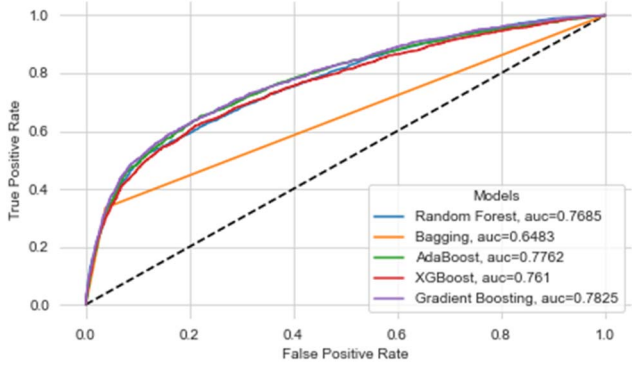

Fig. 7. ROC curves for original dataset

The two other evaluation metrics were ROC and AUC. Fig. 7 and Fig. 8 visualize the performance of all five used algorithms on both samples. The best performance within AUC was achieved by the Gradient boosting, but the best precision for target class 1 reached the Bagging algorithm (Table 1). The results show minimal improvement based on data processing.

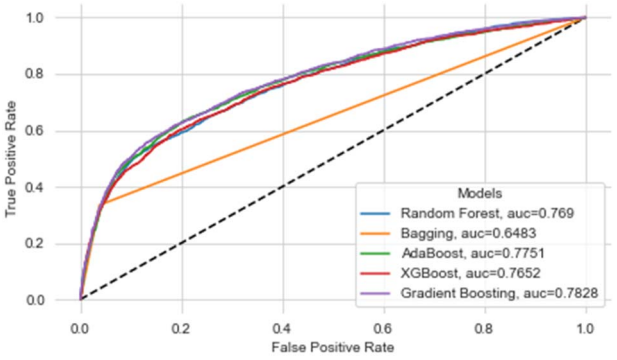

Fig. 8. ROC curves for preprocessed dataset

We also applied 10-fold cross-validation on original data, but it brought no improvements; results remained similar.

Finally, we investigated the feature importance in all generated models. Random forests' (original dataset), three most important attributes: PAY 0 (importance score is 0.0913), AGE, and BILL AMT1, while the least important attribute is defaulters' SEX (0.012) and marital status (Fig.9).

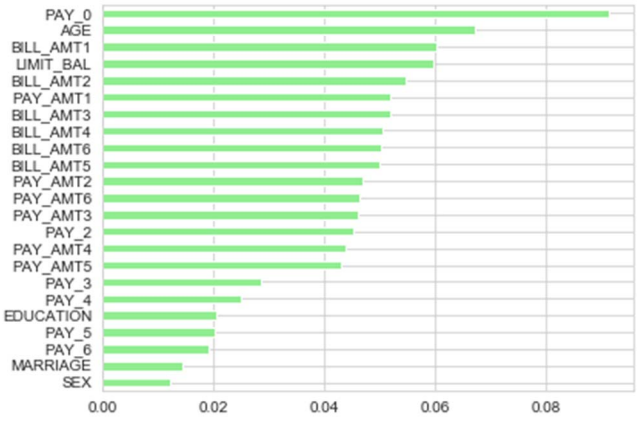

Fig. 9. The importance ranking from the model generated by the Random forest algorithm

XGBoosting classifiers' (processed dataset) three most important features: PAY 0 (importance score 0.3628) was significantly higher than scores of PAY 2 (0.1215) and PAY 3 (0.0697).

Neema and Soibams use the Matthews correlation coefficient (MCC) for model evaluation [14]. This coefficient measures the quality of classification (binary or multiple). The higher correlations' value is, the better classification we reached. In our case, the original data resulted in the following MCC values Gradient Boosting (0.4111), Bagging (0.4044), and Random forest (0.3918). The MCC of XGBoost (original data) was 0.3842 and 0.3972 on preprocessed data. We achieved very similar results within the rest of the models; no model had this metrics lower than 0.3804.

III. Conclusion

Our study aimed to evaluate various machine learning algorithms for the banking sector's selected analytical task. According to the precision score for target class one and ROC, the best algorithms are AdaBoost and Gradient Boosting (defaulter). The study of Neema and Soibam has the MCC value of 0.37 (Random forest) and 0.38 (neural networks). Our best model is Gradient Boosting with 0.4111. However, precision, accuracy, and AUC are comparable to the results of the studies. Sariannidis et al. reached the best accuracy score through the SVC algorithm on original data (81.65%), Chishti, and Awans' neural network performed with 81.83% accuracy. What's positive, results for target class one of Bagging (72%), AdaBoost (69%), and Gradient Boosting (69%) algorithms performed better than the precision of Chishti and Awans' neural network (67%). The study of Çiğşar and Ünal used a different set of data, but results are valid to compare with ours. The feature importance analysis showed which attributes are significant for defaulters' classification, i.e., repayment status and individual age; Sariannidis et al. mentioned the same conclusion. Finally, we can state that our study's algorithms are valuable for identifying clients' default state and producing a good performance. Overall, there is a need for banking industry subjects to use new technologies, such as ML or AI methods, to improve, progress, and be a part of the digital age.

Acknowledgment

The work was partially supported by The Slovak Research and Development Agency under grants no. APVV-16-0213.

References

[1] Accenture Hessens Ambitionen für Künstliche Intelligenz, Ein Beitrag zur nationalen KI-Strategie am Beispiel des Finanzsektors, 2018, available online at: https://wirtschaft.hessen.de/sites/default/files/media/hmwvl/20180925_ki_studie_hessen_report_final_im_auftrag_von_0.pdf

[2] Transforming Paradigms A Global AI in Financial Services Survey, 2020, available online at: http://www3.weforum.org/docs/WEF_AI_in_Financial_Services_Survey.pdf

[3] I-C. Yeh and C. Lien, "The comparison of data mining techniques for the predictive accuracy of probability of default of credit card clients," Expert Systems with Applications, vol. 36, Issue 2, pp. 2473-2480, March 2009.

[4] S. Neema and B. Soibam, "The comparison of machine learning methods to achieve most cost-effective prediction for credit card default," Journal of Management Science and Business Intelligence, vol. 2, No. 2, pp. 36-41, August 2017.

[5] N. Sariannidis, S. Papadakis, A. Garefalakis, C. Lemonakis and T. Kyriaki-Argyro, "Default avoidance on credit card portfolios using accounting, demographical and exploratory factors: decision making based on machine learning (ML) techniques," Annals of Operations Research, pp. 1-25, March 2019.

[6] W. A. Chishti and S. M. Awan, "Deep Neural Network a Step by Step Approach to Classify Credit Card Default Customer," 2019 International Conference on Innovative Computing (ICIC), pp. 1-8, November 2019.

[7] B. Çiğşar and D. Ünal, "Comparison of Data Mining Classification Algorithms Determining the Default Risk," Hindawi Scientific Programming, vol. 9, article ID 8706505, pp. 1-8, February 2019.

[8] C. Shearer, "The CRISP-DM model: the new blueprint for data mining", J Data Warehousing, vol. 5, pp. 13-22, 2000.

[9] L. Breiman, "Random Forests", Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[10] L. Breiman, "Bagging predictors", Machine Learning, vol. 24, pp. 123-140, 1996.

[11] Y. Freund, R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting", Journal of Computer and System Sciences, vol. 55, no. 1, pp.119-139, 1997.

[12] L. Mason, J. Baxter, P. L. Bartlett, M. Frean, "Boosting Algorithms as Gradient Descent". In S.A. Solla and T.K. Leen and K. Müller (ed.). Advances in Neural Information Processing Systems 12, MIT Press. pp. 512–518, 1999.

[13] Lichman, M. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2013.

[14] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme". Biochimica et Biophysica Acta (BBA) - Protein Structure, vol. 405, no., pp. 442-451, 1975.