Fifth International Engineering Conference on Developments in Civil & Computer Engineering Applications 2019 - (IEC2019) - Erbil – IRAQ

Prediction of default payment of credit card clients using Data Mining Techniques

Abdulhamit Subasi
Effat University, College of Engineering
Jeddah, 21478, Saudi Arabia
absubasi@effatuniversity.edu.sa

Selcuk Cankurt
Faculty of Engineering, Ishik University,
Erbil, Iraq
scankurt@ishik.edu.iq

***Abstract*— Recent studies showed that poor people had trouble with their financial decisions. In order to prevent these financial complications including decisions which are easier and more frequent in the unintentional failure for paying monthly credit card balances, we proposed a data mining-based failure prevention system from the view of risk management. This research realized on customers' default payments for the accurate prediction of the probability of default payment. Because the class of default dataset is imbalanced dataset, this study presents the Synthetic Minority Over-Sampling Technique (SMOTE) to deal with the imbalanced dataset. By utilizing SMOTE method, the predicting model produced by Random Forest has the best accuracy and performance with low error rate. Consequently, among the seven data mining algorithms, Random Forest is a good alternative to precisely predict the default payment. The proposed Random Forest model classifies Default of Credit Card Clients in the test data by predicting the target variable 89.01% correctly. This effect also resulted in an improvement of ROC area (AUC=0.947), and F-measure (0.89) of Random Forest, which was higher than other classifiers.**

Keywords—Default of credit card payment, Artificial Neural Networks (ANN), Decision Trees, Random Forest

I. INTRODUCTION

The incidence of recent financial crisis has shown that the general public is not familiar with the elementary financial concepts to make proper financial decisions. In order to understand the relationship between customer income and financial failures, the minimum monthly credit card payment of individuals should be studied deeply. Hence, in this paper we investigate the failing to pay the minimum credit card balance which is a transaction to be regularly carried out every month. In order to increase the consumer finance confidence, to avoid the delinquency it is a big challenge for cardholders and banks as well. In a well-established financial system, risk estimation is more important than crisis management. The significant reason for hazard expectation is to utilize money related data, for example, business financial statement, client transaction and reimbursement records, etc. To predict business execution or individual clients' credit chance, and to lessen the harm and instability.

In order to make accurate customer risk assessment for their credit services department, banks are demanded sophisticated credit scoring systems to automate the credit risk scoring tasks [1]. Management of the credit risk for banking sector and financial organizations have extensively started to gain importance. By developing an automated system to accurately forecast the probability of cardholder's future default, will help not only to manage the efficiency of consumer finance but also effectively handle the credit risk issues encountered in the banking sector [2] [3]. Researches conducted on automated credit-scoring assessments can be categorized into the nine main directions. These directions can be listed as: (1) single-classifier approaches, (2) multiple-classifier approaches, (3) statistical methods, (4) artificial intelligence models, (5) linear and non-linear methods, (6) parametric methods, (7) non-parametric methods, including data mining techniques such as decision trees (like C4.5 and rotation forest and random forest), neural networks (NN), fuzzy NN, k-Nearest Neighbors (KNN), genetic algorithms (GA), support vector machines, (8) ensemble models, and (9) hybrid models [3] [4].

With the evolution of data mining methods, they have been broadly used for credit risk forecasting [5] [6]. Credit risk here represents the delay in the default credit card payment [7]. From the risk control viewpoint, evaluating the likelihood of default credit card payment will be more effective in order to determine whether a cardholder is risky or no risky. Consequently, regardless of whether the assessed likelihood of default payment achieved by data mining techniques can represent the real likelihood of default payment is an imperative issue. To estimate likelihood of default, researchers and practitioners need more study [6], [8]–[13].

Hybrid credit-scoring systems based on ensemble of classifiers models and feature selection is proposed by [14], [15]. Li & Zhong [16] proposed an ensemble learning method which is one of the most recent approaches for credit-scoring. In their studies, the multiple classifiers are combined to yield a final decision. Another hybrid credit-scoring approach based on pooling of classifiers is proposed by Ala'raj and Abbod [17]. They presented the idea of data-filtering for the outliers selection where data points with labels demonstrate weak association with those of their neighbors. Moreover, they proposed a data mining system utilizing both feature selection techniques and combination of classifiers for credit scoring task [17] [18]. Chen and Huang constructed a model by combining the ANN and the rules of CART decision tree and achieved to reduce the risk [2].

978-1-7281-4377-4/19/$31.00 ©2019IEEE

Koutanaei et al. [4] proposed a data mining classifier system using four feature selection (FS) techniques namely principal component analysis (PCA), genetic algorithm, information gain ratio and relief attribute evaluation function in order to obtain a subset of relevant features with higher accuracy. The optimal values of hyper parameters used in FS algorithms are adjusted by considering the performance of SVM classification. Among the four FS algorithms, PCA was ranked the best one in terms of the accuracy. After selecting the best FS method, the classification models were implemented. Finally, by iterating the values of the classifier parameters the best model for each classification method was chosen.

Data mining algorithms usually cannot deal with imbalanced distributions of classes or unequal misclassification costs [19]. Thus, when the data set is imbalanced, data mining algorithms may fail to accurately characterize the distributive features of the data and deliver weak recognition rates between two classes. Numerous methods have been proposed to eliminate the imbalanced classification problem [20]. One of the widely known techniques to treat imbalanced datasets is sampling methods. Sampling can be either by making over-sampling to increase minority class cases or by making under-sampling to reduce the majority class cases or both. For example, the synthetic minority over-sampling technique (SMOTE) [21] is an example for the over-sampling, which generates the synthetic minority class examples; an under-sampling technique called one-side selection was suggested in [22] to decrease the majority class observations by keeping the representative characteristics of the majority class; Batista [23] employed both over-sampling and under-sampling methods to improve the classification rate of the minority class [24] [25]. In regards to over-sampling, this study demonstrates an effective solution to deal with two-class imbalanced classification problems by employing the SMOTE algorithm. Specifically, the SMOTE algorithm is first applied to generate synthetic cases in the minority class to balance the training dataset. In this experimental study, a real imbalanced data set without applying SMOTE, and with SMOTE 100% and SMOTE 200% are used to compare the accuracy performance of the seven data mining algorithms.

In the next section, the description of the datasets and data mining techniques are presented. Section 3 contains the experimental study, results and discussions. The conclusion is provided in Section 4.

II. MATERIALS AND METHODS

A. *Description of the data*

In this study, we employed the dataset involving the payment data in October 2005, which is issued by one of the famous banks from Taiwan. Yeh & Lien [26] constituted 23 explanatory variables and one corresponding dependent variable based on previous studies in the literature [27] [17] [28]. Dataset compromise of one dependent and 23 independents variables. The dependent variable labelled as "default payment for next month" is a binary variable, which is coded as either Yes=1 or No=0. The first five of the explanatory variables are about demographic characteristics. The next six variables are about the status of the past payments. The further six variables are about the amount of pass bill statement and the last six variables are about the amount of paid bills. After a statistical exploratory analysis, it is seen that the credit payment data has a typical characterization of imbalanced datasets in term of the dependent variable, for about 5529 cases (22.12%) are the customers with default payment (associated with 'yes') and 19471 records (77.88%) are associated with 'no'. The imbalanced composition of the dependent variable also can be observed. The following 23 variables are used as an attribute:
X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
X2: Gender (1 = male; 2 = female).
X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
X4: Marital status (1 = married; 2 = single; 3 = others).
X5: Age (year).
X6–X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .; X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: 1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.
X12–X17: Amount of bill statement (NT dollar).
X12 = amount of bill statement in September, 2005;
X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005.
X18–X23: Amount of previous payment (NT dollar).
X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.

B. *Synthetic Minority Over-Sampling Technique (SMOTE)*

When you employee data mining models with imbalanced dataset it could give inaccurate result. The SMOTE algorithm is one of the techniques to deal with the imbalanced classification problem . The minority class is the smallest class in the dataset. Therefore, over-sampling is applied to increase the amounts of data in the minority class, the aim is to generate similar but more descriptive cases to identify region of the minority observations in the feature space . The process of over sampled of the minority class is by considering each sample in the minority class and presenting synthetic cases along the any joining line segments or all of the k minority class nearest neighbors. The k nearest neighbors can be randomly specified depending on the required number of over-sampling [21] . Our study currently uses five nearest neighbors for the Default of Credit Card Clients datasets for one and two times. Data pre-processing techniques are one of the essential phases for developing the data mining methods in the issue of the performance. In this study, we used the SMOTE to balance the given dataset. SMOTE method generates the synthetics cases by over-sampling the minority class in the imbalanced

dataset. Furthermore, we have generated two more datasets beside the original dataset by using the different configuration of the SMOTE model. As a result, we have simulated and assessed our data mining models by using three different datasets: original dataset, dataset with SMOTE 100% (one-pass), and dataset with SMOTE 200% (two-pass).

C. Artificial Neural Networks (ANN)

Neural networks are developed based on the models revealed by the biological neural network of the brain. A single perceptron can classify linearly separable classes and approximate the linear functions. On the other hand, it cannot solve a non-linear classification problem. But a feedforward network, for example, multilayer perceptron (MLP), which has hidden layers has the ability to approximate nonlinear functions and to separate the non-linear classes. The network output is a linear combination of the outputs computed by the hidden neurons using non-linear activation functions. Backpropagation Algorithm is used to train the multilayer perceptron, which is based on the perceptron training technique. The only difference is the output of the hidden layer is inputs of the next hidden or output layer. Outputs of the nodes are calculated using perceptron training techniques and computed errors are propagated back to the previous layers, and then weights are updated according to the calculated gradient to reduce the error term. Back propagations of the errors, which is a technique to train the network is used to name the backpropagation algorithm .

D. Support Vector Machine (SVM)

This soft computing algorithm has proved robustness and accuracy among all state-of-the-art machine learning techniques. In a binary classification task, the task of SVM is maximizing the margin that gives the largest possible distance between members of classes in the training examples, which are linearly separable. In the higher dimensional datasets, SVM finds a hyperplane that separates two classes. There may be many such linear hyperplanes, but here SVM tries to discover the optimal hyperplane to separate the classes in a dataset by maximizing the margin. A solution to this problem is a typical example of the quadratic programming problem, which can be solved by the Lagrangian multiplier technique. The solution of the problem has just a single global minimum. The reason why SVM can find the optimum margin that it provides the best generalization ability. It obtains not only the best accuracy performance in the classification of the training data but also provides much space to accurately separate the future data .

E. k-Nearest Neighbor (k-NN)

It is a non-parametric classifier technique widely used in the machine learning. This technique works based on the simple analogy that is by comparing a given test case with training cases that are close to it. When a new sample is introduced, the K-nearest algorithm finds the pattern space for the k training observations that are similar to the new case. These k training samples are called the k “nearest neighbors” of the new sample. “The degree of the nearness” is measured in terms of a distance metric, for example, Euclidean distance. In this classification technique, the new sample will be classified as the most common class amongst its K-nearest neighbors if it has k-nearest neighbors. If K = 1, then the sample is simply classified as the class of its nearest neighbor. It is a common practice that if there is a larger number of training samples, a larger value of k will be used . But, when you have the large-scale of datasets, K-nearest algorithm requires the labor-intensive effort.

F. C4.5 Decision Tree (DT)

This data mining technique is very popular DT induction model . It is an extension for the ID3 and addresses many issues, which cannot be dealt with by ID3. ID3 uses a statistical test of independence for a stop criterion to avoid generation of the oversized trees, which may cause overfitting the training data. Moreover, C4.5 proposes a different method to gain the generalization ability. It generates (almost) maximally optimum trees and then reduces the size of them by pruning the very detailed nodes that cannot be generalized but just fits to some specific samples in the training dataset. C4.5 has a pre-pruning control parameter. This control avoids the generation of the nodes with a few leaves. Next to the construction of tree, each node is tested with a statistical tool to estimate the probability, which estimates whether a new split improves the error reduction. Any node with the estimated probability below than a given threshold is pruned or grafted. Graphing is adding new nodes either by replacing of a single leave or within the leaves to increase the accuracy .

G. NB Tree

The NBTree, which is introduced by Kohavi [37] works like the traditional recursive partitioning schemes, but only difference on the generation of the leaf, which is generated by the Naive-Bayes classifier instead of nodes predicting. On the other words, NBTree calculates split nodes for each feature using a decision tree, but then the decision of which features will be used differs from the decision tree. To give this decision, NBTree employs 5-fold cross-validation and applies naive Bayes for each subset defined by the splits, to compute classification error. The feature with the smallest classification error is selected to generate the split for the current node .

H. Random Forests

Random forest algorithm employs bagging ensemble technique to implement the classification model. This ensemble approach has two distinctive techniques to construct the base models: bootstrap data sampling and algorithm nondeterminism. In connection with bootstrap data sampling, it improves the diversity of the base models by randomizing the decision tree algorithms. Nondeterminism in the nature of the algorithm is obtained by the employment of the random split selection technique, which is used for tree generation. Stopping criteria for decision tree generation is result in relatively large, accurately generated trees and no pruning is

applied. This tree growing technique generates many splits, which results in a great diversity of base learners. Individual overfitting of the models is cancelled out by the aggregation method, which makes this ensemble of the trees to be highly resistant to overfitting .

I. Rotation Forest

This data mining algorithm generates an ensemble of decision trees using combination of bagging and random subspace methods together with principal components feature generation. In each iteration input features are randomly divided into subsets. Principal components analysis (PCA) is applied to each subset to obtain linear combinations of the features in the subset that are rotations of the original axes. Before the PCA is applied on subset in each iteration a bootstrap sample of data can be employed to further increase diversity. Research shows that rotation forest has similar performance when compared to random forest, using far fewer trees. The analysis of the diversity regarding error for the members of the ensemble indicates a minimal increase in diversity and reduction of error for rotation forest in comparison with bagging. Still it has significantly better performance for the ensemble as whole [39].

II.

III. Results and Discussion

In the experiments, we have used three different versions of the dataset for Default of Credit Card Clients data to assess the accuracies of models. These datasets are Default of Credit Card Clients (Without SMOTE), Default of Credit Card Clients with SMOTE 100%, and with SMOTE 200%, which are explained in previous section. We have employed seven data mining models namely, multilayer perceptron (ANN), Random Forest (RF), Support Vector Machine (SVM), C4.5 decision tree (J48), k-Nearest neighbour (k-NN), Hybrid Naive Bayes/Decision-Tree (NBTree), and ensemble trees of rotation forest. In this work, we used WEKA software which is publicly available JAVA based open source data mining libraries. 10-fold cross validation technique is utilized during the experiments. We compared the efficiency of proposed models without SMOTE, with SMOTE 100%, and with SMOTE 200% (two-pass processing). In the evaluation of the classifier performance, accuracy, F-measure and a ROC area is employed.

A. Results without SMOTE

The experimental results without using any pre-processing technique for Default of Credit Card Clients dataset (Without applying SMOTE) are presented in Table I. We obtain the average accuracies of 80.21% for k- Nearest neighbour (k-NN), 80.56% for DT (C4.5), 81.61% for Random Forest, 81.68% for Multilayer Perceptron (ANN), 81.97% for SVM, 79.68% for NBTree, and 82.14% for Rotation Forest.

B. Results with SMOTE 100%

The experimental results after balancing the dataset with SMOTE 100% technique are shown in Table II. We acquire the average accuracies of 78.41% for k- Nearest neighbor (k-NN), 82.11% for DT (C4.5), 84.59% for Random Forest, 81.70% for Multilayer Perceptron (ANN), 73.52% for SVM, 83.95% for NBTree, and 83.96% for Rotation Forest. After balancing the dataset using SMOTE 100%, the total accuracy has increased up to the 4.27%. On the other hand, by means of the minority class accuracy, more dramatic enhancement can be seen, for example, in the use of Rotation Forest, the accuracy performance of the minority class has increased up to the 33%. The accuracy of the majority class has slightly decreased in every model except NBTree model, which achieved the 5% increment.

C. Results with SMOTE 200%

The experimental results after balancing the dataset with SMOTE 200% technique are presented in Table III. We attain the average accuracies of 81.85% for k- Nearest neighbour (k-NN), 86.53% for DT (C4.5), 89.01% for Random Forest, 84.16% for Multilayer Perceptron (ANN), 69.31% for SVM, 87.55% for NBTree, and 88.32% for Rotation Forest. After balancing the dataset using SMOTE 200%, the total accuracy has increased up to the 8%. On the other hand, the accuracy for the minority class has increased much more than the total accuracy, for example, 33% for NBTree model, 50%, for Rotation forest model. The accuracy of the majority class decreased in every model except NBTree model, which achieved the 9% increment.

D. Discussion

As seen from Table I, the results without applying SMOTE, Rotation Forest achieves the highest accuracy with the percentage of 82.14%, and NBTree has the lowest accuracy rate with 79.68%. While in Default of Credit Card Clients with SMOTE 100% result in Table II, Random Forest achieves the highest obtained the accuracy with 84.59%, and then Rotation Forest and NBTree come after the Random Forest with the accuracy of 83.96% for Rotation Forest and 83.95% for NBTree. Also, in the Default of Credit Card Clients with SMOTE 200% result, Random Forest achieves the highest obtained accuracy with the percentage of 89.01% as seen in Table III.

From the obtained results, three important observations can be seen: (i) applying SMOTE dataset balancing technique in the Default of Credit Card Clients dataset achieved a higher performance in terms of accuracy, (ii) Random Forest outperformed the other models reaching the highest accuracy for two datasets (With SMOTE 100% & With SMOTE 200%) out of the three, (iii) In the use of balancing with SMOTE 100% technique, only decision tree models achieved better accuracy. But in the use of balancing with SMOTE 200%, all data mining models obtained better accuracy except SVM method.

Rotation Forest has achieved 82.14% of the accuracy without using any dataset balancing technique and is ranked the first. After SMOTE 100% technique, it provides slightly better accuracy with the improvement of 1.82%. Using SMOTE 200% technique significantly enhances its accuracy (6%). SVM provides accuracy of 81.97% without employing any dataset balancing technique and is ranked the 2nd place. But after balancing the dataset using SMOTE 100% and SMOTE 200%, it does not obtain any improvement in accuracy. ANN has achieved 81.68% of the accuracy without using any dataset balancing technique and is ranked the 3rd place. After SMOTE 100% technique it does not obtain any promising improvement in accuracy (0.02%). Using SMOTE 200% technique enhances its accuracy only by 2%. After balancing the dataset using both SMOTE 100% and SMOTE 200% techniques, Random Forest model is ranked the first by reaching the accuracy values of 84.59% and 89.01% respectively.

Accurate identification of Default of Credit Card Clients is important for banks evaluation. The proposed Random Forest model classifies Default of Credit Card Clients with SMOTE 200% data in the test data by predicting the target variable 89.01% correctly. This effect also resulted in an improvement of ROC area (AUC=0.947), and F-measure (0.89) of Random Forest, which was higher than other classifiers. After applying several kinds of data mining algorithms on our selected datasets, Rotation Forest also resulted in satisfactory high accuracy of 88.32% and ranked as the second-best classifier.

The main disadvantage of SMOTE is that by producing exact copies of present instances, it results overfitting likely. Actually, with SMOTE it is rather common for a learner to generate a classification rule to cover a single, replicated, example. Another disadvantage of oversampling is to increase the number of training examples, thus increasing the learning time.

IV. CONCLUSION

A wide number of researchers have been studied in the credit card default payment to reach an accurate financial decision. Data mining models have been broadly employed for these tasks. In this research, various data mining models with and without SMOTE have been implemented to efficiently classify the credit card default payment. Random Forest with SMOTE 200% obtained the highest accuracy of 89.01%. In this study, one of the highest classification accuracies was achieved compared to all previous researches done in this field. Also, by using random forest we reached good classification accuracy. Many powerful data mining models in combination with SMOTE have been applied to the Default of Credit Card Clients dataset, which is one of the imbalanced classification problem. It is shown in this study that, instead of applying hybrid methods based on complex algorithms and methods to accomplish high classification rates, a group of more simple data mining models with the pre-processing of dataset balancing technique can be employed as well producing outstanding results. A group of several methods offers us to use advantages of each method in order to obtain high accuracy rates for the Default of Credit Card Clients. We can employ a group of these rather simple methods to classify the clients and to help financial people to make more accurate decision in the Default of Credit Card Clients.

REFERENCES

[1] G. Fan and J. B. Gray, "Regression tree analysis using TARGET," J. Comput. Graph. Stat., vol. 14, no. 1, pp. 206–218, 2005.

[2] C.-C. Lin, S.-C. Chen, and Y.-M. Chu, "Automatic price negotiation on the web: An agent-based web application using fuzzy expert system," Expert Syst. Appl., vol. 38, no. 5, pp. 5090–5100, 2011.

[3] A. Verikas, Z. Kalsyte, M. Bacauskiene, and A. Gelzinis, "Hybrid and ensemble-based soft computing techniques in bankruptcy prediction: a survey," Soft Comput., vol. 14, no. 9, pp. 995–1010, 2010.

[4] F. N. Koutanaei, H. Sajedi, and M. Khanbabaei, "A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring," J. Retail. Consum. Serv., vol. 27, pp. 11–23, 2015.

[5] H. C. Koh and C. K. L. Gerry, "Data mining and customer relationship marketing in the banking industry," Singap. Manag. Rev., vol. 24, no. 2, p. 1, 2002.

[6] L. C. Thomas, "A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers," Int. J. Forecast., vol. 16, no. 2, pp. 149–172, 2000.

[7] P. Giudici, "Bayesian data mining, with application to benchmarking and credit scoring," Appl. Stoch. Models Bus. Ind., vol. 17, no. 1, pp. 69–81, 2001.

[8] B. Baesens, R. Setiono, C. Mues, and J. Vanthienen, "Using neural network rule extraction and decision tables for credit-risk evaluation," Manag. Sci., vol. 49, no. 3, pp. 312–329, 2003.

[9] V. S. Desai, J. N. Crook, and G. A. Overstreet, "A comparison of neural networks and linear scoring models in the credit union environment," Eur. J. Oper. Res., vol. 95, no. 1, pp. 24–37, 1996.

[10] D. J. Hand and W. E. Henley, "Statistical classification methods in consumer credit scoring: a review," J. R. Stat. Soc. Ser. A Stat. Soc., vol. 160, no. 3, pp. 523–541, 1997.

[11] I. Jagielska and J. Jaworski, "Neural network for predicting the performance of credit card accounts," Comput. Econ., vol. 9, no. 1, pp. 77–82, 1996.

[12] T.-S. Lee, C.-C. Chiu, C.-J. Lu, and I.-F. Chen, "Credit scoring using the hybrid neural discriminant technique," Expert Syst. Appl., vol. 23, no. 3, pp. 245–254, 2002.

[13] E. Rosenberg and A. Gleit, "Quantitative methods in credit management: a survey," Oper. Res., vol. 42, no. 4, pp. 589–613, 1994.

[14] C.-F. Tsai, "Feature selection in bankruptcy prediction," Knowl.-Based Syst., vol. 22, no. 2, pp. 120–127, 2009.

[15] P. Yao, "Feature selection based on SVM for credit scoring," presented at the Computational Intelligence and Natural Computing, 2009. CINC'09. International Conference on, 2009, vol. 2, pp. 44–47.

[16] X. Li and Y. Zhong, "An overview of personal credit scoring: Techniques and future work.," pp. 181–189, 2012.

[17] M. Ala'raj and M. F. Abbod, "A new hybrid ensemble credit scoring model based on classifiers consensus system approach," Expert Syst. Appl., vol. 64, pp. 36–55, Dec. 2016.

[18] M. Ala'raj and M. F. Abbod, "A new hybrid ensemble credit scoring model based on classifiers consensus system approach," Expert Syst. Appl., vol. 64, pp. 36–55, 2016.

[19] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263–1284, 2009.

[20] C. Elkan, "The foundations of cost-sensitive learning," presented at the International joint conference on artificial intelligence, 2001, vol. 17, pp. 973–978.

[21] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.

[22] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: one-sided selection," presented at the ICML, 1997, vol. 97, pp. 179–186.

[23] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," ACM Sigkdd Explor. Newsl., vol. 6, no. 1, pp. 20–29, 2004.

[24] F. Cheng, J. Zhang, C. Wen, Z. Liu, and Z. Li, "Large cost-sensitive margin distribution machine for imbalanced data classification," Neurocomputing, vol. 224, pp. 45–57, Feb. 2017.

[25] M. Gao, X. Hong, S. Chen, and C. J. Harris, "A combined SMOTE and PSO based RBF classifier for two-class imbalanced problems," Neurocomputing, vol. 74, no. 17, pp. 3456–3466, 2011.

[26] C. Yeh and C. Lien, "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients," pp. 2473–2480, 2009.

[27] A. Steenackers and M. Goovaerts, "A credit scoring model for personal loans," Insur. Math. Econ., vol. 8, no. 1, pp. 31–34, 1989.

[28] B. Scholnick, N. Massoud, and A. Saunders, "The impact of wealth on financial mistakes: Evidence from credit card non-payment," pp. 27–36, 2012.

TABLE I RESULTS FOR DEFAULT OF CREDIT CARD CLIENTS (WITHOUT SMOTE) USING DIFFERENT DATA MINING TECHNIQUES

	k-NN			ANN			SVM			C4.5 decision tree			RF			Rotation Forest			NBTree	
	ROC Area	F – measure	Accuracy (%)	ROC Area	F - measure	Accuracy (%)	ROC Area	F - measure	Accuracy (%)	ROC Area	F - measure	Accuracy (%)	ROC Area	F - measure	Accuracy (%)	ROC Area	F - measure	Accuracy (%)	ROC Area	F - measure
YES	0.715	0.435	34.50%	0.743	0.462	36%	0.65	0.458	35%	0.656	0.447	36%	0.76	0.485	39%	0.769	0.466	35%	0.751	0.502
NO	0.715	0.88	93.20%	0.743	0.89	94.80%	0.65	0.892	95.50%	0.656	0.882	93.30%	0.76	0.888	93.70%	0.769	0.893	95.50%	0.751	0.872
TOTAL	**0.715**	**0.782**	**80.21%**	**0.743**	**0.795**	**81.68%**	**0.65**	**0.796**	**81.97%**	**0.656**	**0.786**	**80.56%**	**0.76**	**0.799**	**81.61%**	**0.769**	**0.798**	**82.14%**	**0.751**	**0.79**

TABLE II. RESULTS FOR DEFAULT OF CREDIT CARD CLIENTS WITH SMOTE 100% USING DIFFERENT DATA MINING TECHNIQUES

	k-NN			ANN			SVM			C4.5 decision tree			RF			Rotation Forest			NBTree	
	ROC Area	F – measure	Accuracy (%)	ROC Area	F - measure	Accuracy (%)	ROC Area	F - measure	Accuracy (%)	ROC Area	F - measure	Accuracy (%)	ROC Area	F - measure	Accuracy (%)	ROC Area	F - measure	Accuracy (%)	ROC Area	F - measure	Accurac
YES	0.814	0.66	57.70%	0.839	0.71	62%	0.667	0.535	42%	0.831	0.727	66%	0.89	0.765	69%	0.881	0.753	68%	0.875	0.749	66%
NO	0.814	0.842	90.20%	0.839	0.866	93.00%	0.667	0.815	91.40%	0.831	0.867	91.30%	0.89	0.885	93.20 %	0.881	0.881	93.30%	0.875	0.882	94%
TOTAL	**0.814**	**0.776**	**78.41%**	**0.839**	**0.81**	**81.70%**	**0.667**	**0.713**	**73.52%**	**0.831**	**0.816**	**82.11%**	**0.89**	**0.842**	**84.59 %**	**0.881**	**0.835**	**83.96%**	**0.875**	**0.834**	**83.95%**

TABLE III. RESULTS FOR DEFAULT OF CREDIT CARD CLIENTS WITH SMOTE 200% USING DIFFERENT DATA MINING TECHNIQUES

	k-NN			ANN			SVM			C4.5 decision tree			RF			Rotation Forest			NBTree	
	ROC Area	F – measure	Accuracy (%)	ROC Area	F - measure	Accuracy (%)	ROC Area	F - measure	Accuracy (%)	ROC Area	F - measure	Accuracy (%)	ROC Area	F - measure	Accuracy (%)	ROC Area	F - measure	Accuracy (%)	ROC Area	F - measure
YES	0.887	0.823	79.10%	0.902	0.843	80%	0.696	0.695	66%	0.902	0.868	83%	0.947	0.892	85%	0.941	0.883	85%	0.93	0.871
NO	0.887	0.814	85.00%	0.902	0.84	89.10%	0.696	0.691	73.30%	0.902	0.862	90.10%	0.947	0.89	93.10%	0.941	0.886	92.00%	0.93	0.88
TOTAL	**0.887**	**0.819**	**81.85%**	**0.902**	**0.842**	**84.16%**	**0.696**	**0.693**	**69.31%**	**0.902**	**0.865**	**86.53%**	**0.947**	**0.89**	**89.01%**	**0.941**	**0.883**	**88.32%**	**0.93**	**0.875**