ORIGINAL RESEARCH PAPER

Volume-7 | Issue-10 | October-2018 | PRINT ISSN No 2277 - 8179

INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH

# APPLICATION OF MACHINE LEARNING ALGORITHMS IN CREDIT CARD DEFAULT PAYMENT PREDICTION

Information Technology

| | |
|---|---|
| **Admel Husejinovic*** | International Burch University, Faculty of Engineering and Natural Sciences, Department of Information Technologies *Corresponding Author |
| **Dino Keco** | Ph.D., International Burch University, Sarajevo, Bosnia and Herzegovina |
| **Zerina Masetic** | Ph.D., International Burch University, Sarajevo, Bosnia and Herzegovina |

## ABSTRACT

Credit card default payment prediction studies are very important for any financial institution dealing with credit cards. The purpose of this work is to evaluate the performance of machine learning methods on credit card default payment prediction using logistic regression, C4.5 decision tree, support vector machines (SVM), naive Bayes, k-nearest neighbors algorithms (k-NN) and ensemble learning methods voting, bagging and boosting. The performance of the algorithms is evaluated through following performance metrics: accuracy, sensitivity and specificity. The best result among all algorithms for overall accuracy rate was achieved by logistic regression model with a rate of 0.820. The best performing model for default credit card customer detection, with success of 71,3% was naive Bayes model. This approach could improve and ease the process of credit card default, and therefore help the banking system in decision making.

## KEYWORDS

default payment prediction, machine learning methods, ensemble learning

## INTRODUCTION

In bank management, a credit risk is one of the biggest issues. Credit cards are taking a big part of bank's landings. In a modern banks whole division is dealing only with credit cards. Banks use different techniques and score cards to classify customers as potentially good or potentially bad ones.

Banks collect their approved and non-approved requests and by use of machine learning techniques we will test accurately and make the prediction whether a new applicant is going to be default or not. Eight different algorithms are used in this study for credit risk prediction. Naïve Bayes, k-nearest neighbor (k-NN), support vector machines (SVM), Logistic Regression, C4.5 decision tree, Bagging, Boosting and Voting. Each machine learning algorithms' performance is used to compare with other algorithms' performances through accuracy, sensitivity and specificity rates.

## LITERATURE REVIEW

Credit default prediction using machine learning algorithms is widespread topic among researchers. According to Yeh and Lien studied probability of credit card customers default, their finding suggested that artificial neural network was best outperforming machine learning algorithm [1]. Studies performed by Ince and Aktan published in 2009 showed that some traditional scoring models were outperformed by the neural networks in terms of type II errors and predictive accuracy [2]. Khandani, Kim and Lo in 2010 were applying machine learning techniques to make forecasting models for customers' default predictions. They used customers' credit transactions and credit agency information in way that significantly improved the classification rates of defaults and customers misbehaviors, with linear regression $R^2$'s of forecasted misbehaviors [3]. Singh and Aggarwal found out that support vector machines and generic programing are superior models for the determination of classifying the loan applicants. As the result they compare results after using 10 folds cross validation and using 688 folds cross validation whose summary is presented in table 1 below [4]. In 2013 Bellotti and Crook presented models of borrower default for credit cards that include social data about credit card holders combined with macroeconomic parameters. Overall, they found that models containing social data and macroeconomic parameters were statistically more significant [5] than other models. Kraus in 2014 made the empirical evaluation of credit risk based on data from a German bank using dataset with personal information of client [6]. Jacob in 2016 did studies that has been examining for most useful features for prediction and the accuracy of Random Forest Ensemble method showed the highest rate in predicting credit card defaulters [7]. Pasha, Fatima, Dogar and Shahzad in 2017 investigated credit card default risk. In their study six different data mining techniques are modeled with dataset. The results of this research show that the neural network performs best to predict the default of credit card clients and shows the highest accuracy [8].

**TABLE – 1: Singh and Aggarwal's results**

| 10 folds model | Sensitivity | Specificity |
|---|---|---|
| LDA | 0.879 | 0.872 |
| LR | - | - |
| MLP | 0.849 | 0.866 |
| Decision Tree | 0.928 | 0.799 |
| GP | 0.885 | 0.817 |
| SVM | 0.875 | 0.861 |

Source: Singh and Aggarwal [4]

## METHODOLOGY

As methodology part of this research eight machine learning models (logistic regression, C4.5, SVM, naive Bayes, k-NN and ensemble learning methods voting, bagging and boosting) were formed using 10 folds cross validation as testing option. In this research we used Waikato Environment for Knowledge Analysis (weka) as machine learning software. We apply outlier and extreme values elimination based on interquartile ranges. For feature selection we use wrapper method called classifier subset evaluator and as a search method "Best First Method". Data is collected from Taiwan's bank in period of April to September 2005 with 24 attributes like history of payment transactions, bill statements of credit card clients, some personal information and default payments for the next month as binary class attribute (0 class – nondefault, 1 class - default).

## RESULTS

On the original dataset we have 30.000 instances, but after outlier and extreme value elimination we have 23.382 instance left. Ratio of 0 and 1 classes in total instance changed by 2.27 from (0 class) 77.88% – (1 class) 22.12% to (0 class) 75.61% – (1 class) 24.39%. This might me important for further interpretation of model performance. In the table 2 there are accuracy rates for three stages of dataset (initial, no outliers and feature selected dataset) been tested on.

**TABLE – 2: Accuracy**

| Model | Initial dataset | No outliers | Feature selection |
|---|---|---|---|
| Naive Bayes | 0.655 | 0.766 | 0.806 |
| K-NN | 0.812 | 0.795 | 0.792 |
| SVM | 0.819 | 0.803 | 0.804 |
| Logistic | 0.820 | 0.804 | 0.805 |
| C4.5 | 0.808 | 0.789 | 0.791 |
| Bagging | 0.816 | 0.801 | 0.797 |
| Boosting | 0.813 | 0.796 | 0.796 |
| Voting | 0.789 | 0.795 | 0.797 |

**Source: Author's using weka tools**

In the table 3 there are specificity rates for mentioned machine learning algorithms.

**TABLE – 3: Specificity**

| Model | Initial dataset | No outliers | Feature selection |
|---|---|---|---|
| Naive Bayes | 0.713 | 0.528 | 0.396 |
| K-NN | 0.311 | 0.330 | 0.327 |
| SVM | 0.326 | 0.338 | 0.338 |
| Logistic | 0.380 | 0.376 | 0.379 |
| C4.5 | 0.348 | 0.376 | 0.391 |
| Bagging | 0.369 | 0.401 | 0.289 |
| Boosting | 0.278 | 0.289 | 0.289 |
| Voting | 0.429 | 0.433 | 0.420 |

Source: Author's using weka tools

In the table 4 there are sensitivity rates for mentioned machine learning algorithms.

**TABLE – 4: Sensitivity**

| Model | Initial dataset | No outliers | Feature selection |
|---|---|---|---|
| Naive Bayes | 0.638 | 0.843 | 0.938 |
| K-NN | 0.955 | 0.945 | 0.954 |
| SVM | 0.959 | 0.954 | 0.955 |
| Logistic | 0.945 | 0.943 | 0.945 |
| C4.5 | 0.938 | 0.924 | 0.921 |
| Bagging | 0.943 | 0.931 | 0.925 |
| Boosting | 0.965 | 0.960 | 0.960 |
| Voting | 0.925 | 0.913 | 0.920 |

Source: Author's using weka tools

The results of the Pasha et al research indicate that the neural network performs best to predict the default of credit card clients and shows the highest accuracy. In our research we found the best model for predicting default credit card client was SVM and logistic regression model performed highest accuracy for our initial dataset. Singh and Aggarwal found highest sensitivity rate using Decision Tree while we find highest sensitivity rate using Boosting classifier using REPTree as base learner model. Singh and Aggarwal found highest specificity rate using Linear Discriminant Analysis while we find highest specificity rate using naive Bayes model.

The best results among all algorithms for overall accuracy were achieved by logistic regression with an accuracy of 82,00% using initial dataset, 80,40% using dataset after removing outliers and extreme values and naive Bayes model using dataset after feature selection with an accuracy of 80,60%. However, it should be considered that after removing an extreme value and outlier percentage of nondefault payment caring instances in dataset dropped for 2,27 (from 77,88% to 75,61%). Also, it should be pointed that naive Bayes and voting models' accuracy was increasing through tree steps of dataset change (initial dataset, remove of outliers, and feature selection steps) from 65,5%, 76,6% up to 80,6% respectively which is not case for the rest of six models. Voting model's accuracy was also increasing in each step from 78,9%, 79,5% up to 79,7%. The reason for naive Bayes rapid increase in accuracy might be the structure of the naive Bayes algorithm, ignoring of dependencies between attributes, which might exist in the present case.

The best results among all algorithms for overall specificity rate were achieved by naive Bayes model with a rate of 0.713 using initial dataset, 0.528 using dataset after removing outliers and extreme values and voting classifier with majority voting role using naive Bayes, logistic regression, C4.5 as base learners' model using dataset after feature selection with a rate of 0.420. If we remind from confusion matrix that specificity rate is equal to true negatives over total actual negatives that means our model naive Bayes using initial dataset was successful in correctly detecting a fraction of 1 class (defaulters) with more accuracy on initial dataset that other models. This means that 71,3% of all defaulters were correctly classified using Naïve Bayes Model on initial dataset.

The best results among all algorithms for overall sensitivity rate were achieved by Boosting classifier using REPTree as base learner model with a rate of 0.965 using initial dataset, 0.960 using dataset after removing outliers and extreme values and after feature selection with a rate of 0.960. Also, it should be pointed that for all models tested except naive Bayes sensitivity rate on initial dataset is higher than on other two. If we remind that sensitivity rate is equal to true positives over total actual positives that means our models were success in correctly detecting a fraction of 0 class with more accuracy on initial dataset that other two datasets. Except for naive Bayes model where we have an opposite result.

## CONCLUSIONS

Overall, this study suggested that for the modeling of credit card risk accuracy rate of 0.820 achieved by Logistic Regression model has the highest rate compared to other models. Sensitivity rate for Boosting classifier using REPTree as base learner model has the highest rate compared to other models. Specificity rate of 0.713 achieved by naive Bayes model has the highest rate compared to other models.

It must be remained that research was done for only one data set from 2005 year from only one country. For further researches bigger data set from more than one country and more attributes including macroeconomic indicators can be done before drawing general conclusions.

## REFERENCES:

[1] Yeh, I. C., & Lien, C. hui. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications. https://doi.org/10.1016/j.eswa.2007.12.020
[2] Ince, H., & Aktan, B. (2009). A comparison of data mining techniques for credit scoring in banking: A managerial perspective. Journal of Business Economics and Management. https://doi.org/10.3846/1611-1699.2009.10.233-240
[3] Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. Journal of Banking and Finance. https://doi.org/10.1016/j.jbankfin.2010.06.001
[4] Singh, R., & Aggarwal, R. R. (2011). Comparative Evaluation of Predictive Modeling Techniques on Credit Card Data. International Journal of Computer Theory and Engineering. https://doi.org/10.7763/IJCTE.2011.V3.377
[5] Bellotti, T., & Crook, J. (2013). Forecasting and stress testing credit card default using dynamic models. International Journal of Forecasting, 29(4), 563–574.
[6] Kraus, A. (2014). Recent Methods from Statistics and Machine Learning for Credit Scoring.
[7] Jacob, S. G. (2016). Prediction of Credit-Card Defaulters: A Comparative Study on Performance of Classifiers. International Journal of Computer Applications..
[8] Pasha, M., Fatima, M., Dogar, A. M., & Shahzad, F. (2017). Performance Comparison of Data Mining Algorithms for the Predictive Accuracy of Credit Card Defaulters. IJCSNS International Journal of Computer Science and Network Security.
[9] I. Cheng Yeh. (2016). UCI Machine Learning Repository: default of credit card clients Data Set. Retrieved June 17, 2018, from https://archive.ics.uci.edu/ml/datasets/default+ of+credit+card+clients+of+credit+card+clients. [Accessed: 17-Jun-2018].