Alexandru COŞER, PhD Student
Economic Cybernetics and Statistics Doctoral School
The Bucharest University of Economic Studies
E-mail: alexandru.coser@gmail.com
Lecturer Monica Mihaela MAER-MATEI, PhD
The Department of Economic Informatics and Cybernetics
The Bucharest University of Economic Studies
E-mail: matei.monicamihaela@gmail.com
Professor Crişan ALBU, PhD
The Department of Economic Informatics and Cybernetics
The Bucharest University of Economic Studies
E-mail:crisan.albu@csie.ase.ro

PREDICTIVE MODELS FOR LOAN DEFAULT RISK ASSESSMENT

Abstract. *Financial institutions are faced with the need to assess the creditworthiness of a borrower that applies for a loan. In this regard, data scientistscan produce valuable insights that can explain customer profile and behavior. This paper proposes an analysis of a database of customers where a part of them were unable to repay their loans and got into default status. By using the methodology of data mining and machine learning algorithms, a series of predictive models were developedusing classifiers such as LightGBM, XGBoost, Logistic Regression and Random Forest in order to evaluate the probability of a customer's enteringloan default. Three sampling scenarios were created to compare the classification between imbalanced and balanced data sets. Moreover, a model comparison analysis was performed to identify the best classifier by considering the model performance metrics: AUC score, Precision, Recall and Accuracy. The best results were observed for the Random Forest optimal classifier applied on the combined scenario under-over sampling, with a representative AUC of 0.89.*

Keywords:*data mining, machine learning, loan default, AUC score, predictive model, Random Forest, confusion matrix.*

JEL Classification: C52, C55, C63

1. Introduction

The ever increasing interest in data exploration and analysis in the financial-banking field has increased the need for a better understanding of customer behavior and profile. The ability to discriminate between clients – i.e. to identify

DOI: 10.24818/18423264/53.2.19.09

them as good or bad payers – is important for banks and other lending institutions. A slight improvement in the accuracy of prediction may result in a considerable increase in these institutions' profitability. Moreover, the early identification of customers who display a significant risk of falling into default may help lending organizations prevent bad loans and also encourageclients to better manage their personal finance.

At the end of november 2018, the total consumer credit outstanding in USA, as communicated by US Federal Reserve [21], was 3979.2 billions of dollars. However, financial institutions have noticeable opportunities for their businesses to raise revenues but at the same time, risk exposure in consumer lending increases significantly. Moreover, the delinquency rate on consumer loans for all commercial banks (DRCLACBS) [22] was 2.28%in the third quarter of 2018, seasonally adjusted. In contrast, the overall default rate for revolving consumer loans reached a historical maximum of 10.1% in the third quarter of 2009 during the financial crisis (Khandani, A. *et al.*, 2010) [6].

In 2017, 34% of americans took personal loans, which represent almost 83.5 million people. Pureprofile made a research on loan usage in USA by surveying 2245 americans.Their results showed that the top three reasons why people took credits include buying cars for 31% of borrowers, 26% of customers required loans to pay off their bills and debts, whereas 21% of them needed additional money in case of an emergency [23].

Taking into account the increase in consumer spending and consumer lending needs, classification issues and predictive modeling aim to identify patterns in large data sets with many variables. These patterns might be able to explain the occurrence of entering default of a loan, considering that banks and other financial institutions aim to minimize credit risk.

A fundamental problem in the field of credit operations is the assessment of a client's creditworthiness, based on his ability to repay the loan in full.Therefore, the analysis of clients with significant risk of default is essential and reliable, as customers with payment delays may lead to a high probability of default. Thus, companies need to examine their databases in order to discover patterns able to account for customer behavior, in order to reduce the risk of income loss. However, understanding how these processes interact in the prediction of customer non-payment probability remains a major challenge in terms of model performance validation, essential before model deployment so that it can be used in daily decision making process.

The aim of this study is to analyze data from the Lending Club platform [16], which contains a number of clients who could not repay the credit in full, thus entering into default. The study was designed to apply a series of machine learning algorithms to develop four predictive models able to explain the studied event (the entry of customers in default) through classifiers such as: LightGBM, XGBoost, Logistic Regression and Random Forest. The paper is organized in five main sections, as follows: *literature review,data description, research methodology, data analysis and conclusions.*

2.Literature review

Recent research on loan default risk assessment include predictive methods based on machine learning algorithms. The aim is to allocate a particular client a probability of default according to his/her profile and payment behavior.

In his research paper, Bagherpour (2017) conducts a study on predicting mortgage default using machine learning algorithms on a large set of data. The analyzed loans were issuedbetween 2001-2016, at quarterly frequency. A series of classifiers were used to forecast loan default such as K-Nearest Neighbors (KNN), Random Forest (RF), Support-Vector Machines (SVM) and Factorization Machines (FM). The author argues that the performance of non-linear, non-parametric algorithms is substantially better than the traditional logistic regression model. Moreover, the machine learningalgorithms allow one to identify the predictive power of the considered variables. Consequently, Factorization Machines forecast AUC values between 88% and 91% whichwere the highest compared to the other classifiers.

Xiaojun, M. *et al.* (2018) use two novel machine learning algorithms called LightGBM and XGBoost to predict the default of customers based on real-life peer to peer (P2P) transactions from the Lending Club, i.e. transactions from investors to borrowers which offers benefits to the both sides. The authors point out that the reason why they chose these algorithms is that they have a profound theoretical background and a practical applicability proven by numerous studies that reveal the remarkable performance of their application, along with significantly reducing overfitting. The results of the reasearch shown that LightGBM recorded the best performance in comparison with XGBoost, having an error rate of 19.9% and an accuracy of 80.1%.

Kvamme, H.*et al.* (2018) propose a novel approach to predict mortgage default by considering time series data related to customer transactions in current accounts, savings accounts and credit cards. The analytical algorithm was implemented using a type of Deep Learning model called Convolutional Neural Networks (CNN). The research revealed that CNN model obtainedpromising results with an AUC of 0.918 for Neural Networks and 0.926 for Neural Networks in combination with a Random Forest classifier.

In a study conducted by Koutanaei, F.N. *et al.* (2015), a new hybrid credit scoring model was proposed by testing four feature selection algorithms and ensemble learning classifiers. The best choice for feature selection was PCA (Principal Component Analysis). Regarding the classification part, the best model was ANN-AdaBoost, an artificial neural network adaptive boosting algorithm.

However, Kruppa, J. *et al.* (2013) use machine learning methods to estimate the probability of default rather than having a binary classification of good or bad payers. They consider that the probability estimation of these algorithms is based

DOI: 10.24818/18423264/53.2.19.09

on nonparametric regression and compare different approaches using random forests (RF), k-nearest neighbors (KNN) and bagged k-nearest neighbors (bNN) versus the parametric standard method of logistic regression. Lastly, they revealthat random forests model outperforms the other three methods in terms of AUC scores (area under the curve) on the test data.

Khandani, A.E. *et al.* (2010) propose a combination of features starting from the standard credit scoring debt-to-income ratio, to more detailed characteristics such as consumer banking transactions that can be used as input for the model and argue that the latter greatly increases its predictive power. In contrast, Khashman, A. (2011) reveals a novel approach in predicting the credit risk for application scoring by using an emotional neural network that takes into account anxiety and confidence during the learning process and compares the results with the conventional neural network model, in an attempt to simulate the decisions made by a human expert. The author concluded that the emotional neural network model outperformed the other neural networks in terms of speed, simplicity, accuracy and minimum error.

Beque, A., Lessmann, S. (2017) introduce a recent type of feed-forward neural network which is called Extreme Learning Machine (ELM) and compares its performance with other methods such as artificial neural networks, decision trees, support vector machines and regularized logistic regression. They argue that this new approach represents a progress that combines a significant prediction performance with a noticeably computational efficiency.

Harris, T. (2013)conducted a study on prediction of credit risk using a support vector machine algorithm applied for two definitions of default: on the one hand, a broader rule was considered for up to 90 days payment overdue; on the other side, a narrow definition took into account only customers with more than 90 days late payment. He claims that the model used for the broader definition has a higher accuracy than the other one and at the same time, it is a reliable and accurate method to predict credit un-creditworthiness compared to the traditional judgment approach.

Zhang, T. *et al.* (2018)propose a new model that uses Multiple Instance Learning (MIL)in development of a credit scoring model by including not only socio-demographic and loan application data, but also the transaction history data of the applicant. This method allows to extract dynamic features from transactional information and the results showed that all the classifiers that were applied using newly added data had a significant increase in accuracy in comparison with not taking into account transactional data.

Papouskova, M., Hajek, P. (2019) introduce a novel two-stages credit risk model: the first part consists of a model used to predict the probability of default (PD) through ensemble classifiers that discriminate between good and bad payers; the second stage makes an in-depth analysis on customers with a predicted probability of default and a regression ensemble is applied to determine the exposure at default (EAD). Afterwards, the two models are combined to predict the

DOI: 10.24818/18423264/53.2.19.09

expected loss (EL). The researchers claim that this method outperforms other state of the art models used to predict loss given default (LGD) and exposure at default (EAD).

3.Data description

The database we used comes from the Lending Club [16] and consists of 887379 observations with 75 variables for loans issued between 2007 and 2015. One can note that 61176 customers of the total entered into default or registered significant delays in payment, in the case of some of them even more than 120 days, which led to the *"Charged off"*loan closure, since there were no reasonable expectations for those customers to be able to repay the loan in the upcoming period.

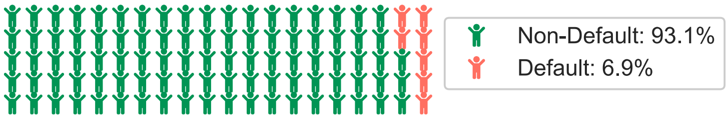
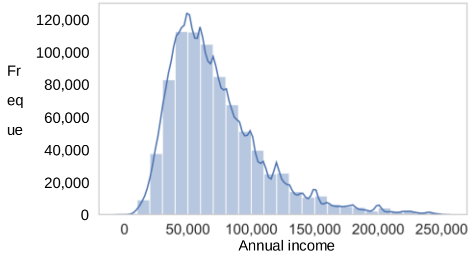


Figure 1. Customers structure by target variable

In the analysis were retained 14 numerical variables and 3 categorical variables. Among the numerical variables we can count those describing the annual income, payment behavior (such as the number of months since the last payment delinquency), the number of open credit lines, the number of public records (i.e. the Credit Bureau), as well as product features such as interest rate or loan amount. Among the considered categorical variables are employment length, house ownership or the degree of risk expressed by the FICO score (A, B, ... , G) [20]. These categorical variables were converted into numerical dummy variables using one-hot encoding (1/0), so that they can be used in the modeling stage.

In the figures below one can observe the distribution of data in terms of *annual income* and *loan amount.*



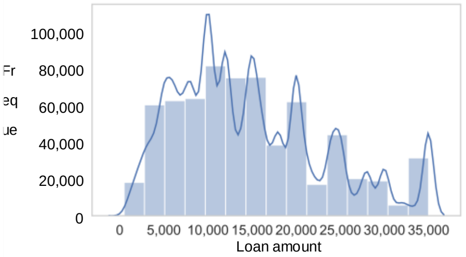


Figure 2.Histogram of annual income expressed in US dollars

Figure 3.Histogram of loan amount expressed in US dollars

DOI: 10.24818/18423264/53.2.19.09

Both *annual income* and *loan amount* series are positively asymmetrical.Additionally, as one can note in Figure 2, the majority of clients had an annual income between $30000 and $60000which accounts for 41.1% of the entire base (364959 customers). Also, Figure 3 highlights the fact that 261796customers (29.5%) applied for a loan between $10000 to $15000.

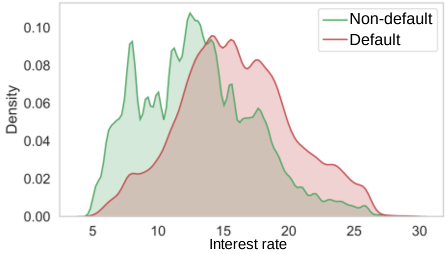


Figure4.Density distribution of interest rate by target

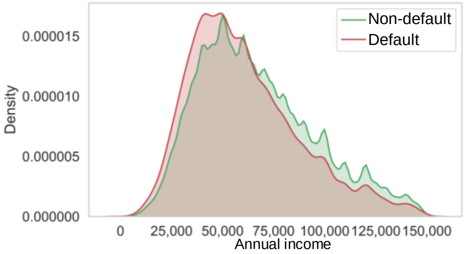


Figure5.Density distribution of annual income by target

In Figure 4 it can be seen that the *interest rate*kernel distribution for default customers is more shifted to the right than the distribution of non-default customers, which indicates that, on average, they had higher interests for loans(16% compared to 13%). At the same time, Figure 5 shows that those who entered default have a slightly smaller distribution of annual income with an average of $66424.5 compared to the non-default clients with an average of $75664.2, which may indicate a greater difficulty in repaying the debt, correlated with higher interest rates.

When looking at the main purpose why people choose to take a loan from Lending Club database, the top three reasons consist of*debt consolidation* for 524215 borrowers (59%), *reimbursment of debts oncredit card* for 206182 people (23%) and *home improvement* for 51829 customers (6%), which represent a cumulative percentage of 88% from the entire base.

People can take credits even though they may find themselves in certain difficult financial situations and may own properties or not. Therefore, when looking at home ownership, 50% of borrowers (443557) have *mortgage*, whereas 40% of them have *rent* (356117), while only 10% *own a house*. When slicing the distribution of home ownership between the target variable it can be seen that people who entered default have a smaller share of mortgage 44% versus 50% for non-default, whilst the former have a higher share of rent (47%) compared to the latter (40%), which might represent a greater exposure at risk that can lead to a higher probability of default.

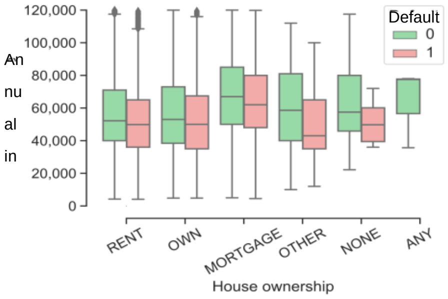


Figure6.Boxplot charts for annual income versus property type

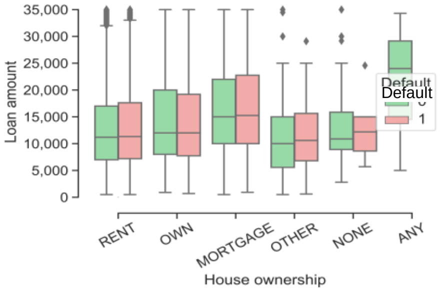


Figure7.Boxplot charts for loan amountversus property type

The boxplot diagrams (Figures 6 and 7) show that there are a series of outliers (extreme values) of the data series corresponding to annual income and loan amount respectively, by type of property. At the same time, default customers with a *rent* or *mortgage* type of property have a slightly higher distribution of loan amounts than non-default customers, while for the same segments, the annual income series is slightly lower for default than for the non-default clients.

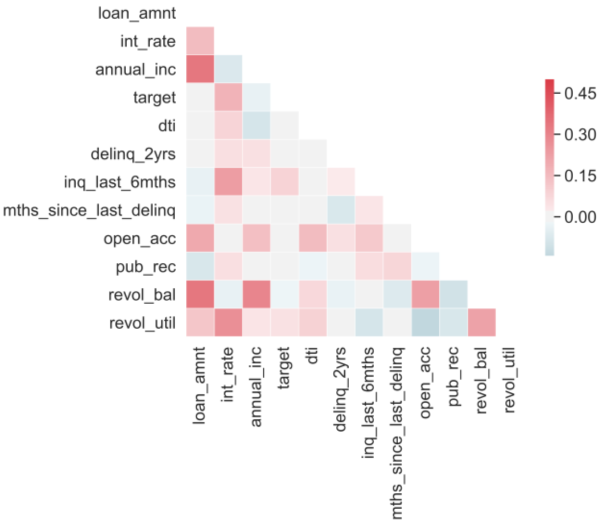


Figure8.Heatmap for the correlation matrix applied to numerical variables

Judging by the values of the Pearson correlation coefficients of the correlation matrix graphically represented in Figure 8, most of the variables are weakly correlated. The highest correlation coefficients are 0.33 for the link between *loan_amnt* (loan amount) and *annual_inc* (annual income), 0.33 for *revol_bal* (automatic renewal credit line balance) and *loan_amnt*, 0.29 for

DOI: 10.24818/18423264/53.2.19.09

revol_bal and *annual_inc* and 0.22 for *revol_bal* and *open_acc* (the number of open credit lines).

4. Research methodology

Nowadays, the interest in machine learning algorithms and artificial intelligence is increasing all over the world. Below we shall briefly present the methods used to predict default risk.

LightGBM is a relatively new algorithm developed by Microsoft, which is based on *gradient boosting* framework and uses a learning algorithm consisting of decision trees. LightGBM is different from similar tree-based gradient boosting algorithms (such as *XGBoost*), as the former builds trees vertically, starting from the leaves, while the other algorithms build symmetricalbranches in depth at each level. In the process of tree development, this type of algorithm *(leaf-wise)* can reduce leaf entropy much more than algorithms based on deep tree construction *(level-wise)*[15].

The process is illustrated below:

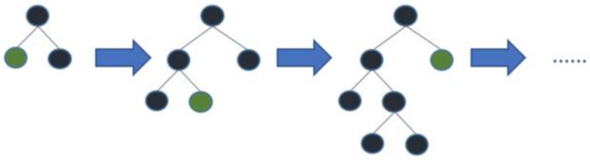


Figure 9.The tree-building process through LightGBM

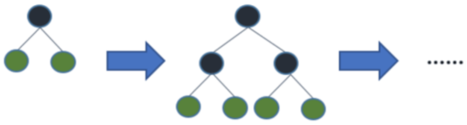


Figure 10.The tree-building process through XGBoost

Source:*Mandot, P. (2017) [15]*

The LightGBM algorithm has grown in popularity due to the high processing power of huge data volumes. Using grid search, it finds the optimal point after a number of iterations when the model's performance is maximum at a certain level of the parameters. It has also been proved that it significantly reduces overfitting.

Logistic Regression is the adequate regression analysis when the dependent variable is dichotomous (binary). Logistic Regression is used to describe data and explain the relationship between a binary dependent variable and one or more independent variables. This generates the coefficients and standard errors for the significance levels of the formula used to predict a logit transformation of the probability of event occurrence[17][18].

Random Forest is a supervised learning algorithm and is based on a collection of decision trees. Several trees are built, then their results are combined,

DOI: 10.24818/18423264/53.2.19.09

so that the final prediction can be more stable and more accurate than theclassification result of a singledecision tree [19].

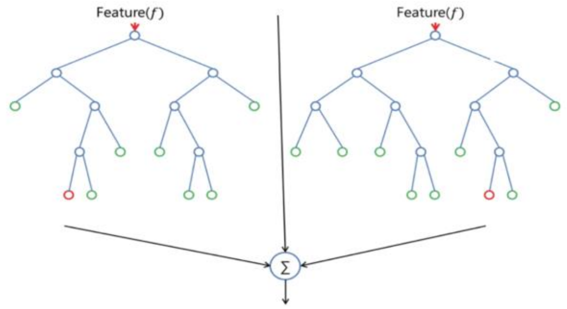


Figure 11.Classification process of the Random Forest algorithm

Source: *https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd [19]*

Because the best performances have been recorded for these four classifiers in terms of accuracy and AUC scores, we decided to test them on the analyzed data set.

Data analysis and predictive modeling were carried out using Python version 3.5 and the following libraries: *numpy, pandas, matplotlib, seaborn, scikit-learn, scipy, shap, lightgbm, xgboost, pywaffle*.

5. Data analysis

To determine a customer's probability of entering default, three sampling scenarios were elaborated, in which four classifiers were applied using the following machine learning algorithms: *LightGBM, XGBoost, LogisticRegression*and *Random Forests.*

The first considered scenario was the modeling on the original database, where the distribution of non-default and default values of the target variable was 93.1% - 6.9%, i.e. a proportion of 13.5 non-target cases to 1 default case.

One can observe that the analyzed database is *strongly imbalanced*, as the minority class has significantly fewer observations than the other category, which may lead to difficulties in the modeling stage. This situation might happen as the prediction of the default event occurrence in the predictive model may be biased. Due to this fact, the model might not capture sufficient information from the available data, thus favoring the prediction of the non-default class.

In a series of studies it has been shown that modelling on imbalanced data sets can lead to a very high specificity or local accuracy for the majority class but to poor results of the same measures on the minority class (Fernandez, A.*et al.*, 2018). Furthermore, a difficulty that arises from this problem is the effectiveness of both accuracy and error rate when assessing the performance of the classifiers. For example, if the majority class represents 99% of the data and the minority class only 1%, then the most naive classifier that always predicts only the majority class would still have an accuracy of 99%. In this regard, such a classifier would be useless in predicting the minority event of interest. Moreover, the results of the traditional classifiers are affected when are applied on imbalanced datasets because they tend to optimize the accuracy while developing a model that is similar to the naive classifier described previosuly (Hoens, T.R., 2013).

In order to minimize this inconvenience, the second scenario aims to reduce the discrepancy between the proportion of non-target and target cases by applying an *under-sampling* on the majority class to reach an 80% to 20% ratio between non-target and target values.

Starting from the second scenario, the third scenario tries to balance even more the imbalanced data set, by combining the two methods of *under-sampling*on the majority class and *over-sampling* on the minority class by replicating 50% of the default cases (sampling with replacement).

In order to evaluate the performance of the classifiers, the data sets corresponding to each scenario were divided in 75%- *training set*and 25% - *validation data* (test set). The models weretrained only on the training sets, so that the other observations can be used to evaluate the likelihood of entering default, as if the models were implemented into production *(unseen data)*.

In the following, we tested the stability and distribution of all variables between the trainingand test data sets after random sampling, in order to ensure the representativeness of both data sets.

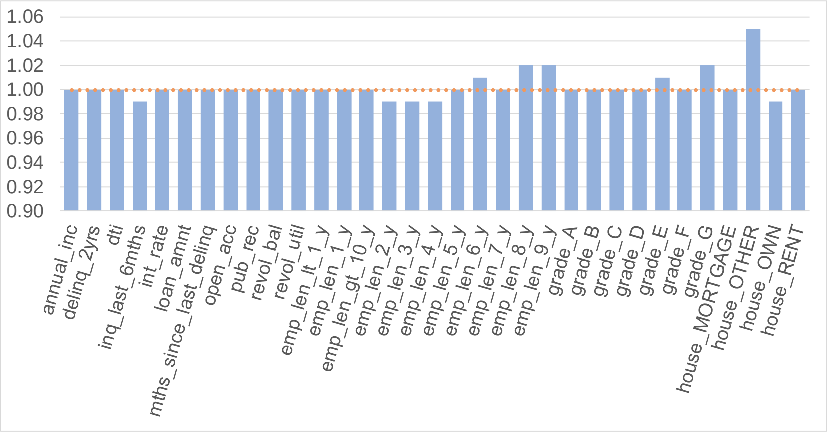


Figure12.Ratio between the mean of the variables between the training set and the test set

As it can be observed in Figure 12, the ratio between the mean of the training set variables and the mean of the test set variables is very close to the unit value for most features. This implies that the two data sets are homogeneous and balanced in terms of customer profile.

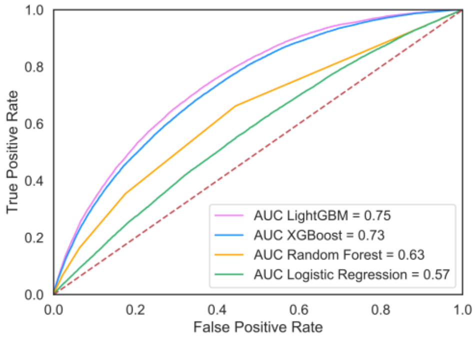


Figure13.AUC scores of the models on the initial database

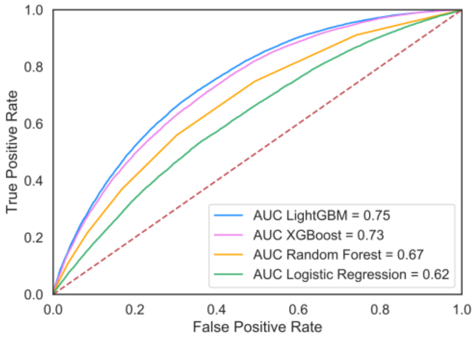


Figure14.AUC scores of the models for under-sampling scenario

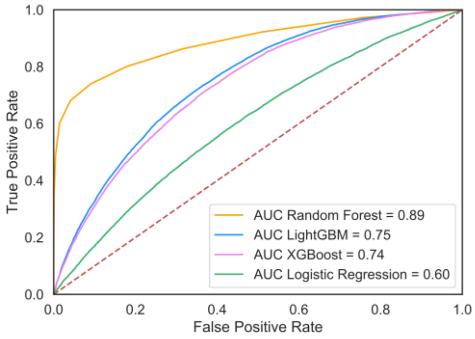


Figure 15.AUC scores for under-sampling combined with over-sampling scenario

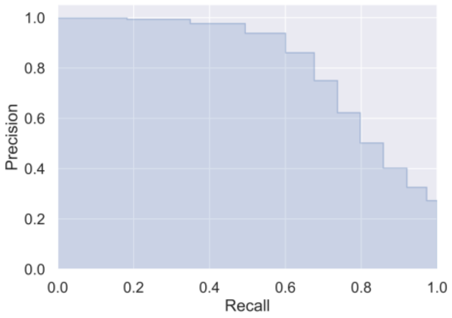


Figure 16.Precision-Recall Curve for the optimal classifier - Random Forest

The performance assessment of a classifier can be done by following the AUC *(Area Under the Curve)* maximization. The ideal classifier has an area under the curve equal to 1, while the graph is as close as possible to the top left corner. This case corresponds to a perfect classification, where the rate of *True Positive* classification is equal to 1, and the rate of *False Positive* classification is equal to 0.

Figures 13, 14 and 15 illustrate that, after applying all classifiers on the three sampling scenarios, the optimal model with the highest value of the area under the ROC *(Receiver Operating Characteristic)* curve was recorded for the combined *under-over sampling* scenario for the **Random Forest** classifier→$AUC = 0.89$on the test set (the yellow curve in Figure 15).

DOI: 10.24818/18423264/53.2.19.09

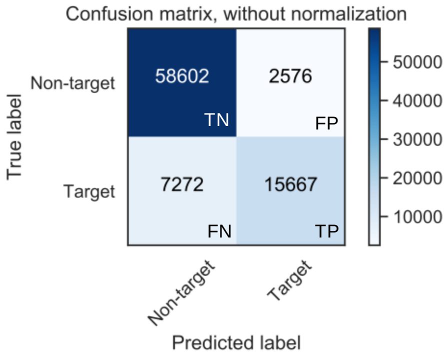


Figure 17.Random Forest confusion matrix

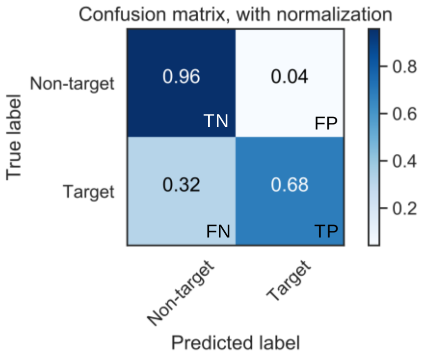


Figure 18.Normalized Random Forest confusion matrix

Confusion matrices show the number of cases classified by the model as target/non-target *(Predicted label)* compared to the real category of the evaluatedobservations *(True label)* [12]. Therefore, if the model predicts that a client will enter default, and he/she really enters into default, this is a true positive situation (TP). On the other hand, when the model classifies a case as positive, but it is actually negative, this is a false positive observation (FP). But then, if the model predicts that a customer will not enter into default, and the customer didn’tentered into default in the test set, then the case is true negative (TN). Finally, the prediction of the model may be negative, while the client may actually enter into default, which is equivalent to a false negative classification (FN).

Since the initial data set is highly imbalanced in terms of the target distribution with default/non-default cases, it is not enough to simply verify the accuracy of the model. Contrariwise, one needs to calculate the *precision* and *recall* indicators which are fundamental for the performance assessment of any binary classification model.

The key performance indicators of the Random Forest optimal model are:

$$precision = \frac{TP}{TP+FP} = \frac{15513}{15513+2504} = \frac{15513}{18017} = 0.861$$

and

$$recall = \frac{TP}{TP+FN} = \frac{15513}{15513+7426} = \frac{15513}{22939} = 0.676\ .$$

Previous indicators show that the obtained results are very good and can be interpreted as follows: 86.1% of the positive predictions of the model are truly positive in reality *(True Positive)* and 67.6% of the real positive cases are captured by the model by classifying customers in the positive class*(target default = 1)*.Ideally, the graphin Figure 16 should have a curve as close as possible to the top

DOI: 10.24818/18423264/53.2.19.09

right corner, in which the *precision* and *recall* measures show that the model effectively distinguishes between default and negative cases.

Model accuracy is the ratio between the sum of the elements on the diagonal of the confusion matrix and the sum of all the classifications:

$$accuracy = \frac{TP+TN}{TP+FN+TN+FP} = \frac{15513+58674}{15513+7426+58674+2504} = \frac{74187}{84117} = 0.882$$

The following figure illustrates the most important variables used in the model for making predictions.These characteristics used as predictors in estimating the model are plotted in decreasing order of the importance value.

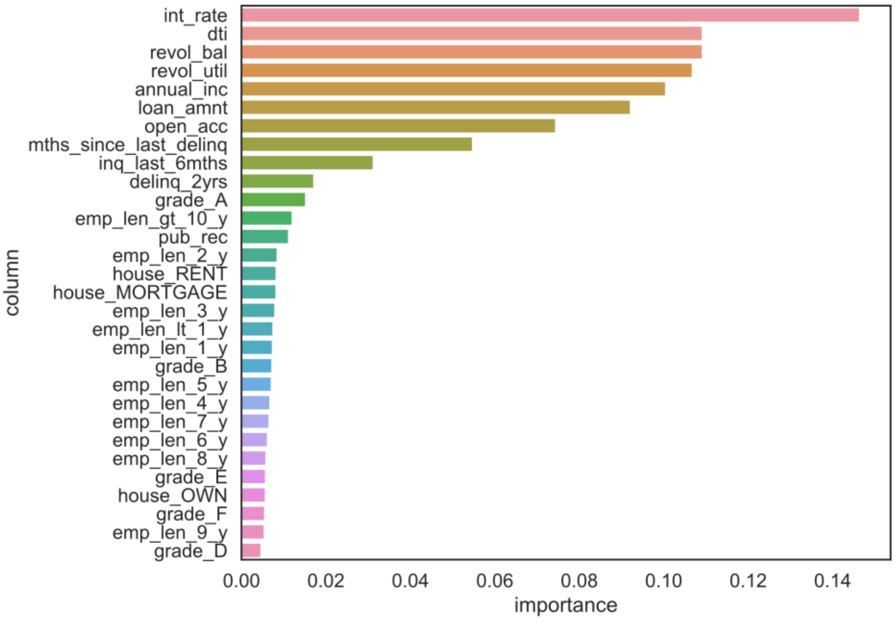


Figure 19.Importance of variables used in the Random Forest model, top 30

Figure 19 highlights the most relevant 10 characteristics: *int_rate* (interest rate), *dti* (the degree of indebtedness), *revol_bal* (the total balance of the revolving credit line), *revol_util* (the amount used from the revolving credit line), *annual_inc* (annual income), *loan_amnt* (loan amount), *open_acc* (the number of open credit lines), *mths_since_last_delinq* (the number of months since the last payment delay), *inq_last_6mths* (the number of queries at the Credit Bureau in the last 6 months) and *delinq_2yrs* (the number of delays in payment over the past 2 years).

6. Conclusions

The purpose of the present research was twofold: to understand the patterns that can lead to a significant risk for a customer to enter default; and to build an

DOI: 10.24818/18423264/53.2.19.09

accurate predictive model able to effectively classify observations in the two classes, i.e. good paying customers and bad payers (those likely to enter default).

Because imbalanced data sets can affect the performance of the classifiers in modeling stage with a reduced specificity or local accuracy for the minority class, three sampling scenarios were created so that the data set could be balanced and four classifiers were applied: LightGBM, XGBoost, Random Forest and Logistic Regression. The best results were observed for the Random Forest optimal classifier applied on the combined scenario ofunder-sampling withover-sampling, having a representative AUC of 0.89.

By analyzing the distribution of main characteristics in the data set, a few interesting insights were revealed. These explained that customers who entered default had higher interests for loans on average (16% compared to 13%). Also, those who entered default had a slightly smaller annual income when compared to the non-default clients, which could indicate a greater difficulty in repaying the debt, correlated with higher interest rates.

A loan strategy department of a bank or other lending institution may use such a model in production to identify customers with an increased risk of default. This will ultimately reduce potential revenue losses due to non-performing loans, whichcan be caused by customers who cannot fully repay the amounts due.Also, the implementation of an early warning system maydetectchanges in customerbehavior that can indicate aprobability of a customer to have future difficulties in repaying the debt. This means that the bank can advise clients to better manage their personal finances. Thus, customers can refinance the current loan to reduce the burden of a too big monthly installment by extending the loan period and lowering the installment amount.

Moreover, further research directions that might be considered refer to applying balancing techniques by generating artificial (synthetic) data using either ROSE (Random Over Sampling Examples) or SMOTE (Syntethic Minority Over-sampling Technique) methods. It has been shown that these methods have produced very good results in other similar research based on imbalanced data sets.

Appendix A. Descriptive statistics

Table 1. Descriptive statistics for whole database

Variable	mean	std	min	25%	50%	75%	99%	max
Annual income	75027	64698	0.0	45000	65000	90000	250000	9500000
Delinquencies last 2 years	0.3	0.9	0.0	0.0	0.0	0.0	4.0	39.0
Degree of indebtedness	18.2	17.2	0.0	11.9	17.7	24.0	37.5	9999.0
Inquiries last 6 months	0.7	1.0	0.0	0.0	0.0	1.0	4.0	33.0

Interest rate	13.2	4.4	5.3	10.0	13.0	16.2	25.0	29.0
Loan amount	14755	8436	500	8000	13000	20000	35000	35000
Months since last delinquencies	16.6	22.9	0.0	0.0	0.0	30.0	79.0	188.0
Number of open accounts	11.5	5.3	0.0	8.0	11.0	14.0	28.0	90.0
Number of public records	0.2	0.6	0.0	0.0	0.0	0.0	2.0	86.0
Revolving balance	16921	22427	0.0	6443	11875	20829	93411	2904836
Revolving line utilization rate	55.0	23.9	0.0	37.6	56.0	73.6	98.5	892.3

Source: Authors' computation

Table 2. Descriptive statistics for customers with default (target = 1)

Variable	mean	std	min	25%	50%	75%	99%	max
Annual income	66424	54887	2000	41000	57400	80000	215000	8706582
Delinquencies last 2 years	0.3	0.9	0.0	0.0	0.0	0.0	4.0	21.0
Degree of indebtedness	18.7	8.1	0.0	12.7	18.6	24.5	36.7	40.0
Inquiries last 6 months	1.0	1.2	0.0	0.0	1.0	2.0	5.0	33.0
Interest rate	16.0	4.3	5.3	13.0	15.6	18.8	25.8	29.0
Loan amount	14741	8440	500	8000	13000	20000	35000	35000
Months since last delinquencies	15.7	22.6	0.0	0.0	0.0	28.0	79.0	152.0
Number of open accounts	11.2	5.0	0.0	8.0	10.0	14.0	27.0	76.0
Number of public records	0.2	0.5	0.0	0.0	0.0	0.0	2.0	12.0
Revolving balance	15251	19671	0	6144	11227	19206	73196	1746716
Revolving line utilization rate	59.1	23.6	0.0	42.7	61.0	77.7	98.7	148.0

Source: Authors' computation

Table 3. Descriptive statistics for non-default customers (target = 0)

Variable	mean	std	min	25%	50%	75%	99%	max
Annual income	75664	65321	0.0	46000	65000	90000	250000	9500000
Delinquencies last 2 years	0.3	0.9	0.0	0.0	0.0	0.0	4.0	39.0
Degree of indebtedness	18.1	17.7	0.0	11.9	17.6	23.9	37.5	9999.0
Inquiries last 6 months	0.7	1.0	0.0	0.0	0.0	1.0	4.0	31.0
Interest rate	13.0	4.3	5.3	9.8	12.7	15.6	24.5	29.0
Loan amount	14756	8435	500	8000	13000	20000	35000	35000
Months since last delinquencies	16.7	22.9	0.0	0.0	0.0	30.0	79.0	188.0

Number of open accounts	11.6	5.3	0.0	8.0	11.0	14.0	28.0	90.0
Number of public records	0.2	0.6	0.0	0.0	0.0	0.0	2.0	86.0
Revolving balance	17044	22613	0	6467	11926	20961	94747	2904836
Revolving line utilization rate	54.7	23.9	0.0	37.3	55.6	73.2	98.5	892.3

Source: Authors' computation

REFERENCES

[1] **Bagherpour, A. (2017),*Predicting Mortgage Loan Default with Machine Learning Methods***; University of California, Riverside;

[2] **Xiaojun, M.*et al.* (2018), *Study on a Prediction Of P2P Network Loan Default Based on the Machine Learning Lightgbm and Xgboost Algorithms According to Different High Dimensional Data Cleaning***; Electronic Commerce Research and Applications, 31, pp.24-39;

[3] **Kvamme,H. *et al*. (2018), *Predicting Mortgage Default Using Convolutional Neural Networks***; Expert Systems With Applications, 102, pp.207-217;

[4] **Koutanaei, F.N. *etal.* (2015), *A Hybrid Data Mining Model of Feature Selection Algorithm and Ensemble Learning Classifiers for Credit Scoring***;Journal of Retailing and Consumer Services, 27, pp.11-23;

[5] **Kruppa, J. *et al.* (2013), *Consumer Credit Risk: Individual Probability Estimates Using Machine Learning***;Expert Systems with Applications, 40, pp.5125-5131;

[6] **Khandani, A.E. *et al.* (2010), *Consumer Credit-Risk Models via Machine-Learning Algorithms***;Journal of Banking & Finance, 34, pp.2767-2787;

[7] **Khashman, A. (2011), *Credit Risk Evaluation Using Neural Networks: Emotional versus Conventional Models***;Applied Soft Computing, 11, pp.5477-5484;

[8] **Beque, A., Lessmann, S. (2017), *Extreme Learning Machines for Credit Scoring: An Empirical Evaluation***;Expert Systems with Applications, 86, pp.42-53;

[9] **Harris, T. (2013), *Quantitative Credit Risk Assessment Using Support Vector Machines: Broad versus Narrow Default Definitions***; Expert Systems with Applications, 40, pp.4404-4413;

[10] **Zhang, T. *et al.* (2018), *Multiple Instance Learning for Credit Risk Assessment with Transaction Data***;Knowledge-Based Systems, 161, pp.65-77;

[11] **Papouskova, M., Hajek, P. (2019), *Two-stage Consumer Credit Risk Modelling Using Heterogeneous Ensemble Learning***;Decision Support Systems, 118, pp.33-45;

[12] **Larose, D.T., Larose, C.D. (2015), *Data Mining and Predictive Analytics***;*John Wiley & Sons,* USA, pp.359-383, 451-464;

[13] **Fernandez, A. *et al.* (2018), *SMOTE for Learning from Imbalanced Data: Progress and Challenges***; Journal of Artificial Intelligence Research 61, pp.863-905;

[14] **Hoens, T.R., Chawla, N.V. (2013), *Imbalanced Learning: Foundations, Algorithms and Applications, First Edition***;*John Wiley & Sons,* pp.43-59;

[15] **Mandot, P. (2017), *What is LightGBM? How to Implement it?***URL: https://medium.com/@pushkarmandot/https-medium-com-pushkarmandot-what-is-lightgbm-how-to-implement-it-how-to-fine-tune-the-parameters-60347819b7fc;

[16] **Lending Club database**, URL: https://www.lendingclub.com/info/download-data.action

[17] https://www.statisticssolutions.com/what-is-logistic-regression/

[18] https://www.medcalc.org/manual/logistic_regression.php

[19] https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd

[20] **FICO Score**, https://www.myfico.com/credit-education/credit-scores/

[21] **USA Federal Reserve**:https://www.federalreserve.gov/releases/g19/current/

[22] **Board of Governors of the Federal Reserve System (US), *Delinquency Rate on Consumer Loans***, All Commercial Banks [DRCLACBS], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/DRCLACBS, February 4, 2019

[23] https://www.finder.com/personal-loans-statistics

[24] **Python Software Foundation (2018)**.*Python Language Reference,* version3.5. http://www.python.org.

DOI: 10.24818/18423264/53.2.19.09