İzmir Yönetim Dergisi
İzmir Journal of Management

E-ISSN: 2757-637X **YIL:** 2024 **Cilt:** 5 **Sayı:** 1 **Sayfa:** 55-69
Geliş Tarihi: 13.02.2024 **Kabul Tarihi:** 27.06.2024 **Online Yayın:** 27.06.2024 **Doi:** 10.56203/iyd.1436742
ÖZGÜN ARAŞTIRMA

Investigating the Effect of Class Balancing Methods on the Performance of Machine Learning Techniques: Credit Risk Application*

Migraç Enes Furkan MİLLİ[1], İpek DEVECİ KOCAKOÇ[2], Serkan ARAS[3]

Abstract

Credit risk arises as a result of the failure of the loans given by banks to the customers to fulfill their obligations at the end of the specified term. Technological advances allow the use of machine learning methods in various sectors. These methods aim to facilitate the identification of customers at risk with the system adapted to the creditworthiness processes of banks. For this purpose, in order to make the most appropriate evaluation in the lending process of banks, re-sampling techniques to eliminate the problem of class imbalance encountered in unbalanced data sets were made balanced and their effects on machine learning were investigated. During the implementation phase, German, Australian and HMEQ credit data sets were used. Different machine learning classification methods such as Logistic Regression (LR), K-Narest Neighbor (KNN), Naive Bayes (NB), Support Vector Machines (SVM), Multilayer Perceptron (MLP), Decision Trees (DT), Random Forests (RF), Gradient Boosting Decision Trees (GBDT), Extremely Randomized Trees, Hard and Soft Voting were used to detect risky customers. The problem of class imbalance was balanced with resampling and hybrid techniques such as Random Oversampling (ROS), Random Undersampling (RUS), Balanced Bagging Classifier (BBC), SMOTE-Tomek Links and SMOTE-ENN. In this context, the performances of three different data sets were examined in four different scenarios. As a result of the study, the hybrid method, in which oversampling and undersampling methods are used together for the class balancing problem, showed the best classification performance among machine learning techniques.

Keywords: *Credit Risk, Machine Learning, Ensemble Learning, Classification Algorithms, Class Balancing, Resampling.*

Jel Kodu: C10, C38, C55, G21.

Sınıf Dengeleme Yöntemlerinin Makine Öğrenmesi Tekniklerinin Performansları Üzerindeki Etkilerinin Araştırılması: Kredi Riski Uygulaması

Özet

Bankalar tarafından müşterilere verilen kredilerin belirlenen vade sonunda yükümlülüklerini yerine getirememesi sonucu kredi riski ortaya çıkmaktadır. Teknolojik gelişmeler, çeşitli sektörlerde makine öğrenmesi yöntemlerinin kullanılmasına olanak tanımaktadır. Bu yöntemler, bankaların kredibilite süreçlerine uyarlanan sistem ile risk altındaki müşterilerin saptanmasını kolaylaştırmayı amaçlamaktadır. Bu amaçla, bankaların kredi verme sürecinde en uygun değerlendirmenin yapılabilmesi için dengesiz veri setlerinde karşılaşılan sınıf dengesizliği probleminin ortadan kaldırılması için yeniden örnekleme teknikleri ile veri setleri dengeli bir hâle getirilerek makine öğrenmesi üzerindeki etkileri araştırılmıştır. Uygulamada, Alman, Avustralya ve HMEQ kredi veri setleri kullanılmıştır. Riskli müşterilerin belirlenmesinde Lojistik Regresyon (LR), K-En Yakın komşu (KNN), Naive Bayes (NB), Destek Vektör Makineleri (SVM), Çok Katmanlı Algılayıcı (MLP), Karar Ağaçları (DT), Rassal Ormanlar (RF), Gradyan Artırma Karar Ağaçları (GBDT), Extremely Randomized Trees, Sert ve Yumuşak Oylama olmak üzere farklı makine öğrenmesi teknikleri kullanılmıştır. Sınıf dengesizliği sorunu; Random

* This study is derived from the author's Master's thesis.

ATIF ÖNERİSİ (APA): Milli, M.E.F., Deveci Kocakoç, İ., & Aras, S. (2024). Investigating the Effect of Class Balancing Methods on the Performance of Machine Learning Techniques: Credit Risk Application. *İzmir Yönetim Dergisi*, 5(1), 55-69. Doi: 10.24988/ije.2020351XXX

[1] Ekonometri Doktora Programı Öğrencisi, İstanbul Üniversitesi, Sosyal Bilimler Enstitüsü, Beyazıt / İSTANBUL, **EMAIL:** migracenesfurkan.milli@ogr.iu.edu.tr **ORCID:** 0000-0003-2516-7723

[2] Prof.Dr., Dokuz Eylül Üniversitesi, İktisadi ve İdari Bilimler Fakültesi, Buca / İZMİR, **EMAIL**: ipek.deveci@deu.edu.tr **ORCID:** 0000-0001-9155-8269

[3] Doç. Dr., Dokuz Eylül Üniversitesi, İktisadi ve İdari Bilimler Fakültesi, Buca / İZMİR, **EMAIL:** serkan.aras@deu.edu.tr **ORCID:** 0000-0002-6808-3979

İzmir Yönetim Dergisi
İzmir Journal of Management

E-ISSN: 2757-637X **YIL:** 2024 **Cilt:** 5 **Sayı:** 1 **Sayfa:** 55-69
Geliş Tarihi: 13.02.2024 **Kabul Tarihi:** 27.06.2024 **Online Yayın:** 27.06.2024 **Doi:** 10.56203/iyd.1436742
ÖZGÜN ARAŞTIRMA

Oversampling (ROS), Random Undersampling (RUS), Balanced Bagging Classifier (BBC), SMOTE-Tomek Links ve SMOTE-ENN gibi yeniden örnekleme ve hibrit teknikler ile sınıflar dengeli hâle getirilmiştir. Bu kapsamda, üç farklı veri kümesinin performansları dört farklı senaryo üzerinde incelenmiştir. Çalışmanın sonucunda, sınıf dengeleme problemi için aşağı ve yukarı örnekleme yöntemlerinin bir arada kullanıldığı hibrit yöntem, makine öğrenme teknikleri arasında en iyi sınıflandırma performansı göstermiştir.

Anahtar Kelimeler: Kredi Riski, Makine Öğrenmesi, Topluluk Öğrenmesi, Sınıflandırma Algoritmaları, Yeniden Örnekleme, Sınıf Dengeleme.
Jel Codes: C10, C38, C55, G21.

1. INTRODUCTION

Today, loans are among the most important sources of financing for the development of the economic structure and the realization of development objectives. Loans are provided by banks and other lending financial institutions. Banks evaluate customers loan requests under certain conditions and grant loans according to their suitability. If this assessment is not done correctly, some risk factors emerge and lead to credit risk. One of these risks arises when the information and documents received from customers applying for a loan are incomplete or untrue. In this case, customer information during the credit decision will cause decision makers to make the wrong decision and thus inappropriate customers will receive credit and create credit risk. Credit risk does not only originate from the customer. At the same time, credit allocation officers should comprehensively evaluate the information provided by customers within the framework of established policies and procedures and identify the most suitable customers for credit. Otherwise, an incorrect assessment by the credit allocation officers will result in the granting of credit to customers who are not eligible for credit and credit risk will be in question. If banks identify customers with credit risk manually and non-automatically, the workload, cost, time and resource utilization increase and the process can become quite challenging. In order to identify risky customers more easily and to improve the process, machine learning techniques have been frequently utilized in recent years. These techniques can help banks increase their profits by reducing credit risk.One of the problems encountered in modeling the credit risk problem with machine learning techniques is class imbalance. Class imbalance occurs when one class is too small in the dataset compared to another class and can lead to modeling techniques over-learning the majority class. In this study, the performance effects of various class balancing methods developed in recent years on the machine learning techniques used are analyzed. For this purpose, class balancing methods, which are used by banks and financial institutions to evaluate the creditworthiness of customers requesting loans in the most appropriate way, are run on different machine learning techniques. With this method, it is aimed to identify the most suitable customers among the customers requesting loans and at the same time to increase profitability by reducing credit risk for banks. Credit datasets for real life problems labelled as German, Australian and Home Mortgage (HMEQ) were used from an open source website. These datasets were balanced with the Balanced Iterative Bagging Classifier and two different hybrid methods, SMOTE-Tomek Links and SMOTE-ENN, based on random undersampling and random oversampling, in addition to the standard random undersampling and random oversampling methods. These datasets were tested on four different scenarios. These scenarios were created to determine which type of machine learning and which balancing method should be used to identify the most suitable customers for lending.Analysing machine learning techniques in two categories as single models and ensemble models constitutes the first criterion in creating scenarios. The method to be followed in balancing the data was analysed as the second criterion. Thus, the scenario that produces the best prediction performance for each data set was determined. In the second part of the study, machine learning techniques and resampling methods are mentioned, in the third part, the application and the results obtained are evaluated and in the last part, the study is concluded by giving information about the results of the study.

2. LITERATURE REVIEW

When the literature on machine learning is analysed, Malekipirbazari and Aksakalli (2015) used data sets containing information such as credit and financial characteristics of

approximately 350,000 people in order to identify good borrowers in terms of creditworthiness. In the study, non-standard financial characteristics are included to improve the reliability of credit risk scoring and machine learning techniques such as random forest (RF), support vector machine (SVM), k-nearest neighbour (k-NN) and logistic regression (LR) are used to identify borrowers with good creditworthiness.Random forests were found to give the best results. Dahiya et al. (2016) developed three models for credit risk assessment. The first of these models is the MLP model. Second, a variable selection technique is used to improve the predictive accuracy of this MLP model. Finally, the Bagging-hybrid MLP method is applied to further improve the accuracy values. The accuracies of the developed models were compared on Australian and German datasets. The results have accuracy values of 90.50% and 80% respectively. Khemakhem et al. (2018) used a Tunisian bank loan application dataset. In order to balance the unbalanced data in accordance with the purpose of the study, Random oversampling (ROS) and Synthetic minority extreme learning technique (SMOTE) sampling techniques were used and their effects on classification performances were investigated. By looking at the effects of the combination of random oversampling techniques and artificial intelligence combinations on performance, an important contribution has been made in terms of predicting the repayment of loans. Shen et al. (2019), credit risk assessment was performed with back-propagation neural network (BP) model on Australian and German credit data set. A classifier-based optimisation method is proposed for individual credit risk assessment with SMOTE-based ensemble model. Particle swarm optimisation (PSO) algorithms are used to determine the best weights and search for deviations with BP neural networks. The developed model gave more significant results compared to classical methods. Hou et al. (2020) proposed a new unified Dynamic ensemble selection (DES) model called META-DESKNN-MI. It was applied on the P2P loan dataset. They combined their framework using META-DES and DES-KNN to make the performance of ensemble learning classifiers more effective. They found that the proposed model improved the performance of DES. Niu et al. (2020) proposed an ensemble learning method based on resampling with unbalanced data distribution (REMDD) method used in P2P credit data. The proposed model, REMMD, gave better results in the evaluation of unbalanced credit data of P2P. Jin et al. (2021) proposed a new ensemble model consisting of several stages based on a hybrid genetic algorithm to make credit forecasting accurate and consistently predictable. The proposed models are tested on German, Poland-1 and Poland-2 real data sets. The effect of the proposed model outperformed the classical models. Xiong and Huang (2021) developed a correlation-based classifier to improve the classification power of ensemble learning models using the maximum information coefficient (MIC-CCS). The proposed model was tested on Australian, Taiwanese and German datasets. In total, 8 classification models and 4 ensemble learning methods are compared. The proposed MIC-CCS technique gave the most effective result among the ensemble learning models according to AUC. Linear Discriminant Analysis (LDA) showed the best performance among the classification models. As a meta-classifier, the best results were obtained with support vector machine for Australian data, random forest and naive Bayes models for Germany and Taiwan data, respectively. Dumitrescu et al. (2022) proposed a penalised logistic tree regression (PLTR) technique based on information from decision trees to improve the effectiveness of logistic regression. This model was tested on "Housing", "Australian dataset" and "Taiwan dataset" datasets. PLTR was found to be more competitive compared to random forests, while out-of-sample performance was found to be more effective than non-linear and linear logistic regression.

3. MACHINE LEARNING TECHNIQUES AND RESAMPLING METHODS

3.1. Classification Algorithms

3.1.1. K-Nearest Neighbour

The k-Nearest Neighbour (k-NN) algorithm, which is a supervised machine learning algorithm, is a non-parametric method widely used for classification when there is insufficient information about the distributions of the data (Peterson, 2009). Using the data of the classes in the sample dataset, the current distance is calculated for each sample to be included in the datasets according to the existing data. In the k-NN algorithm, the correct selection of the function to calculate the distance measurement is important for the results (Duda et al., 1973; Weinberger and Lawrence, 2008). The k value is generally preferred as an odd number value in order to minimise the level of complexity between two close neighbours.

3.1.2. Logistic Regression

Logistic regression is a model in which the dependent variable has one of the binary values depending on the values in the independent variables (Le & Eberly, 2016: 351). In addition, response functions are asymptotes to the X and Y axes at values between 0 and 1 (Hosmer -Lemeshow, 1980: 1043-1069). The logistic regression model, unlike the linear regression model, is concerned with maximising the probability of an event occurring rather than minimising deviations (Hair et al., 1998).

3.1.3. Naïve Bayes

Naïve Bayes is a probability-based algorithm based on Bayes theorem. It is a method in which the probabilities of occurrence of events are calculated based on the condition that one event is known to occur while another event is known to occur. Although Naïve Bayes classifiers are simple, they can be easily applied to high-dimensional data sets and generally perform better than other alternative classifiers in complex classification techniques (Domingos and M. Pazzani, 1997: 121).

3.1.4. Multilayer Perceptron (MLP)

The network consists of a number of artificial neural cells (neurons or nodes) forming the input layer, one or more hidden layers and an output layer. A network in which the inputs move through the network layer by layer is called a multilayer perceptron (MLP) (Maciel and Ballini, 2008: 7). For each input value, an output value is produced by the network. The weights of the network connections are arranged to minimise this output value and the expected output value. These processes continue until the desired result is reached (McClelland et al., 1986: 533-535).

3.1.5. Support Vector Machine

Support Vector Machine (SVM) is a machine learning model used in non-linear classification, density or function estimation and based on kernel function (Li et al., 2006: 11). The main purpose of support vector machines is based on determining the hyperplane that maximally divides two classes for classification (Vapnik, 1995: 290-291). A support vector machine is a linear discriminator that uses the optimal dividing hyperplane. This hyperplane, also known as the maximal margin hyperplane, is calculated by a quadratic optimisation (Martin, 2001: 5).

3.1.6. Decision Trees Algorithm

Decision trees are widely used for classification and regression problems (Gehrke, 2003: 3-4). Decision trees are classifiers called as a recursive partition of the sample space. Decision trees consist of a tree structure consisting of roots, branches and leaves (Sharma et al., 2011: 191). Classification of the dataset using the decision tree algorithm is performed in a two-stage process. The first stage is the learning stage. The trained model is represented as a decision tree. The second stage is the classification

stage. In this stage, test data is used to test the accuracy of the decision tree. If the accuracy rates are found appropriate, the existing rules are then used to classify the included data

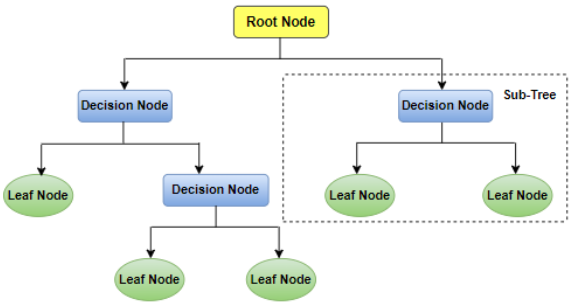

Figure 1: Example of decision tree structure (Prepared by the author)

One of the most important stages of decision trees is the determination of the root variable. While the number of branches according to the determined optimal root variable will be at a minimum level, it will allow the algorithm to make strong predictions. (Kumari and Godara, 2011: 305; Tu et al., 2009: 184).

3.1.7. Random Forests Algorithm

The Random Forests Algorithm was developed by Leo Breiman (2001). Random Forests (RF) is a supervised machine learning method that focuses only on ensembles of decision trees. This method is a combination of multiple decision trees (Breiman, 2001: 5). Random forests are an ensemble learning model in which results from multiple different models are used to compute a response (Horning, 2010: 2). Each decision tree in random forests is trained with samples drawn from the training set. At each decision node, either random selection is made among all variables or the variable that optimises the desired criterion is selected (Breiman, 2011; Akman et al., 2011: 37).

3.1.8. Gradient Boosting Decision Trees (GBDT)

GBDT produces a robust forecasting model by combining weak forecasting models, especially decision trees. The term regularisation is used to minimise the complexity of the model structure and the overfitting problem.
The GBDT algorithm, due to the combination of various number of trees, may face the risk of overlearning if the dataset overfits the data set allocated for training. In order to reduce the effects of overlearning, various methods such as subsampling, shrinkage and early stopping are used (Natekin and Knoll, 2013: 8).

(Han and Kamber, 2011: 18; Özekes, 2003: 69). Figure 1 shows the general structure of the decision tree.

3.1.9. Extra Trees Classifier

The Extra Trees Algorithm, also known as Extremely Randomized Trees, generates an ensemble of unpruned decision trees. Compared to the randomised forests algorithm, it has two important differences. Firstly, the cut points are chosen completely randomly and divided into nodes. Second, the entire learning sample is used to grow the tree. In addition, when extra trees are evaluated in terms of variance and bias, using the entire learning sample instead of using recursive copies will minimise the bias (Geurts et al., 2006: 5-6).

3.1.10. Voting Classifier

The ensemble voting classifier is a meta-classifier that is similar to majority voting classification and utilises different machine learning classifiers. The training dataset is taken as input and classification models are constructed as (C_1,C_2,...,C_m). Then, the prediction values (P_1,P_2,....,P_m) obtained from each classifier are taken and the voting process is performed. Finally, the final probability (Pf) value is obtained after the voting process is completed (Mahabub and Habib, 2019: 3).

3.1.10.1. Hard Voting

Hard Voting is the simplest form of majority voting. Here, the class with the highest vote value is taken as the final prediction value to be used by the classifier. The class label is $\hat{y}$ and each classifier is Cj. Hard voting is as in equation 1 (Mahabub et al., 2019: 808).

$$\hat{y} = mod\ \{C_1(x), C_2(x), \ldots, C_m(x)\} \quad (1)$$

3.1.10.2. Soft Voting

In soft voting, each class label is predicted based on the predicted probability values P of the classifier. The probability values of the predictions from each model are summed and the class label with the maximum probability value is selected as the final prediction. Soft voting is calculated by the equation given in 2 (Mahabub, 2020: 4).

$$\hat{y} = \underset{i}{argmax} \sum_{j=1}^{m} W_j P_{ij} \quad (2)$$

3.2. Resampling Techniques

3.2.1. Random Undersampling

The Random Undersampling (RUS) method, which is based on the undersampling method, is a method that aims to balance the number of samples in the minority class by reducing some of the samples in the majority class in unbalanced datasets. The majority class samples in the training dataset continue until they reach the specified level between the minority and majority classes.

3.2.2. Random Oversampling

The Random Oversampling (ROS) method, which is based on the oversampling method, is a method that aims to balance the number of samples in the majority class and minority class by randomly duplicating the samples in the minority class in unbalanced datasets. In this method, only the samples in the minority class are copied and used in the training data (Alam et al., 2020: 201179; Kotsiantis et al., 2006: 4).

3.3. Ensemble of Samplers

3.3.1. Balanced Bagging Classifier (BBC)

In the Balanced Bagging Classifier method, new subsets are created from the original dataset sampling by using subsampling to balance the number of samples in the minority class and the number of samples in the majority class (Barros et al., 2019: 3). Then, these subsets contain ensembles of binary classification models by fitting different tree-based classifiers to various numbers of recursive training examples. Finally, the prediction values are obtained independently from each ensemble with the help of majority voting (Schlögl, 2020: 1).

3.4. Combination of Over- and Under-Sampling Methods

3.4.1. SMOTE-Tomek Links

The SMOTE-ENN method, known as the combination of Synthetic Minority oversampling Technique (SMOTE) and Edited Nearest Neighbour (ENN) methods, is a hybrid method used to reduce the number of samples in the majority class and increase the number of samples in the minority class in order to equalise the classes. The SMOTE method is used to add minority class instances, new artificial synthetic minority class instances and the Tomek Links method is used to reduce the majority class instances (Batista et al., 2003: 5).

3.4.2. SMOTE-ENN

The SMOTE-ENN method, known as the combination of the Synthetic Minority oversampling Technique (SMOTE) and the Edited Nearest Neighbor (ENN) methods, is performed by the SMOTE method for the addition of minority class samples, new artificial synthetic minority class samples, and the ENN method for the removal of observations that are determined to belong to the different class between the observation class and the K-nearest neighbour majority

from the two classes. The ENN method is a technique that is effectively used to remove noisy data in the dataset (Muaz et al., 2020: 481).

3.5. Performance Evaluation

Accuracy

The accuracy measure is mathematically as shown in Equation (3) (Vakili et al., 2020: 5).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

Precision

It shows the ability to predict the proportion of positives that are actually correct out of all positive predictions predicted by the model (Bradley et al., 2006: 2; Kumar, 2022). It is as shown in Equation (4).

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4)$$

Recall

It helps to measure how well the classifier performs in detecting all values that are actually positive from positive values (Kumar, 2022). It is as shown in Equation (5).

$$\text{Recall} = \frac{TP}{TP + FN} \quad (5)$$

F1 Score

The F1 score is a combination of precision and accuracy statistics and is also defined as the harmonic mean (Wood and Joshi, 2016). The F1 score is in the range [0,1]. The F1 score is as shown in Equation (6) (Chicco and Jurman, 2020: 5).

$$F_1 \text{ Score} = 2 * \frac{precision * recall}{precision + recall} \quad (6)$$

Matthews Correlation Coefficient (MCC)

MCC is expressed as a measure of the correlation between actual values and predicted values. The MCC measure takes values between [-1, +1]. MCC = -1 indicates a negative correlation between actual values and predicted values, while MCC = 1 indicates a positive correlation. MCC = 0 means that the forecasts are randomly generated (Chicco and Jurman, 2020: 5). The MCC criterion is as shown in Equation (7) (Al-Abbasi et al., 2020: 3126).

$$\text{MCC} = \frac{\text{TP} * \text{TN} - \text{FP} * \text{FN}}{\sqrt{(\text{TP} + \text{FP}).(\text{TP} + \text{FN}).(\text{TN} + \text{FP}).(\text{TN} + \text{FN})}} \quad (7)$$

4. DATASETS and VISUALIZATION

4.1. Datasets

The aim of the study on credit risk analysis is to identify risky customers and reduce credit risk by solving the class imbalance problem encountered in unbalanced datasets in the realisation of the lending process, thus contributing positively to the increase in the profit of banks. For this purpose, "Statlog German Credit Data Set" and "Statlog Australian Credit Approval Data Set" from the University of California Irvine (UCI) Machine Learning Repository and "HMEQ" datasets from the Kaggle website were taken (Hofmann,1994; Markelle et al.; Quinlan ; Vallala,2017). The datasets for Germany, Australia, and HMEQ consist of 1000 observations and 21 variables, 690 observations and 15 variables, and 5960 observations and 12 variables, respectively. The dependent variable is identified by a binary categorical value of "able or unable to receive a loan". The datasets for Germany, Australia, and HMEQ depict imbalances in the form of 700:300, 307:383, and 4771:1189, respectively.

In order to balance the unbalanced data, Resampling Methods; Random undersampling, Random oversampling, Ensemble of Samplers; Balanced Bagging Classifier and under-and-oversampling Combinations; SMOTE-Tomek Links and SMOTE-ENN methods were used. The data were classified using individual machine learning techniques such as K-nn, Logistic Regression, Naïve Bayes, Artificial

Neural Networks, Support Vector Machine, Decision Trees Algorithm and ensemble learning techniques such as Random Forests Algorithm, Gradient Boosting Decision Trees, Extra Trees Classifier and Voting Classifier (Hard Voting and Soft Voting) methods. Before the analysis, an experimental design was created in order to observe the effect of the results obtained. The most important advantage of the experimental design is to determine the level at which the inputs should be kept in order to optimise the final output and to facilitate the analysis to reach the optimal solution quickly. Four different scenarios were prepared to be used in the application of this study. These scenarios and their content information are shown in Table 1.

Table1: Scenarios in the Implementation Phase

Application Scenarios	Information on the Scenarios Used in Implementation
Scenario 1	Estimation of Singular Models with Random Undersampling and Random Oversampling Techniques
Scenario 2	Estimation of Ensemble Learning Models with Random Undersampling and Random Oversampling Techniques
Scenario 3	Estimation of Ensemble of Samplers and Singular Models by Combination of Under and Oversampling Methods
Scenario 4	Estimation of Ensemble of Samplers and Ensemble Learning Models by Combination of Under and Oversampling Methods

In this study, both undersampling methods and oversampling methods and their combining techniques such as "random oversampling, random undersampling, Balanced Bagging Classifier, SMOTE-ENN Combine, SMOTE-Tomek Links" were used in balancing three different unbalanced data sets; German loan data set, Australian loan data set and HMEQ loan data. 11 different machine learning techniques were used to evaluate the prediction performance of balanced datasets. These techniques are divided into different types within themselves. While techniques such as K-nearest neighbour, Naive Bayes, Logistic Regression, Support Vector Machines, Multilayer Perceptron and Decision Trees are called individual classification models, techniques such as Random Forests, Gradient Boosting Decision Trees, Extra Trees, Hard and Soft voting are called ensemble learning models. In addition, 5 different measurement statistics were utilised to evaluate the performance of these techniques. These performance measures are; Accuracy, Precision, Sensitivity, F1-Score and Matthews Correlation Coefficient.

In the data pre-processing process in the application phase; mean values were used to complete the missing observations of quantitative variables and mode values were used to complete the missing observations for qualitative data containing categorical structure. The application was implemented in Python using the scikit-learn library (Pedregosa et al., 2011). With the OneHotEncoder function of the library, the conversion of qualitative data with categorical properties into a new variable with numerical properties is performed. LabelEncoder function was used to digitise binary categorical variables and OrdinalEncoder function was used to digitise ordinal variable values if the variable values in the dataset are ordinal. Within the scope of this study, only information about the HMEQ Home loan dataset is mentioned.

4.2. Information on the HMEQ Loan Dataset

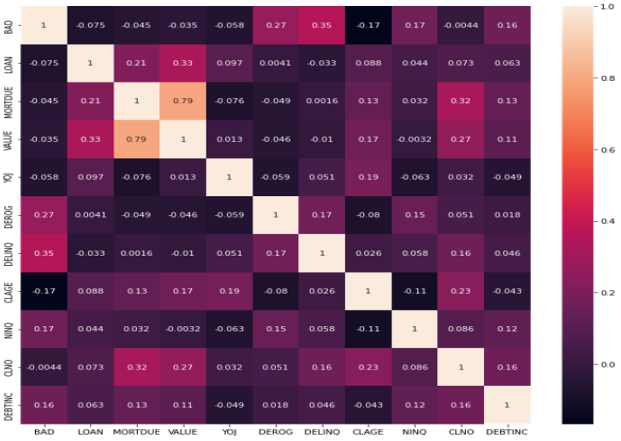

Figure 2: HMEQ home loan correlation matrix

When the correlation matrix obtained from the HMEQ housing loan dataset in Figure 1 is analysed, it is observed that the correlation value between the loan amount requested by the borrower and the value of the borrower's property is 0.33. This value means that there is a low strong positive linear relationship between the requested loan amount and the borrower's property value. The correlation value between the amount of the mortgage to be paid by the borrower (Mortdue) and the value of the borrower's property is 0.79. This value means that there is a strong positive linear relationship between the amount of the mortgage to be repaid by the borrower and the property value of the borrower. The correlation between the amount of mortgage to be paid by the borrower (Mortdue) and the number of loan instalments (Clno) is 0.32. It is possible to make similar interpretations among other variables.

Figure 3: Credit risk by customer's occupation and loan request amount

The swarmplot graph representing the credit risk status of the customers who requested a loan according to the amount of the loan requested and their occupational status is given. When the Other occupational group, which is not included in the occupational groups in Figure 3, is analysed, it is seen that the loans of the customers requesting loans in the range of approximately USD 1,000 to 3,000 ($) and in the range of approximately USD 40,000 to 42,000 ($) are not accepted. After the data discovery phase, the raw dataset was divided into two parts as training and test datasets in order to train the prediction models to be created for machine learning. While the majority of the raw dataset is used for training
the models, the remaining part is reserved for testing the model.
Seventy per cent of the data is allocated for training and 30 per cent of the data is allocated for testing. After this stage, machine learning classification algorithms were used.

5. RESULTS

In this section, a selection process is performed among the sampling methods and model combinations determined for the application scenarios. This selection process is based on selecting the method and scenario that performs the best in terms of both accuracy and effectiveness. In the process of determining the best method, a comparison was made between the performance values obtained from the methods in the applied scenarios and the method with the maximum value was determined as the best method. However, the choice of the optimal method varies in the performance measure according to the problem of interest. F1-score and MCC, which are widely used for unbalanced datasets, are evaluated in terms of performance measurement statistics. The performance measurement values of the resampling method and model combination for the best results obtained from four different scenarios performed on three different datasets are shown in Table 2 and Table 3 according to both F1-score and MCC criterion. Then, the final sampling method and model combination are selected from the optimal values selected in terms of F1-score and MCC, and this selection is represented in bold colour in the tables.

Tablo 2: Best scenarios in terms of F1-score

IMPLEMENTATION SCENARIOS				
Dataset	**Scenario 1**	**Scenario 2**	**Scenario 3**	**Scenario 4**
From the German Credit Dataset Sampling Combination of Method and Selected Model	0,5762	0,5810	0,5801	**0,5897**
	RUS-DT	RUS-XTrees	Balanced Bagging Classifier-MLP	**SMOTE-ENN Combine-GBDT**
From the Australian Credit Dataset Sampling Combination of Method and Selected Model	**0,8872**	0,8730	0,88	**0,8799**
	RUS-DT	RUS-XTrees	Balanced Bagging Classifier-MLP	**SMOTE-ENN Combine-Soft**
From the HMEQ Home Loan Dataset Sampling Combination of Method and Selected Model	0,9680	0,9456	**0,9699**	**0,9678**
	ROS-KNN	ROS-RF	**SMOTE-Tomek Links-KNN**	**SMOTE-Tomek Links- XTrees**

Note: Those giving the best results are shown in bold.

Table 3: Best scenarios in terms of MCC

IMPLEMENTATION SCENARIOS				
Dataset	**Scenario 1**	**Scenario 2**	**Scenario 3**	**Scenario 4**
From the German Credit Dataset Sampling Combination of Method and Selected Model	0,3701	0,4045	0,3867	**0,4397**
	RUS-DT	ROS-RF/ ROS-Hard	Balanced Bagging Classifier – SVM	**Balanced Bagging Classifier - RF**
From the Australian Credit Dataset Sampling Combination of Method and Selected Model	**0,7831**	0,7678	0,7830	0,7830
	RUS-DT	RUS-XTrees	Balanced Bagging Classifier -MLP	SMOTE-ENN Combine-Soft
From the HMEQ Home Loan Dataset Sampling Combination of Method and Selected Model	0,8301	0,7714	**0,8409**	0,8331
	ROS-KNN	ROS-XTrees	**SMOTE-Tomek Links-KNN**	SMOTE-Tomek Links- XTrees

Note: Those giving the best results are shown in bold.

When the best scenario values of the F1-score among the four scenarios applied to the German loan dataset are analysed, the most optimal result in terms of classification performance is measured in the SMOTE-ENN Combine-Gradient Boosting Decision Trees combination in scenario 4 with 58.97%. In addition, when the best scenario values of MCC among the four scenarios applied to the German loan dataset are analysed, the Balanced Bagging Classifier-Random Forests combination in

scenario 4 was measured with 43.97% in terms of classification performance. In other words, it can be said that the model success is at a moderate level. When the best scenario values of the F1-score among the four scenarios applied to the Australian loan dataset are analysed, the most optimal result in terms of classification performance is measured in the random undersampling-decision trees combination in scenario 1 with 88.72%. In addition, when the best scenario values of MCC among the four scenarios applied to the Australian loan dataset are analysed, it is seen that Balanced Bagging Classifier-MLP and SMOTE-ENN Combine-Soft give the same result and the random oversampling-decision tree combination in scenario 1 is selected with 78.31% in terms of classification performance with a small difference. In other words, it shows that the success of the model is at a high level.

When the best scenario values of F1-score among the four scenarios applied to the HMEQ Home loan dataset are examined, the most optimal result in terms of classification performance is measured in the SMOTE-Tomek Links-K-nearest neighbour combination in scenario 3 with 96.99%. In addition, when the best scenario values according to the MCC criterion among the four scenarios applied to the HMEQ Home loan dataset are examined, the SMOTE-Tomek Links-K-nearest neighbour combination in scenario 3 with 84.09% in terms of classification performance. In other words, it can be said that the model success is at a high level. If a general evaluation is made in terms of all scenarios and models of machine learning methods applied within the scope of the study, ensemble learning models showed better classification performance than single models. In this context, the identification of risky customers with the help of hybrid method and ensemble learning models using a combination of under-and-oversampling methods will have a positive impact on banking and other lending institutions in the financial sector by reducing workload, time, cost and resource utilisation costs, while at the same time increasing profits by reducing credit risk.

6. CONCLUSION

As a result, the scenario formed by the hybrid method based on oversampling and undersampling used in data balancing and machine learning models based on ensemble learning showed the best classification performance and it is suggested that it can be used in similar problems that banks and financial institutions may face in the future.

In future studies, in addition to the machine learning algorithms and classification performance measures used in the application phase, different methods and performance measures such as AUC-ROC curve can be used for evaluation. At the same time, the scope of the methods used in this study can be extended and compared with the existing findings. Within the scope of future studies, variable selection methods can be applied for the methods in the current study. These methods can be variable selection methods such as Lasso, Ridge and Elastic Net. Thus, by applying variable selection methods on the relevant data sets and scenarios used, the effect of variable selection on performance can be investigated. Furthermore, in addition to the methods used in the current study to eliminate class imbalance, ensemble sampler methods such as Bagging Classifier, Balanced Random Forest Classifier, RUS Boost, Easy Ensemble Classifier can be used to re-evaluate the performance of these datasets over the current scenarios.

REFERENCES

Akman, M., Genç, Y. ve Ankarali, H. (2011). Random Forests Yöntemi ve Saglik Alaninda Bir Uygulama/Random Forests Methods and an Application in Health Science. Türkiye Klinikleri Biyoistatistik. 3(1): 36.

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

Barros, T. M., Souza Neto, P. A., Silva, I. ve Guedes, L. A. (2019). Predictive models for imbalanced data: A school dropout perspective. Education Sciences. 9(4): 275.

Batista, G. E., Bazzan, A. L. ve Monard, M. C. (2003, December). Balancing Training Data for Automated Annotation of Keywords: a Case Study. In WOB (ss. 10-18).

Bradley, A. P., Duin, R. P. W., Paclik, P. ve Landgrebe, T. C. W. (2006). Precision-Recall Operating Characteristic (P-ROC) Curves in Imprecise Environments. In 18th International Conference on Pattern Recognition (ICPR'06) (pp.123-127). Cambridge , United Kingdom.

Breiman, L. (2001). Random forests. Machine learning. 45(1): 5-32.

Boughorbel, S., Jarray, F. ve El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PloS One. 12(6): 0177678.

Chicco, D. ve Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC genomics. 21(1): 1-13.

Chicco, D., Warrens, M. J. ve Jurman, G. (2021). The Matthews Correlation Coefficient (MCC) is More Informative Than

Dahiya, S., Handa, S. S. ve Singh, N. P. (2016). Impact of Bagging on MLP Classifier for Credit Evaluation. In 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), IEEE (pp. 3794-3800).

Domingos, P. ve Pazzani, M. (1997). On the Optimality of the Simple Bayesian Classifier Under Zero-One Loss. Machine learning. 29(2): 103-130.

Duda, R. O. ve Hart, P. E. (1973). Pattern Classification and Scene Analysis. New York: Wiley.

Dumitrescu, E., Hue, S., Hurlin, C. ve Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. European Journal of Operational Research. 297(3): 1178-1192.

Gehrke, J. (2003). Decision Trees. The Handbook of Data Mining (pp. 3-24). Editors Nong Ye. New Jersey: Lawrence Erlbaum Associates Inc.

Gupta, S., Kumar, D. ve Sharma, A. 2011. Data Mining Classification Techniques Applied for Breast Cancer Diagnosis and Prognosis. Indian Journal of Computer Science and Engineering. 2(2): 188-195.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E. ve Tatham, R. L. (1998). Multivariate data analysis . Uppersaddle River. Multivariate Data Analysis (5th ed) Upper Saddle River. 5(3): 207-219.

Han, J., Pei, J. ve Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

Hofmann, Hans. (1994). Statlog (German Credit Data). UCI Machine Learning Repository.

Horning, N. (2010). Random Forests: An Algorithm for Image Classification and Generation of Continuous Fields Data Sets. International Conference On Bioinformatics for Spatial Infrastructure Development in

Earth andAllied Sciences 2010 (pp.1–6). Osaka, Japan.

Hosmer, D. W. ve Lemesbow, S. (1980). Goodness of Fit Tests for the Multiple Logistic Regression Model. Communications in Statistics-Theory and Methods. 9(10): 1043–1069.

Hou, W. H., Wang, X. K., Zhang, H. Y., Wang, J. Q. ve Li, L. (2020). A novel dynamic ensemble selection classifier for an imbalanced data set: An application for credit risk assessment. Knowledge-Based Systems. 208: 106462.

Jin, Y., Zhang, W., Wu, X., Liu, Y. ve Hu, Z. (2021). A Novel Multi-Stage Ensemble Model With a Hybrid Genetic Algorithm for Credit Scoring on Imbalanced Data. IEEE Access. 9: 143593-143607.

Khemakhem, S., Ben Said, F. ve Boujelbene, Y. (2018). Credit Risk Assessment for Unbalanced Datasets Based on Data Mining, Artificial Neural Network and Support Vector Machines. Journal of Modelling in Management. 13(4): 932-951.

Kotsiantis, S., Kanellopoulos, D. ve Pintelas, P. (2006). Handling imbalanced datasets: A review. GESTS international transactions on computer science and engineering. 30(1): 25-36.

Kumar, A. (20.01.2022). Accuracy, Precision, Recall and F1-Score–Python Examples.

https://vitalflux.com/accuracy-precision-recall-f1-score-python-example/ #What _ is_ Recall_ Score, (26.02.2022).

Kumari, M. ve Godara, S. (2011). Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction. I. International Journal of Computer Science and Technology. 2: 304-308.

Le, C. T. ve Eberly, L. E. (2016). Introductory Biostatistics (2nd ed.). New Jersey: John Wiley & Sons.

Li, Y., Zhang, W. ve Lin, C. (2006). Simplify support vector machines by iterative learning. Neural Information Processing: Letters and Reviews. 10(1): 11-17.

Maciel, L. S. ve Ballini, R. (2008). Design a Neural Network for Time Series Financial Forecasting: Accuracy and Robustness Analysis. Anales do 9º Encontro Brasileiro de Finanças. Sao Pablo, Brazil.

Mahabub, A. (2020). A robust technique of fake news detection using Ensemble Voting Classifier and comparison with other classifiers. SN Applied Sciences. 2(4): 1-9.

Mahabub, A., Mahmud, M. I. ve Hossain, M. F. (2019). A robust system for message filtering using an ensemble machine learning supervised approach. ICIC Express Letters, Part B: Applications. 10(9): 805-812.

Malekipirbazari, M. ve Aksakalli, V. (2015). Risk assessment in social lending via random forests. Expert Systems with Applications. 42(10): 4621-4631.

Markelle Kelly, Rachel Longjohn, Kolby Nottingham. The UCI Machine Learning Repository, https://archive.ics.uci.edu.

Marksfeld, A. E. (2018). Strojové učení v sociodemografické segmentaci zákazníků telekomunikační společnosti. (Unpublished Bachelor's Thesis,). Czech Republic: Czech Technical University Computer and Information Center.

Martin, S. B. (2001). Techniques in Support Vector Classification. (Yayınlanmış Doktora Tezi). USA: Colorado State University.

McClelland, J. L., Rumelhart, D. E. ve Hinton,G. E. (1986). The appeal of parallel distributed processing. MIT Press, Cambridge MA, 3-44.

Muaz, A., Jayabalan, M. ve Thiruchelvam, V. (2020). A Comparison of Data Sampling Techniques for Credit Card Fraud Detection. International Journal of Advanced Computer Science and Applications (IJACSA).11(6).

Natekin, A. ve Knoll, A. (2013). Gradient boosting machines, a tutorial. Frontiers in neurorobotics. 7:21.

Niu, K., Zhang, Z., Liu, Y. ve Li, R. (2020). Resampling ensemble model based on data distribution for imbalanced credit risk evaluation in P2P lending. Information Sciences. 536: 120-134.

Özekes, S. (2003). Veri Madenciliği Modelleri ve Uygulama Alanları. İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi. 2(3): 65-82.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

Peterson, L. E. (2009). K-nearest neighbor. Scholarpedia. 4(2):1883.

Quinlan,Ross. Statlog (Australian Credit Approval). UCI Machine Learning Repository.

Shen, F., Zhao, X., Li, Z., Li, K. ve Meng, Z. (2019). A novel ensemble classification model based on neural networks and a classifier optimisation technique for imbalanced credit risk evaluation. Physica A: Statistical Mechanics and its Applications. 526: 121073.

Tu, M. C., Shin, D. ve Shin, D. (2009). A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms. In 2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing . (pp.183-187).

Vakili, M., Ghamsari, M. ve Rezaei, M. (2020). Performance Analysis and Comparison of Machine and Deep Learning Algorithms for Iot Data Classification. arXiv preprint arXiv:2001.09636.

Vallala, A. (2017). HMEQ_Dataset. Kaggle. https://www.kaggle.com/datasets/ajay1735/hmeq-data/data, (10.12.2021).

Vapnik, V. (1982). Estimation of Dependences Based on Empirical Data. New York: Springer Science+ Business Media, Inc.

Weinberger, K. Q., & Saul, L. K. (2008). Fast solvers and efficient implementations for distance metric learning. In Proceedings of the 25th international conference on Machine learning (ss. 1160-1167).

Wood, T. (2020). What is the F-score?. https://deepai.org/machine-learning-glossary-and-terms/f-score#:~:text=The%20F%2 Dscore%2C%20also% 20called, positive '% 20or %20'negative', (27.02.2022).