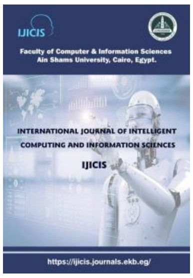
IJICIS, Vol.22, No.3, 92-108
DOI: 10.21608/ijicis.2022.105654.1138

International Journal of Intelligent
Computing and Information Sciences

https://ijicis.journals.ekb.eg/


IJICIS
Faculty of Computer & Information Sciences
Ain Shams University, Cairo, Egypt.
INTERNATIONAL JOURNAL OF INTELLIGENT
COMPUTING AND INFORMATION SCIENCES
IJICIS
https://ijicis.journals.ekb.eg/

Intelligent Model for Enhancing the Bankruptcy Prediction with Imbalanced Data Using Oversampling and CatBoost

Samar Aly*
Computer Science Department,
Faculty of Computer and
Information Science, Ain Shams
University, Cairo, Egypt
samar.aly@cis.asu.edu.eg

Marco Alfonse
Computer Science Department,
Faculty of Computer and
Information Science, Ain Shams
University, Cairo, Egypt
marco_alfonse@cis.asu.edu.eg

Abdel-Badeeh M. Salem
Computer Science Department,
Faculty of Computer and
Information Science, Ain Shams
University, Cairo, Egypt
absalem@cis.asu.edu.eg

Received 2021-11-12; Revised 2022-06-06; Accepted 2022-06-09

Abstract: *Bankruptcy prediction is one of the most significant financial decision-making problems, which prevents financial institutions from sever risks. Most of bankruptcy datasets suffer from imbalanced distribution between output classes, which could lead to misclassification in the prediction results. This research paper presents an efficient bankruptcy prediction model that can handle imbalanced dataset problem by applying Synthetic Minority Oversampling Technique (SMOTE) as a pre-processing step. It applies ensemble-based machine learning classifier, namely, Categorical Boosting (CatBoost) to classify between active and inactive classes. Moreover, the proposed model reduces the dimensionality of the used dataset to increase predictive performance by using three different feature selection techniques. The proposed model is evaluated across the most popular imbalanced bankrupt dataset, which is the Polish dataset. The obtained results proved the efficiency of the applied model, especially in terms of the accuracy. The accuracies ofthe proposed model in predicting bankruptcy on the Polish five years datasets are 98%, 98%, 97%, 97% and 95%, respectively.*

Keywords: *Bankruptcy prediction, Imbalanced dataset, Machine learning, CatBoost, Classification*

1. Introduction

Prediction of bankruptcy is a remarkable research point, which helps decision makers to monitor the health of financial institutions. In the recent years, financial institutions became exposed to a higher probability of being bankrupt. Which is due to the latest changes in the global economy after the 2007

* Corresponding author: Samar Aly
Computer Science Department, Faculty of Computer and Information Science, Ain Shams University, Cairo, Egypt
E-mail address: samar.aly@cis.asu.edu.eg

financial crisis [1]. The early prediction of the bankruptcy problem aims to ensure more stability in the economy [2]. The bankruptcy prediction models aim to help in bankruptcy problem to assess if the institution will go bankrupt or not [3]. Developing an efficient model for predicting bankruptcy to estimate the financial crisis before it happens is very critical point for both society and economy [4].

Bankruptcy prediction models perform a binary classification process. The classified output of classification models represents two types either bankrupt class or non-bankrupt class. In general, models of bankruptcy prediction can be categorized by statistical techniques and Artificial Intelligent (AI) techniques. Statistical techniques were used earlier as predicting models for the bankruptcy problem. The widely used statistical techniques are multiple discriminant analysis [5–7], Logistic Regression (LR) [8–11]. After that, the AI techniques proved an outstanding performance in predicting bankruptcy more than statistical techniques [6], [7], [8], [1]. In binary classification, machine learning techniques are the most significant branch of AI especially in the finance field [12].

The most popular machine learning techniques are ensemble-based classifiers. Ensemble-based classifiers integrate more than one single-based classifier to construct a robust ensemble classifier [9], [10]. Examples for using ensemble classifiers in predicting bankruptcy are Bagging (BA) [12], [1], Boosting (BO) [11], [13], Adaptive Boosting (AdaBoost) [14], [15], [16] and GBoost [17], [18]. The most recent ensemble classifiers are Extreme gradient boosting (XGBoost) [19–21], Light Boosting (LBoost) [22] and CatBoost [23–25].

Most of the used bankrupt datasets suffer from the problem of imbalanced data. The imbalanced data problem separates the dataset into majority (negative) class and minority class (positive). The classification models based on imbalanced datasets may cause bias to the majority class in construction models by classifying most of the institutions as non-bankrupt [26]. Therefore, most of the recent studies intended to solve the problem of imbalanced data. Recently, the most popular re-sampling technique for handling imbalanced data is oversampling technique [3], [27], [28]. Thus, this research paper aims to handle imbalanced dataset by applying SMOTE oversampling technique to increase the efficiency of proposed model.

This paper is structured as follows: Section 2 presents the related work of some previous studies; Section 3 presents the proposed methodology for predicting bankruptcy; Section 4 presents the performance evaluation of the proposed model; Section 5 presents the experimental results; and Section 6 shows our conclusions and future work.

2. Related Work

This section summarizes some relevant studies that are based on using the machine learning models to help in predicting bankruptcy problem across the Polish dataset. The relevant studies were from 2019 to 2021. We collected the relevant studies from Web of Science (WoS) as we care more about the highly citation journals.

Vicente García et al. [29] applied seven ensemble-based machine learning classifiers to predict bankruptcy problem across fourteen real-life datasets. The main objective of García et al.'s study is to determine the relation between ensemble-based classifiers' performance and the type of bankrupt instances. The applied classifiers are BA, AdaBoost, random subspace, DECORATE, rotation forest, Random Forest (RF), and stochastic GBoost. The used datasets contain various types of bankrupt instances. García et al. [29] proved that the performance of ensemble-based classifiers is affected by type of bankrupt instances.

Tsai [30] presented a hybrid machine learning technique consists of two stages in predicting bankruptcy problem. The first stage depends on a clustering technique to ignore the irrelevant instances from training set. In the second stage, the output from the first stage is used in constructing classification models to predict bankruptcy problem. Tsai [30] applied the LR as a statistical technique and Support Vector Machine (SVM) as a machine learning technique to compare between them and show the importance of machine learning techniques. He used five varied datasets which are Australian [31], German [32], Japanese [33], Polish [34] and Taiwanese [35] datasets. His experimental results showed that the SVM with using Affinity Propagation (AP) and k-means to select relevant instances prove the best performance. Moreover, Tsai' study proved that performance of SVM with AP and K-means depends on the used dataset.

Smiti and Soui [2] applied five different machine learning classifiers for predicting bankruptcy problem and compared between their performance. The applied classifiers are k- nearest neighbor, Decision Tree (DT), SVM, and Artificial Neural Network (ANN), C5.0. The authors were interested to solve imbalanced dataset problem by applying borderline synthetic minority oversampling technique. They evaluated their applied model across Polish dataset [34]. Moreover, they applied stacked autoencoder feature selection technique as a pre-processing step to filter out the irrelevant features from training set. Their obtained results showed that the C5. classifier outperformed the other applied machine learning classifiers in predicting bankruptcy problem.

Zhang et al. [14] proposed a new multi-stage approach to predict bankruptcy problem with advanced outlier adjustment. He applied BA technique with an algorithm of local outlier factor to adapt outliers which exists in the noise datasets and caused a bad consequence. The adapted outliers are boosted again in the training dataset to construct a good classifier. The authors proposed a new feature selection, which uses the chi-square to reduce features dimension and ignore unrepresentative features. To adapt parameters of base ensemble classifier, the self-adaptive parameter technique is used with stacking-based ensemble technique. The proposed model was evaluated across ten datasets which include Australian [31], German [32], Polish [34], Japanese [33] and Taiwan datasets [35]. To indicate the performance and effectiveness of the proposed model, the obtained results contain a statistical analysis outcome.

3. Proposed Methodology for predicting bankruptcy

3.1 Problem Description

The proposed bankruptcy prediction model consists of three main stages: re-sampling the used dataset, relevant feature selection and CatBoost based data classification. The proposed classification model is evaluated on the highly imbalanced Polish dataset. The used dataset contains some missing values and may lose useful information if instances with missing values are removed. Hence, the major pre-processing step before any stage of the proposed model is filling the missed cells. The re-sampling stage is applying the SMOTE oversampling technique to have a balanced distribution ratio between bankrupt and no-bankrupt classes. It ensures more accurate classification results. In the second stage, the filter-based and wrapper-based feature selection methods were applied to filter out the irrelevant data. The applied filter-based method is Correlation Feature Selection (CFS) and the applied wrapper-based methods are Sequential Feature Selection (SFS) and Recursive Feature Elimination (RFE). Then, the used dataset is split into training and testing sets. In the third stage, the CatBoost classifier is used in constructing the classification model based on the training and testing datasets. The overall stages of the proposed model are shown in Figure1.

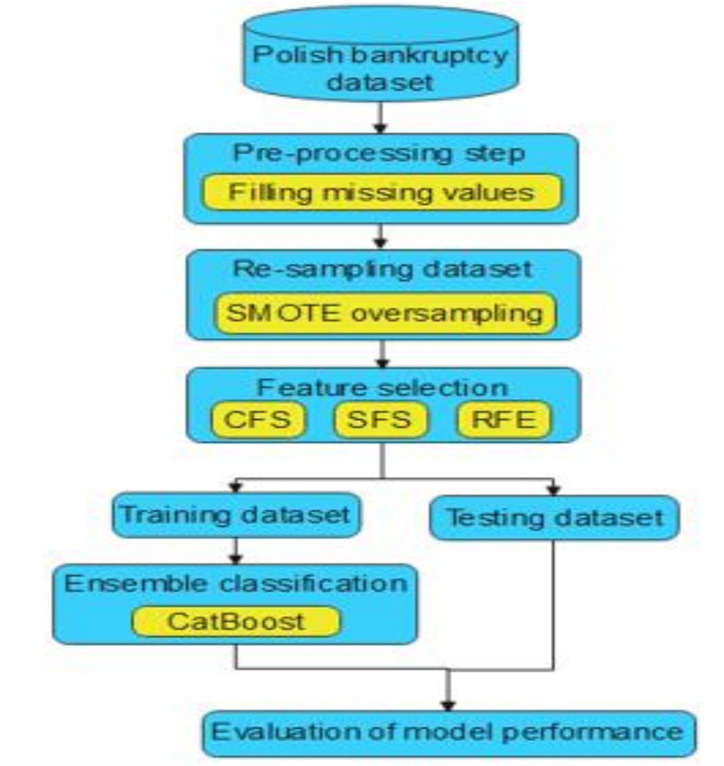


Figure 1: The overall steps of the proposed model

The following subsections discuss the investigated techniques to solve the bankruptcy prediction problem. It shows the main stages in the proposed model to predict bankruptcy problem on the Polish dataset.

3.2 Re-sampling Input Dataset with SMOTE

One of the main problems that leads to misclassification results with machine learning models is imbalanced class distribution of the input dataset. Most of the credit datasets have imbalanced class distribution problem since number of majority class (non-bankrupt institutions) is much more than number of minority class (bankrupt institutions). The oversampling technique is responsible for handling imbalanced dataset problem by creating new synthetic values in the minority class to balance input dataset. It chooses a random value from the minority class to be the new synthetic feature which could lead to overfitting problem. In 2002, Chawla [36] proposed SMOTE to overcome the drawbacks of random oversampling technique [37].

The SMOTE depends on the similarity between instances of the minority feature space to create the new synthetic values. For each instance $\boldsymbol{x_i} \in$ the minority class, the k-nearest neighbors $\boldsymbol{K_{x_i}}$are calculated based on Euclidean distance of $\boldsymbol{x_i}$ using K input variable from user. The SMOTE selects a random value from nearest neighbors $\boldsymbol{K_{x_i}}$ to calculate the new synthetic value in the minority class. Algorithm 1 shows the creation of the new synthetic value by SMOTE technique according to [3].

Algorithm 1. Oversampling SMOTE

```
Input: M is the number of instances in minority class; R: the ratio of SMOTE
M%; k: number of nearest neighbors; minority data D = $x_i$ ϵ X, where i = 1, 2, 3,
…, T.
Output: New synthetic values N
R= (int)(R/100)
    for i = 1 to M do
    1.    Find $K_{x_i}$of $x_i$
    While R ≠ 0 do
    1.        Select a random value from $K_{x_i}$; $x_{\hat{\imath}}$
    2.        Select a random number ɣ ϵ [0,1]
    3.        New synthetic value = $x_i$+($x_{\hat{\imath}}$  -$x_i$)×ɣ
    4.        Append New synthetic value to N
    5.        R – –
        end while
    end for
```

3.3 Feature Selection

In bankruptcy prediction problem, feature selection is an important pre-processing step with machine learning based models and high dimensional datasets [28]. The high dimensional dataset with some machine learning classifiers may lead to wasting memory storage and consuming long running time. Feature selection methods help in removing redundant and irrelevant features from datasets while keeping the most effective features. Choosing a good feature selection method improves prediction performance and minimizing the complexity of classification models [38]. In the proposed model, the filter-based and wrapper-based feature selection methods were applied in the pre-processing step to select relevant and dependent variables from the Polish dataset.

3.3.1 Filter based feature selection

The filter-based feature selection method is independent of any machine learning technique. By using univariant statistics, the filter method can analyze the features properties [39]. The filter method generates the best subset of features with the highest score by calculating scores for each feature [19]. The CFS method is the most commonly used filter-based methods. The proposed work applied the CFS as a pre-processing feature selection step.

CFS is a linear method to measure the relevance between features and target class. It selects a good relevance subset of features which contain the highest correlated features with target class. However, the features in the selected subset are uncorrelated to each other [3]. CFS uses a heuristic method to evaluate the importance of features and if features can highly contribute to the decision of the target class [40].

3.3.2 Wrapper-based feature selection

Wrapper method depends on a machine learning predictor. It relies on the greedy search algorithm and generates all possible combination of features, then the best subset of features is selected based on the high-performance scores of the machine learning predictor. The wrapper method requires very high computational complexity more than the filter method [41]. However, the wrapper methods have a better performance than filter methods in some cases [42].

The proposed model applied two wrapper feature selection methods which are Sequential Feature Selection (SFS) and Recursive Feature Elimination (RFE) based on CatBoost predictor with accuracy as a performance metric.

- **Sequential Forward Selection**

  SFS is a widely used wrapper method. It is based on the greedy search algorithm to select the perfect subset of features based on the importance score of each feature. It calculates the score of each feature based on the predictive technique CatBoost to filter out redundant features from initial dataset. Initially, the SFS adds the feature with the highest score to the initial subset of features. After that, the SFS sequentially calculates the score of the remaining features to add the most significant features to the subset of selected features until the predetermined number of features by the user [17].

- **Recursive Feature Elimination**

RFE is a popular and flexible wrapper feature selection method. It is based on backward selection strategy to select the most optimal subset of features. It begins by constructing the CatBoost predictor with initial set of features to score the importance of each feature based on their coefficients. After that, RFE eliminates the lowest importance features from the initial set of features to obtain a new subset of features. RFE recursively reconstructs the CatBoost predictor to score importance of features with the new subset of features until a predetermined number of features [43].

3.4 CatBoost Classifier

CatBoost classifier is an enhancement of the GBoost technique [43]. In 2018, CatBoost was developed by [23], [24]. It showed a robust performance in many machine learning applications especially in finance [22]. The CatBoost technique provides high performance with its default parameters without the need to waste time in tuning parameters. It depends on a symmetric DT as a base ensemble classifier. In the training step, the GBoost uses the same samples to estimate the gradients boosting which could lead to prediction shift problem. CatBoost technique attempts to solve the prediction shift problem that may lead to overfitting as in GBoost [25]. Most of the bankruptcy datasets contain categorical data. Converting categorical features to numeric features is a very important step, this step may lead to overfitting problem. Various training datasets are needed in the training step to avoid overfitting in converting categorical features, but this process is not available. CatBoost technique has two outstanding functions to overcome overfitting problem more than any other GBoost technique. First, it applies order boosting technique to overcome prediction shift problem. Second, it can handle categorical features without need for any pre-processing step to encode categorical features [22]. CatBoost technique performs various random permutations to ensure a robust choosing for the leaf node to build a powerful structure of the based DT classifier without overfitting problem. The predicted label for each instance is calculated according to [24] as shown in Eq.(1)

$$\mathrm{L}(x) = m_i \mathbb{1}_{\{x \in \mathcal{R}_i\}} \qquad \text{Eq.(1)}$$

Such that $\mathrm{L}(x)$ represents the function of DT, and $\mathcal{R}_i$is the disjoint region compatible with the leaves of the DT. The CatBoost techniques shows a high performance as a feature selection predictor to determine the importance score of each feature [22]. The proposed model applied CatBoost in the pre-processing step as a feature selector and also in constructing the main model as a machine learning classifier. Table 1 presents the tuned parameters of the CatBoost as a feature selector and machine learning classifier.

Table 1: Meta-parameters of CatBoost technique

	Verbose	iterations	estimators	random_state	l2_leaf_reg
Feature selector	0	1000	-	42	5
Classifier	0	-	100	42	0.01

For the parameters in Table 1, the number of estimators represents how many trees will be used in the algorithm, selecting too many trees increases the complexity of the model as well as the training time while selecting too few trees may lead to low accuracy, here we selected 100 trees which allows the model to gain high accuracy without increasing the model complexity. The number of iterations and the coefficient of the l2 regularization of the cost function are selected by try and error, here we needed 1000 iterations to gain reasonable accuracy. Other hyper-parameters were selected as the default value of the parameter which are tuned by the python package.

4. The evaluation of Performance

This section introduces the used dataset to construct the proposed model. Also, it describes the performance measures that are used to evaluate the prediction performance of the proposed model.

4.1 Polish Enterprises Dataset

The prediction performance of the proposed model is evaluated based on the Polish enterprises' dataset. The Polish dataset is available online on the University of California Irvine (UCI) Machine Learning Repository [34]. It contains five real-life classification datasets with different periods [2]. Table 2 contains the basic information of the used Polish dataset. The bankrupt enterprises were evaluated from 2000 to 2012 and the non-bankrupt enterprises were evaluated from 2007 to 2013. The Polish dataset has missing values in its five periods.

In most cases, missing values reduce the accuracy of machine learning models because they may bias classification results. The basic approaches for handling missing values are deleting rows with missing values or imputation of missing values. Drop missing values is the worst solution since it causes the loss of some valuable data. Imputing the missing values is the best solution to replace missing values with the most approximate values based on each feature. It also provides the ability for machine learning models to train with a large number of records to enhance prediction performance. The missing values within instances in Polish dataset were filled by the mean value of each feature independently. The Polish dataset has an imbalanced class distribution ratio between the classification classes as many credit datasets [1].

Table 2: The basic information of the Polish dataset.

Dataset	Features	Non-bankrupt institutions	Bankrupt institutions	Records
Polish-1st	64	6756	271	7027
Polish-2nd	64	9773	400	10173
Polish-3rd	64	10503	495	10008
Polish-4th	64	9277	515	9792
Polish-5th	64	5500	410	5910

Table 3 shows a sample of the selected features that have high correlation to the target class for the datasets used in evaluation where the correlation is greater than 0.8.

Table 3: The highly correlated features to the target class

Dataset	Number of Highly Correlated Features	Highly Correlated Features
Polish-1st	12	['Attr1', 'Attr4', 'Attr13', 'Attr21', 'Atrr27', 'Attr28', 'Attr29', 'Attr37', 'Attr45', 'Attr47', 'Attr55', 'Attr57']
Polish-2nd	64	All the features are correlated
Polish-3rd	48	All the features are correlated except ['Attr2', 'Attr3', 'Attr6', 'Attr7', 'Attr10', 'Attr11', 'Attr14', 'Attr18', 'Attr22', 'Attr25', 'Attr35', 'Attr36', 'Attr38', 'Attr48', 'Attr51', 'Attr59']
Polish-4th	64	All the features are correlated
Polish-5th	52	All the features are correlated except ['Attr19', 'Attr23', 'Attr30', 'Attr31', 'Attr39', 'Attr42', 'Attr43', 'Attr44', 'Attr49', 'Attr56', 'Attr58', 'Attr62']

4.2 Performance Evaluation of The Proposed Model

To evaluate the classification performance of the proposed model in predicting bankruptcy problem with more reliable results, three performance measurements are used. The performance measurements are accuracy, Area Under the Curve (AUC) and F-score. The evaluation indicators of the used performance measures are represented by $2x2$ confusion matrix as shown in Table 4. Based on the bankruptcy prediction problem, True Positive (TP) and True Negative (TN) are the correctly classified bankrupt and non-bankrupt classes, respectively. Moreover, False Negative (FN) and False Positive (FP) are the misclassified bankrupt and non-bankrupt classes, respectively.

Table 4: The indicators of the performance measurements

Actual class	Predicted bankrupt	Predicted non-bankrupt
Bankrupt	TP	FN
non-bankrupt	FP	TN

Accuracy: Accuracy is the most widely used performance measure to represent efficiency of machine learning models. Accuracy is estimated by the ratio between the correctly predicted instances and the total number of instances as show in Eq.(2) [26].

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \qquad \text{Eq. (2)}$$

The accuracy cannot differentiate between bankrupt and non-bankrupt instances, so the proposed model depends on two other performance measures [2].

AUC: AUC ratio is the most robust performance measure to evaluate the overall performance of binary classification problems. It shows a high ability in bankruptcy prediction with machine learning models to differentiate between bankrupt and non-bankrupt classes. The Receiver Operating Characteristic (ROC) curve is used to calculate the AUC, which measures the efficiency of the applied model to balance between FP and TP rates [44].

F-score: F-score combines between precision and recall and is defined as the harmonic average between them as shown in Eq.(5). The precision and recall ratio are presented in equations Eqs.(3) and (4), respectively. It cares more about FP and FN. Its high ratio is close to 1 [45].

$$\text{Precision} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP}) \text{Eq. (3)}$$

$$\text{Recall} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN}) \text{Eq. (4)}$$

$$\text{F-score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (5)$$

5. Experimental Results

This section presents the experimental results of the proposed model to predict bankruptcy problem. Table 5 shows the performance of the proposed model based on CFS, SFS and RFE feature selection methods, respectively. Thirty out of the 64 features were selected based on the used wrapper methods. The proposed model combined the SMOTE oversampling technique with tuned CatBoost machine learning classifier on the Polish dataset. The used Polish dataset were split into 80% for constructing the proposed machine learning model, while 20% of the Polish dataset were used to evaluate prediction performance. The experiments were evaluated on NVIDIA GEFORCE GPU with Intel CORE i7, 2.6 GHz and 8 GB RAM using python language.

Table 5: The performance evaluation of the proposed model on the Polish dataset with the three used feature selection methods

Feature selection method	Measure	Polish-1st	Polish-2nd	Polish-3rd	Polish-4th	Polish-5th
CFS	Accuracy	0.96515	0.98133	0.96906	0.96682	0.94924
	AUC	0.90658	0.92110	0.93450	0.93394	0.94209
	F-score	0.57	0.68	0.67	0.64	0.68
SFS	Accuracy	0.98293	0.97936	0.96811	0.96069	0.94670
	AUC	0.95744	0.91917	0.93222	0.92556	0.93897
	F-score	0.76	0.66	0.66	0.62	0.66
RFE	Accuracy	0.98435	0.97543	0.96716	0.96069	0.95347
	AUC	0.97077	0.91974	0.93093	0.93088	0.95163
	F-score	0.78	0.61	0.65	0.61	0.70

Regarding the Polish-1st dataset, the proposed model shows a better performance with wrapper-based feature selection methods (SFS and RFE) more than the CFS. The RFE could efficiently determine the relevant features and prove a higher performance more than the SFS method. In general, the difference in performance between used wrapper methods is very slight. However, the SFS requires a high computation and consume running time. The SFS spent much time to extract the most important features in the training step. The CFS could only determine 19 features which are correlated with the target class. It cannot determine the accurate correlation between features, so with 19 features the CFS could not prove a high-performance measure.

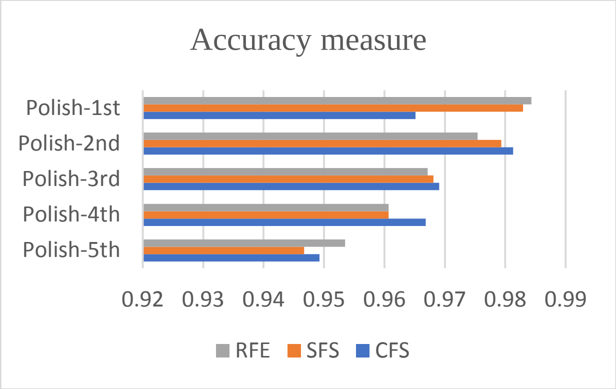
Regarding the Polish-2nd dataset, the CFS method proved the better performance measures. The CFS selected all the features as correlated with the target class, so it did not miss any feature that might be correlated with the target class. The SFS follows the CFS method in the performance measures. The F-score difference between the CFS and RFE is about 0.07.

Regarding the Polish-3rd dataset, the difference between the three used feature selection methods is very slight is about 0.001, but the CFS method proved the better performance followed by the SFS. The CFS selected 48 features correlated with the target class.

Regarding the Polish-4th dataset, the CFS selected all the 64 features as correlated with target class. The CFS proved the better performance followed by RFE then SFS. It showed the best F-score measure which indicates its high ability to distinguish between bankrupt and non-bankrupt classes.

Regarding the Polish-5th dataset, the RFE proved the better performance measures followed by the CFS then SFS. Despite of the large number of features which is 53, the RFE conquered the CFS method.

Figure 2-Figure 4 present the performance measures of the proposed model in terms of accuracy, AUC and F-score, respectively.



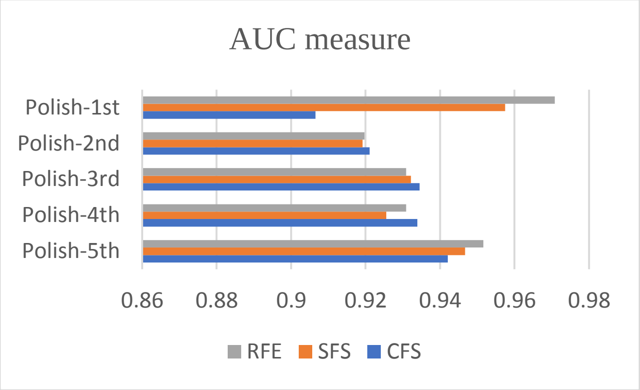


Figure 2: Accuracy performance measure of proposed model

Figure 3: AUC performance measure of proposed model

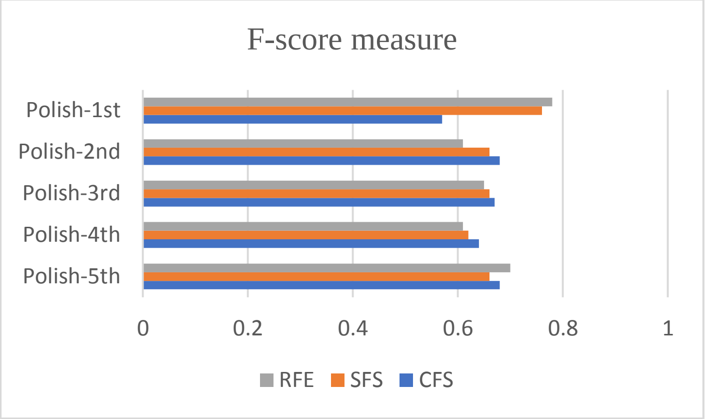


Figure 4: F-score performance measure of proposed model

Table 6 presents the experimental results of the previous studies on the five years of Polish dataset.

Table 6: The performance evaluation of the previous studies on the Polish dataset

Author	Measure	Polish-1st	Polish-2nd	Polish-3rd	Polish-4th	Polish-5th
Vicente García et al. 2019 [29]	Accuracy	-	-	-	-	-
	AUC	0.923	0.879	0.900	0.901	0.936
	F-score	-	-	-	-	-
Tsai 2020 [30]	Accuracy	-	-	-	-	-
	AUC	0.917	0.959	0.958	0.950	0.892
	F-Score	-	-	-	-	-
Smiti and Soui 2020 [2]	Accuracy	-	-	-	-	-
	AUC	0.965	0.950	0.968	0.969	0.950
	F-score	-	-	-	-	-
Zhang et al. 2021 [14]	Accuracy	0.98172	0.97621	0.97583	0.97136	0.96863
	AUC	0.93662	0.89494	0.89905	0.95528	0.95439
	F-score	0.74510	0.57692	0.66265	0.76272	0.77120

In what follows, the comparison between the proposed model and previous benchmark studies presented by [29], [30], [2] and [14] are discussed. These previous studies used only the AUC as a performance measure. However, the proposed model and the previous study by [14] depends on three performance measures (accuracy, AUC and F-Score).

Regarding the Polish-1st, the AUC ratio of the proposed model with used wrapper feature selection methods (SFS and RFE) is much better than the proposed models by [29], [30] and [14]. The proposed model with RFE has a better AUC ratio better than [2]. Also, the proposed model with wrapper feature selection has a better accuracy and F-score more than [14].

Regarding the Polish-2nd, the AUC ratio of the proposed model with the three used feature selection methods (CFS, SFS and RFE) is much better than the proposed models by [29] and [14]. The proposed model with CFS and SFS feature selection methods has a better performance accuracy more than [14]. However, the difference between the proposed model with RFE and [14] according to the accuracy is only 0.001. The proposed model with the three used feature selection methods outperforms the performance of [14] according to F-score.

Regarding the Polish-3rd, the AUC ratio of the proposed model with the three used feature selection methods (CFS, SFS and RFE) is much better than the proposed models by [29] and [14]. The proposed model with CFS and SFS feature selection methods has a better performance according to F-score more than [14]. The proposed model care more about AUC ratio as the most accurate measure with classification problems especially when previous models used imbalanced dataset as [14].

Regarding the Polish-4th, the AUC ratio of the proposed model with the three used feature selection methods (CFS, SFS and RFE) is much better than the proposed model by [29].

Regarding the Polish-5th, the AUC ratio of the proposed model with the three used feature selection methods (CFS, SFS and RFE) is much better than the proposed models by [29] and [30]. The AUC ratio of the proposed model with RFE is much better than [2]. The difference between the proposed model and the proposed model by [14] according to the AUC ratio is only 0.003.

6. Conclusion and Future Work

This research paper aims to solve the bankruptcy prediction problem using machine learning techniques. One of the major challenges facing the machine learning techniques is dealing with highly imbalanced dataset. Hence, the proposed research aims to develop an efficient model, throughout combining the CatBoost ensemble classifier for predicting bankruptcy with the SMOTE oversampling technique and several feature selection techniques. The proposed model focused on SMOTE re-sampling technique to avoid drawbacks of random oversampling technique. Based on evaluating the results, it became appeared that the proposed model with CatBoost classifier as an auxiliary method in the wrapper feature selection process, showed promising results in predicting bankruptcy.

The proposed model was applied on five different datasets of Polish enterprises across different periods. The missing values of these datasets are handled by replacing the missing value with the mean value of each feature, this handling method when combined with the CatBoost classifier achieved a high classification accuracy. We observed that CFS method can show a better performance relative to wrapper methods in dealing with large number of instances. This is because the CFS method has the ability to extract the features that are highly correlated with target class. On the other hand, the wrapper feature selection methods showed a better performance relative to the CFS method in dealing with small number of instances. The future work is to apply the proposed methodology to larger datasets to ensure the reliability of the obtained results.

References

[1] Z. Chen, W. Chen, and Y. Shi, “Ensemble learning with label proportions for bankruptcy prediction,” Expert Systems with Applications, vol. 146, p. 113155, 2020, doi: 10.1016/j.eswa.2019.113155.

[2] S. Smiti and M. Soui, “Bankruptcy Prediction Using Deep Learning Approach Based on Borderline SMOTE,” Information Systems Frontiers, vol. 22, no. 5, pp. 1067–1083, 2020, doi: 10.1007/s10796-020-10031-6.

[3] N. Ghatasheh, H. Faris, R. Abukhurma, P. A. Castillo, N. Al-Madi, A. M. Mora, A. M. Al-Zoubi, and A. Hassanat, “Cost-sensitive ensemble methods for bankruptcy prediction in a highly imbalanced data distribution: a real case from the Spanish market,” Progress in Artificial Intelligence, vol. 9, no. 4, pp. 361–375, 2020, doi: 10.1007/s13748-020-00219-x.

[4] S. K. Shrivastav and P. Janaki Ramudu, “Bankruptcy prediction and stress quantification using support vector machine: Evidence from Indian banks,” Risks, vol. 8, no. 2, p. 52, 2020, doi: 10.3390/risks8020052.

[5] F. Antunes, B. Ribeiro, and F. Pereira, “Probabilistic modeling and visualization for bankruptcy prediction,” Applied Soft Computing Journal, vol. 60, pp. 831–843, 2017, doi: 10.1016/j.asoc.2017.06.043.

[6] D. Zhao, C. Huang, Y. Wei, F. Yu, M. Wang, and H. Chen, “An Effective Computational Model for Bankruptcy Prediction Using Kernel Extreme Learning Machine Approach,” Computational Economics, vol. 49, no. 2, pp. 325–341, 2017, doi: 10.1007/s10614-016-9562-7.

[7] M. Wang, H. Chen, H. Li, Z. Cai, and X. Zhao, “Engineering Applications of Artificial Intelligence Grey wolf optimization evolving kernel extreme learning machine : Application to bankruptcy prediction,” Engineering Applications of Artificial Intelligence, vol. 63, pp. 54–68, 2017, doi: 10.1016/j.engappai.2017.05.003.

[8] H. Son, C. Hyun, D. Phan, and H. J. Hwang, “Data analytic approach for bankruptcy prediction,” Expert Systems With Applications, vol. 138, pp. 112–816, 2019, doi: 10.1016/j.eswa.2019.07.033.

[9] D. Liang, C. Lu, C. Tsai, and G. Shih, “Financial ratios and corporate governance indicators in bankruptcy prediction : A comprehensive study,” European Journal of Operational Research, vol. 252, no. 2, pp. 561–572, 2016, doi: 10.1016/j.ejor.2016.01.012.

[10] M. Ala’raj and M. F.Abbod, “A new hybrid ensemble credit scoring model based on classifiers consensus system approach,” Expert Systems with Applications, vol. 64, pp. 36–55, 2016, doi: 10.1016/j.eswa.2016.07.017.

[11] J. Abellán and J. G. Castellano, “A comparative study on base classifiers in ensemble methods for credit scoring,” Expert Systems with Applications, vol. 73, pp. 1–10, 2017, doi: 10.1016/j.eswa.2016.12.020.

[12] F. Barboza, H. Kimura, and E. Altman, “Machine learning models and bankruptcy prediction,” Expert Systems with Applications, vol. 83, pp. 405–417, 2017, doi: 10.1016/j.eswa.2017.04.006.

[13] P. Carmona, F. Climent, and A. Momparler, “Predicting failure in the U.S. banking sector: An extreme gradient boosting approach,” International Review of Economics and Finance, vol. 61, pp. 304–323, May 2019, doi: 10.1016/j.iref.2018.03.008.

[14] W. Zhang, D. Yang, S. Zhang, J. H. Ablanedo-Rosas, X. Wu, and Y. Lou, “A novel multi-stage ensemble model with enhanced outlier adaptation for credit scoring,” Expert Systems with Applications, vol. 165, no. December 2019, pp. 113–872, 2021, doi: 10.1016/j.eswa.2020.113872.

[15] I. Chaabane, R. Guermazi, and M. Hammami, Enhancing techniques for learning decision trees from imbalanced data, vol. 14, no. 3. Springer Berlin Heidelberg, 2020. doi: 10.1007/s11634-019-00354-x.

[16] N. O. S. S. Al Abdouli, “Handling the Class Imbalance Problem in Binary Classification,” 2014.

[17] Y. Xia, C. Liu, Y. Y. Li, and N. Liu, “A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring,” Expert Systems with Applications, vol. 78, pp. 225–241, 2017, doi: 10.1016/j.eswa.2017.02.017.

[18] N. Chen, B. Ribeiro, and A. Chen, “Financial credit risk assessment : a recent review,” Artificial Intelligence Review, vol. 45, no. 1, pp. 1–23, 2016, doi: 10.1007/s10462-015-9434-x.

[19] X. Dastile, T. Celik, and M. Potsane, “Statistical and machine learning models in credit scoring : A systematic literature survey,” Applied Soft Computing Journal, vol. 91, pp. 106–263, 2020, doi: 10.1016/j.asoc.2020.106263.

[20] Y. Xia, J. Zhao, L. He, Y. Li, and M. Niu, “A novel tree-based dynamic heterogeneous ensemble method for credit scoring,” Expert Systems with Applications, vol. 159, p. 113615, 2020, doi: 10.1016/j.eswa.2020.113615.

[21] T. Le, L. H. Son, M. T. Vo, M. Y. Lee, and S. W. Baik, “A cluster-based boosting algorithm for bankruptcy prediction in a highly imbalanced dataset,” Symmetry, vol. 10, no. 7, pp. 1–12, 2018, doi: 10.3390/sym10070250.

[22] S. Ben Jabeur, C. Gharib, S. Mefteh-wali, and W. Ben Arfi, “Technological Forecasting & Social Change CatBoost model and artificial intelligence techniques for corporate failure prediction,” Technological Forecasting & Social Change, vol. 166, pp. 120–658, 2021, doi: 10.1016/j.techfore.2021.120658.

[23] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, “CatBoost : unbiased boosting with categorical features,” pp. 1–11, 2018.

[24] A. V. Dorogush, V. Ershov, and A. Gulin, “CatBoost : gradient boosting with categorical features support,” 2018. doi: https://doi.org/arXiv:1810.11363.

[25] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, A comparative analysis of gradient boosting algorithms, vol. 54, no. 3. Springer Netherlands, 2021. doi: 10.1007/s10462-020-09896-5.

[26] H. Al Majzoub, I. Elgedawy, Ö. Akaydın, and M. K. Ulukök, “HCAB ‑ SMOTE : A Hybrid Clustered Affinitive Borderline SMOTE Approach for Imbalanced Data Binary Classification,” Arabian Journal for Science and Engineering, vol. 45, no. 4, pp. 3205–3222, 2020, doi: 10.1007/s13369-019-04336-1.

[27] H. Kim, H. Cho, and D. Ryu, “Corporate Default Predictions Using Machine Learning : Literature Review,” sustainability, pp. 1–11, 2020.

[28] G. Kou, Y. Xu, Y. Peng, F. Shen, Y. Chen, K. Chang, and S. Kou, “Bankruptcy prediction for SMEs using transactional data and two-stage multiobjective feature selection,” Decision Support Systems, vol. 140, p. 113429, 2021, doi: https://doi.org/10.1016/j.dss.2020.113429.

[29] V. García, A. I. Marqués, and S. Sánchez, “Exploring the synergetic effects of sample types on the performance of ensembles for credit risk and corporate bankruptcy prediction,” information Fusion, vol. 47, pp. 88–101, 2019, doi: 10.1016/j.inffus.2018.07.004.

[30] C. F. Tsai, “Two-stage hybrid learning techniques for bankruptcy prediction*,” Statistical Analysis and Data Mining, vol. 13, no. 6, pp. 565–572, 2020, doi: 10.1002/sam.11482.

[31] A. Asuncion and D. Newman, “Statlog (Australian Credit Approval) Data Set,” 2007. https://archive.ics.uci.edu/ml/datasets/statlog%0A+(australian+credit+approval) (accessed Sep. 01, 2021).

[32] A. Asuncion and D. Newman, “Statlog (German Credit Data) Data Set,” 2007. https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data) (accessed Sep. 04, 2021).

[33] S. Research, “Japan Stock Market,” 2015. https://siblisresearch.com/data/japan-shiller-pe-cape/ (accessed Sep. 07, 2021).

[34] A. Asuncion and D. Newman, “Polish companies bankruptcy data,” 2007. https://archive.ics.uci.edu/ml/datasets/Polish%0A+companies+bankruptcy+data (accessed Aug. 28, 2021).

[35] A. Asuncion and D. Newman, “Taiwanese Bankruptcy Prediction Data Set,” 2007. https://archive.ics.uci.edu/ml/datasets/Taiwanese+Bankruptcy+Prediction (accessed Sep. 05, 2021).

[36] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE : Synthetic Minority Over-sampling Technique,” Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[37] T. Le and S. W. Baik, “A Robust Framework for Self-Care Problem Identification for Children with Disability,” Symmetry, vol. 11, no. 1, p. 89, 2019, doi: https://doi.org/10.3390/sym11010089.

[38] S. Zeng, Y. Li, W. Yang, and Y. Li, “A Financial Distress Prediction Model Based on Sparse Algorithm and Support Vector Machine,” Mathematical Problems in Engineering, vol. 2020, p. 11, 2020, doi: https://doi.org/10.1155/2020/5625271.

[39] E. C. Blessie and E. Karthikeyan, “Sigmis : A Feature Selection Algorithm Using Correlation Based Method,” Journal of Algorithms & Computational Technology, vol. 6, no. 3, pp. 385–394, 2012, doi: 10.1260/1748-3018.6.3.385.

[40] A. Wosiak and D. Zakrzewska, “Integrating Correlation-Based Feature Selection and Clustering for Improved Cardiovascular Disease Diagnosis,” Complexity, vol. 2018, p. 11, 2018, doi: https://doi.org/10.1155/2018/2520706.

[41] W. C. Lin, Y. H. Lu, and C. F. Tsai, “Feature selection in single and ensemble learning-based bankruptcy prediction models,” Expert Systems, vol. 36, no. 1, pp. 1–8, 2019, doi: 10.1111/exsy.12335.

[42] Y. B. Wah, N. Ibrahim, H. A. Hamid, and S. A. Rahman, “Feature selection methods : Case of filter and wrapper approaches for maximising classification accuracy SCIENCE & TECHNOLOGY Feature Selection Methods : Case of Filter and Wrapper Approaches for Maximising Classification Accuracy,” Pertanika Journal of Science and Technology, vol. 26, no. 1, pp. 329–340, 2018.

[43] M. Luo, Y. Wang, Y. Xie, L. Zhou, J. Qiao, Q. Sun, and Y. Siyu, “Combination of Feature Selection and CatBoost for Prediction : The First Application to the Estimation of Aboveground Biomass,” Forests, vol. 12, no. 2, p. 216, 2021, doi: https://doi.org/10.3390/f12020216.

[44] Y. Cao, X. Liu2, J. Zhai, and S. Hua, “A two-stage Bayesian network model for corporate bankruptcy prediction,” International journal of Finance & Economics, pp. 1–18, 2020, doi: 10.1002/ijfe.2162.

[45] D. Delen, C. Kuzey, and A. Uyar, “Measuring firm performance using financial ratios : A decision tree approach,” Expert Systems With Applications, vol. 40, no. 10, pp. 3970–3983, 2013, doi: 10.1016/j.eswa.2013.01.012.