Research Article

Bankruptcy Prediction using Robust Machine Learning Model

Amer Tabbakh[a], Jitendra Kumar Rout[a], Kshira Sagar Sahoo[b], NZ Jhanjhi[c,*]

,Mudassar Hussain Shah[d], Minakhi Rout[a]
[a]School of Computer Science & Engineering, Kalinga Institute of Industrial Technology
University, Bhubaneswar, Odisha 751024, India
[b]Department of CSE, SRM University, AP, India, 522502
[c]School of Computer Science and Engineering, Taylor's University, Subang Jaya, 47500, Malaysia.
[d]Department of Communication and Media Studies, University of Sargodha, Sargodha

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 28 April 2021

Abstract
The prediction of bankruptcy is the job of forecasting bankruptcy and different financial crisis measures for businesses. It is an enormous area in business and accounting. The significance of the field is partially attributed to its value for creditors and investors in determining the likelihood of business bankruptcy. A predictive model that combines different economic parameters that enable the financial status of a business to be foreseen is the purpose of predicting financial distress. There were various approaches in this area focused on predictive tests, statistical modeling (e.g. generalized linear models), and in addition, artificial intelligence (e.g. Neural Networks, SVM, Decision Trees). In this work, we record our remarks by designing, experimenting, and evaluating some of the classification models used in most cases i.e. Gradient Boosting, Decision Trees, Balanced Bagging, Random Forests, SVM, and Ada Boost which are applicable to expected bankruptcy. The bankruptcy data is collected from Polish firms, in which synthetic features are used to represent statistics of a higher order. The dataset has outliers and is imbalanced. Synthetic Minority Over-sampling Technique (SMOTE) is used to over-sample minority class labels and tackle the data imbalance issue. The feature selection technique is an important step in the preprocessing in which three techniques were applied i.e. PCA, Select Percentile, and Sequential Feature Selection. To evaluate the models, the results are compared using four matrices i.e. accuracy, F1-score, recall, root-mean-square error (RMSE). The simulation studies reveal that the Ada Boost classifier with SFS as a feature selection method is giving the better result of 98.7% in terms of accuracy.

Keywords: Bankruptcy Prediction; Machine Learning; Outlier Values; Imbalanced Data; Ada Boost; Sequential Feature Selection

1. Introduction

In economic decision-making, the anticipation of business collapse is of great importance. An organization's business situation affects the community, market players, and consumers but also impacts policymakers and the financial system. Hence, the high social and economic costs due to corporate bankruptcies, have drawn researchers' attention for a better understanding of the causes of bankruptcy and eventually prediction of company distress [2]. The research on this topic also depends on data availability i.e., whether a firm went bankrupt or not, account rations that might indicate the possibility of bankruptcy, and other potential factors [3]. The history of prediction of bankruptcy includes the application of various statistical tools that have increasingly become available and requires a growing understanding of different pitfalls in early analyses.

The bankruptcy forecast aims to determine the financial condition of a company and its prospects in the sense of a long-term business operation [4]. It is a large area of business and banking that blends expert knowledge about the trend of prosperous and unsuccessful companies with historical data. Usually, business is analysed by various measures that define their market conditions. Further, these are used to trigger a statistical model using recent conclusions [5, 24]. The bankruptcy prediction issue can be classified as a two-class classification problem, companies either go bankrupt at a given period or survive during that period [16,22,23].

Our objective is to figure out an effective scheme for bankruptcy prediction by addressing the following key issues:

- Domain experts suggest the econometric measures representing the state of the business, but how to incorporate these into a successful model is rather unclear.
- The statistical findings used to train the model are typically affected by the imbalanced data effect, as there are generally far more successful companies than bankrupt ones. As a result, a trained model appears to predict that businesses are successful (majority class) even when some of them are troubled companies.

These two problems also concern the model's final predictive capabilities. Speaking about new approaches to the field of bankruptcy prediction, it is interesting to note that survival strategies have been implemented. Choice valuation methods have been developed, involving volatility in stock prices. According to structural models, a default occurrence is considered to develop for a corporation while its resources reach a level that is sufficiently low compared with its obligations. The new methods are used by business intelligence firms transcend the annual report material and take into report recent trends such as age, judgments, negative news, payment events, and creditors' payment knowledge.

The main contributions of the work are as follows:

- Different machine learning models are used for data-set of Polish companies [1]. The key focus is on boosting algorithms(Ada-Boost) which combine many low-precision models to construct a high-precision model.
- Use of appropriate preprocessing techniques to deal with specific issues:
  - Missing value: mean strategy
  - Outliers: omission approach
  - Imbalance nature of dataset: Oversampling Technique (SMOTE)
- Analyzing different feature selection techniques (PCA, select percentile, sequential feature selection) on different classifiers.

The rest of the paper is organized as follows: literature review is presented in Section 2. Section 3 provides details of the Polish company's data-set used in empirical analyzes. The preprocessing dataset can be found in Section 4. Data modeling is mentioned in Section 5. Results were discussed in Section 6. Finally, Section 7 concludes the work.

2. Related Work

The first attempts of the systematic prediction of bankruptcy date back to the early 20th century, while first standard economic measures were introduced to define the predictive abilities of financial disaster [6,7,8]. The year 1960s produced a tipping point in the study of early identification of the signs of financial catastrophe. Second, Beaver's dissertation (1966) [9]. introduced the application of statistical models to the prediction of bankruptcy. Following that line of thought, Altman [10] has suggested a complex modeling technique to foresee company bankruptcy which was subsequently used by others [11,12,13,14]. To achieve understandable models of information representation that were easy to understand the first-order logic decision-making rules were induced using different methods, with only a few naming, such as rough sets [15] or evolutionary programming [2]. Nevertheless, the precision of the description of the decision-making laws is very often inadequate, and therefore more reliable approaches have been used to forecast bankruptcy. Support vector machines (SVM) were one of the most successful models [17]. The drawbacks of SVM are that the kernel feature must be carefully crafted and it is impossible to achieve an intelligible layout. The ensemble classification has shown that the bankruptcy forecast can be successfully applied [18, 19] and that other approaches can be considerably better adopted.

Recently some researchers used deep learning algorithms for implementation on finances. Zelenkov et al. [20] proposed a two-step classification system (TPCM) based on genetic algorithms that enables both the selection of relevant factors and the model itself to be implemented. The first move is to train classifier of different model types after selecting the significant features, At the second stage, the voting group with majority voting rule is made up of the classification trained at the first stage. A genetic algorithm is used for both steps (selection of features and weight determinations of the ensemble). Fischer and Krauss [21] introduced LSTM networks to forecast out-of-sample behavioral trends for the component S &P 500 securities from 1992 to 2015. They presented the results by three stages, First: Pre-similarity returns are evaluated and a transaction cost of five bps per half-turn is found in the effectiveness of the LSTM network against the random forest, the deep neural net, and logistic regression. Second: In top and flop stocks, common patterns are derived. third: On the basis of such returns, thus, a simplified trading strategy is advanced. LSTM networks are an innovative series learning methodology, it is less often used in forecasts of financial time series but is fundamentally appropriate for this field. LSTM networks are found to outperform memory-free classification systems, namely a RAF, a Deep Neural Net and a logistic control classifier (LOG), with average sales of 0.46% and a Sharpe ratio of 5.8 percent before transaction costs.

The most famous describing variables of bankruptcy prediction models are ratio-style financial indicators. Sometimes due to the presence of outliers, these tests display a highly skewed distribution. The literature seems to have an agreement on the need to handle outliers although, at the same time, it is not explicit how severe values are described to optimize model predictive power. The discrimination arising from outliers is minimized by two possible ways: omission and winsorization. Nyitrai and Vir{\'a}g [22] applied in the fields of discriminant processing, logistic regression, (CHAID and CART) and the ANN to the most common classification methodologies. They measured the predictive power of the models as part of the tenfold stratified cross-validation and the region under the ROC curve. They evaluated the influence of winsorization at 1, 3 and 5 percent and the standard deviations at 2 and 3; however, they separated the distribution of each indicator by the CHAID approach and used the ordinal metrics thus obtained instead of the original financial ratios. They found that the latter preprocessing method for data-set is the most successful. The machine learning approaches are [23-25] are used in different applications as well in the same manner.

3. Dataset Used

To address the issue of bankruptcy prediction, the Polish bankruptcy dataset UCI repository have been considered. The dataset deals with Polish firms ' bankruptcy prediction. EMIS is a Polish bankruptcy database. In the span 2000-2012, the troubled firms have been studied, and the businesses currently active from 2007- 2013. The dataset includes a large number of samples from Polish firms evaluated in five separate time frames.

i. 1st Year: The data includes the first year of financial prediction and the related class label reflecting 5 years of bankruptcy status.

ii. 2nd Year: The data includes the second year of financial prediction and the related class label reflecting 4 years of bankruptcy status.

iii. 3rd Year: The data includes the third year of financial prediction and the related class label reflecting 3 years of bankruptcy status.

iv. 4th Year: The data includes the fours year of financial prediction and the related class label reflecting 2 years of bankruptcy status.

v. 5th Year: The data includes the fifth year of financial prediction and the related class label reflecting 1 year of bankruptcy status.

Table 1. Summary of the Polish bankruptcy dataset.

Data	Total Instances	Bankrupt Instances	Non-bankrupt Instances
1st Year	7027	271	6756
2nd Year	10173	400	9773
3rd Year	10503	495	10008
4th Year	9792	515	9227
5th Year	5910	410	5500

Table indicates a cumulative amount of data collection attributes, instances and the number of examples in all 5 databases of each class (bankrupt or non-bankrupt).

4. Data preprocessing

Preprocessing is the most important thing to build a model with high accuracy. Initially, it has been observed that some of the instances are duplicated, so they are handled by simply eliminating them. Subsequently, following steps were followed:

4.1. Missing Data

The data set contain a lot of missing values, which is evident from Fig. 1 for the first year of the data in the dataset. A naive method of handling missing values is to drop in all instances that have NaN values, but it will lead to loss of lots of data. Table 2 displays the number of instances in each data set in the second column, and the third column indicates the number of instances or rows absent for at least one of the functions, column 4 represents the number of cases that will stay in-data set if all rows were removed with missing values, the percentage of data loss in column 5 is seen if all rows of missing data values are actually dropped. Ascertain data sets have data losses

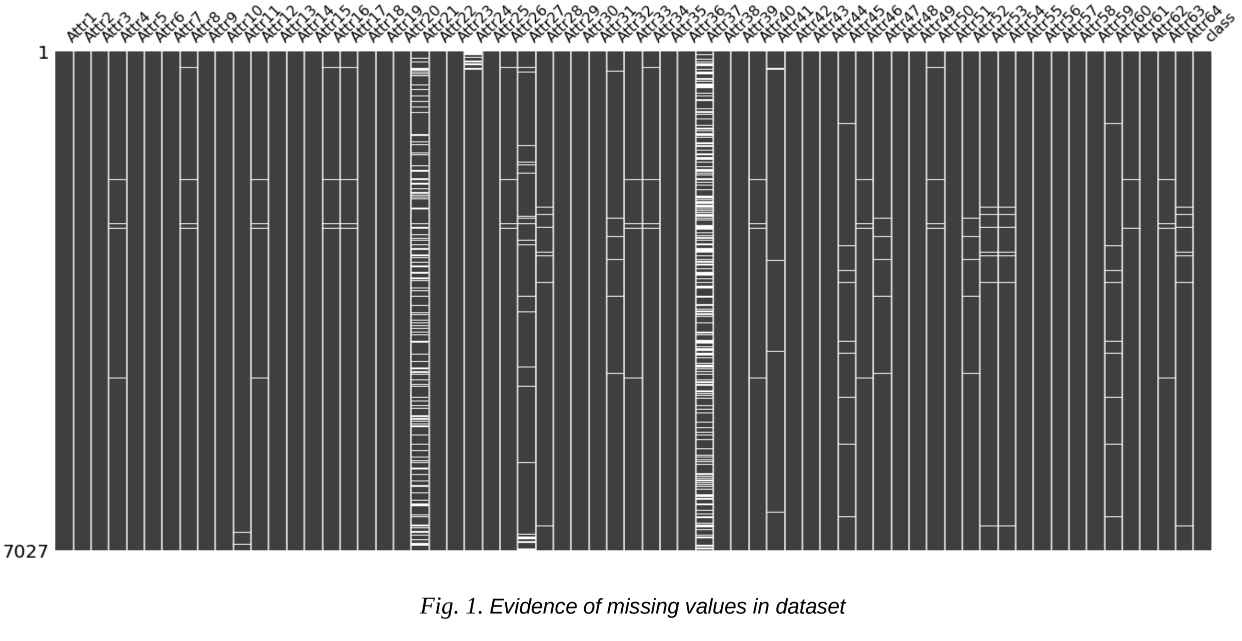

Fig. 1. Evidence of missing values in dataset

of over 50%, it is now obvious that the rows of NaN values cannot be discarded as it contributes to a significant reduction of data representativeness. So in the first step to handle missing values, we dropped some instances which have 50% of features as NaN values. In the next step, mean strategy is used to handle the Nan values. Mean strategy is a method in which every missed value is substituted by the mean value of that attribute.

Table 2. Assessing all databases with the missing information

Data Set	Total # of instances	Instances with missing values	Instances after dropping missing value rows	Data loss if missing values were dropped
Year 1	7027	3833	3194	54.54 %
Year 2	10173	6085	4088	59.81 %
Year 3	10503	5618	4885	53.48 %
Year 4	9792	5023	4769	51.29 %
Year 5	5910	2879	3031	48.71

4.2. Handling Outliers

Since the outliers are data points that are far away from the other data points, outliers do not require identification of data that are usually not distributed. As most statistic assessments presume that data is spread normally, outlining detection will precede data analysis. In the normal distribution, various approaches may be used to classify outliers. Throughout the literature, there is agreement on the need to treat outliers before the prediction of bankruptcy. Two different approaches are used to handling outliers:

i. Omission: Observations of outlier values from the analysis and research samples are excluded.
ii. Winsorization: is the mathematical transformation by reducing extreme data values to decrease the influence of potential spurious outliers.

In this work, the omission approach is used for handling the outliers, where quantile of Q1 equal 0.02 and quantile of Q3 equal 0.98 as a border to drop the outliers.

4.3. Normalization

Normalization is a methodology commonly used in machine learning data planning. The purpose of standardization is to shift the quantitative attribute value in the dataset to a single dimension, without distorting values, without distortion. Each data set does not need standardization for machine learning.

4.4. Imbalance dataset

The imbalance of data generally represents an unfair class representation within a dataset: if two classes are in data-set, then the balance data-set must be 50 percentage points for each class.

Table 3. Data imbalance assessment for each data-sets.

Year	Total # of instances	Before using SMOTE			After using SMOTE		
		Bankrupt instances	Non-Bankrupt instances	Percentage of minority class	Bankrupt instances	Non-Bankrupt instances	Percentage of minority class
Year 1	7027	271	6756	3.85%	5838	5838	50%
Year 2	10173	400	9773	3.93%	8177	8177	50%
Year 3	10503	495	10008	4.71%	8504	8504	50%
Year 4	9792	515	9277	5.25%	7787	7787	50%
Year 5	5910	410	5500	6.93%	4776	4776	50%

Table 3 presents a description of class label populations for all data-sets. When the imbalance nature of dataset, if not handled properly, the input samples of the minority class will not be adequate enough to be applied on the model. As a result, the over-fitting situation may occur. Here in this work, Synthetic Minority Oversampling Technique (SMOTE) is used to handle an imbalanced dataset, it adjusts unequal data classes and constructs equal datasets. If the volume of data is insufficient, the oversampling technique tries to balance the unique sample by growing the scale.

4.5. Features Selection

It is necessary to pick a group of features that provide further prediction knowledge in many fields of study, such as machine modeling, pattern recognition, etc. Reducing the quantity of unnecessary and redundant features significantly reduce the time of a test algorithm and gives a wider definition. Feature selection has various benefits and can help to imagine and interpret results, reduce calculation and storage requirements, reduce training, and usage times, defy the dimensional curse to improve prediction efficiency, etc. This adds to a deeper comprehension of the fundamental definition of a classification. In this work, three feature selection technique i.e. PCA, Select Percentile, and Sequential Feature Selection Techniques were used and compared. It has been observed that the SFS technique outperforms other two.

5. Model Selection and Result Analysis

In this work, six well known classifiers such as Gradient Boosting, Decision Trees, Balanced Bagging, Random Forests, SVM, and Ada Boost were used along with three feature selection techniques i.e. PCA, Select Percentile, and Sequential Feature Selector to classify bankruptcy samples efficiently All experiments have been performed on Windows 10 64-bit OS, Intel(R) Core(TM)-i5-7200U@ 2.50GHz Processor, 8GB RAM. All the simulations have been carried out on Windows 10 (64-bit OS), Intel(R) Core(TM)-i5-7200U@ 2.50GHz Processor, 8GB RAM.

To evaluate the classifiers F1-score, accuracy, recall, RMSE measures were used. The implementation result of different models for each feature selection technique are obtained year wise and presented in the tabular format and plots.

5.1. 1st Year dataset

5.1.1. Using Sequential Feature Selector Technique: Table 4 presents the results of implementation by using the SFS Technique of features section, we observed that the Ada Boost classifier model is the best model for our data-set in the first year with 99.5 % accuracy and Gradient Boosting classifier was the worst one with 85.6 % accuracy.

Table 4. Results of first-year using SFS

Model	F1-score	Accuracy	RMSE	Recall
Decision Tree	0.953	0.953	0.216	0.957
Random Forest	0.991	0.991	0.096	0.994
SVM	0.988	0.988	0.109	0.999
Balanced Bagging	0.992	0.992	0.089	0.994
Gradient Boosting	0.856	0.856	0.379	0.854
Ada Boost	0.995	0.995	0.068	0.998

5.1.2. Using Select Percentile: Table 5 presents the results of implementation by using the Select Percentile of Technique features section, it is observed that that the Ada Boost classifier model is the best model for our data-set in the first year with 98.5 % accuracy and Gradient Boosting classifier was the worst one with 86.2 % accuracy.

Table 5. Results of first-year using Select Percentile

Model	F1-score	Accuracy	RMSE	Recall
Decision Tree	0.951	0.951	0.222	0.959
Random Forest	0.978	0.978	0.148	0.984
SVM	0.955	0.955	0.212	0.966
Balanced Bagging	0.983	0.983	0.13	0.989
Gradient Boosting	0.862	0.862	0.371	0.871
Ada Boost	0.986	0.986	0.119	0.992

5.1.3. Using PCA: Table 6 presents the results of implementation for first year by using the PCA Technique of features section, the Ada Boost classifier model is observed as the best model for our data-set in the first year with 99.1% accuracy and Gradient Boosting classifier was the worst one with 84.7% accuracy.

Table 6. Results of first-year using PCA

Model	F1-score	Accuracy	RMSE	Recall
Decision Tree	0.935	0.935	0.254	0.954
Random Forest	0.985	0.985	0.124	0.985
SVM	0.984	0.984	0.126	0.988
Balanced Bagging	0.989	0.989	0.103	0.991
Gradient Boosting	0.847	0.847	0.391	0.818
Ada Boost	0.991	0.991	0.093	0.992

Fig. 2 represents a comparison of the results of implementation by using three features selection techniques (PCA, Select Percentile, and SFS) and used accuracy as measures to evaluate. It has been observed that for First year data SFS technique outperforms others w.r.t all the models.

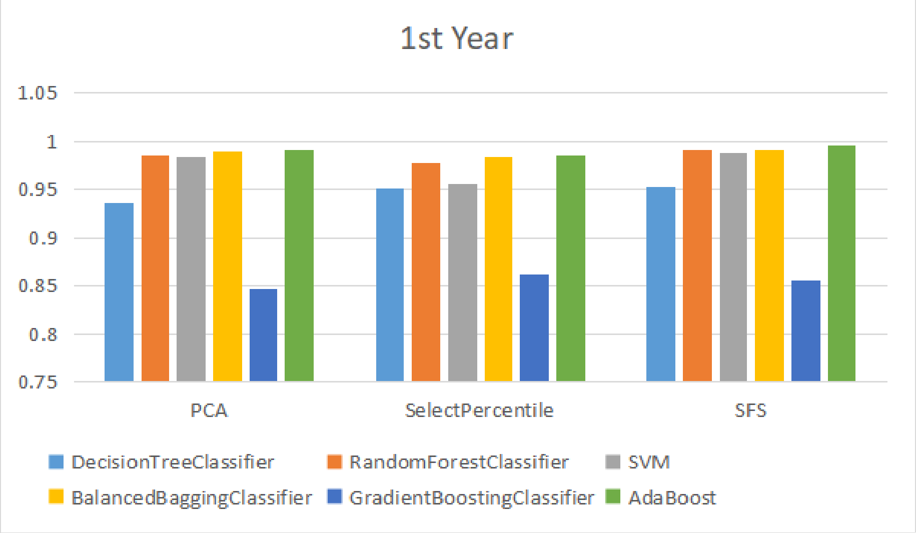

Fig. 2. Year 1 dataset accuracy assessment

5.2. 2nd Year dataset

5.2.1. Using Sequential Feature Selector Technique: Table 7 presents the results of implementation by using the SFS Technique of features section, the Ada Boost classifier model is observed as the best model for our data-set in the second year with 99 % accuracy and Gradient Boosting classifier was the worst one with 81.7 % accuracy.

Table 7. Result of second year using SFS

Model	F1-score	Accuracy	RMSE	Recall
Decision Tree	0.945	0.945	0.235	0.953
Random Forest	0.982	0.982	0.133	0.984
SVM	0.967	0.967	0.182	0.989
Balanced Bagging	0.985	0.985	0.124	0.987
Gradient Boosting	0.817	0.817	0.427	0.807
Ada Boost	0.99	0.99	0.101	0.993

5.2.2. Using Select Percentile: Table 8 presents the results of implementation by using the Select Percentile of Technique features section, it is observed that the Ada Boost classifier model is the best model for our data-set in the second year with 96.6% accuracy and Gradient Boosting classifier was the worst one with 78.2% accuracy.

Table 8. Results of second-year using Select Percentile

Model	F1-score	Accuracy	RMSE	Recall

Decision Tree	0.912	0.912	0.297	0.925
Random Forest	0.959	0.959	0.203	0.964
SVM	0.919	0.92	0.284	0.94
Balanced Bagging	0.964	0.964	0.191	0.968
Gradient Boosting	0.782	0.782	0.467	0.796
Ada Boost	0.966	0.966	0.185	0.973

5.2.3. Using PCA: Table 9 presents the results of implementation by using the PCA Technique of features section, we observed that the Ada Boost classifier model is the best model for our data-set in the second year with 98.3% accuracy and Gradient Boosting classifier was the worst one with 79.1% accuracy.

Table 9. Results of second-year using PCA

Model	F1-score	Accuracy	RMSE	Recall
Decision Tree	0.909	0.909	0.301	0.926
Random Forest	0.974	0.974	0.16	0.979
SVM	0.975	0.975	0.157	0.999
Balanced Bagging	0.983	0.983	0.13	0.988
Gradient Boosting	0.791	0.791	0.457	0.79
Ada Boost	0.983	0.983	0.129	0.99

Fig. 3 represents a comparison of the results of implementation for second year by using three features selection techniques (PCA, Select Percentile, and SFS) and used accuracy as measures to evaluate. It is observed that the SFS technique is the best technique for all models which used in the second year of our data-set.

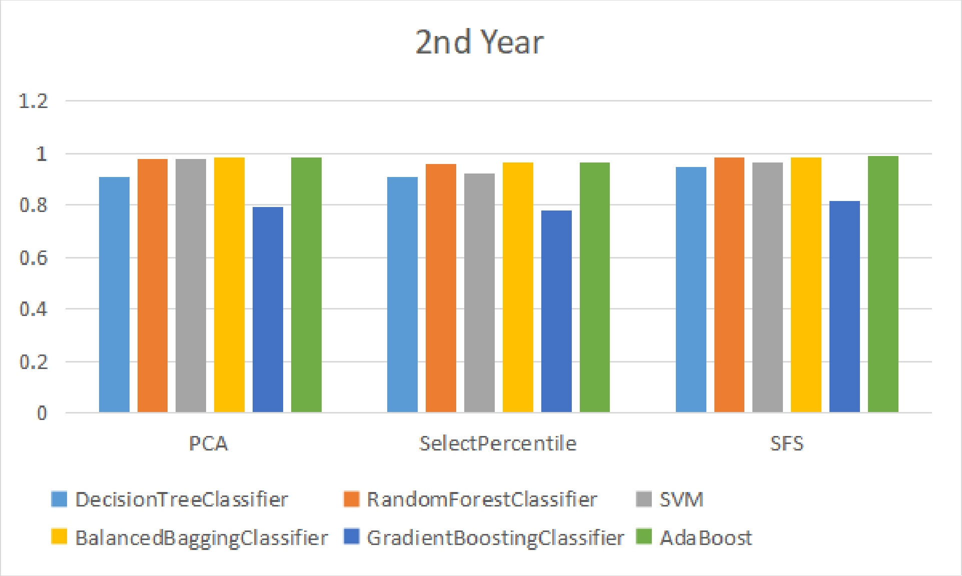

Fig. 3. Year 2 dataset accuracy assessment.

5.3. 3rd Year dataset

5.3.1. Using Sequential Feature Selector Technique: Table 10 presents the results of implementation by using the SFS Technique of features section, the Ada Boost classifier model is observed as the best model for our data-set in the third year with 98.8% accuracy and Gradient Boosting classifier was the worst one with 84.3% accuracy.

Table 10. Results of third-year using SFS.

Model	F1-score	Accuracy	RMSE	Recall
Decision Tree	0.941	0.941	0.244	0.959
Random Forest	0.978	0.978	0.149	0.989

SVM	0.968	0.968	0.178	0.99
Balanced Bagging	0.984	0.984	0.128	0.996
Gradient Boosting	0.843	0.843	0.396	0.853
Ada Boost	0.988	0.988	0.108	0.994

5.3.2. Using Select Percentile: Table 11 presents the results of implementation by using the Select Percentile of Technique features section, it is observed that the Balanced Bagging Classifier model is the best model for our data-set in the third year with 96.6% accuracy and Gradient Boosting classifier was the worst one with 81.7% accuracy.

Table 11. Results of third-year using Select Percentile.

Model	F1-score	Accuracy	RMSE	Recall
Decision Tree	0.916	0.917	0.289	0.938
Random Forest	0.964	0.964	0.191	0.977
SVM	0.934	0.934	0.256	0.97
Balanced Bagging	0.968	0.968	0.179	0.983
Gradient Boosting	0.817	0.817	0.428	0.851
Ada Boost	0.966	0.966	0.184	0.985

5.3.3. Using PCA: Table 12 presents the results of implementation by using the PCA Technique of features section, the Ada Boost classifier model is observed as the best model for our data-set in the third year with 98.1% accuracy and Gradient Boosting classifier was the worst one with 83% accuracy.

Table 12. Results of third-year using PCA.

Model	F1-score	Accuracy	RMSE	Recall
Decision Tree	0.912	0.912	0.297	0.935
Random Forest	0.971	0.971	0.171	0.985
SVM	0.973	0.973	0.163	0.995
Balanced Bagging	0.976	0.976	0.155	0.991
Gradient Boosting	0.83	0.83	0.412	0.825
Ada Boost	0.982	0.982	0.136	0.995

Fig. 4 represents a comparison of the results of implementation for third year by using three features selection techniques (PCA, Select Percentile, and SFS) and used accuracy as measures to evaluate. We observed that the SFS technique is the best technique for all models which used in the third year of our data-set.

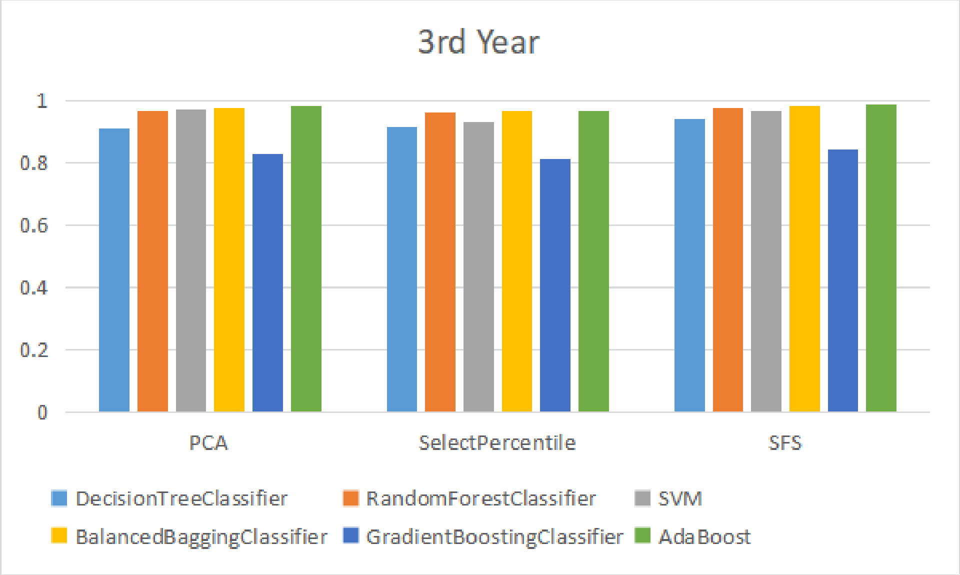

Fig. 4. Year 3 dataset accuracy assessment.

5.4. 4th Year dataset

5.4.1. Using Sequential Feature Selector Technique: Table 13 presents the results of implementation by using the SFS Technique of features section, the Ada Boost classifier model is observed as the best model for our data-set in the fourth year with 98.2% accuracy and Gradient Boosting classifier was the worst one with 81.1% accuracy.

Table 13. Results of fourth-year using SFS.

Model	F1-score	Accuracy	RMSE	Recall
Decision Tree	0.926	0.926	0.272	0.942
Random Forest	0.972	0.972	0.166	0.985
SVM	0.959	0.959	0.202	0.984
Balanced Bagging	0.977	0.977	0.152	0.99
Gradient Boosting	0.811	0.812	0.434	0.842
Ada Boost	0.982	0.982	0.133	0.992

5.4.2. Using Select Percentile: Table 14 presents the results of implementation by using the Select Percentile of Technique features section, we observed that the Ada Boost Classifier model is the best model for our data-set in the fourth year with 96.1% accuracy and Gradient Boosting classifier was the worst one with 79.1% accuracy.

Table 14. Results of fourth-year using Select Percentile.

Model	F1-score	Accuracy	RMSE	Recall
Decision Tree	0.91	0.91	0.299	0.93
Random Forest	0.954	0.954	0.215	0.969
SVM	0.924	0.924	0.276	0.954
Balanced Bagging	0.955	0.955	0.212	0.975
Gradient Boosting	0.791	0.792	0.457	0.807
Ada Boost	0.961	0.961	0.196	0.982

5.4.3. Using PCA: Table 15 presents the results of implementation by using the PCA Technique of features section, it is observed that the Ada Boost classifier model is the best model for our data-set in the fourth year with 97.38% accuracy and Gradient Boosting classifier was the worst one with 78.8% accuracy.

Table 15. Results of fourth-year using PCA.

Model	F1-score	Accuracy	RMSE	Recall
Decision Tree	0.905	0.905	0.308	0.94

Random Forest	0.962	0.962	0.195	0.976
SVM	0.967	0.967	0.18	0.994
Balanced Bagging	0.973	0.973	0.163	0.987
Gradient Boosting	0.789	0.789	0.46	0.777
Ada Boost	0.974	0.974	0.162	0.989

Fig. 5 represents a comparison of the results of implementation for fourth year by using three features selection techniques (PCA, Select Percentile, and SFS) and used accuracy as measures to evaluate. We observed that the SFS technique is the best technique for all models which used in the fourth year of our data-set

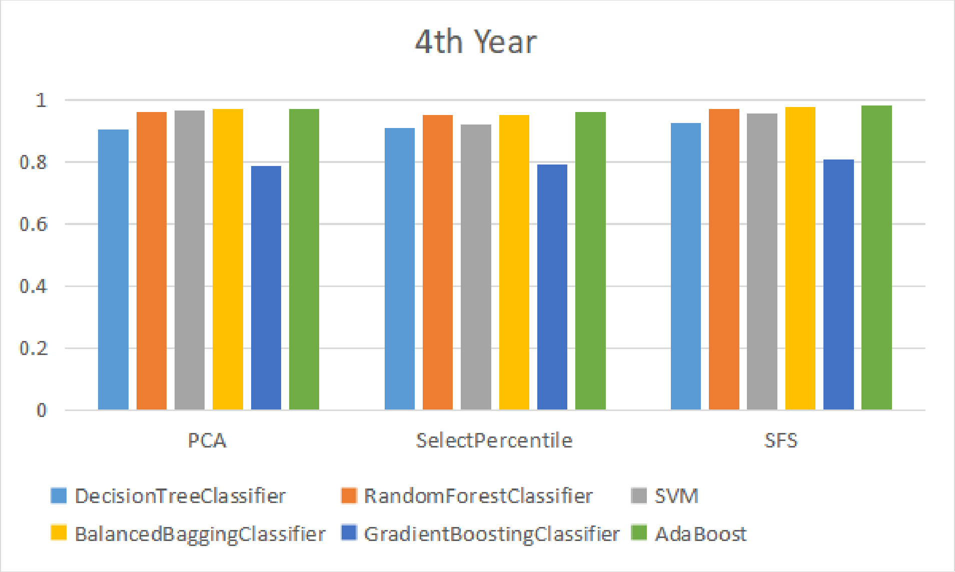

Fig. 5. Year 4 dataset accuracy assessment.

5.5. 5th Year datase

5.5.1. Using Sequential Feature Selector Technique: Table 16 presents the results of implementation by using the SFS Technique of features section, the Ada Boost classifier model is observed as the best model for our data-set in the fifth year with 98.3% accuracy and Gradient Boosting classifier was the worst one with 86.5% accuracy.

Table 16. Results of fifth-year using SFS.

Model	F1-score		Accuracy	RMSE	Recall
Decision Tree	0.932	0.932	0.261	0.941	0.932
Random Forest	0.973	0.973	0.165	0.98	0.973
SVM	0.968	0.968	0.179	0.982	0.968
Balanced Bagging	0.973	0.973	0.165	0.979	0.973
Gradient Boosting	0.865	0.865	0.367	0.856	0.865
Ada Boost	0.983	0.983	0.129	0.989	0.983

5.5.2. Using Select Percentile: The Table 17 presents the results of implementation by using the Select Percentile of Technique features section, we observed that the Ada Boost Classifier model is the best model for our data-set in the fifth year with 97% accuracy and Gradient Boosting classifier was the worst one with 85.1% accuracy.

Table 17. Results of fifth-year using Select Percentile.

Model	F1-score	Accuracy	RMSE	Recall
Decision Tree	0.929	0.929	0.267	0.947
Random Forest	0.959	0.959	0.203	0.959
SVM	0.951	0.952	0.22	0.965

Balanced Bagging	0.96	0.96	0.201	0.961
Gradient Boosting	0.851	0.851	0.386	0.83
Ada Boost	0.97	0.97	0.172	0.973

5.5.3. Using PCA: Table 18 presents the results of implementation by using the PCA Technique of features section, the Ada Boost classifier model is observed as the best model for our data-set in the fifth year with 98.2% accuracy and Gradient Boosting classifier was the worst one with 84.2% accuracy.

Table 18. Results of fifth-year using PCA.

Model	F1-score	Accuracy	RMSE	Recall
Decision Tree	0.903	0.903	0.312	0.918
Random Forest	0.967	0.967	0.181	0.975
SVM	0.974	0.974	0.162	0.994
Balanced Bagging	0.971	0.971	0.171	0.977
Gradient Boosting	0.842	0.842	0.398	0.827
Ada Boost	0.982	0.982	0.133	0.987

Fig. 6 represents a comparison of the results of implementation for fifth year by using three features selection techniques (PCA, Select Percentile, and SFS) and used accuracy as measures to evaluate. It is observed that the SFS technique is the best technique for all models which used in the fifth year of our data-set.

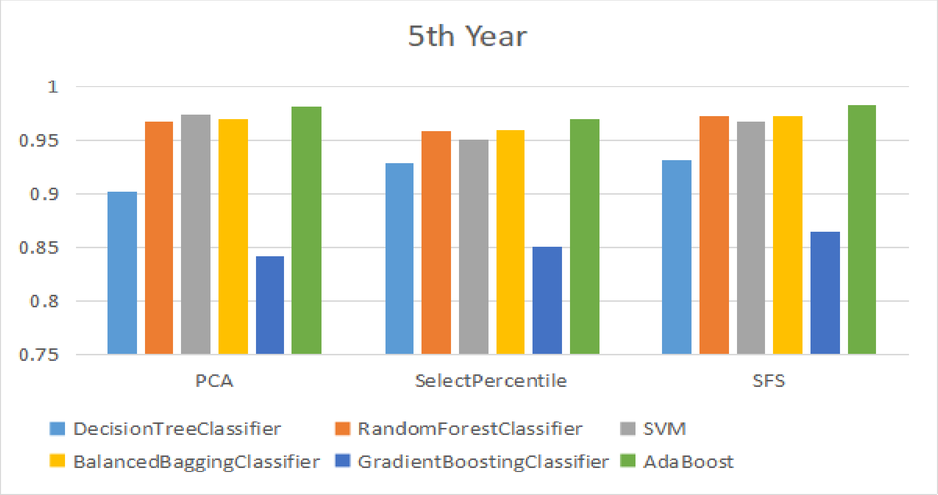

Fig. 6. Year 5 dataset accuracy assessment.

5.6. The mean of all years’ data-set

5.6.1. Using Sequential Feature Selector Technique: Table 19 presents mean of the results of implementation for all years by using the SFS Technique of features section, we observed that the Ada Boost classifier model is the best model with 98.8% accuracy and Gradient Boosting classifier was the worst one with 83.9% accuracy.

Table 19. Results of the mean of all years using SFS.

Model	F1-score	Accuracy	RMSE	Recall
Decision Tree	0.939	0.939	0.246	0.95
Random Forest	0.979	0.979	0.142	0.986
SVM	0.97	0.97	0.17	0.989
Balanced Bagging	0.982	0.982	0.132	0.989
Gradient Boosting	0.839	0.839	0.401	0.842
Ada Boost	0.988	0.988	0.108	0.993

5.6.2. Using Select Percentile: Table 20 presents mean of the results of implementation for all years by using the Select Percentile Technique of features section, the Ada Boost classifier model is observed as the best model with 96.6% accuracy and Gradient Boosting classifier was the worst one with 82.1% accuracy.

Table 20. Results of the mean of all years using Select Percentile.

Model	F1-score	Accuracy	RMSE	Recall
Decision Tree	0.924	0.924	0.275	0.94
Random Forest	0.963	0.963	0.192	0.971
SVM	0.937	0.937	0.25	0.959
Balanced Bagging	0.966	0.966	0.183	0.975
Gradient Boosting	0.821	0.821	0.422	0.831
Ada Boost	0.97	0.97	0.171	0.981

5.6.3. Using PCA: Table 21 presents mean of the results of implementation for all years by using the PCA Technique of features section, it is observed that the Ada Boost classifier model is the best model with 98.2% accuracy and Gradient Boosting classifier was the worst one with 82% accuracy.

Table 21. Results of the mean of all years using PCA.

Model	F1-score	Accuracy	RMSE	Recall
Decision Tree	0.913	0.913	0.294	0.935
Random Forest	0.972	0.972	0.166	0.98
SVM	0.975	0.975	0.158	0.996
Balanced Bagging	0.979	0.979	0.144	0.987
Gradient Boosting	0.82	0.82	0.424	0.807
Ada Boost	0.982	0.982	0.131	0.991

Fig. 7 represents a comparison of the mean of the results of implementation for all five years against using three features selection techniques (PCA, Select Percentile, and SFS) and used accuracy as measures to evaluate. We observed that the SFS technique is the best technique for all models used in all five years of our data-set.

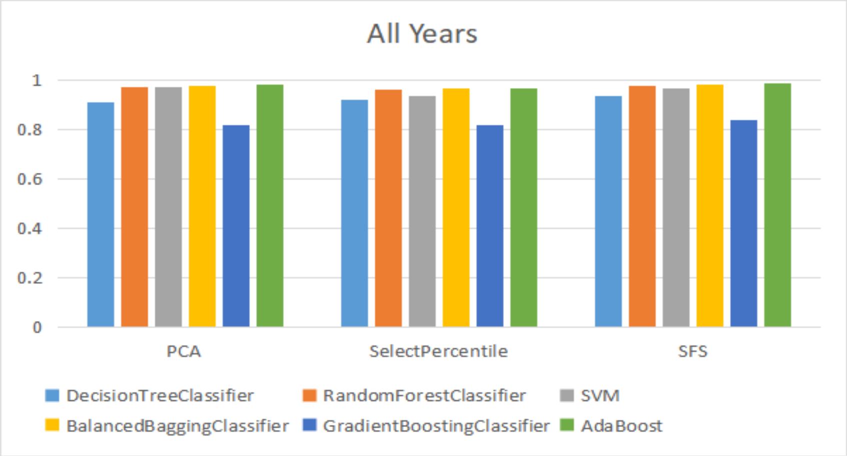

Fig. 7. The mean of all years' dataset accuracy assessment.

6. Conclusion and Future Work

The bankruptcy prediction issue can be classified as a two-class classification problem, companies either go bankrupt at a given period or survive during that period. In this work, our focus is basically on data preprocessing phase of the bankruptcy prediction. The first step was to deal with missing values by mean strategy. Subsequently, to handle outliers omission technique was used which may have a large impact on the accuracy of a model. The third step dealt with the imbalance nature of the dataset using an oversampling technique i.e. SMOTE. Finally, three different feature selection techniques (PCA, SFS, and Select Percentile) have been applied in order to select relevant features. The selected features from each selection technique applied to the specified classifiers individually in order to carry out the bankruptcy prediction. From the result analysis, it is observed that Ada Boost Classifier along with SFS feature selection technique provides better result as compared to other combinations. The work carried out in this paper can be extended using Deep Learning models as well as performance of the model can further be evaluated by applying it on other different available bankruptcy datasets.

References

[1] UCI machine learning, https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data. (Accessed 15 11 2019)

[2] Y. Zhang, S. Wang and G. Ji, A rule-based model for bankruptcy prediction based on an improved genetic ant colony algorithm, Mathematical Problems in Engineering, Hindawi.

[3] Wikipedia: Bankruptcy prediction, https://en.wikipedia.org/wiki/Bankruptcy_prediction. (Accessed 15 12 2019).

[4] R. L. Constand and R. Yazdipour, Firm failure prediction models: a critique and a review of recent developments, Advances in Entrepreneurial Finance, Springer, (2011) 185--204.

[5] E. I. Altman and E. Hotchkiss, Corporate financial distress and bankruptcy: Predict and avoid bankruptcy, analyze and invest in distressed debt, John Wiley & Sons, **289** (2010).

[6] P. Patrick, A comparison of ratios of successful industrial enterprises with those of failed firms, Certified Public Accountant, **2** (1932) 598--605.

[7] A. Winakor and R. Smith, Changes in the financial structure of unsuccessful industrial corporations, Bulletin, **51** (1935) 44.

[8] C. L. Merwin, Financing small corporations in five manufacturing industries, National Bureau of Economic Research, New York, (1942) 1926--1936.

[9] W. H. Beaver, Financial ratios as predictors of failure, Journal of accounting research, JSTOR, (1966) 71--111.

[10] E. I. Altman, Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, The journal of finance, Wiley Online Library, **23(4)** (1968) 589--609.

[11] P. A. Meyer and H. W. Pifer, Prediction of bank failures, The journal of finance, Wiley Online Library, **25(4)** (1970) 853--868.

[12] E. B. Deakin, A discriminant analysis of predictors of business failure, Journal of accounting research, JSTOR, (1972) 167--179.

[13] H. C. Koh and L. N. Killough, The use of multiple discriminant analysis in the assessment of the going-concern status of an audit client, Journal of Business Finance & Accounting, Wiley Online Library, **17(2)** (1990) 179--192.

[14] M. A. Rujoub, D. M. Cook and L. E. Hay, Using cash flow ratios to predict business failures, Journal of Managerial Issues, JSTOR, (1995) 75--90.

[15] A. I. Dimitras, R. Slowinski, R. Susmaga, and C. Zopounidis, Business failure prediction using rough sets, European Journal of operational research, Elsevier, **114(2)** (1999) 263--280.

[16] J. Alzubi, A. Nayyar, and A. Kumar, Machine learning from theory to algorithms: an overview, in Journal of physics: conference series, **1142(1)** (November 2018) 012012.

[17] K. S. Shin, T. S. Lee, and H. J. Kim, An application of support vector machines in bankruptcy prediction model, Expert systems with applications, Elsevier, **28(1)** (2005) 127--135.

[18] L. Nanni, and A. Lumini, An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring, Expert systems with applications, Elsevier, **36(2)** (2009) 3028--3033.

[19] M. Zięba, S. K. Tomczak, and J. M. Tomczak, Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction, Expert systems with applications, Elsevier, **5** (2016) 93--101.

[20] Y. Zelenkov, E. Fedorova, and D. Chekrizov, Two-step classification method based on genetic algorithm for bankruptcy forecasting, Expert Systems with Applications, Elsevier, **88** (2017) 393--401.

[21] T. Fischer and C. Krauss, Deep learning with long short-term memory networks for financial market predictions, European Journal of Operational Research, Elsevier, **270(2)** (2018) 654--669.

[22] T. Nyitrai, and M. Virág, The effects of handling outliers on the performance of bankruptcy prediction models, Socio-Economic Planning Sciences, Elsevier, 67 (2019) 34--42.

[23] Usmani, Raja Sher Afgun, et al. "A spatial feature engineering algorithm for creating air pollution health datasets." International Journal of Cognitive Computing in Engineering 1 (2020): 98-107.

[24] Saeed, Soobia, et al. "Performance Analysis of Machine Learning Algorithm for Healthcare Tools with High Dimension Segmentation." Machine Learning for Healthcare: Handling and Managing

Data (2020): 115.

[25] Kok, S. H., Abdullah, A., Jhanjhi, N. Z., & Supramaniam, M. (2019). A review of intrusion detection system using machine learning approach. International Journal of Engineering Research and Technology, 12(1), 8-15.