International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)
doi : https://doi.org/10.32628/IJSRSET

An Intelligent Agent System for Bankruptcy Analysis and Prediction

N. Saipriya[1], N. Harshitha[1], Sunil Bhutada[2]
[1]B. Tech 4th year Student, [2]Head of the Department
Department of IT, Sreenidhi Institute of Science and Technology Yamnampet, Hyderabad, Telangana, India

Article Info
Volume 9, Issue 3

Page Number : 364-370

Publication Issue :
May-June-2022

Article History
Accepted : 01 June 2022
Published: 10 June 2022

ABSTRACT

The term Bankruptcy can be interpreted as a legal proceeding in which any person or organization is unable to repay the loans. Bankruptcy is one of the crucial problems for both organizations and banks. Throughout the world, academic literature and professional researchers have discussed the possibility of business insolvency. Successful prediction at the initial stage of bankruptcy may help the banks reduce their financial losses and assist them to make correct decisions. We used a bankruptcy data set from Polish companies, where synthetic characteristics were utilized to depict higher-order statistics. This study focuses on the analysis of bankruptcy using different Machine learning algorithms. Among them, Random Forest has shown the highest accuracy. This model helps us to detect whether any person or organization will go bankrupt or not.

Keywords :- Bankrupt Analysis, Logistic Regression, Random Forest, Ensemble Methods, Machine Learning

I. INTRODUCTION

Bankruptcy prediction is the problem of detecting financial distress in business, leading to eventual bankruptcy. The Study of bankruptcy prediction started at least in the 1940s. The previous works on bankruptcy prediction and analysis involved various methodologies such as using univariate statistical approaches. Machine Learning has advanced over the past few decades and the usage of machine learning models has proliferated in all the fields. A Study by Altman had also proven that machine learning models can defeat classical statistical models considering the performance. We initiate by carrying out data pre-processing and exploratory analysis where we impute the missing data values using some of the data imputation techniques like Mean, k-Nearest Neighbours, and Expectation-Maximization. To solve the data imbalance issue, we apply Synthetic Minority Oversampling Technique (SMOTE) to oversample the minority class labels. Later, we model the data using K-Fold Cross Validation on the said models, and the imputed and re-sampled datasets. In the end, we analyze and evaluate the performance of the models on the validation datasets using several metrics such as accuracy, precision, recall, etc., and rank the models accordingly.

II. LITERATURE REVIEW

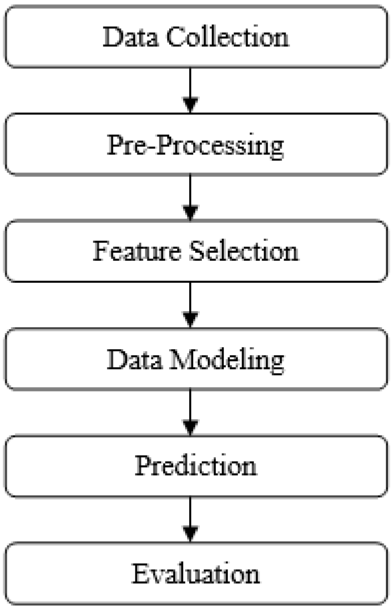
In [1] the author proposed the use of multidimensional analysis to predict corporate bankruptcy which was further developed by others. Later, the boosting method was introduced to develop a Taylor expansion of the loss functions, an approach known as Extreme Gradient Boosting. Because ratios are now widely used as indicators of failure, the goal chosen in [2] was failure prediction. This isn't the only way to use ratios, but it's a good place to start if you want to create an empirical case for ratio analysis. The work in [3] suggested a new method for predicting bankruptcy that uses Extreme Gradient Boosting to learn an ensemble of decision trees along with a novel concept known as synthetic features to reflect higher-order statistics. A synthetic feature is a set of econometric measurements that have been combined using arithmetic procedures. The study in [4] provides a more efficient version of a fruit fly optimization (FOA) technique called LSEOFOA to develop and harmonize the penalty and the kernel parameter in KELM, which is based on kernel extreme learning machine (KELM). With satisfactory performance in bankruptcy prediction, the suggested LSEOFOA-KELM prediction model is regarded as a viable warning tool for financial decision making in the direction of more evolutionary and efficient prediction models. Using various data balancing strategies, the problem of correcting the challenges caused by the imbalance between the two classes is handled. To overcome the problem of uneven data, [5] proposes using random undersampling and the Synthetic Minority Over Sampling Technique. In [6] author discovered that modifying the undersampling rate of the cluster centroid-based technique affects the performance of the Linear Discriminant Analysis (LDA) and Naive Bayes (NB). For the first time, the author applies logic regression analysis to this field [7]. Because of the advancements in economics and computer technology, machine learning models are increasingly being applied in this discipline, in addition to classic statistical models. Machine learning classification models, on the other hand, are based on the assumption of data equilibrium in two classes. Machine learning algorithms can be hampered if there is a disparity in the amount of data available between categories. Author explains this in [8]. [9] is a study that was conducted. There are three types of data imbalance processing methods: data-level, algorithm-level cost-sensitive learning, and hybrid techniques. The algorithmic change improves the original model's suitability for category predictions. The most often used algorithm is allocating misclassification costs to correct class prediction. Deep learning of picture data was also used to forecast bankruptcy in a publication [10].

III. METHODOLOGY

Firstly, we use the Polish bankruptcy dataset and displayed the details of the dataset like features, instances, data organization, etc. The next step is preprocessing the data, where we assert the problems such as missing data values and data imbalance conditions and gave an explanation on how to solve the issues. Then we introduced the classification models and explained how we train our data using these models. Eventually, In order to analyze and evaluate the performance of these models we used certain metrics like accuracy, precision, and recall. Data Imbalance can be reduced with Oversampling and/or Undersampling. Oversampling and Undersampling are opposite and roughly equivalent techniques for dealing with Data Imbalance, where they adjust the class distribution of a data set (i.e. the ratio between the different classes/categories represented). Oversampling increases the minority class label's class distribution, while undersampling decreases the majority class label's class distribution. In our project, we explored Synthetic Minority Oversampling Technique SMOTE.

3.1 ARCHITECTURE DIAGRAM

The prediction model was built using this methods. As shown below, the model defines the various steps of implementing a machine learning project.



3.2 DATA

The dataset we have chosen to assert the bankruptcy prediction problem is the Polish bankruptcy data, provided by the University of California Irvine (UCI) Machine Learning Repository—an open source repository consists of datasets for research and learning purposes to be made for the Machine Learning/Data Science community. This dataset is very much useful in predicting the bankruptcy. There are five datasets and the description of the datasets is as follows

- 1st year: The data contains financial rates from the 1st year of the forecasting period and the corresponding class label that indicates bankruptcy status after 5 years.
- 2nd year: The data contains financial rates from the 2nd year of the forecasting period and the corresponding class label that indicates bankruptcy status after 4 years.
- 3rd year: The data contains financial rates from the 3rd year of the forecasting period and the corresponding class label that indicates bankruptcy status after 3 years.
- 4th year: The data contains financial rates from the 4th year of the forecasting period and the corresponding class label that indicates bankruptcy status after 2 years.
- 5th year: The data contains financial rates from the 5th year of the forecasting period and the corresponding class label that indicates bankruptcy status after 1 year.

In this modeling method, the processing steps are as follows:

a. Phase 1: Data collection
b. Phase 2: Data Analysis & pre-processing
   i. Step 1: Missing Data Analysis
   ii. Step 2: Data Imputation
      A. Mean
      B. K-NN Algorithm
      C. Expectation Maximization (EM) Algorithm
   iii. Step 3: Dealing with imbalanced data using SMOTE
c. Phase 3: Model Building
d. Phase 4: Results and Evaluation

Phase 1: Data Collection:

The data is about the likelihood of a Polish company going bankrupt. The information was gathered from the Emerging Markets Information Service, a database that contains information on emerging markets all over the world. The insolvent companies were studied from 2000 to 2012, while the enterprises that were still running were assessed from 2007 to 2013.

Phase 2: Data Analysis & Preprocessing

In this segment we will be dealing with missing data, and replace using different imputation techniques. The imbalanced data is treated with SMOTE. Data imbalance has a significant impact on modeling since the models will not have enough data from minority classes to train on, resulting in biased models and poor performance on test data.

STEP 1: MISSING DATA ANALYSIS

- First, we look into missing values statistics. For example, the plot of the nullity matrix for the 1st year dataset explains the sparsity of 1st Year data. This was plotted using the library “missingno”. The nullity matrix provides a data-dense display which is helpful for us to clearly distinguish the missing data patterns in the dataset. Observation has proven that features X21 and X37 have the highest number of missing values in the given dataset.
- Summarize the population of class labels in each dataset. consider the population percentage of the minority class, i.e., the Bankruptcy class label of the dataset. These numbers show us that there is a huge data imbalance.
- Missing data leads to three main problems. First, Missing data can introduce a substantial amount of bias. Second, makes the handling and analysis of the data more difficult. Third, reduce efficiency

STEP 2: DATA IMPUTATION

In our project we explored 3 techniques of imputation, and we will see them in the subsequent sections.

I. Mean Imputation

Mean imputation technique to replace any missing value in the data with the mean of that variable in context. The missing value of the feature in the dataset is replaced by the mean of the other non-missing values of that feature. Mean imputation reduces any correlations involving the variable(s) that are imputed. Hence, the univariate analysis uses mean imputation but becomes problematic while dealing with multivariate analysis. Hence we chose Mean Imputation as a standard method. We achieved mean imputation using sci-kit-learn’s Imputer class.

II. k-Nearest Neighbors Imputation

The k-nearest neighbors' algorithm or k-NN, is a method or non-parametric approach used for classification and regression. In both cases, the input is composed of the closest training examples summed up to k in the feature space. k-NN imputation replaces missing values in Data with the corresponding value from the nearest-neighbor row or column depending upon the requirement. The nearest neighbor is computed using Euclidean distance. The next nearest neighbor is used only if the corresponding value from the nearest neighbor is also null. We used the fancyimpute library to do k-NN imputation, and we used 100 as K value for the process.

III. Expectation-Maximization Imputation

EM Imputation is the method of replacing missing values using Expectation-Maximization. It replaces missing values of variables by their expected value calculated using the Expectation-Maximization (EM) algorithm. In practice, a Multivariate Gaussian distribution is assumed. As a whole, EM imputation is far better than mean imputations because they maintain the relationship with other variables. We achieved EM Imputation using the impyute library.

STEP 3: DEALING WITH IMBALANCED DATA

Synthetic Minority Oversampling Technique (SMOTE) is one of the extensively used oversampling techniques. To understand how this technique works consider some training data which has n samples, and f features in the feature space of the data. For simplicity, assume the features are continuous. Let's have a look at a dataset of birds as an example. The minority class’s feature space in which we want to oversample could be anything such as beak length, wingspan, and weight. To perform this, consider a sample from the dataset, and take its k nearest neighbors in the feature space. Take the vector between one of those k neighbors, and the current data point to generate synthetic data point. Multiply this vector by a number x that is between 0 and 1 at random. Add this to the current data point. It results in the new synthetic data point. SMOTE was implemented from the imbalanced learn library.

Phase 3: Model Building

In this segment, we examine various classification models that are considered for training on the Polish bankruptcy datasets to attain the task of approaching with a predictive model that would predict the bankruptcy status of a given (unseen) company with an appreciable accuracy.

Since the Polish bankruptcy dataset does not have a separate 'unlabeled' test dataset, it is obvious that we need to split the training data to obtain a validation dataset (for each year's data). If the split was done in a simple way, we end up with just one validation dataset and the inherent difference in the class label distributions for training and validation datasets would lead to poor performance of the model on the training and hence on validation sets. On the other hand, in K-Fold Cross Validation, K bins are created from the training datasets . In each iteration (total = K iterations), one bin is retained as a validation dataset and the other bins of data are used for training the model. The performance metrics (like accuracy, precision, recall, etc) are noted for each validation set. After all the iterations, each of the bins will have served as validation dataset at least once (depending on K). The metrics are averaged over all the K iterations and the final metrics are output.

Hence, towards the end of the modeling step, we obtain 15 different results (5 models × 3 imputer datasets). In each of the sub-sections that follow, we first explain the model briefly, explain the experiment by specifying the hyperparameters of the model. Later, in the Results section we report the (Cross-Validation-Average) performance of the model on the validation data.

A. NAIVE BAYES CLASSIFIER

A Naive Bayes classifier is a supervised learning method based on Bayes' theorem and the "naive" assumption of independence between every pair of attributes. Given a class variable yy and a dependent feature vector $xx1$ through $xxnn$, Bayes' theorem states the following relationship:

P(y | x1,..., xn) = P(y)P(x1,..., xn)/P($x1, \dots , xn$)

B. LOGISTIC REGRESSION CLASSIFIER

Logistic regression is a linear model for classification. The log-linear classifier is also known as logit regression, maximum-entropy classification (MaxEnt), or logit regression.

In this model, a logistic function is used to predict the probability of the likely outcomes of a single experiment.

We implemented the model with λ = 1 and equal weights are given for all the features, using L1 regularization.

C. DECISION TREES CLASSIFIER

Decision Trees (DTs) are a supervised learning method that is non-parametric and can be used for classification and regression. For our classification task, we create a model that predicts the value of a target variable (y = will a firm go bankrupt?) by learning simple decision rules inferred from the data features ($x1$, $x2$…. $x64$ - all the financial distress variables of a firm). While building a decision tree, the data comes in records in the form:

(x, Y) = ($x1$, $x2$, ………, $x64$, Y)

Our model considers all features and gives equal weights to each of them while looking for the best split during the construction of a decision tree. We have considered the 'Gini' index as a measure of the quality of a split.

D. RANDOM FOREST CLASSIFIER

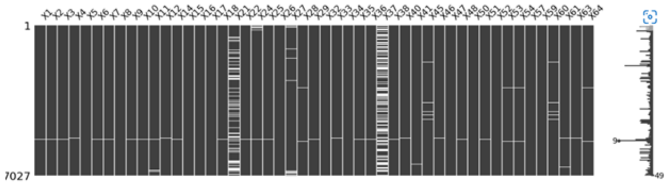
A random forest is a meta estimator that employs averaging to increase predicted accuracy and control over-fitting by fitting numerous decision tree classifiers on various sub-samples of the dataset. In random forests, each tree in the ensemble is built from a sample drawn with replacement from the training set. Furthermore, when dividing a node during tree construction, the chosen split is no longer the optimal split across all features. Instead, the best split among a randomly selected subset of the features is chosen. As a result of this randomness, the bias of the forest usually slightly increases but, due to averaging, its variance also decreases, usually more than compensating for the increase in bias, hence yielding an overall better model. In our model, the number of estimators used are 5 and we have considered ‘Entropy’ as a measure of the quality of a split.

E. EXTREME GRADIENT BOOSTING CLASSIFIER

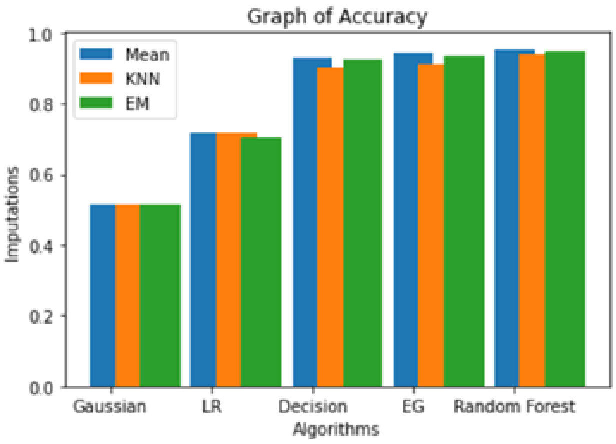
Extreme Gradient Boosting (XGBoost) is built on the principles of the gradient boosting framework. Gradient boosting is a machine learning technique for regression and classification problems that generate a prediction model from an ensemble of weak prediction models, usually decision trees. It constructs the model in the same stage-by-stage manner as other boosting approaches, but it broadens the scope by allowing optimization of any differentiable loss function. To control over-fitting, XGBoost employs a more regularised model formalization, which improves performance. In our model, the number of estimators used is 100. The model internally uses a log-linear classifier for regularizing the model with λ = 1.

IV. EXPERIMENTAL SETUP

The present model is intended to forecast whether or not a company will go bankrupt. The data comes from the University of California, Irvine, and includes features X1,X2,....,X64, as well as a class label Y. The model is built using several libraries, including numpy and pandas, which are basic tools for data management and statistical calculations. We used impyute and fancy impyute libraries for imputation approaches, and smote for smote analysis from the imbalanced learn module. All model classifiers were imported from the sklearn package. Surely, there are gaps in the information gathered. We created a sparsity matrix to detect relationships between missing variables.



We can see a lot of sparsity for the feature in the above plot of sparsity for one year. Among all the features, X37 has the highest sparsity. However, the sparsity matrix failed to reveal any association between the variables. As a result, we used the missingno library's heatmap function to create a heatmap. Following the discovery of the correlation, we performed a smote analysis to ensure that the data was balanced and ready for the next step, data modeling. After examining the data, we created a figure to show the accuracies of utilizing this model with three different imputation strategies.



V. RESULTS

Our results are organized as follow: Firstly, we report the accuracy score of the 6 models we have experimented with, using a plot of the accuracy score against each of the imputation method (Mean, k-NN, and EM), and internally, on each of the 5 datasets (Year 1 – Year 5). Later, we also report the accuracy scores by years’ datasets, i.e., the plot of accuracy scores for each year’s dataset, plotted against the 3 imputation techniques, and internally, the 5 models

Models	Imputation Techniques		
	Mean	K-NN	EM
Guassian Naïve Bayes	51.48	51.64	51.48
Logistic Regression	71.58	71.71	70.21
Descision Tree	92.97	90.42	92.37

Extreme Gradient Boosting	94.26	91.34	93.38
Random Forest	95.39	93.81	94.59

Figure 1: Mean accuracies across all years' datasets for various models and imputation methods

VI. CONCLUSION

In this segment, we go over the highlights of our work on this project thus far. We have successfully modeled 6 classification models: Gaussian Naïve Bayes, Logistic Regression, Decision Trees, Random Forests, and Extreme Gradient Boosting. The training sets were made sure to have balanced sets of class labels, by oversampling the minority class labels using the Synthetic Minority Oversampling technique. Also, we have imputed the missing values in the data using 4 imputer techniques: Mean, k-Nearest Neighbors (k-NN), and Expectation-Maximization (EM) The biggest challenge was dealing with the missing/sparse data. It's tough to collect and organize significant data because all of the companies being considered for bankruptcy don't work on the same schedule. The characteristics that are used to predict bankruptcy are not as simple as the financial statistics seen on a company's balance sheet, and they must be thoroughly investigated and evaluated. In our effort, we successfully documented our findings and proposed the finest bankruptcy prediction model we've seen.

VII. REFERENCES

[1]. Altman, E.L, "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy", The Journal of Finance, Vol. 23(4), pp.589-609, 1968

[2]. Beaver, W, "Financial Ratios as Predictors of Failure, Empirical Research in Accounting : Selected Studied", Journal of Accounting Research, Vol.4(3), pp. 71-111,1966

[3]. Maciej Zieba, "Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction", Expert Systems with Applications, Vol. 58, pp. 93-101, 2016

[4]. Yanan Zhang, "Towards augmented kernel extreme learning models for bankruptcy prediction: Algorithmic behavior and comprehensive analysis", Neurocomputing, Vol. 430, pp. 185-212, 2021

[5]. Talha Mahboob Alam, "Corporate Bankruptcy Prediction: An Approach Towards Better Corporate World", The Computer Journal, Vol. 64(11), pp.1731–1746, 2021

[6]. Wang, Haoming and Xiangdong Liu, "Undersampling bankruptcy prediction: Taiwan bankruptcy data." PLoS ONE, vol. 16, 2021

[7]. Ohlson JA, "Financial Ratios and the Probabilistic Prediction of Bankruptcy", Journal of Accounting Research, vol.18(1), pp.109, 1980

[8]. Kubat M, Matwin S, "Addressing the Curse of Imbalanced Training Sets: One-Sided Selectio", International Conference on Machine Learning, vol. 4, pp.186–197, 1997

[9]. Singh A, Purohit A, "A Survey on Methods for Solving Data Imbalance Problem for Classification", International Journal of Computer Applications, vol. 127(15), pp.37–41, 2015

[10]. Hosaka T, "Bankruptcy Prediction Using Imaged Financial Ratios and Convolutional Neural Networks", Expert Systems with Applications, vol. 117, pp.287–299, 2019

Cite this Article

N. Saipriya, N. Harshitha, Sunil Bhutada, "An Intelligent Agent System for Bankruptcy Analysis and Prediction", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 9 Issue 3, pp. 364-370, May-June 2022.
Journal URL : https://ijsrset.com/IJSRSET2293153