Project Summary

Batch details	PGP-DSE December 2023 Pune
Team members	• Anushree Ladhake • Vaibhav Kadam • Niteen Mane • Gaurav Chavan • Vivek Bhangre
Domain of Project	Finance and Risk Analytics
Proposed project title	Predicting the bankruptcy of polish companies
Group Number	3
Team Leader	Anushree Ladhake
Mentor Name	Vaibhav Kulkarni

Date: 18-05-2024

Signature of the Mentor

Signature of the Team Leader

TABLE OF CONTENTS

OVERVIEW

Bankruptcy prediction is an important problem in finance, since successful predictions would allow stakeholders to take early actions to limit their economic losses. In recent years many studies have explored the application of machine learning models to bankruptcy prediction with financial ratios as predictors. This study extends this research by applying machine learning techniques to a quarterly data set covering financial ratios for a large sample of public Polish firms from 2000–2013.

The dataset is about bankruptcy prediction of Polish companies. The data was collected from Emerging Markets Information Service (EMIS), which is a database containing information on emerging markets around the world. The bankrupt companies were analyzed in the period 2000-2012, while the still operating companies were evaluated from 2007 to 2013.

Bankruptcy prediction is the problem of detecting financial distress in businesses which will lead to eventual bankruptcy. Bankruptcy prediction has been studied since at least 1930s. The early models of bankruptcy prediction employed univariate statistical models over financial ratios. The univariate models were followed by multi-variate statistical models such as the famous Altman Z-score model. The recent advances in the field of Machine learning have led to the adoption of Machine learning algorithms for bankruptcy prediction. Machine Learning methods are increasingly being used for bankruptcy prediction using financial ratios. A study by Barboza, Kimura and Altman found that Machine Learning models can outperform classical statistical models like multiple discriminant analysis (MDA) by a significant margin in bankruptcy prediction (Barboza et al., 2017).

BUSINESS PROBLEM STATEMENT

1. Business Problem Understanding:

   Bankruptcy prediction is an important for modern economies because early warnings of bankrupt help not only the investor but also public policy makers to take proactive steps to minimize the impact of bankruptcies.

2. Business Objective:

   Build a ML model to predict wether the polish company will go bankrupt or not so we can provide early warning to business and investors.

3. Approach:
   a. Defining the objective of the problem statement.
   b. Understanding the dataset.
   c. Exploring the dataset and doing initial analysis.
   d. Performing double classification on the dataset to predict.
   e. Model evaluation and fine tuning to get better results.

4. Conclusions:

   We find that two different Machine Learning algorithms, Random Forest (RF) and Extreme Gradient Boosting (XGBoost) produce accurate predictions of whether a firm will go bankrupt within the next 30, 90, or 180 days, using financial ratios as input features. The XGBoost based models perform exceptionally well, with 99% out-of-sample accuracy. Our training dataset uses a large database of public US firms over a period of 2000-2013. This study has used a substantially larger training dataset as compared to previous studies.

TOPIC SURVEY IN BRIEF

Bankruptcy prediction is an important for modern economies because early warnings of bankrupt help not only the investor but also public policy makers to take proactive steps to minimize the impact of bankruptcies. The reasons that add to the significance of bankruptcy prediction are as follows:

(1). Better allocation of resources

Institutional investors, banks, lenders, retail investors are always looking at information that predicts financial distress in publicly traded firms. Early prediction of bankruptcy helps not only the investors and lenders but also the managers of a firm to take corrective action thereby conserving scare economic resources. Efficient allocation of capital is the cornerstone of growth in modern economies.

(2). Input to policy makers

Accurate prediction of bankruptcies of businesses and individuals before they happen gives law makers and policy makers some additional time to alleviate systemic issues that might be causing the bankruptcies. Indeed, with bankruptcies taking center stage in political discourse of many countries, the accurate prediction of bankruptcy is a key input for politicians, bureaucrats and in general for anyone who is making public policy.

(3). Corrective action for business managers

The early prediction of bankruptcy is likely to highlight business issues thereby giving the company's manager additional time to make decisions that will help avoid bankruptcy. This effect is likely to be more profound in public companies where the management has a fiduciary duty to the shareholders.

(4). Identification of sector wide problems

Bankruptcy prediction models that flag firms belonging to a certain sector are likely to be a leading indicator of an upcoming downturn in a certain sector of an economy. With robust models, the business managers and government policy makers would become aware and take corrective action to limit the magnitude and intensity of the downturn in the specific sector. Industry groups in turn has been shown to significantly effect forecasting models (Chava and Jarrow, 2004).

(5). Signal to Investors

Investors can make better and more informed decisions based on the prediction of bankruptcy models. This not only forces the management of firms to take corrective action but also helps to soften the overall economic fallout that results from the bankruptcies. Empirical studies have shown that investment opportunities are significantly related to likelihood of bankruptcy (Lyandres & Zhdanov, 2007).

(6). Relation to adjacent problems

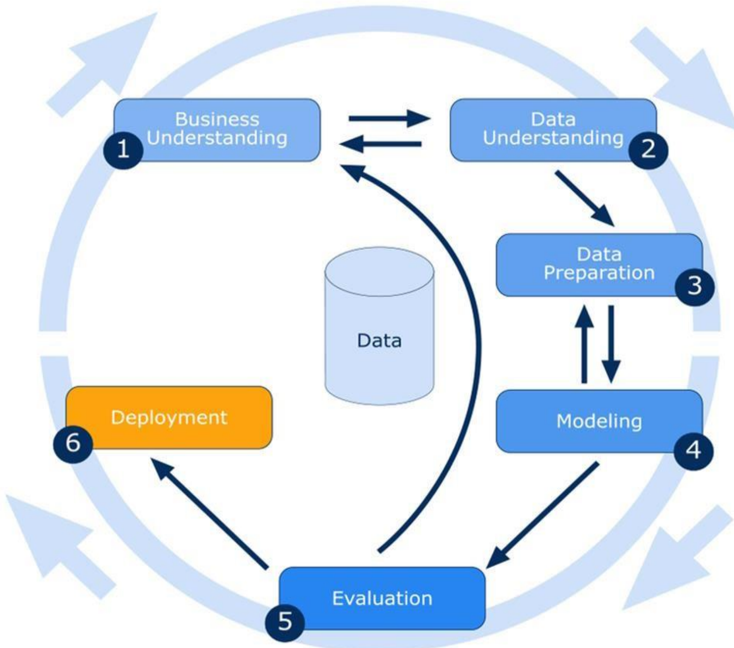
Bankruptcy prediction is often the first step used by ratings agencies to detect financial distress in firms. Based on the predictions of bankruptcy models, ratings agencies investigate and assess credit risk. Getting flagged by bankruptcy prediction models is often the first step that triggers the process of revising credit ratings. A literature survey covering 2000–2013 demonstrates the close relation between bankruptcy prediction and credit risk (García et al., 2015).

CRITICAL ASSESSMENT OF TOPIC SURVEY

Bankruptcy prediction models prior to 1990s were primarily statistical models employing univariate, multivariate and logit & probit techniques. In 1966, Beaver applied univariate analysis in which the predictive ability of 30 financial ratios was tested one at a time to predict bankruptcy (Beaver, 1966). Altman in 1968 performed a multi-variate discriminant analysis (MDA) using 5 ratios to create a linear discriminant function of 5 variables (Altman, 1968). Several variants of MDA were developed in the following years. Edmister used 19 financial ratios to build a linear model for bankruptcy prediction (Edmister, 1972). Deakin found that a linear combination of the 14 ratios could be used to predict bankruptcy five years prior to failure (Deakin, 1972). Ohlson studied the shortcomings of MDA models and built a conditional logit model using maximum likelihood estimation (Ohlson, 1980). The datasets used in all these studies were quite small as compared to modern standards. Ohlson's study for example used a dataset of 2058 firms out of which 105 firms represented the bankrupt class.

The next phase in the evolution of bankruptcy models started in the 1990s with several machine learning algorithms outperforming the older statistical models. Machine learning models such as Random Forests, Support Vector Machines (SVM) and Gradient Boosted Trees were found to be particularly effective for bankruptcy prediction. Barboza, Kimura and Altman compared statistical models with machine learning (ML) models. They found the Random Forests outperformed Alman's Z-score model by a significant margin (Barboza et al., 2017). These results were corroborated by studies (Joshi et al., 2018; Rustam and Saragih, 2018; Gnip and Drotár, 2019). Support Vector Machine (SVM) was also found to be a very effective machine learning algorithm in several studies. Hang et al. (2004) and Chen et al. (2008) achieved superior results for credit rating classification problem by using SVM. Song et al. (2008) used SVM to predict financial distress. Some studies also found boosted trees-based algorithms to outperform SVM. Wang, Ma and Yang proposed a new boosted tree-based algorithm for bankruptcy prediction which they found to be more effective than SVM (Wang et al., 2014). Heo and Yang (2014) used Adaboost algorithm to predict bankruptcy for Korean construction firm. They found Adaboost to have better accuracy than SVM (Heo and Yang, 2014). A more recent study in 2021 has used XGBoost and Random Forest algorithms to predict bankruptcies over 12 months. This study used a medium sized training dataset containing data for 8959 firms registered in Italy (Perboli and Arabnezad, 2021). Another recent study uses a database of Taiwanese firms to predict bankruptcy. This study used data set contain 96 attributes for 6819 firms to train machine learning models (Wang and Liu, 2021). One common attribute shared by all the forementioned studies is the relatively small size of their training data sets. The datasets used by these studies are small as compared to datasets used in the big data era. The largest training dataset in these studies had just 2600 samples which is quite small.

METHODOLOGY TO BE FOLLOWED



Business understanding: In this phase, we defined what problem we are trying to solve. So, we can define clear business problems like:

- Classify the bankrupt companies.

Data understanding: This phase involves understanding the data considered for finding the solution.

- Rudimentary inspection of application data.
- Information about the columns with missing values.
- Statistical summary of numerical variables.
- Check the skewness of each feature.

Data preparation: This phase involves cleaning and processing the data to be in a format suitable for the model used to solve the problem. We can prepare the data as follows:

- Treat the missing values.
- Changing the column names to lower-case for convenience.
- The categorical variables need to be dummy encoded.

- Check for correlation among variables.
- List of columns with at least one values missing

Modelling: This phase involves finding the model that captures the solution to the business problem using available data. We may have to try multiple models and go back and forth between data preparation and modelling to choose the correct model.

Evaluation: Once the model is built, we will check how good the model performs on unseen data. This process is done during the evaluation phase.

Deployment: If we are satisfied with the performance of the model from the previous phase, we deploy it in the deployment phase.

Deployment can be achieved using streamlit and Heroku as follows: -

- Streamlit quickly turns data scripts into shareable web applications
- Once you have developed web application using streamlit

There can be deployed on two types of servers: -

- It can be either deployment on local server
- It can be deployed on web server using Heroku server

REFERENCE

1. https://www.aimspress.com/article/doi/10.3934/DSFE.2021010?viewType=HTML#:~:text=Bankruptcy%20prediction%20is%20the%20problem,statistical%20models%20over%20financial%20ratios.

2. https://archive.ics.uci.edu/dataset/365/polish+companies+bankruptcy+data

3. https://www.kaggle.com/datasets/nitindantu/polish-bankruptcy-data

Original owner of data	The data was taken from a hackathon.
Data set information	Prediction of Genetic disorder
Any past relevant articles using the dataset	No article currently present.
Reference	Kaggle
Link to web page	https://www.kaggle.com/aibuzz/predict-the-genetic-disorders-datasetof-genomes

****