CPS 853 Project Report
Company Bankruptcy Prediction

Andy Liang
500833782
Computer Science
Toronto Metropolitan University (Formerly known as Ryerson University)
Toronto, Ontario

Tusaif Azmat
500660278
Computer Science
Toronto Metropolitan University (Formerly known as Ryerson University)
Toronto, Ontario

Abstract

The objective for this project is to test big data to build machine learning models to classify or identify companies that will file for bankruptcy as per their financial situations. We want to predict if the company will go bankrupt based on the input attributes and identify it as bankruptcy yes or no. To tackle this problem, the application of machine learning techniques is considered to be very useful in helping to predict Bankruptcy. Specifically, we apply seven different machine learning methodologies in this project: Gradient-boosted tree classifier, decision tree, support vector machine, logistic regression, naive Bayes, and random forest. We also further analyze the difference among these methodologies by looking at the logic behind each method and their performance results. We then conclude the paper with the technique that has the highest accuracy rate and lowest false negative rate.

I. Introduction

A. Overview

The possibility of bankruptcy is always the main concern for an enterprise. Not only because it reflects the current financial condition of a company, but also influences the financial players to make decisions for the company in the future. Moreover, for others companies, it can be treated as a sign of whether or not to cooperate with it. Finally, for the governments, they could take measures to prevent financial crisis if too many companies have high possibility of bankruptcy.

The rest of this paper is structured as follows: Section II discusses the problem statement and an overview of the dataset that we use in this project. Section III describes all the methods and techniques that were used in data preparation. Section IV goes over model planning, building, and results of our project. Section V describes our acknowledgments for the project. Finally, a list of references is listed in Section VI.

B. Significance and Challenge

Due to insufficient public Company Bankruptcy databases, limited computer power, and limited time frame to run this experiment, we are only able to run this experiment on a dataset consisting of 23,995 transactions with a total of 65 features. In addition, because of confidentiality issues, most features are not in their original form but have been transformed using principal component analysis (PCA). Based on a realistic scenario, the bankruptcy ratio in this dataset is minuscule - presenting only 0.172% of all transactions.

C. Prior Work

We choose this project from the dataset provided on Kaggle[1] website. There is some work done but we will use this dataset and apply the machine learning algorithms to further enhance the accuracy of our predictions.

II. Problem Statement and Data Discovery

A. Dataset

The chosen data set contains bankruptcy information about Polish companies, which can be obtained from Kaggle.com. The dataset is about bankruptcy prediction of Polish companies. The data was collected from Emerging Markets Information Service, which is 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

Note that the dataset is highly imbalanced towards not bankruptcy which has the label of "0". Following is the result from running "*df2.groupBy ('class').count ().to Pandas ()"*:

	class	count
0	0.0	23995
1	1.0	1126

B. Problem Statement

Generally, deciding whether the company would go bankrupt or not in a short of time for the given data set is what we want to solve for this project. According to Wagenmans[2], financial environments in which the company operates and decision makers of a company are the key factors for causing the bankruptcy. In the economy domain, there are many factors influencing the possibility of bankruptcy such as ROA (a financial ratio that shows the percentage of profit a

company earns in relation to its overall resources), COGS (the total revenue minus the direct costs of producing that good or service). This information is all given to us in the dataset and make up the features that we will train our model on to get our results.

There is an increasingly interests for researchers to apply different machine learning models for predicting the possibility of bankruptcy due to its importance to financial area. Sonam Guptahas[3] uses bagging and Adaboost classifier to train the model and the precision rate is up to 91.64% According to its paper "Systematic review of bankruptcy prediction models: towards a framework for tool selection" published in 2017.

We use all the techniques which we had learned during the course of machine learning. We had applied the supervised learning techniques for the classification algorithms. Following are the algorithms we use to predict the class results:

1. Decision Tree.
2. Random Forest.
3. Gradient-boosted tree classifier.
4. Logistic Regression.

We use the above machine learning algorithms to predict using both predictors and target values.

III. Methods and Data Prep

We will start our project with data collection and for that we will use the dataset provided to us through Kaggle[1] dataset. After downloading the dataset we will be performing some functions such as preprocessing, analysis and evaluation based on our experiments. We will be performing the following methods:

A. Methods

1). Imbalance Learning:

Standard decision trees use information gain as the splitting criterion for learning which results in rules biased towards the majority. Research also shows that imbalanced datasets pose a problem for Gradient-boosted tree classifier, and support vector machines (SVM). This problem is most pronounced when the two classes overlap as in the case of the Kaggle dataset; the majority of machine intelligence algorithms are not suited to handle both unbalanced and overlapped class distributions.
Fortunately, particular algorithms exist that can take class imbalance into account. In addition there are techniques at the data level and algorithm level which can reduce the negative effects of these biases.

2). Concept Drift:

This project was carried out in various steps such as Exploratory Data Analysis, Data Pre-processing, Model Fitting, and finally, selecting the best model. Data Pre-processing consisted of splitting the data-set into train, test and validation sets, taking care of missing values, trying out several imputation techniques, taking care of data imbalance, and taking care of correlation among the features. The next step consisted of fitting various supervised machine learning classification models on the processed training data, and using the validation set for k-fold cross-validation for hyper-parameter tuning. Finally, we compared the various models based on their performance on the part of the data set we had kept aside for testing. All these methods are described in detail below.

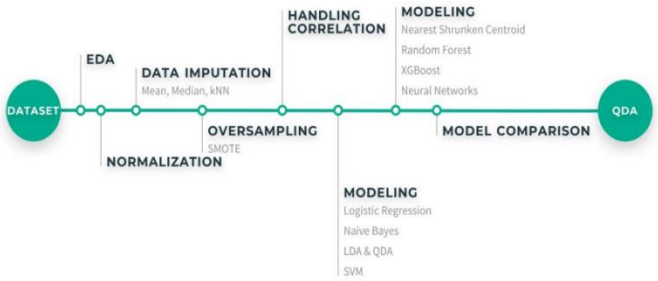

Fig. 1. Timeline

3). Sampling:

Sampling methods are used to compensate for the un-balanceness of the dataset by reducing the classes to near equivalence in size. We use the oversampling technique which uses a bias to achieve this purpose. Complex algorithms such as synthetic minority oversampling technique (SMOTE) actually create new data points based on known samples and their features instead of simply replicating the minority class [2]. However this algorithm relies on assumption of the minority class and is generally computationally expensive. Particularly, the created data generally is an interpolation of prior data which may not actually provide a realistic approximation of if the classes were in fact, balanced. Despite this, sampling methods can provide a more robust approach to imbalance learning than other methods, for example cost-based techniques which penalize errors differently depending on class such that the minority class is favored [2]. For example, what is the cost to use?

Finally, the results will be compared to cost-based balancing methods, if applicable.

4). Data Interpolation Techniques

As mentioned above, one of the main challenges in this dataset is missing data for certain rows. This is especially problematic as it also is not consistent in which columns the data is missing. One row could be missing data in C53 whereas another row could be missing data in C2 - C20, adding to the complexity of the issue. As we are using all of the columns available, missing data has a very large effect on our model and can completely the prediction based on how we interpolate the missing data values. We have taken the following approaches to this problem:

i) Mean Interpolation

A basic interpolation method that uses the mean average value of each column to replace any missing data values.

ii) Median Interpolation

This is the same interpolation method as above, except that we are using the median value of each column instead of the mean value to approximate the missing data values.

Both interpolation techniques make use of the imputer function for calculation. These methods provide a very approximate estimation of a missing value but are may not be representative of the actual value at all. As both these methods are based on other existing data in the dataset, it will be heavily influenced by sampling bias. As this dataset is 95% skewed towards being non-bankrupt, the mean and median values are also heavily biased towards giving a non-bankrupt result. This means that no matter how bankrupt-skewed the existing numbers are, the missing values will skew it more towards non-bankrupt, even completely changing the prediction in some cases.

One way we could mitigate this is by changing the training set to have an equal amount of bankrupt and non-bankrupt companies. This way, the interpolated averages would be more non-influential in the overall prediction, putting more weight onto the other numbers, which would give us a fairer result.

Another way to potentially mitigate this is to use percentiles instead of means and medians. By using 95^{th} percentile or 5^{th} percentile values depending on whether the column has higher values as more bankrupt or visa versa, we would be able to make the model fairer when replacing missing values instead of heavily skewing towards non-bankrupt values. Unfortunately, we were unable to get a working version of our project with this interpolation technique as our implementation kept throwing an error we were unable to fix. Nonetheless, we were still able to achieve very good results with our other techniques.

IV. Model Planning, Building, and Review

A. Ideologies

1). Model Training:

In this project, we implemented our classification algorithms both by using scratch coding and also a Python machine learning library called PySpark. PySpark is considered to be efficient, useful, robust, and consistent on DataBricks.

2). Dataset Split Strategy

We will use the Holdout method where 80% sample size will be used for training and 20% will be tested on models.

3). Feature Extraction

We will use default parameters for the learning curve of the model and the feature extraction methods to find the best fit for our model.

In a dataset the features C1, C2, to C64 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. Feature 'Class' is the target variable with value 1 in case of Bankruptcy and 0 otherwise.

B. Models (Classifications)

1). Model Evaluation and Results:

Strategy	Score
All columns, Logistic Regression, mean interpolation	50.00%
All columns, Decision Tree, mean interpolation	35.96%
All columns, random forest, drop all rows with any missing values	90.64%
First 3 columns, random forest, replace missing values with 0	70.15%
All columns, random forest, replace missing values with 0	87.27%
All columns, random forest, mean interpolation	90.31%
All columns, random forest, median interpolation	89.01%
All columns, random forest, replace missing values with percentiles representing even sampling	N/A
Top 10 Important columns, random forest, mean interpolation	80.58%
Top 10 Important columns, random forest, median interpolation	80.57%
All columns, Gradient Boosted Tree, Mean interpolation	90.31%
All columns, Gradient Boosted Tree, Median interpolation	89.01%

We tried using logistic regression and decision tree models but as it is shown, they were very ineffective in predicting so we stopped using them. Additionally, we also experimented with limited the feature-set to a limited number of columns as according to what the random forest deemed as important using the full data-set, but this also gave us less accurate predictions.

Overall, the accuracy rates for the models are extremely high. However, looking at this metric alone does not indicate that the predicted results are perfect. We should also inspect the metrics to have a better view and deeper understanding of how these results are truly interpreted by each model.

With the Holdout method, it seems that Gradient-boosted tree classifier and Random Forest are the best two models for bankruptcy detection on the test dataset in this case. Logistic Regression, on the other hand, does not predict bankruptcy transactions well because the precision value was quite low.

C. *Conclusions*

By comparing performance metrics, we can see Random Forest and Gradient Boosted Tree models have the highest accuracy with 90.31% each, with mean interpolation having marginally better performance than median value interpolation.

The contest dataset also agreed with our evaluation as we scored a whopping 0.97025 which is good enough to land us at number 10 on the leaderboards at the time of this report, proving to us that our model is incredibly effective at identifying companies on the verge of bankruptcy.

V. Acknowledgments

We would like to take this opportunity to thank Dr. Manar Alalfi as well as Mr. Chang Nian Chuy of Toronto Metropolitan University (formerly known as Ryerson University). It is only thanks to their help and guidance that we were able to learn the necessary skills to undertake and succeed in this contest.

VI. References

[1] “Company Bankruptcy Prediction.” https://www.kaggle.com/competitions/company-bankruptcy-prediction/data.

[2] Frank Wagenmans. Machine learning in bankruptcy prediction. Master’s thesis, 2017.

[3] Hafiz A Alaka, Lukumon O Oyedele, Hakeem A Owolabi, Vikas Kumar, Saheed O Ajayi, Olugbenga O Akinade, and Muhammad Bilal. Systematic review of bankruptcy prediction models: Towards a framework for tool selection. Expert Systems with Applications, 94:164–184, 2018.

[4] “Scikit Learn Tutorial.” https://www.tutorialspoint.com/scikit_learn/index.htm

[5] “Global Payment Fraud Statistics, Trends & Forecasts.” https://www.merchantsavvy.co.uk/payment-fraud-statistics/