International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 08 | Aug 2020 www.irjet.net p-ISSN: 2395-0072

Machine Learning based Analysis of Industry Finances Subjected to Bankruptcy

Parth Trada[1], Himanshu Gajera[2]

[1]*Student, Dharmsinh Desai University (DDU), Nadiad, India*
[2]*Himanshu Gajera, Babaria Institute of Technology, (BITS edu campus), Vadodara, India*

Abstract - *Bankruptcy prediction is the art of predicting bankruptcy and various measures of financial distress of public firms.The rationale for developing and predicting the financial distress of a company is to develop a predictive model used to forecast the financial condition of a company by combining several econometric variables of interest to the researcher. It is a vast area of finance and accounting research. The importance of the area is due in part to the relevance for creditors and investors in evaluating the likelihood that a firm may go bankrupt. The quantity of research is also a function of the availability of data and for that matter here the public data of five polish companies up to five years of financial records have been used to train the model.*

Key Words: (Machine Learning, Industry Finance, Data Science, Statistics, Bankruptcy

1. INTRODUCTION

The most widely accepted theory on the origin of the word “BANKRUPTCY” comes from a mixing of the ancient latin words “BANCUS" (Bench or Table) and “RUPTUS" (Broken). Bankruptcy is a state of insolvency wherein the company or the person is not able to repay the creditors the debt amount. Bankruptcy is a legal process through which people or other entities who cannot repay debts to creditors may seek relief from some or all of their debts. In most jurisdictions, bankruptcy is imposed by a court order, often initiated by the debtor. Bankruptcy fraud is a white-collar crime. While difficult to generalize across jurisdictions, common criminal acts under bankruptcy statutes typically involve concealment of assets, concealment or destruction of documents, conflicts of interest, fraudulent claims, false statements or declarations, and fee fixing or redistribution arrangements. Falsifications on bankruptcy forms often constitute perjury. Multiple filings are not in and of themselves criminal, but they may violate provisions of bankruptcy law. Bankruptcy fraud is a federal crime in the United States. It is necessary that we develop methods to identify firms that might run a risk of going bankrupt and more so in an environment such as the current one which is of recession.

2. PREFACE

There is a sharp rise in personal bankruptcy filings between 1994 and 1998, a period of economic expansion. Bankruptcy is subjected to different scenarios like the failure process of unsuccessful startups, the failure process of ambitious growth companies, the failure process of dazzled growth companies, and the failure process of a pathetic established company. Market conditions, Financing, poor decision making and other factors like poor business location, loss of key employees, lawsuits raised by competitors and personal issues like illness or divorce. Unforeseen disasters and criminal activity like floods, storms, fires, theft and fraud can also cause hardships that lead to bankruptcy. It is seen that the term "bankruptcy" is associated with several humanly activities and natural disasters. However, to ensemble this poser by machine learning approach provides deeper level insights of the crux. Keeping the natural causes aside, use of a machine learning approach could help prejudge the causes that reside at the core level within companies and organisations and can help build the appropriate rectification strategies.

3. PROPOSED MODEL

In this section, we explain our step-by-step solution of how we achieved benchmark results for bankruptcy prediction. Firstly, we introduce the Polish bankruptcy dataset and explain the details of the dataset like features, instances, data organization, etc. Next, we excavate into data preprocessing steps, where we state the problems present with the data like missing data and data imbalance, and explain how we dealt with them. Next, we introduce the classification models we have considered and explain how we train our data using seven machine learning models. Later, we analyze and evaluate the performance of these models using certain metrics like accuracy, precision, recall and F1 score and AUC_ROC. After the training we observed different results for mean imputation and kNN imputation and concluded which imputation is best for this dataset.

3.1. Data

The dataset we have considered for addressing the bankruptcy prediction problem is the Polish bankruptcy data, hosted by the University of California Irvine (UCI)

Machine Learning Repository—a huge repository of freely accessible datasets for research and learning purposes intended for the Machine Learning/Data Science community. The dataset is about bankruptcy predictions 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. The dataset is very apt for our research about bankruptcy prediction because it has highly useful econometric indicators as attributes (features) and comes with a huge number of samples of Polish companies that were analyzed in 5 different timeframes.

	Data	Total Instances	Bankrupt Instances	Non bankrupt Instances
Number of 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 -1: Summary of the Polish bankruptcy dataset.

Table 1 shows the total number of features and instances in the dataset, and the number of samples in each class (bankrupt or not-bankrupt) of all the 5 datasets. The features are explained in Table 2. As shown in the table, there are 64 features labelled X1 through X64, and each feature is a synthetic feature. A synthetic feature is a combination of the econometric measures using arithmetic operations (addition, subtraction, multiplication, division). Each synthetic feature is a single regression model that is developed in an evolutionary manner. The purpose of the synthetic features is to combine the econometric indicators proposed by the domain experts into complex features. The synthetic features can be seen as hidden features extracted by the neural networks but the fashion they are extracted is different.

3.2. Dealing with Missing Data

Missing data causes 3 problems:

1. Missing data can introduce a substantial amount of bias.
2. Makes the handling and analysis of the data more difficult.
3. Create reductions in efficiency.

Dropping all the rows with missing values or Listwise deletion, introduces bias and affects representativeness of the results. The only viable alternative to Listwise deletion of missing data is Imputation. Imputation is the process of replacing missing data with substituted values and it preserves all the cases by replacing missing data with an estimated value, based on other available information. In our project we explored 2 techniques of imputation, and we will see them in the subsequent sections.

1. Mean Imputation
2. k-Nearest Neighbors Imputation

3.2.1. Mean Imputation

Mean imputation technique is the process of replacing any missing value in the data with the mean of that variable in context. In our dataset, we replaced a missing value of a feature, with the mean of the other non-missing values of that feature. Mean imputation attenuates any correlations involving the variable(s) that are imputed. This is because, in cases with imputation, there is guaranteed to be no relationship between the imputed variable and any other measured variables. Thus, mean imputation has some attractive properties for univariate analysis but becomes problematic for multivariate analysis. Hence we opted Mean Imputation as a baseline method. We achieved mean imputation using scikit-learn's **Imputer** class.

3.2.2. k-Nearest Neighbors Imputation

The k-nearest neighbors algorithm or k-NN, is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. It can also be used as a data imputation technique k-NN imputation replaces NaNs in Data with the corresponding value from the nearest-neighbor row or column depending upon the requirement. The nearest-neighbor row or column is the closest row or column by Euclidean distance. If the corresponding value from the nearest-neighbor is also NaN, the next nearest neighbor is used. We used the **fancyimpute** library to perform k-NN data imputation, and we used 100 nearest neighbors for the process.

Here we proposed a detailed description of all the features in the list below as X1 to X64.

ID	Description	ID	Description
X1	net profit / total assets	X33	operating expenses / short-term liabilities
X2	total liabilities / total assets	X34	operating expenses / total liabilities
X3	working capital / total assets	X35	profit on sales / total assets
X4	current assets / short-term liabilities	X36	total sales / total assets
X5	[(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation)] * 365	X37	(current assets - inventories) / long-term liabilities
X6	retained earnings / total assets	X38	constant capital / total assets
X7	EBIT / total assets	X39	profit on sales / sales
X8	book value of equity / total liabilities	X40	(current assets - inventory - receivables) / short-term liabilities
X9	sales / total assets	X41	total liabilities / ((profit on operating activities + depreciation) * (12/365))
X10	equity / total assets	X42	profit on operating activities / sales
X11	(gross profit + extraordinary items + financial expenses) / total assets	X43	rotation receivables + inventory turnover in days
X12	gross profit / short-term liabilities	X44	(receivables * 365) / sales
X13	(gross profit + depreciation) / sales	X45	net profit / inventory
X14	(gross profit + interest) / total assets	X46	(current assets - inventory) / short-term liabilities
X15	(total liabilities * 365) / (gross profit + depreciation)	X47	(inventory * 365) / cost of products sold
X16	(gross profit + depreciation) / total liabilities	X48	EBITDA (profit on operating activities - depreciation) / total assets
X17	total assets / total liabilities	X49	EBITDA (profit on operating activities - depreciation) / sales
X18	gross profit / total assets	X50	current assets / total liabilities
X19	gross profit / sales	X51	short-term liabilities / total assets
X20	(inventory * 365) / sales	X52	(short-term liabilities * 365) / cost of products sold)
X21	sales (n) / sales (n-1)	X53	equity / fixed assets
X22	profit on operating activities / total assets	X54	constant capital / fixed assets
X23	net profit / sales	X55	working capital
X24	gross profit (in 3 years) / total assets	X56	(sales - cost of products sold) / sales
X25	(equity - share capital) / total assets	X57	(current assets - inventory - short-term liabilities) / (sales - gross profit - depreciation)
X26	(net profit + depreciation) / total liabilities	X58	total costs /total sales
X27	profit on operating activities / financial expenses	X59	long-term liabilities / equity
X28	working capital / fixed assets	X60	sales / inventory
X29	logarithm of total assets	X61	sales / receivables
X30	(total liabilities - cash) / sales	X62	(short-term liabilities *365) / sales
X31	(gross profit + interest) / sales	X63	sales / short-term liabilities
X32	(current liabilities * 365) / cost of products sold	X64	sales / fixed assets

3.3 Data Modeling

In this section, we will look at the various classification models that we have considered for training on the Polish bankruptcy datasets to achieve the task of coming up with a predictive model that would predict the bankruptcy status of a given (unseen) company with an appreciable accuracy. We have considered the following 8 models:

1. Logistic Regression
2. Decision Tree
3. Support Vector Machine
4. Linear Discriminant Analysis
5. Quadratic Discriminant Analysis
6. Random Forest
7. K-Nearest Neighbors
8. Naive Bayes

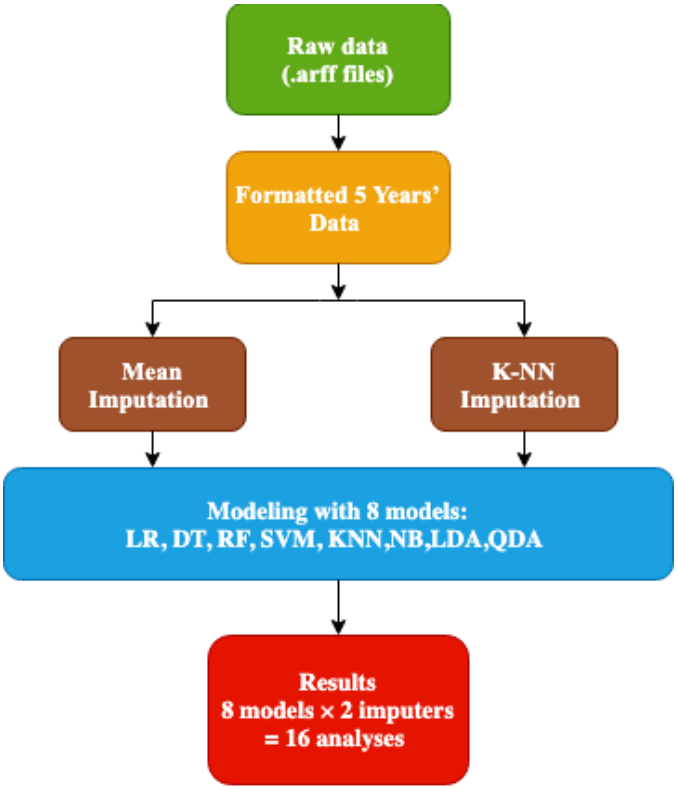

Fig -1: Pipeline for data modeling

Figure 1 shows the pipeline of data modeling for our project. After having obtained the formatted datasets from the raw data (.arff files), we have imputed the missing values via 2 different independent imputer methods (mean, k-NN). Later, We model each of these 2 datasets with the 8 models listed above. While modeling, we use the K-Fold Cross Validation technique for validation.

4. Code

The programming environment used for the project is Python v3.6. Our code workflow exactly mimics the data modeling pipeline shown in Figure 1.We used the libraries listed in Table 3 to run our experiments and achieve our results.

Library	Description
numpy	Data organization and statistical operations
pandas	Data manipulation and analysis. Storing and manipulating numerical tables.
matplotlib	Plotting library
scipy.io	Loading .arff raw data
fancyimpute	Perform k-NN and imputation
sklearn.preprocessing.Imputer	Perform Mean imputation
sklearn.ensemble.RandomForestClassifier	Random Forest Classifier
sklearn.discriminant_analysis.LinearDiscriminantAnalysis	Linear Discriminant Analysis
sklearn.discriminant_analysis.QuadraticDiscriminantAnalysis	Quadratic Discriminant Analysis
sklearn.neighbors.KNeighborsClassifier	k-NN Classifier
sklearn.svm.SVC	SVM Classifier
sklearn.linear_model.LogisticRegression	Logistic Regression Classifier
sklearn.naive bayes.GaussianNB	Gaussian Naive Bayes Classifier
sklearn.tree.DecisionTreeClassifier	Decision Tree Classifier
sklearn.metrics	Performance evaluation metrics like accuracy score, recall, precision, ROC curve, etc

Table -3: Libraries used for model

1. Firstly, we imported all the libraries we listed in Table 3.
2. Then we load the raw data (.arff files) as panda dataframes and assign the new column headers to them. Although the features are numeric and class labels are binary, in the dataframes, all the values were stored as objects. So we converted them to float and int values respectively.
3. Now we start the data analysis. Firstly, we see how much data is missing in each data frame and look at the nullity (sparsity) by generating the nullity matrix and nullity heat maps respectively.
4. Then we perform imputation of the missing data using Mean, k-NN imputation techniques and generate fresh dataframes of imputed data.
5. We create (instantiate) the 8 classifier models (GNB, LR, DT, RF, LDA, QDA, SVM, kNN) and store them in a dictionary.
6. We iterate over all the models. In each model, we iterate over all the 2 imputed-oversampled dataset collections. Each collection has 5 data frames corresponding to 5 years' data. On each of these years' datasets, we train the model using K-Folds Cross Validation and store results.

5. RESULTS

Our results are organized as follow: Firstly, we report the different score of the 8 models we have experimented with, using a score of the fitting time, scoring time, accuracy, precision, recall and F1 score against each of the imputation method (Mean, k-NN), and internally, on each of the 5 datasets (Year 1 – Year 5).Here we presented result of all the ML models with mean and kNN imputation.

Model	Accuracy	Precision	Recall	F1 score	AUC_ROC
RF	0.975706	0.954172	0.700446	0.970462	0.902829
SVM	0.961481	0.480741	0.500000	0.942600	0.541742
LR	0.960911	0.480729	0.499704	0.942314	0.498500
K-NN	0.960341	0.635943	0.527314	0.945715	0.675286
DT	0.959575	0.734851	0.750496	0.960497	0.750496
LDA	0.959395	0.675182	0.529678	0.945520	0.673523
QDA	0.292990	0.507217	0.537192	0.411584	0.618026
NB	0.066033	0.491097	0.490751	0.059554	0.490249

Table -4: Results using mean imputation

Model	Accuracy	Precision	Recall	F1_score	AUC_ROC
SVM	0.960155	0.480078	0.500000	0.940639	0.558935
RF	0.959396	0.580501	0.510929	0.941916	0.863929
LR	0.958254	0.492626	0.501294	0.939997	0.483832
K-NN	0.957306	0.492609	0.500692	0.939526	0.590196
LDA	0.953888	0.501053	0.503488	0.938310	0.700300
DT	0.941364	0.644045	0.665791	0.943946	0.665791
QDA	0.329591	0.512281	0.566364	0.453580	0.631389
NB	0.086709	0.501296	0.499812	0.095685	0.504739

Table -5: Results using kNN imputation

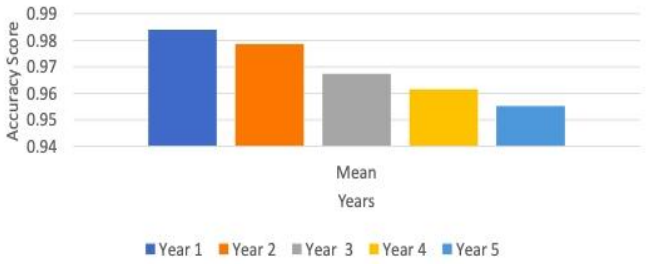

Fig -2: 5 years analysis of Polish bankruptcy dataset

6. CONCLUSION

Knowledge of an upcoming bankruptcy is a crucial aspect of the decision-making process of the imperilled company itself, as well as of other institutions interacting with the company. There is a need to conduct a comparative study in companies that are not listed but publish their financial statements. This will help in developing a robust model that can be used in the country when making investment decisions. For the researcher to improve on model construction, there is a need to construct an industry-based model. This will help in selecting effective models applicable in a sector.

7. AUTHORS REVIEW

Prejudgement of bankruptcies can help save lot of investments and market values and surely using the machine learning based approach can help save time and provide accurate results. However, some humanly factors and natural calamities that are out of the reach of this digitized system can also lead to bankruptcy. So, one should do its own analysis while using such a systematic approach for better understanding.

8. REFERENCES

[1] Wilcox, J. W. (1973). A prediction of business failure using accounting data. Journal of Accounting Research, 11, 163–179.

[2] Chen, T., & He, T. (2015b). xgboost: extreme gradient boosting. R package version 0.3-0. Technical Report.

[3] Friedman JH (2001). "Greedy function approximation: a gradient boosting machine." Annals of Statistics, pp. 1189–1232.

[4] Bache K, Lichman M (2013). "UCI Machine Learning Repository." URL http://archive.ics. uci.edu/ml. Friedman J, Hastie T, Tibshirani R, et al. (2000). "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)." The annals of statistics, 28(2), 337–407. Friedman JH (2001). "Greedy function approximation: a gradient boosting machine." Annals of Statistics, pp. 1189–1232.

[5] Friedman J, Hastie T, Tibshirani R, et al. (2000). "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)." The annals of statistics, 28(2), 337–407.

[6] Kittler, J., Hatef, M., Duin, R. P., & Matas, J. (1998). On combining classifiers. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 20, 226–239.

[7] Merwin, C. L. (1942). Financing small corporations in five manufacturing industries. NBER Books p. 1926-1936. New York: National Bureau of Economic Research, Inc.

[8] Sinkey, J. F. (1975). A multivariate statistical analysis of the characteristics of problem banks. The Journal of Finance, 30, 21–36.

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