econometrics_model

Grzegorz Krochmal and Katarzyna Kry?ska

1. Introduction

In the business world it is important to know if the company you are dealing with is now and is going to be in the future in a good financial condition. Therefore, the way to predict the company's default has been the subject of interest for researchers for many years. The first important model to predict corporate bancruptcy was Altman's Z-Score model from 1968 (Altman, 1968). Since than many papers occured which were focusing this issue with different approaches. The techniques has evolved from Multidimensional Analysis proposed by Altman, through Logit (Ohlson, 1980) and Probit (Zmijewski, 1984) models, to modern solutions which incorporate highly developed Machine Learning techniques like Ensemble Boosted Trees(Tomczak et al., 2016). In our work we want to focus on Logit/Probit strategy of modelling company's default. Before the modelling part there is an obvious need to define main and the secondary hypothesis. In our case they are:

H0: *It is possible to predict company's default with specific financial indicators achieved by particular company*

H1: *Financial indicators can't be used as default predictors*

As it was stated before, it is very important for a market's health to be able to predict particular company's default with available financial indicators. It is crucial not only for investors who surely do not want to invest in companies which are going to go bankrupt but it is also really important for the whole country's economy. With the knowledge what financial indicators can predict company's default it would allow the policy makers to concentrate on this indicators and so on be able to prevent bankruptcies which would be especially important in case of major companies.

Some imporant works in the area of bankruptcy prediction were already mentioned. However this issue is deeply analysed in **Chapter 2**. In **Chapter 3** there is a description of a data set we have chosen for our modelling. Also in this part we introduce data processing and feature selection to properly conduct our study. In **Chapter 4** we make a choice between Probit and Logit model before running the model.

2. Literature Review

In the introduction few important works which challenged the problem of prediction of corporate bankruptcy were briefly mentioned. It all started with Altman's Multidemsional Analysis (Altman, 1968). Later many different ideas were introduce to solve the task of a proper prediction of bankruptcies. The works which are important for our analysis are ones which incorporate Logit (Ohlson, 1980) and Probit (Zmijewski, 1984). Obviously, it is important to state that those approaches are rather old and they do not keep up to the newest strategies. Currently, where the computing power is huge few modern techniques of bankruptcy prediction were proposed:

1. Rough Sets (Dimitras et al., 1999)
2. Evolutionary Programming (Zhang et al., 2013)
3. Ensemble Boosted Trees with Synthetic Features Generation (Tomczak et al., 2016)
4. Support Vector Machines (Shin et al., 2005)

However those methods can overcome the shortcomings of older approaches there are surely much more demanding. For example the SVM method which prediction power seems to be worse only than the Ensamble Boosted Trees method requires the function hand-tuning which makes it a tough tool for everyday business applications (Tomczak et al., 2016).
Coming back to the models which are the most important for this paper we are focusing on Ohlson's And Zmijewski's works. The Ohlson model is based on Logit. Beneath we present the list of variables Ohlson used in his analysis (Grice & Dugan, 2003):

- Y = The probability of membership in the bankrupt group based on a logistic function
- X1 = log (total assets/GNP price-level index)
- X2 = total liabilities/total assets
- X3 = working capital/total assets
- X4 = current liabilities/current assets
- X5 = one if total liabilities exceed total assets, zero otherwise
- X6 = net income/total assets
- X7 = funds provided by operations/total liabilities
- X8 = one if net income was negative for the last two years, zero otherwise
- X9 = measure of change in net income,1 and Y = overall index.

It is important to indicate that Ohlson selected mentioned variables in an arbitrary way, just choosing those which were most frequently mentioned in literature. It is some limitation which we will try to deal with in modelling part. This limitation is also the case with model proposed by Zmijewski. The researcher applied such variables:

- Y = overall index
- X1 = net income/total assets
- X2 = total debt/total assets
- X3 = current assets/current liabilities

They were all selected because of their importance in similar previous works. Additionaly, what has to be said about both models' shortcomings is the fact that the coefficients coming from the models are highly dependable on the data. So basically the models have to be retrained for every time period which is measured to obtain realiable results (Grice & Dugan, 2003). In our case it is not a significant problem but might be with huge datasets in business aplications. What might also be an important issue with those models is the fact that the dependent variables are selected arbitraty and their number is seriously limited. Nevertheless mentioned issues, both models still seem to be interesting and valuable. Now it is all the matter of the dataset how well those models are going to perform and eventually how we can improve them.

3. Data description and analysis

The analysed Data Set - Polish companies bankruptcy data - was created by Sebastian Tomczak and is about bankruptcy prediction of Polish companies. In our analysis we used financial rates from 5th year of the forecasting period that holds information about bankruptcy status after 1 year. We chose this dataset as it is the least imbalanced of all provided. This dataset contains 64 attributes and 5910 instances (financial statements) where 410 represent bankrupted firms and 5500 companies that did not bankrupt.
This dataset contains missing values. Firstly, we will look at columns to see which contain the most missing values.

	Number of missing values	Variable
(current assets - inventories) / long-term liabilities	2548	(current assets - invento
profit on operating activities / financial expenses	391	profit on operating act
net profit / inventory	268	net profi
sales / inventory	268	sales /
gross profit (in 3 years) / total assets	135	gross profit (in 3
working capital / fixed assets	107	working capi

Attribut #37 contains 2548 values, which is almost 50% of number of rows, so we will dispose this variable. After that, we used Multivariate Imputation by Chained Equations to impute missing data where possible. MI imputes the missing values multiple times, resulting in

multiple full datasets. Then each dataset is analyzed and the results are combined.MI gives approximately unbiased estimates of all the estimates from the random error.

```
imputed_Data <- mice(data, m=5, maxit = 5, method = 'pmm', seed = 500)
new_data <- complete(imputed_Data)
write.csv2(new_data, "imputedData.csv", row.names = FALSE)
```

```
## [1] 40
```

After these operations we have only 4.81% (284) of rows which contain missing data, so we will delete them.