Welcome to the 'Predicting Bankruptcy with Financial Statements' Wiki!

Introduction and Motivation

Bankruptcy prediction is the task of predicting whether a firm will go bankrupt given information from the firm’s financial statements. This analysis is important to creditors and investors as they evaluate bankruptcy risk.

Data Description

The dataset was created by Sebastian Tomczak and contains bankruptcy information for 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.

Pre-Processing and Data Cleaning

The data is clean already, except for a few missing values, 'dfSummary' function from the SummaryTools Library is used to represent the NA's and VIM library is used to visualize the missing values.

Exploratory Data Analysis

Correlation of Variables is represented using 'ggcorrplot' library, and the most correlated variables are dropped. This updated list have few missing data cells, therefore, Data Imputation is done by using **'Multivariate Imputation by Chained Equations (MICE)'** method.

Data Science Model

The classification models used in this project are:

1. Gaussian Naïve Bayes
2. Random Forest

Model Validation

The main technique used to validate the above 4 models is using Classification Accuracy through Confusion Matrix. The first two models randomly split the dataset into 70% training data and 30% testing data. The last two models train on the first four years of data and test on the 5th year of data.

Bayes Classifier: Accuracy:0.9799 I Precision:0.9720 I Recall:0.7126

Random Forests: Accuracy: 0.9827 I Precision: 0.9115 I Recall: 0.7671

Bayes Classifier (5th Year Test): Accuracy:0.9809 I Precision:0.9578 I Recall:0.8798

Random Forest (5th Year Test): Accuracy:0.9780 I Precision:0.9920 I Recall:0.8065

Results

Analysis and evaluation of the performance of the models is decided using several metrics such as accuracy, precision, recall, etc., and rank the models accordingly. Since the data is highly imbalanced, i.e., it response is mainly towards companies being non-bankrupt. Therefore, In Model Validation segment, Precision and Recall are also calculated along with Accuracy of the classifier.

Best Model: Each of the models performed similarly when observing model accuracy. The Naïve Bayes Model is simple and computationally inexpensive while the Random Forest Model is a supervised learning method and is computationally expensive. The Naïve Bayes Model performed worse than Random Forest Model on recall, but better on precision with randomly generated training and testing sets. The Naïve Bayes 5 Year Model performed worse on precision but better on recall than the Random Forest 5 Year model. The last two models test training data from years 1-4 on the 5th year testing set. These models performed better than the original models as they included time series related information. The best model depends on whether recall or precision is most important.