Project McNulty

Build Classification Model to Predict Corporate Bankruptcy

Anjali - October 17, 2018

Goals

- Develop a supervised machine learning classification model to understand and predict corporate bankruptcy.
- Identify features that can be used to predict bankruptcy and carry out feature engineering
- Use multiple classification algorithms (Logistic Regression, SVM, …) and see which is best as predicting
- Use different time periods (1 year, 2 years … upto 5 years) for each model to find what is the best period for predicting bankruptcy.
- Collect sufficient data to be able to train the model (at least 5000 points)
- Quantify how good/accurate the model is.

Data

Polish Companies Bankruptcy Dataset - available on UCI Machine Learning Repository. - This has mostly financial statement data for each company. The data has five years worth of information. Each year is split into a different data file. The bankrupt companies were analyzed in the period 2000-2012, while the still operating companies were evaluated from 2007 to 2013.

Note: Each year about 500 companies go bankrupt out of 4000 - 5000 listed in the dataset set this is not a bad ratio for training.

Data Features
Target = y = Bankrupt or not (in a certain specified time frame)

Part 1 : Financial Statement Information
Information typically available in the annual report on the balance sheet, income statement and cash flow statement. The dateset of the Polish companies works with his financial information and the features are variety of ratios. There are over 60+ features in this dataset.

net profit	short-term liabilities	EBIT	inventory
total assets	cash	book value of equity	EBITDA
total liabilities	receivables	sales	Etc
working capital	operating Expenses	equity	
current assets	retained earnings	gross profit	

Note: these features do not involve competitiveness, current industry health, or the country level economic indicators. This implies that predictions based on these features would be relevant for any set of companies in any country. The model should have applicability for US companies as well.

Part 2 : Try to construct a similar dataset for US companies (if time permits)
Use data from the SEC/ data.gov to re-construct a similar dataset with American Companies. Note the dataset will need to have a good proportion of companies that will bankrupt in the given forecast period for proper training.

Part 3 : Supplementary Information (if time permits)
- Supplement with data on Health of the sector that company belongs to make a "score"
- Supplement with data from stock markets - eg. Stock price, stock volatility - measure variance of stock prices over the period, …

Tools

Will need to use API for data gathering
Use Pandas, Pickling, SQL for handling data and storing
Use at least one of the following for visualization - D3, Tableau, flask
Use scikit-learn for developing classification models