Data Science Honors Thesis: Data Source Documentation

Rodrigo Palmaka
University of California, Berkeley, California, USA.

1 Data Collection

1.1 Features

The data used to predict credit ratings will come from several different sources and it is likely that each feature will require its own data collection process. There are many different publicly available metrics that can influence prediction: from traditional variables under the scope of hard, financial data (balance sheets, income statements, cash flow statements) to more speculative metrics. Focusing on financial data, I will be using the Securities and Exchange Commission's Financial Statement Data Sets. These data sets are published on a quarterly basis and include numerical data extracted from 10-K, 10-Q and other filings companies submit to the SEC. This data contains strictly public information on filing companies that can be found in respective SEC filings, there are no licensing restrictions.

1.2 Labels

The target variables for this prediction are the corporate credit ratings issued by Standard and Poors, Moody's or Fitch – the "big three" credit rating agencies. Currently, I am exploring S&P's Compustat database for historical ratings data. Unfortunately, this data set is not freely available so as a backup, I will also consider implementing a basic scraping tool to query credit ratings from the Standard and Poors/Moody's website. Once the target variables are collected at quarterly intervals, I will then label each company at a given time with a respective credit rating.

2 Bias

Regarding the reliability of the data, understanding bias will be necessary. The SEC filings I will work with in the Financial Statement Data Sets include data that was submitted by publicly traded companies to the SEC per regulatory demands. This data is publicly available for all to analyze and it is scrutinized by the regulatory agency itself to detect discrepancies and fraud. Therefore, I believe it is fair to trust the data published by the SEC and assume their due diligence process would detect fraudulent data. The credit ratings themselves, however, aren't as impartial. During the great recession, for example, the "big three" rating agencies faced a barrage of criticism for propping up credit ratings of institutions in distress and downplaying the risk exposure of products such as Mortgage-Backed Securities. The questionable validity of ratings given by the ratings organizations misled investors and may have contributed greatly to the 2008 recession.