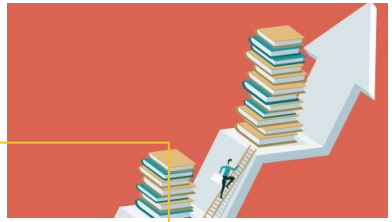
My Project Data: Annual financial accounting data from 4000 companies *(2016,2017,2018,2019)*

Financial Accounting is often called the language of business, that is, the language that managers use to communicate the firm's financial and economic information to external parties such as shareholders and creditors (Coursera: Accounting, Principles of Finance course)

Balance Sheet– describes the assets and liabilities of the company and the capital that the business has (31 variables)

Cash Flow– describes the liquidity of the business, it shows the amount of cash or cash equivalents that enter or leave the country (22 variables)

Income Statement– contains the indicators of the profitability of the business (20 variables)



Say I'm company X. How can I compare my financial performance in relation to the market, i.e. other companies?

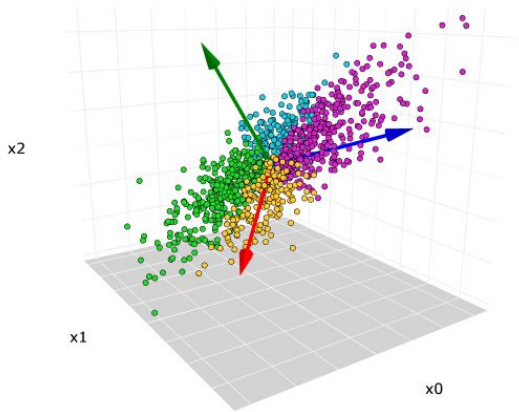
Which financial variables should I focus on to improve my financial performance?

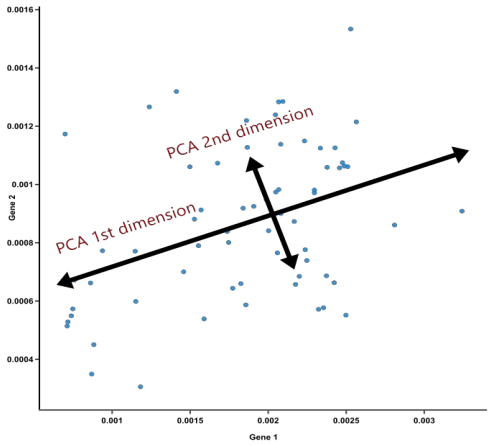
Financial data is high-dimensional… Can we reduce complexity?

Balance sheet, income statement and cash flow statement have 70 variables between them! This is too many to comprehend or plot.

Can we quantify the correlations between the variables in the data and use this information to reduce dimensionality? (I ran a preliminary PCA analysis where 10 variables from the cash flow data explained 90% of the variance in the data)

Idea: Use Principal Components Analysis to extract linear combinations of data columns that explain the most variability across firms in the data set :-)

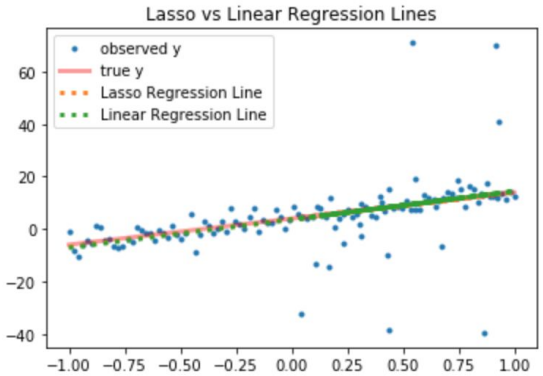


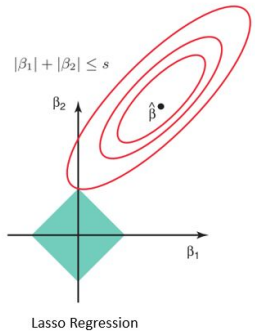


Regression analyses with sparsity penalties– quantifying which variables have an association

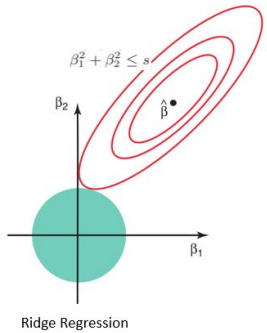
We can predict one financial variable as a target and use all of the other variables to predict it using linear regression. To find out which variables had the most impact we can analyze their significance or use a sparsity penalty to shrink the variables that are not predictive.

Using regression analyses we can quantify which variables explain the most variance in the regression target and use these relationships to make recommendations for planning the financial trajectory of future years (for example, is it better for profitability to increase or decrease debt?)

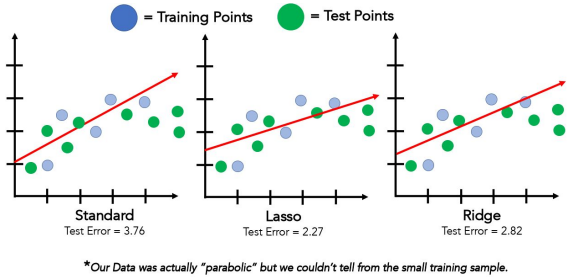




Lasso Regression



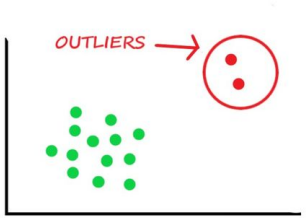
Ridge Regression

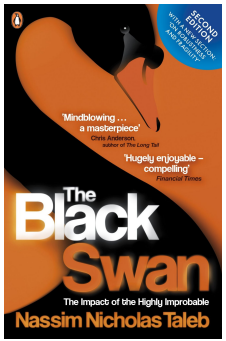


Detecting Black Swans

Financial distributions are known to have fat tails (i.e. things that shouldn't happen tend to happen). Normal distribution is an example of a distribution that is predictable. Market crashes obey the Cauchy distribution, where extremely rare events tend to appear.

Can we identify financial Outliers in the data? How do they skew the data analysis? What do these companies do?





Assessment

Data availability: 5/5

Data value: 4.5/5

This data set is publicly posted on Kaggle It is easy to access the data and it is in a shape that is good for analysis.

In consequence, this data can easily be analyzed with the methods outlined in this presentation using the Python programming language and the sklearn library.

Thus, the priority is to analyze the data and see what insights it can unlock.

The analysis is of potential relevance for many companies as it aims to give insight into some general patterns in balance sheet, income statement and cash flow financial data.

Additional data that could be collected includes the industry and operation area annotations for each company. This can be done programmatically using web-scraping, but here we focus on the existing data.