Metis Regression Project – Aug 2021

Sam Reiff

Abstract

The objective of this regression analysis is to identify and quantify the relationship between the number of Wall Street analysts covering a public company and that company's fundamental attributes pertaining to size, profitability, and trading metrics. My data set was created via web scraping Yahoo! Finance's ("YF") company-specific pages. Specifically, over 3,000 unique YF web pages were scraped to aggregate over 10 features for my initial data set. I then supplemented this data with sector identifiers and growth characteristics via the Yahoo! Finance API. Across multiple regression models, Ridge Regression performed best in cross-validation, and produced an R^2 value of 0.4598 on the holdout set (MAE = ~4). While this certainly does not explain all of the variance in the underlying data set, the Ridge Model may be a starting point for identifying companies with coverage mismatching their fundamentals.

Design

Traditional academic finance supposes that markets are at least weak-form efficient (stock prices reflect historical pricing data, but do not always fully reflect public/non-public information relevant to valuation), and many finance/banking professions are predicated on capturing inefficient security pricings, e.g. buying 'undervalued' stocks. Sell-side research analysts at Wall Street firms issue investment research on thousands of publicly traded stocks, publicly communicating their opinions on the investment prospects for individual stocks and sectors, usually including their estimates for financial metrics such as revenue and earnings. My hypothesis is that stocks lacking adequate Wall Street research coverage may have more investing opportunities due to reduced information dissemination. This project serves as a starting point for identifying undercovered stocks for further investing due diligence, with the assumption that an investor with time and skill can make informed investing decisions to generate investing alpha.

Data

My data set was constructed by joining web-scraped data from YF and two additional features I generated separately using the YF API:

- Data scraped from YF included;
  - **Enterprise Value.** Market valuation of the company's total assets.
  - **Market Capitalization.** Market valuation of the company's traded equity.
  - **Average Trading Volume.** Average number of shares traded daily for the latest 3 months.
  - **Float.** Percent of shares actively traded compared to total shares outstanding.
  - **ROA.** Return on the company's assets.
  - **ROE.** Retrun on the company's equity.
  - **LTM Revenues.** Latest twelve months of revenue.
  - **LTM Gross Profit Margin.** Gross profit margin, based on LTM P&L.
  - **LTM Operating Profit Margin.** Operating profit margin, based on LTM P&L.
  - **LTM EBITDA.** Latest twelve months of Earnings Before Interest, Taxes, Depreciation, and Amortization.
  - **LQ Assets.** Latest quarter assets.
- Supplementary data included company industry/sector and growth; I ultimately disregarded growth as a viable feature

My Ridge model identified market capitalization and average daily volume as the two most influential factors in analyst coverage. There were disparities in sector coefficients that also suggest technology, consumer cyclicals, and industrials are favored by sell-side analysts, while real estate and utilities are less likely to be covered (this comports with my domain knowledge of the sell-side).

Algorithms/tools

- Python BeautifulSoup for scraping
- Python Pandas and Numpy for arithmetic and data cleaning
- Python SKLearn for regression, specifically standard linear regression, LASSO regression, and Ridge regressions were examined
- Python Seaborn and Matplotlib for visualization

Communication

The findings of this exploratory analysis are principally communicated in the presentation associated with this document.