CS 490 Project Proposal: Prospective Firm Evaluation using SEC Filing Data
By Michael Byrnes

Advisors: Prof. James Aspnes (official) and Alex Hetherington

Introduction

The Yale Investments Office (YIO) manages Yale’s Endowment of over $27.2 billion, which provides for the university’s continual operation year-to-year. The YIO generates superior performance by choosing a sensible asset allocation based on modern portfolio theory and selecting world-class managers to steward the university’s capital. That is, instead of directly investing in assets like stock or bonds, the YIO develops partnerships with top-tier firms who manage money on behalf of the Endowment. Each manager has expertise in a specific sub-field of finance, such as venture capital or natural resources. By developing long-term partnerships with like-minded investment firms, the university has been able to produce consistent outperformance across all asset classes. Over the 20 years ending June 30th, 2017, the endowment returned 12.1% per annum. Ultimately, effective manager selection is one of the most integral components of the YIO’s success.

The process of identifying new firms requires extensive effort. The YIO employs a number of strategies to find potential partners, such as attending conferences, sourcing referrals from current managers, and accepting cold submissions through the YIO’s website. However, the vast majority of these processes are fully manual, requiring significant time and energy on the part of the office’s relatively small staff of 25-30 investment professionals. I propose a system for pre-emptively tracking and evaluating new firms using SEC Form D filing data. By combining the automated efficiency of web scraping with an effective classification algorithm, this system could greatly improve the office’s ability to hone in on the most promising new firms.

Background and System Design

Investment firms are required to submit SEC Form D when they raise capital (e.g. when starting a new firm, raising a new fund). Form D contains various pieces of information about a firm, such as the firm’s asset class, the amount of capital being raised, and names of the firm’s executive officers. Automatically monitoring these submissions would allow the YIO to identify start-up firms early and take proactive steps to evaluate their attractiveness as a partner.

My proposed system will consist of the following parts/phases, explored in greater depth below:

1. Data Scraping and Parsing
2. Firm Evaluation
3. Workflow Integration

Data Scraping and Parsing:
Form D filing data is publicly available through the SEC’s Electronic Data Gathering, Analysis, and Retrieval platform, commonly known as EDGAR. New filings are published on a daily basis. Though EDGAR does not have a user-friendly API, Form D data is accessible in a relatively well-structured format. I will build a crawler to scrape both historical data for testing
purposes and new data as it is published online. As the data is collected, I will compare it against an internal database of managers that the YIO has already interacted with in order to separate known managers from new ones. While we are primarily interested in discovering new managers through this method, data for managers that we have already considered could prove useful for building a training dataset.

In addition, I will implement a relational database to store Form D data, which will be necessary given that 100-200 Form Ds are filed per day. Though many of these forms will be filtered prior to insertion, the total number of entries would grow quickly as the program runs for an extended time. Furthermore, a relational database will provide an efficient means of storing the data for search queries and further analysis.

**Firm Evaluation:**

Attempting to evaluate the relative attractiveness of firms found through Form D will present interesting challenges and opportunities. The first step of data processing will involve classification based on well-structured quantitative and qualitative factors. For example, some Form D filers are not investment funds that accept outside capital. By assessing the industry group for each firm, I will be able to classify certain firms as high interest, other as potential interest, and some as entirely irrelevant. In addition, I will be able to classify firms further based on numerical attributes such as the total offering amount of securities, minimum investment accepted, and number of investors who have already invested in the offering. Each of these inputs will factor into the attractiveness score assigned to a firm.

Since the majority of the firms under consideration will be relatively new, there will likely be little quantitative data available through other filings, such as Form 13-F. However, I will explore the potential to gain further information by doing searches based on keywords from Form D. Using the name of the firm and its industry group, I could locate a description of the firm via Bloomberg’s search API. Using natural language processing, I could compare this description with descriptions of high-quality managers to provide a rough metric of similarity.

Similarly, Form D contains the names of people related to the new firm. Searching for information on each individual could provide insight into that person’s prior work history, which could in turn influence the attractiveness of the new firm. If, for instance, a portfolio manager worked at a well-regarded firm, then the new firm would receive a score boost.

The most challenging aspect of this project will be tuning the parameters of this scoring model. My goal is to explore several angles, including traditional classification techniques and machine learning where applicable. If the problem ultimately appears well-suited to machine learning, I will need to build a training dataset of past firms. This will likely require either human input or a mapping function that relates information in the YIO’s manager database to a numerical attractiveness score for each manager. That is, hired managers would receive extremely high scores, while firms that were interesting, but that did not pan out, would map to a lower score. The accuracy of machine learning techniques hinges on the quality of the training data, thus requiring a careful treatment of this design aspect of the project.
Workflow Integration:
To make this tool effective for production use, I intend to integrate it as seamlessly as possible into the workflow of the office. After identifying the most promising new firms, my system could email a weekly briefing to an investment analyst (or several to divvy up the work, or partition by asset class) for human evaluation. This could include information such as each firm’s score, contact information to set up a meeting, and relationships between each firm and our existing managers. Alternatively, the system could simply provide a notification that a new list of firms is ready for evaluation. Then, an analyst could access the information through either a stand-alone GUI or a web app interface.

Deliverables

The totality of the aforementioned design components represents the key deliverable for this project. In addition to a functional application, I will produce a written report detailing the system design process. The report will focus on key difficulties and challenges faced in using SEC filing data for classification and learning. Furthermore, I will include an evaluation of the program’s accuracy and efficiency, both from an algorithmic perspective and a real-time savings perspective. Finally, I will include ample documentation so that future users can build upon the work and use it as a framework for more generalized company/firm analysis based on SEC filing data.

Conclusion

The ultimate goal of this project is to develop an effective system to aid in prospective firm evaluation. As a future employee of the YIO, I hope to delve more deeply into the manager selection process and meaningfully augment the workflow of YIO analysts with machine evaluation. In addition to designing an effective system, I hope to develop a deeper understanding of software engineering principles by applying skills from the controlled, classroom environment to a real-life, messier problem.