Capstone Proposal

Predicting Sector/Industry of a firm

Domain Background

This problem comes from the world of finance. For financial analysis of a company, it is useful to assign the company to appropriate sector/industry (i.e. its peers). This assignment helps the financial analysts to better understand various metrics of the company and put them in perspective. E.g. Cost of Goods Sold for a "Consumer Goods" company would be a very high percentage of revenue as compared to, say, a "Technology" company.

The question then arises, how does the analyst go about assigning this sector? In United States and other financially developed countries, this assignment is done by the company itself (or by other analysts) if the company is sufficiently large. However, for very small companies (e.g. unlisted) or for other non-public firms (e.g. funds), this assignment is not readily available.

An analyst looking at a universe of small non-public companies would struggle to identify the ones relevant to their own sector. In this project, I will develop a model to assign sector/industry to a company based on available information.

Problem Statement

Identify the correct sector/industry of a firm based on its returns, brief description and some other available factors.

Datasets and Inputs

Most the dataset related to the actual problem is proprietary. In its stead, I'll get data for listed companies from NYSE/NASDAQ. Companies/analysts have already assigned Industry/Sector for these firms. Our dataset will come from Yahoo finance[1]/MorningStar[2] or a similar source and will include various features typically available for such firms. A manually created sample dataset (3 rows) accompanies this proposal. The dataset is in two files:

1) Fundamental Data (which contains the fundamental data for the firm)
2) Stock Price Data (which contains the closing price for each of the firms included in Fundamental Data)

Solution Statement

I will model this problem as a multiclass classification problem. We need to create features that will help us identify classify each firm. The model will be built in python and use one or more of commonly used classification techniques.

[1] https://finance.yahoo.com/

[2] http://financials.morningstar.com/

The solution will provide an estimated classification (sector/industry) for each of the firms in our dataset. The estimated classification will be compared with the actual sector/industry to analyze the usefulness of the solution.

Benchmark Model

Currently, there isn't a machine learning solution to this problem. Based on interviews with portfolio managers and analysts, I estimate that it would take a portfolio manager (PM) approximately 10 minutes to correctly identify the appropriate sector/industry of a firm. In contrast, it would take around 2 minutes to determine whether the industry/sector allocation is correct or not. This gives us an approximation of current manual model which is presumed to be 100% accurate (since there is no objectively true assignment).

Consider a universe of 100 firms. Under current methodology, it would take a PM 1,000 minutes to correctly assign the industry/sector to all firms.

Now, consider a new model which assigns a sector/industry correctly (as judged by a PM) 50% of the times. To review this, we expect the PM to spend around 200 minutes (2 minutes per firm) and without further effort, have 50 correctly identified firms (a process that otherwise would have taken around 500 minutes).

Based on this, a model that is accurate 20% of times would breakeven in terms of marginal time spent in review. A higher accuracy would be an improvement over current methodology.

There may be additional benefits to even an "incorrect" assignment as, depending on the assignment, it may provide the analyst important insights into the firm.

Evaluation Metrics

We will use "exact match" to evaluate our model. This metric calculates number of correct assignments/total number of assignments.

Project Design

Step 1: Download relevant data. I will use yahoo finance and/or Edgar online (or a similar source). Specifically, I will try to get the following data:

a) Brief Description (A paragraph of text describing the activities of the firm)
b) Total daily returns (or, as a proxy, stock price adjusted for dividends/splits etc.)
c) Address (if available)
d) Inception date (if available)
e) Website (if available)
f) Currency (if available)

For training/testing purposes, I'll also get the sector/industry assignment.

I will use BeautifulSoup[3]/Selenium[4] to download the data. Alternatively, if that doesn't work, I'll use data from Bloomberg.

Step 2: Clean the data.

a) Convert the price data to returns data.
b) Convert the string data into usable numbers (e.g. for Net Income, Market Cap etc.)
c) Fix issues in data (such as inconsistent usage of 'United States', 'USA' etc for country)
d) Decide what to do with the missing data.

Step 3: Transform the data.

a) Decide how to extract features from the description text
b) Create new features
c) Decide if we need to prune features

Step 4: Use one or more classification algorithms to classify the firms. Currently, I am considering SVMs as prior research has supported their use in text classification[5]. Other potential alternatives include neural networks and decision trees.

Step 5: Iterate through the models to improve results.

[3] https://www.crummy.com/software/BeautifulSoup/
[4] http://selenium-python.readthedocs.io/
[5] http://www.cs.cornell.edu/people/tj/publications/joachims_98a.pdf