CMPT 733 Big Data Programming – Final Project

Call for Project Idea

Title of the Project:

Machine learning to detect misstated financial statements.

Description:

Corporations regularly prepare financial statements that reflect their performance. The release of these statements to the public strongly affects the corporation's share price. A small percentage of the statements are ultimately found to be incorrect, which is very disruptive to the firm and to the stock market in general. This research project is aimed at identifying (at the time of their release) which financial statements will contain misstatements, either due to error or fraud. A good algorithm could predict which statements will have misstatements, while a better algorithm would identify where the misstatements actually are. The findings would be useful to investors, securities regulators, and external auditors, and could ultimately improve the functioning of our capital markets.

The domain knowledge for this project is accounting and finance, which is a relatively new and important area for machine learning. Students who undertake this project will be trained for 2-3 hours in order to gain sufficient proficiency in financial reporting.

Unsupervised learning through cluster analysis may be the most effective method of evaluating patterns, by first analyzing a dataset of non-misstated reports. An example of a typical pattern is that corporations with more revenues will typically have higher accounts receivable unless they receive more cash. The results from the cluster analysis would be applied to a new dataset of reports (some of which have errors deliberately imbedded into them) to predict which ones are misstated.

Datasets

This project requires 2 datasets, one with archival (non-misstated) financial statements and one with simulated financial statements that have embedded misstatements. The archival dataset will provide an opportunity for machine learning of patterns, clusters, and correlations. The simulated dataset will be used to assess whether machine learning took place.

<u>Archival financial statements</u>

This dataset is publicly available structured text. All corporations provide the same 20,000 pieces of information, which is mainly numerical (some may be missing or set to zero) along with some categorical data in text format. There are some interesting labels that allow for certain groupings of corporations, such as the industry they belong to. The data covers 500-

1000 corporations over 9 years (2009 - 2017), so comprises approximately 135 million pieces of information.

The data is available through a database called EDGAR, which contains the financial statements of all corporations listed on US stock exchanges. The format used by EDGAR is called XBRL: https://www.sec.gov/dera/data/financial-statement-data-sets.html

Simulated financial statements

This dataset will be created in the next few months. It will comprise somewhere between 10-30 financial statements, some of which are free from misstatements, and some which will have been deliberately misstated.

Contact person:

This project is being proposed by Professor Kim Trottier from the Beedie School of Business. You may contact Professor Trottier at kim.trottier@sfu.ca.

Contributor of the Project Idea:

http://beedie.sfu.ca/profiles/KimTrottier