Disruptor Identification

Franco Ho Ting Lin, Xiaoxia Hao(CPPIB), Jonathan Briggs(CPPIB), Sebastian Jaimungal (UofT)

Introduction

Disruptors are forward-thinking, innovative and ambitious companies that are disrupting marketplaces and industries. The perfect example of these companies are the FAANGs, they are easy to recognize ex-post. The value lies in predicting which companies will be disruptors in the future. The information we use to identify disruptors include traditional financial statement data, stock market data, as well as features extracted by NLP from relevant texts and web scraped data. Due to the high dimensional feature space, we use Machine Learning techniques such as Logistic Regression as baseline and Neural Networks to dynamically forecast disruptor companies within each industry in the the US economy. We then build a trading strategy on top of the identified disruptors

Disruptors

- Companies that are risk takers using the latest technologies to disrupt the industry
- Aggressive investment strategies that achieve significant growth
- Well known examples:

  Facebook, **A**pple, **A**mazon, **N**etflix, **G**oogle (FAANGs)

Target Labeling

- We consider companies to be disruptors given they have high growth and high sales
- We quantify high growth as companies that invest heavily in four different channels (Capital Expenditure, R&D, Acquisition and Hiring)
- Additional characterization is to specify the high growth and under performers
- We narrow our disruptor selection to the specific Technology, Healthcare and Retailing sector

Perfect Foresight Analysis

- Given we have labelled disruptors, we will run a perfect foresight analysis to see what the maximum returns we get given we are able to accurately forecast all the disruptors

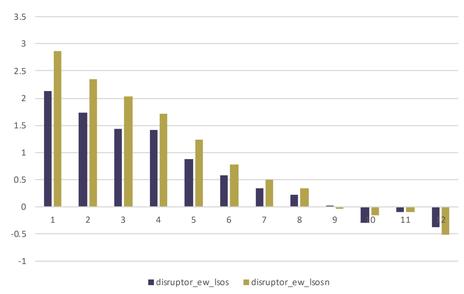

Figure 1: Perfect Foresight IRs, purple is when we take long-short positions on over spenders, gold is when we take long,short and neutral positions

Features

- We use features related to company decisions, value, quality, growth, momentum, volume, short interests, analysts or risk
- Additional text data from Conference Calls, SEC 10K reports is used to generate document level embeddings (using doc2vec [1] or SIF [2])

Model

- Due to the data being time dependent, we adopt a sliding window cross validation schema shown below
- We use Gradient Boosting Tree and Neural Net to test and Logistic Regression as baseline

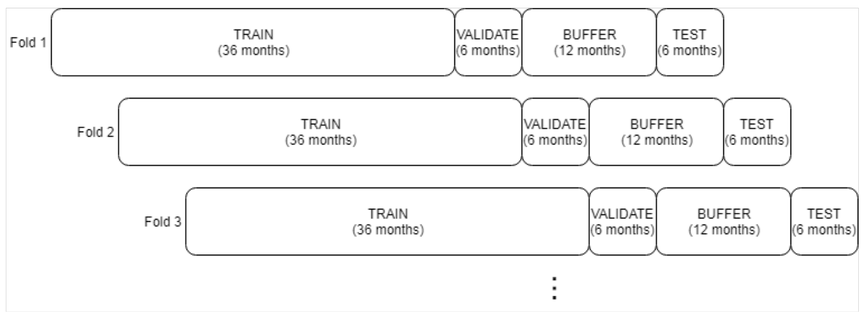

Figure 2: Cross Validation Schema - to prevent any look ahead bias, we adopt a sliding window instead of the regular CV

Results

- We focus specifically on Precision, we do not want to be selecting too many companies to be disruptors
- The scores seem low in Machine Learning space, but given the noisy space we are working in, they are quite reasonable

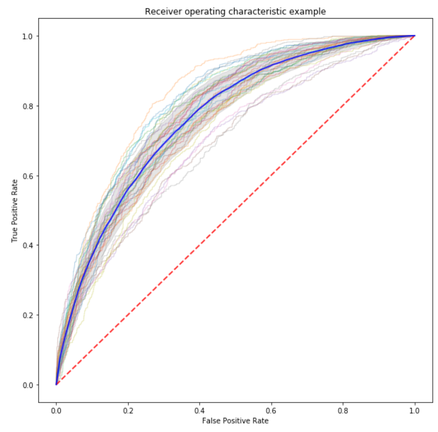

Figure 3: ROC Curve on Test set

	Validate	Test
Precision	0.6709	0.5776
Recall	0.4669	0.3935
F1	0.5443	0.4545

Table 1: Precision, Recall and F1 on Validation/Test

Trading Signals

- We can use the prediction probabilities as raw signals to build a standalone disruptor factor, the results are displayed below

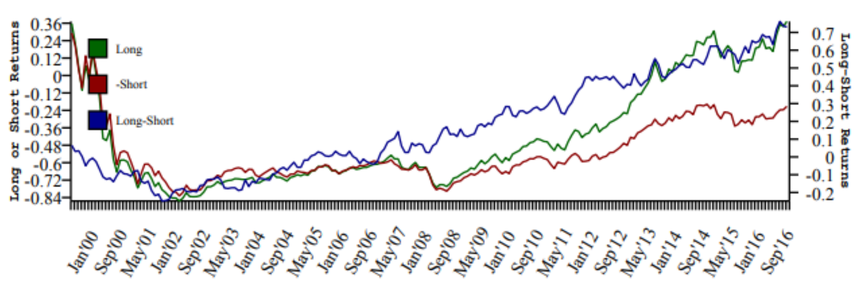

Figure 4: Returns for back test over time, the blue represents our combined returns, the green is returns from long positions, red from short positions

Ongoing Work

- Additional work is needed for better document level embeddings
- Incorporation of alternative data
- Improving the definition of a disruptor

[1] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32 (ICML'14), Eric P. Xing and Tony Jebara (Eds.), Vol. 32. JMLR.org II-1188-II-1196.
[2] Sanjeev Arora and Yingyu Liang and Tengyu Ma. 2017. A Simple but Tough-to-Beat Baseline for Sentence Embeddings. International Conference on Learning Representations 2017