Capstone 2 Proposal for Springboard Data Science Career track
Version: 1.0
Author: Jiaqi Xu
Coach: Kevin Glynn
Date: 09/16/2020

Predict Stock Price Using Natural Language Processing (NLP)

Background

Besides by only using the daily price with machine learning methods (this is what I did in capstone 1) to predict the short movements of stock price, analysts are dedicated to analyzing and attempt to quantify qualitative data from news and SEC mandated reporting. There are huge amounts of reports or articles related to different kinds of stocks, it is impossible to read all articles manually by one person. With the help of NLP, we can use machine learning methods to analysis the opinion and ideas from many reports to help us to predict the stock's movement and make a trading decision.

Previous Work

Ryan (2017) developed a model with CNN-Glove to predict the future stock market performance of public companies in categories where a financial services partner of her company invested. She used SEC (Securities and Exchange Commission) 10-K reports and got 62% prediction accuracy. Yusuf (2018) compared machine learning methods MLP, CNN, RNN, CNN-RNN to predict stock movement based on SEC 8-K reports. He found CNN-RNN reached 64.5% accuracy on the validation data. Following Yusuf, Babbe et. al (2019), they trained a BERT + single layer MLP model and produced an out of sample accuracy of 71%. Joseph (2020) defined a stock price formula with the sentiment signal from 267 journal articles related to APPL to predict one day price of it.

Goal

This project will follow Yusuf (2018) and Babbe et.al (2019)'s work at first, collect the SEC 8-K and 10-k reports (or journal articles) related to S&P 500 companies up to the Aug of year 2020, use CNN-RNN, BERT model to train at first for stocks with most articles, then use transfer learning to predict the movements of stocks with least articles. The baseline will be all stocks be trained with CNN-RNN and BERT model. The goal is to find the best model with recent reports (especially for the COVID-19 pandemic period) and

check the hypothesis that “if the transfer learning helps to improve the prediction accuracy compared to baseline approach”.

Data

The dataset I will use for this project:

Same as my capstone 1:

- Download historical S&P500 components on Wharton Research Data Services (WRDS) and Wikipedia, clean them and get a mapping table indicates which stock will be included in S&P500 at the end of the specific time.
- Based on the components from above resources, I will use Python yahoo finance module yahoo-finance-1.0.4 to download daily data of all components and S&P 500 index from 2005/01/01 to 2020/09/01 (15.75 years data). Since yahoo finance will not keep the delisted stock’s information, several delisted stock’s data in historical S&P 500 will not be available. This will not affect the prediction too much because most of the stock’s information will be arrived.

More data:

- SEC 8-k report (>20,000 documents)
- SEC 10-k report
- Other related stock analysis journal articles

Approach

(1) Data Collection & Preprocessing

Follow my capstone 1 to clean the stock prices. Follow Yusuf (2018) to download and clean all reports or journal articles

(2) Data Exploration

Data was explored for distributions of document lengths, and category and class imbalances. A little less than half of all samples had signals of "up", which intuitively makes sense considering how stock prices rise over time, especially in the past five years. This means steps such as over or under sampling will have to take place before machine learning.

(3) Machine Learning models

All or parts of the following machine learning models will be used in prediction process:

→ GloVe 100
→ CNN + RNN, or CNN + LSTM

→ Bidirectional Encoder Representations from Transformers (BERT) + transfer learning

(4) Trading

Do one day stock trading based on the prediction from step (3) and compare the daily return.

Deliverables

- Jupyter notebook with relative codes on GitHub
- Final Report includes data acquisition, data transformation, model implementation and analysis, results comparation and conclusions
- Presentation Slides
- PDF report

Reference

Patty Ryan (2017). Stock Market Predictions with Natural Language Deep Learning.

Yusuf Aktan (2018). Using NLP and Deep Learning to Predict Stock Price Movements.

Mark Babbe et al. (2019). BERT is the Word: Predicting Stock Prices with Language Models.

Trist'n Joseph (2020). An investigation into NLP using sentiment analysis to predict Apple stock price movements.