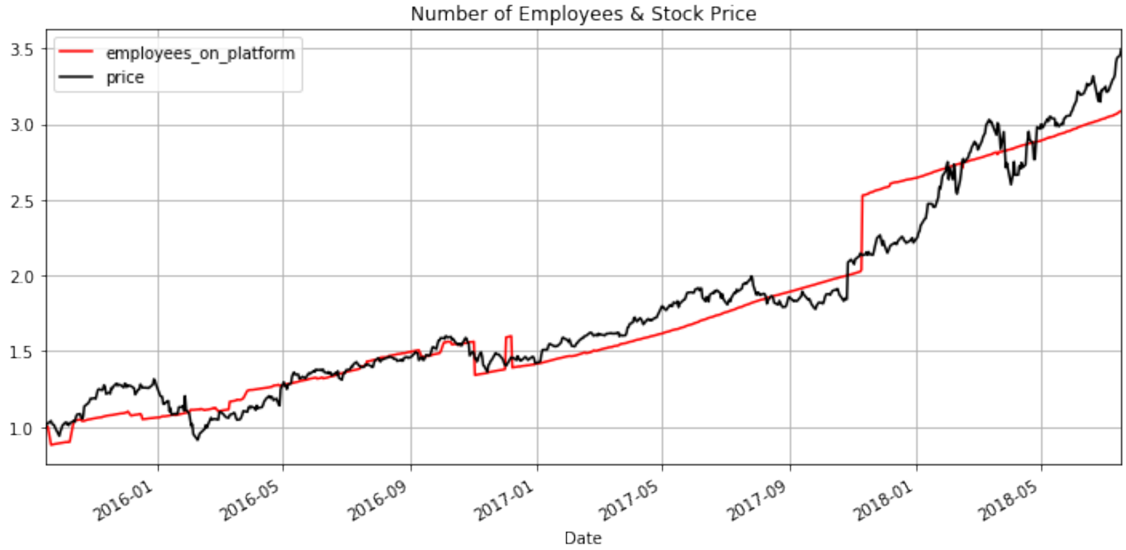
Predict Stock Price From LinkedIn Profiles

Yichao Zhang

Motivation

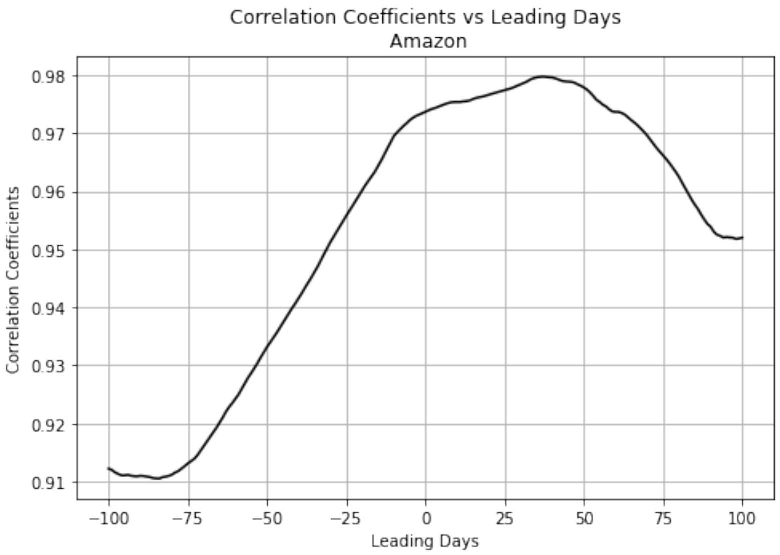
Data Analysis

Here is a simple example of comparing the number of employees and stock price for company Amazon. For the purpose of generalization, we scale the number of employees and stock prices by dividing the first value (shown in the figure below):



We can see that the stock price of Amazon is highly correlated with the number of employees from 2015 to 2018.

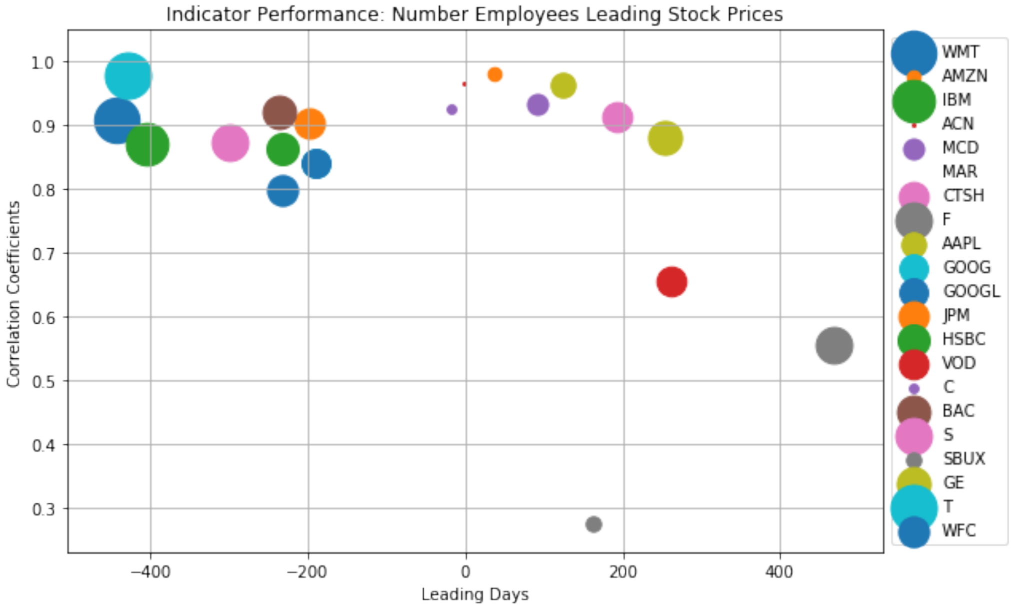
However, high correlation coefficient is not enough for a market indicator. Because we want the predictor in front of the market, thus we can use it to predict the market. The figure below calculate the correlation coefficients on different the leading days of the number of employees for Amazon. We can find that the number of employees is a quite good indicator (correlation coefficient = 0.97972886) to predict the stock price of Amazon 37 weekdays in the future.



Leading: If the date of indicator is leading the date of stock (as the figure of Amazon shown above), we say the leading days > 0, and high correlation leads to a good indicator to predict the stock price. Based the efficient-market hypothesis (EMH), if less people use this indicator, the indicator is more likely a private information, and can be used to get profit in semi-strong or weak efficient markets.

Lagging: If the date of indicator is lagging the date of stock, we say the leading days < 0, and high correlation might be useless to predict the stock price. However, the stock price can be used to predict the company's expansion plan. A current employee knows that he/she may need to find another employer when the stock price goes down, while a job seeker knows that it is a good chance to join the company when the stock price goes up.

The figure below shows the indicator performance of the stocks of top 20 expansion companies.



Leading Indicator Companies: on the upper right of the graph, we can see that the numbers of employees lead the stocks of { AMZN, MCD, APPL, CTSH, GE } with high correlation coefficients (>0.85). People can use this indicator for trading among these stocks.

Lagging Indicator Companies: on the upper left of the graph, we can see that the stock prices of { T, WMT, IBM, BAC, S, HSBC, JPM, GOOGL, WFC } lead the number of employees of corresponding companies, with high correlation coefficients (>0.8). The employees and job seekers can use this indicator to prepare to find new positions.

Data

- LinkedIn profile data set (954 MB, 2426196 records, in 2015/09/14 - 2018/07/17). It covers 5028 companies, and 141 industry types. Each record includes 14 variables. We mainly focus on 5 variables: [date, company_name, followers_count, employees_on_platform, industry]
- Stock price data from Yahoo Finance (https://finance.yahoo.com)

  From a company name of the previous dataset, we can find a stock name of that company. We download the stock price data in 2015/09/14 – 2018/07/17 from Yahoo Finance. We use the daily Adj Close price as our target.

Approach

- Exploratory Data Analysis: calculate the expansion amount, speed, and ratio of each company for feature selection. Focus on companies with long date range for machine learning modeling.
- Split the date range into in sample period (for training) and out of sample period (for testing)
- Predictive modeling (Naive Bayes, Random Forest, LSTM, etc.). Based on the expansion of a company is leading or lagging the stock price, we build 2 different type of models:
  - Leading model: select companies with high correlation coefficients leading indicators, and build a machine learning model to predict the stock price from the number of employees
  - Lagging model: select companies with high correlation coefficients lagging indicators, and build a machine learning model to predict the number of employees from the stock price
- Explore the followers_count, and add it as an additional indicator into the models
- Backtesting in sample and out of sample, fine tune the model
- The performance metric:
  - For regression model (e.g. predict the stock price values), we use Root Mean Square Error (RMSE)
  - For classification model (e.g. predict of the stock price goes up or down), we use Area Under the Receiver Operating Characteristics (AUROC)

Experimental Setup

The hardware for our experimental setup consists of a workstation with the following characteristics:

- CPU: 2.2 GHz Intel Core i7
- RAM: 6 GB 1600 MHz DDR3

The software environment includes:

- macOS Mojave 10.14.4
- Python 3.6 with current versions of the following libraries:
  - Numpy
  - Pandas
  - Matplotlib
  - Scikit-learn
  - Scipy
  - Deep Learning libraries: PyTorch, Keras

Timelines

The project timelines are listed below.

Table 1. Project timeline.

Date	**Milestone**
05/03	Data Selection and Preprocessing
05/05	Exploratory Data Analysis
05/06	**Submit Proposal**
05/10	Explore the indicator for all the 5028 companies in this dataset. Focus on companies with long date range to build predictive models
05/11	Leading modeling: select companies with high correlation coefficients leading indicators, and build a machine learning model to predict the stock price from the number of employees
05/12	Lagging modeling: select companies with high correlation coefficients lagging indicators, and build a machine learning model to predict the number of employees from the stock price
05/20	Explore the followers_count, and add it as an additional indicator into the models
05/25	**Report Draft**
06/01	Backtesting in sample and out of sample, fine tune the model
06/15	Make the final version of the project code; make video presentation
06/25	**Final report**