**Financial Performance Prediction Project (FP3)**

FINAL CAPSTONE REPORT

30 June 2018

The Cohort 11 Financial Performance Prediction Project (FP3) Team submits this capstone report in partial completion of the requirements to earn a Certificate in Data Science from the Georgetown University, Center for Continuing and Professional Education (CCPE). Team FP3 includes:

* M.D. Alam
* Riley Back
* Melissa Burn
* Michael Iapalucci
* Muralidharan Kannan

The Jupiter Notebooks and data files used in the IPO Performance Predictor, as well as this final report and the capstone presentation slides are available in the Team’s CCPE GitHub Repository[[1]](#footnote-1). Presentation slides will also be posted to Slideshare.

**Introduction**

When a company “goes public” through an initial public offering (IPO), it signals they are making stock in the company available to the general investor. This is a significant event for the company, the stock market, and individual investors. Prior to the IPO, companies are privately held, owned by one or more people, or with ownership conveyed through stocks held by a limited number of shareholders. There are several reasons for going public:

* the company may need to raise cash
* being listed on a stock exchange can improve the company’s reputation as it generally signals adherence to industry standards for transparency and good management as well as “buy-in” from underwriters
* an IPO allows shareholders to sell their stock on the open market
* offering publicly-traded stock in a benefits package can help attract new talent

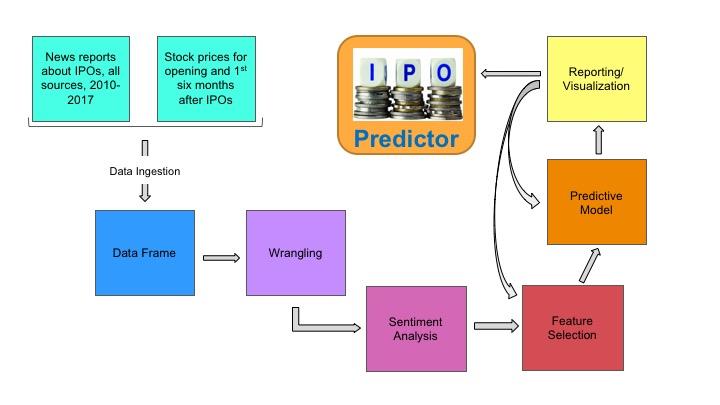
Newspapers and other media outlets, ranging from local news groups to national business-oriented papers like the *Wall Street Journal* (WSJ), frequently publish information about IPOs. The WSJ and similar high-profile outlets report financial news considered internationally significant while smaller papers focus on companies of interest to their local community.

What impact might these stories have on the stock price of an IPO? Do positive stories increase the trading price of the stock? Do skeptical or negative stories depress the stock price? This is the question Team FP3 tried to answer with this project. If the Team can create a data product that uses the sentiment of published news headlines to predict the trajectory of the stock price after an IPO, that would be a very useful tool for potential investors.

Project Hypothesis: ***Sentiment analysis of media headlines about an IPO can be used to predict the trajectory of the stock price over the first three months.***

**Project Overview**

To create the IPO Performance Predictor, the team collected headlines published on the Internet concerning 16 different IPOs spanning four different industries – Technology, Food and Drink, Automotive, and Retail. This effort produced a dataset of 5395 instances. Next, FP3 collected associated stock prices: the initial IPO “offer price”, the actual opening price and the price each day for six months after the IPO. FP3 used natural language processing by two different algorithms to assess the sentiment of each headline, and those scores, together with the changes in stock price data, were used to train a variety of machine learning classifier models to develop a price trajectory prediction system. This report describes FP3’s processes for ingesting and wrangling data, performing the sentiment analysis, selecting features, modeling and testing, to develop the IPO performance predictor. The report ends with an assessment of the IPO performance predictor and lessons learned from the project.



**Data Ingestion and Cleaning**

FP3 team members initially scraped the Internet for news articles about IPOs going back two years, discovering only three well-publicized IPOs. Subsequently, the team went farther back to identify additional IPOs for the dataset. One team member accessed the Dow Jones Factiva product (at work) and also used a Chrome Google add-on called Data Miner, collecting 5395 headlines on 16 IPOs spanning 2010 through 2017[[2]](#footnote-2). FP3 stored a cleaned version of these data including the headline, publication source, date, and day of the week. Another team member pulled data from [www.iposcoop.com](http://www.iposcoop.com) to collect ticker symbols, IPO release dates, managers, and initial prices. The team employed a Jupiter Notebook to read the stored headlines, ticker symbols and release dates for each IPO and used that information to scrape market price data from a Morningstar API. This yielded daily prices for all 16 IPOs from inception through May 31, 2018. The market prices represented the largest dataset used for the project; for Facebook, for example, which went public in 2012, there were 17,658 records. The table below identifies the IPOs used for training the predictive model, which covered four distinct industries and represented some of the most anticipated releases of the period.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Company | Ticker | Industry | IPO Date | Instances |
| Alibaba | BABA | Technology | 2014-09-19 | 544 |
| Blue Apron | APRN | Food/Drink | 2017-06-29 | 125 |
| Etsy | ETSY | Retail | 2015-04-16 | 104 |
| Facebook | FB | Technology | 2012-05-18 | 1977 |
| Ferrari | RACE | Automotive | 2015-10-21 | 46 |
| Fitbit | FIT | Technology | 2015-06-18 | 114 |
| General Motors | GM | Automotive | 2010-11-18 | 368 |
| GoPro | GPRO | Technology | 2014-06-26 | 91 |
| Groupon | GRPN | Retail | 2011-11-04 | 426 |
| LinkedIn | LNKD | Technology | 2011-05-19 | 193 |
| Shake Shack | SHAK | Food/Drink | 2015-01-30 | 103 |
| Snapchat | SNAP | Technology | 2017-03-02 | 185 |
| Stitch Fix | SFIX | Retail | 2017-11-17 | 33 |
| Tesla | TSLA | Automotive | 2010-06-29 | 94 |
| Twitter | TWTR | Technology | 2013-11-07 | 902 |
| Workday | WDAY | Technology | 2012-10-12 | 90 |

**Data Wrangling**

The team used a Jupiter Notebook to perform an internal join of the two initial datasets (headlines and IPO information), keying on date and IPO ticker. After that, much of the data wrangling was completed together with the ingestion as some data had to be processed and filtered to guide collecting the market price data from the Morningstar API.

The team speculated that the stature and circulation of the source publication might have a bearing on how much influence the article would have on stock prices. FP3 developed a code to distinguish first tier publications, such as the Wall Street Journal and Barron’s, from less well-known sources such as the Augusta Chronicle, defining four tiers based on readership size, saved as one of the wrangled features.

Team FP3 also created other features that could be used in the prediction. The headline publication date was compared to the IPO release date so the article could be characterized as either before the IPO or after it, also saved as an integer. The offer price, which the Company publishes with the IPO, was saved as well as the actual opening price (a response by the market), the price at closing on the first day, and the price after one month, three months, and six months. From these, change percentages were computed and stored as potential features. In addition to price changes features, the team identified the industry for each IPO and several of the managers involved in the IPO price valuation, publicity, and release. Once the sentiment scores were separated into columns for each category of sentiment (i.e. positive, negative, neutral, uncertain) the large list of all human-generated features included 52 items, some of which proved not useful as the project moved forward.

The “target” was defined as whether the stock price increased or decreased 90 days from the IPO release date, compared to the offer price.

**Sentiment Analysis**

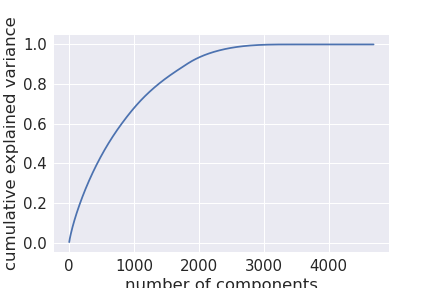
A key part of the data product was the sentiment analysis of the headlines about the IPOs. However, the team really struggled with this component of the data product stream. One team member identified a well-respected lexicon for financial sentiment analysis[[3]](#footnote-3) and the team expected this to be helpful in the sentiment analysis but found this not to be the case. Both sentiment analysis programs the team used seemed to work better with their internal lexicon o”r word-generation systems.

The Team exercised three different sentiment analysis algorithms – Empath[[4]](#footnote-4), OpinionFinder[[5]](#footnote-5), and a simple word count routine developed by a team member for preliminary testing. Initial results from Empath and the member-scripted word counter were not conclusive, and in fact, seemed not useful for the model. While working to understand possible problems with Empath, another team member started using OpinionFinder to generate sentiment scores. Work continued on both Empath and OpinionFinder in parallel to generate at least one set of useful scores so the work on feature selection and model testing could proceed while the sentiment analysis was corrected and refined. Eventually, the team guessed that perhaps Empath was struggling with the headlines because they are so much shorter than text chunks typically analyzed. Once more reasonable scores were obtained, higher numbers represented positive sentiment and lower numbers represented less positive (or negative) sentiment. In addition, it seemed that Empath worked better with its own built-in lexicon, rather than using that developed specifically for financial industry text analysis. FP3 eventually ran Empath with just a seed word instead of the custom financial lexicon and results seemed to improve. OpinionFinder, a trained algorithm, was also used with its own internal lexicon. (One team member also experimented with a Google sentiment analyzer, but this did not prove superior to either Empath or OpinionFinder).

**Feature Selection**

While some members worked on the sentiment analysis, others used existing acceptable sentiment scores[[6]](#footnote-6) to conduct preliminary runs of the variety of models available in Scikit-Learn. Tested models included LinearSVC (Support Vector Machine), NuSVC, SVC, KNeighbors, LogisticRegressionCV, LogisticRegression, SGDClassifier (Stochastic Gradient Descent), BaggingClassifier, ExtraTreesClassifier, RandomForestClassifier, and MultinomialNB (Naive Bayes). At first, all models gave F1 scores, measuring precision and recall, between 0.963 and. 1.0, which was deemed incorrect. Consultation with the Capstone Coordinator suggested the result was likely due to the fact that so many of the features being used related to the IPO rather than the distinctive headlines, creating too much similarity across the instances (there were only 16 distinct IPOs across more than 5000 instances).

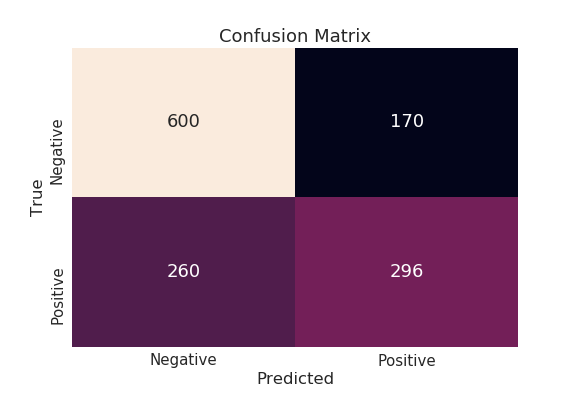
When the original features list was pared down to just headline text, date, sentiment scores, IPO ticker, IPO date and price differential after 90 days, the training F1 scores dropped below 0.96 for the first time. Subsequently, the team added CountVectorization to the processing of the headlines, which increased the number of features to over 4,600. Then, using principal component analysis (PCA) and k set to retain 95% of the variance, the number of features was reduced to 2100-2500 (depending on the testing run).

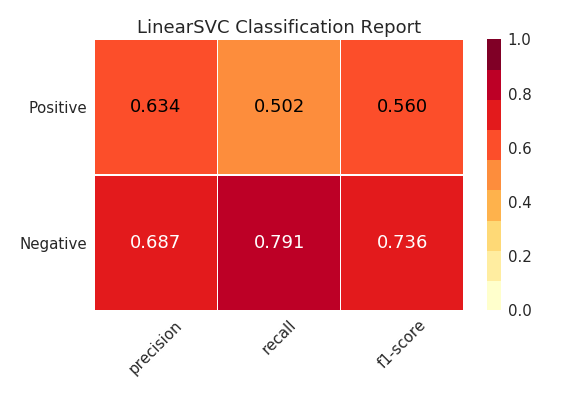


PCA analysis with k = 0.95

Using improved sentiment scores from both Empath or OpinionFinder, the F1 scores were 0.74 for LinearSVC, 0.57 for SVC and 0.65 for SGD, a distinct improvement though the team was still not confident the models were working as desired. The Random Forest results did not change. Of these models, the LinearSVC seemed the best as it was able to predict both positive and negative stock price changes over 90 days.

Once the sentiment analysis was refined and the team had more confidence in the scores, retaining the CountVectorization on the headline text, FP3 reran the PCA with the SVC, LinearSVC, SGD, and Naive Bayes models to finalize the features list and prediction algorithms. Again, LinearSVC provided the best predictions of both positive and negative price changes.





**Predictive Model**

FP3 focused attention on LinearSVC modeling for the 90-day IPO stock performance prediction based on the headline sentiment.

IPO Performance Predictor

1. **Predictor Release Date**: 30 June 2018
2. **Purpose**: To predict whether stock prices will rise or fall within the first three months after an IPO
3. **Methodology**: Assess the sentiment of news headlines about the subject IPO for the period leading up to and immediately after the IPO release (NLP = Empath and OpinionFinder). Then, using a classification-focused machine learning algorithm trained on 5500 headlines covering 16 IPOs between 2010 and 2017, and a LinearSVC model, predict whether the stock will rise or fall in the first 3 months.
4. **Required Input**:
   1. FOR THE IPO: IPO ticker, release date, and offer price
   2. FOR EACH HEADLINE: Headline text, date and IPO ticker
5. **Predictive Output**: "Up" or "Down" indicator of stock performance over first 90 days
6. **Intended Users**: The curious and casual.

**Model Demonstration**

To test the model prediction and demonstrate it during the Capstone Brief, FP3 collected additional IPO and headline data for two new IPOs: Dropbox and DocuSign, as given in the table below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Company | Ticker | Industry | IPO Date | Instances |
| Dropbox | DBX | Technology | 2018-03-23 | 299 |
| DocuSign | DOCU | Technology | 2018-04-27 | 68 |

The team encountered new problems after the new IPOs were captured for testing. In particular, some of the data in the first 16 IPO dataset was not captured for the next two IPOs. Even more serious, running the CountVectorizer and PCA on the newer dataset yielded a very different set of features from those gleaned for the first 16 IPOs, making it impossible to feed the dataset into the trained prediction model.

Through this, the team realized that there was no need to re-compute the CountVectorizer. The training vectorizer was the appropriate mechanism for processing the headline text into features. When the new headlines were processed this way, and all features were fed through the PCA and the prediction models, much better results were achieved. The results were comparable to those for the original, larger dataset except that for the two available IPOs, most of the sentiment scores were positive rather than being as mixed as the training set. Also, while the target was a 90-day price change trajectory, DocuSign has not yet been out that long so a shorter time-frame prediction was achieved and validated.

**Product Assessment**

What Worked – After the shift to examining IPOs, the team was fortunate to find enough headlines and a broad cross-section of IPOs. The data scraping and wrangling steps proved quite successful. And, after much trial and error, the team successfully analyzed the sentiment of the headline and article text for each news article about the IPOs. Both Empath and OpinionFinder provided similar results, though Empath used a built-in lexicon while OpinionFinder was exercised using the Loughran-MacDonald lexicon for financial text.

Challenges – It must be noted that the team originally embarked on a very different project, one focused on using mutual fund manager report text sentiment to predict mutual fund performance. The team pursued this project for almost three months before finally accepting defeat when it proved impossible to collect the text for enough manager fund reports. So, the first challenge is to get sufficient data of the kind required for the model. Once the team shifted focus to IPOs, the most persistent problems occurred around two things: running the sentiment analysis and feature selection. First, working with Empath for the first time, it took some time to realize the algorithm would work better using a built-in lexicon than with our financially-oriented lexicon. Fortunately, we were able to gain traction with OpinionFinder and the financial lexicon and when the problems with Empath were resolved, both sentiments scores were used in the model. Regarding feature selection, the features that seemed to make sense proved either irrelevant (type of publication) or too similar across instances (features specific to the IPOs). It was only through a process of trial and error that the team settled on the need for CountVectorization and PCA processing to achieve a productive set of features.

Tasks on the Back Burner – Team FP3 recognizes that the predictor needs a better user interface and more streamlined set of processes so users just enter an IPO name and the Data Product pulls headlines, IPO information, and price data from known and stable online sources. Then, the wrangling and modeling processes should also be streamlined and “bullet-proofed” with better exception handling routines as well as a better database management system.

Github Repo Location: <https://github.com/georgetown-analytics/MFP3>

Major work performed by team members (for email to Karen):

* Riley Back collected the initial 16 IPO headlines and associated data and performed preliminary cleaning. She also collected the two additional IPOs, along with the associated headlines and stock data, to be used for the Capstone demonstration brief.
* Mike Iapalucci identified and collected all stock price data for the IPOs and also created and exercised the main collection and wrangling Notebook which merged the headline and IPO data.
* Murali Kennan and Riley Back collaborated on the initial Empath sentiment analysis set-up. Murali continued with the Empath work, identified and fixed problems in the process and provided a final set of sentiment scores for each headline that the team had confidence in.
* MD Alam created a simple sentiment analyzer that counted lexicon words found in the headlines. He then exercised OpinionFinder to generate sentiment scores that could be used to refine the features list and were also used for comparison with those from Empath. He also conducted some regression analysis on the stock prices to see if they were sufficiently predictable to be independent of sentiment.
* Mike Iapalucci did the work running all the various models, refined the features list ,and identified the optimum model for the predictor.
* X, Y, and Z created visualizations used in the analysis and the briefing/report.
* Melissa Burn wrote and submitted all progress reports as well as the final report (submitted to the Team for comment and correction) and briefing slides.

1. The Team’s CCPE GitHub Repository can be found at: <https://github.com/georgetown-analytics/MFP3>. The team took the name MFP3 because the project originally focused on mutual fund performance. The team abandoned that effort due to insufficient data and pursued the IPO Predictor under the revised team name, FP3. [↑](#footnote-ref-1)
2. The Factiva website describes it as a,” global news database of nearly 33,000 premium sources, including licensed publications, influential websites, blogs, images and video. 74% of Factiva’s premium news sources are not available on the free web and thousands more are available via Factiva on or before the date of publication by the source.” URL: <https://www.dowjones.com/products/factiva/> accessed May 2018.

   . [↑](#footnote-ref-2)
3. Loughran, Tim and Bill McDonald. 2011. “When is a liability not a liability? Textual analysis, dictionaries, and 10-ks.” *The Journal of Finance*, 66(1): 3-65, referenced in Wang, Chuan-Ju, Ming-Feng Tsai, Tse Liu and Chin-Ting Chang. “Financial Sentiment Analysis for Risk-Prediction.” *International Joint Conference on Natural Language Processing*, pages 802-808, Nagoya, Japan, 14-18 October 2013. [↑](#footnote-ref-3)
4. Ward, Charles B., Yejin Choi, Steven Skiena and Eduardo C. Xavier. “Empath: A framework for evaluating entity-level sentiment analysis. Published in 2011 8th International Conference & Expo on Emerging Technologies for a Smarter World, 2-3 Nov, 2011. URL: <https://ieeexplore.ieee.org/document/6135866/> [↑](#footnote-ref-4)
5. OpinionFinder was developed by the University of Pittsburgh, Cornell University and the University of Utah. URL: <http://mpqa.cs.pitt.edu/opinionfinder/> [↑](#footnote-ref-5)
6. The “acceptable” set of sentiment scores consisted of those that at least gave an F1 score lower than 0.96 when run through the numerous models being tested. [↑](#footnote-ref-6)