Investor sentiment and the prediction of the IPO Underpricing - Based on the Random Forest-LSTM Model

Yanwei Pan 2020.05.26

Abstract

IPO underpricing has become one of the most famous market anomalies in the modern financial market. This phenomenon, which was called the "New Issue Puzzle", stimulated a hot debate within the economists and scholars. As many of them focused on explaining the cause of this anomaly and using linear regression to examine the issuer-based determinants of this phenomenon, I would like to stand on investors' side, busing investor sentiments to predict whether the IPO would be underpriced or not based on the behavioral finance theories. Besides, inspired by previous research, I would like to use a random Forests-LSTM model to make the prediction.

Key words: IPO; Underpricing; Random Forests; LSTM; Prospectus

1. Data

1.1. Dataset and Sample Construction

The data includes two categories- numerical/categorical data (direct variables) and the text data of IPO prospectus (indirect variables), both of which can reflect the investor sentiments. My first dataset gathered from IPOScoop.com[1], which is the list of IPOs in the US since 2000. This dataset contains 3486 IPOs, including their issue dates, issuer names, symbols (tickers), leading underwriters, offer prices and their first day closed prices. The second dataset is the google trend data that scraped from Google Trend website using Pytrends library in Python, which includes the daily searching trend of certain companies in the 31 days before the issue date. From Yahoo Finance, I scraped down and downloaded the industry/section data for each company and the

[1] https://www.iposcoop.com/scoop-track-record-from-2000-to-present/

Nasdaq monthly trading volume data. As Google Trend only provides the data from 2004, I obtained 2993 pieces of data that contain all of these direct variables. For the text data of IPO prospectus, I gathered the S-1 filings and Form 424 from the Electronic Data Gathering, Analysis, and Retrieval system (EDGAR) based on the Python script developed by javedqadruddin on Github[2]. Due to the comprehensive and different structure of the filings of companies, I only successfully extracted the text from 597 IPOs' prospectuses, which requires improvement in the subsequent research. The last dataset is the Loughran and McDonald (2013)[3] word lists[4]. As they have upgraded the wordlists in 2018, I used their 2018 version, which includes the wordlist of the uncertain, positive, negative, weak modal and strong modal.

Unlike Butler, Keefe, and Kieschnick (2014)[5] focused a lot on the company financial data, like their annual sales, EBITDA, etc. in their research, here I focus on the market data, which can reflect the general patterns of investor sentiments. Besides, based on the prospect theory, investors are loss aversion. Therefore, they care more about losses than gains. In other words, investors would pay more attention to the parts related to risks over other factors. Thus, the text data of the prospectuses in this research only contain the 'Risk Factors' part, which is quite different from previous research.

1.2. Data Summary Statistics

The summary statistics for the numerical variables are shown in Table 1. The statistics of market variables come from 2993 pieces of data from the IPOs in 2004/01/01 – 2020/05/15. And the statistics of text variables come from 597 IPO prospectuses. As we can from Table 1, the mean and median of first-day close price are higher than the mean and median of the offer price, reflecting that most of the IPOs in this sample are underpriced. In the 'Risk Factors' part in the prospectuses, the sentiment of texts is more likely to be negative and uncertain, which corresponds to our common sense.

---

[2] https://github.com/javedqadruddin/EDGAR

[3] Loughran, T., & Mcdonald, B. (2013). IPO first-day returns, offer price revisions, volatility, and form S-1 language. Journal of Financial Economics, 109(2), 307–326. doi: 10.1016/j.jfineco.2013.02.017

[4] https://sraf.nd.edu/textual-analysis/resources

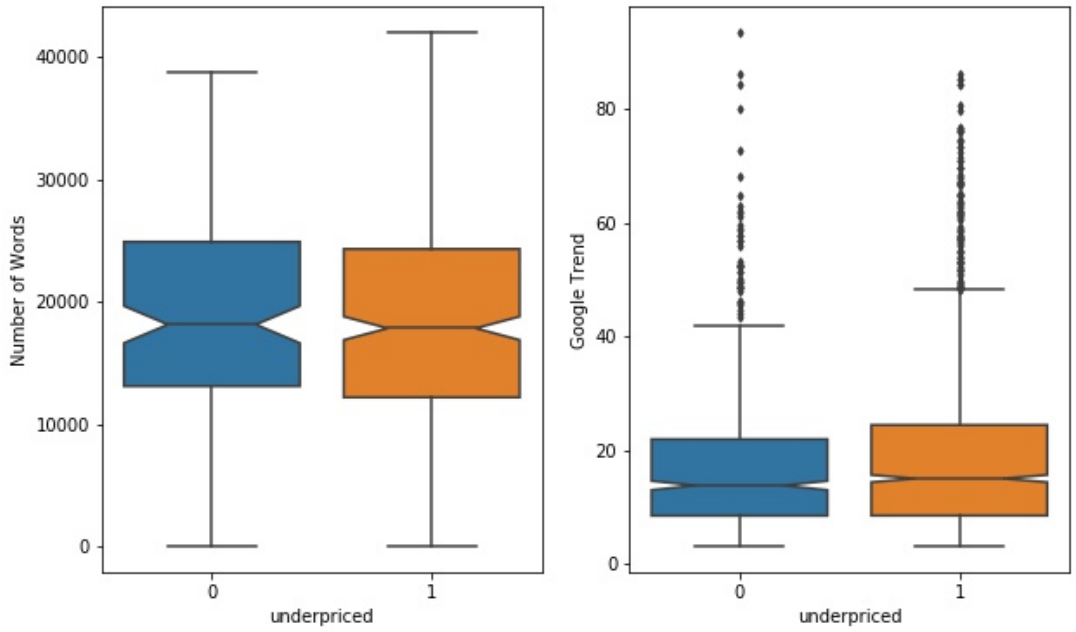
[5] Butler, A. W., Keefe, M. O., & Kieschnick, R. (2014). Robust determinants of IPO underpricing and their implications for IPO research. Journal of Corporate Finance, 27, 367–383. doi: 10.1016/j.jcorpfin.2014.06.002

Table 1

Variable Category	Variable	Mean	Std. dev	25th	Median	75th
Market Variables	Offer price	14.385991	6.328609	10.000000	14.000000	18.000000
	First-day close price	16.386545	9.220454	10.040000	14.600000	20.160000
	First-day price change	11.5732%	23.9400%	0.0000%	3.1350%	17.3975%
	Average daily searching trend	18.648337	15.049424	8.531250	14.625000	23.718750
	Monthly share volume	413.7920	79.0970	366.8631	401.5312	445.0541
Text Variables	Word count	19011	8493	12456	17982	24599
	Positive%	1.7747%	0.4317%	1.5133%	1.7353%	2.0162%
	Negative%	5.7203%	1.0809%	5.2144%	5.8367%	6.3629%
	Uncertain%	4.8847%	1.1878%	4.5235%	4.8639%	5.2438%
	Weak Modal%	3.6772%	0.6426%	3.4177%	3.7573%	4.0386%
	Strong Modal%	1.0030%	0.2309%	0.8614%	0.9889%	1.1305%

Figure 1 shows the box plots of word counts and average daily searching trends of underpriced and non-underpriced IPOs (0 = non-underpriced and 1 = underpriced). It demonstrates that there are differences between these two kinds of IPOs, and we can classify whether the IPO would be underpriced based on these variables.

Figure 1. Boxplot of Number of Words & Google Trend



When we see the top 20 industries that have the most IPOs from 2014 to 2020, shows in Figure 2, we can find that there are some industries that have a high proportion of underpriced IPOs, e.g. Software-Application, Banks-Regional, Software-Infrastructure, Internet Content & Information, etc. However, industries like Asset Management, REIT-Mortgage have a relatively low proportion of underpriced IPOs. The industries that are more likely to generate underpriced IPOs are related to the internet and high-tech, and the industries that seldom generate underpriced IPOs are related to financial area. Therefore, industry could be a good classifier for us to classify the underpriced IPOs.

Figure 2. Underpriced IPO Industry

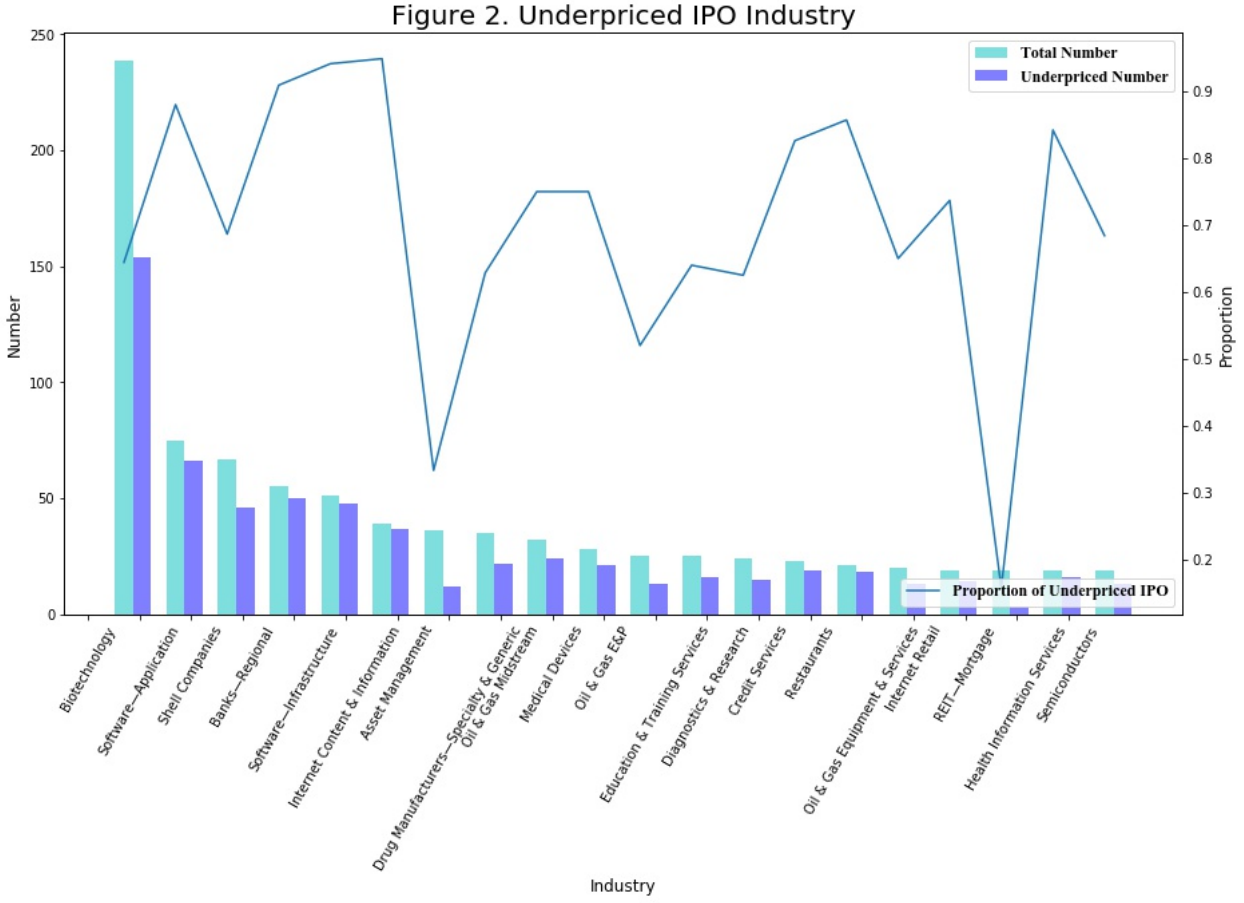
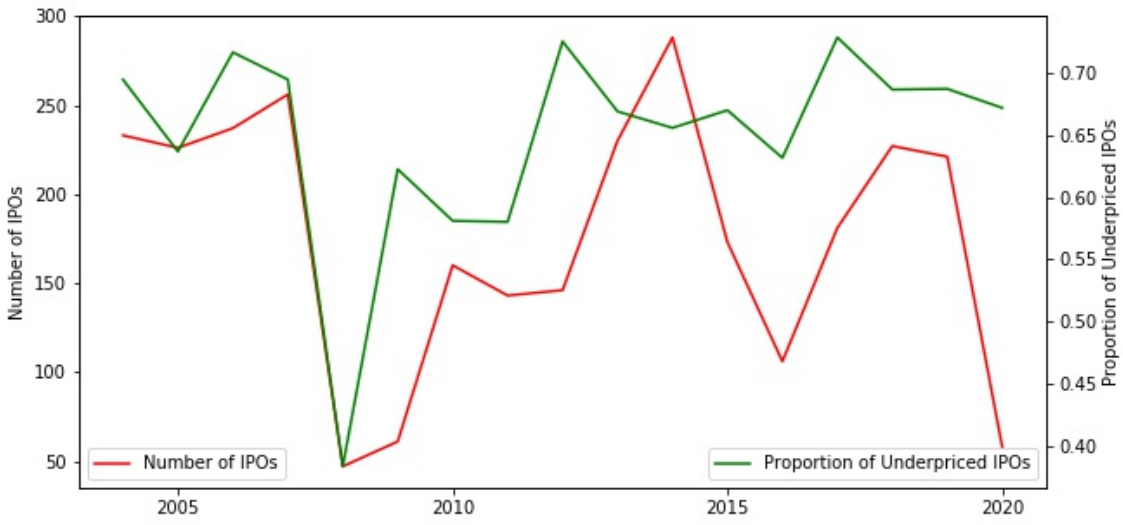


Figure 3 shows the change in the number of IPOs and the proportion of underpriced IPOs from 2004 to 2020 (the data in 2020 only includes the IPOs issued before May 15th, 2020). We can identify that there is a sharp decrease in about 2008 when the number of IPOs and the proportion of underpriced IPOs reached an extremely low rate. As in 2008, there was a serious financial crisis in the world, which stroked the financial markets all over the world, especially the US stock market. As the investors lost their confidence towards the stock market and became more conservative in

investment during this period, we could see there were less IPO underpriced. Thus, whether the IPO would be underpriced or not is closely related to investor sentiments, which can also be reflected by the share volume in the market.

Figure 3. Number of IPOs and Proportion of Underpriced IPOs in Years



2. Models

The models I used include decision tree, random forests, and the LSTM model. As I split the variables into two categories, the direct one and the indirect one, I would use a different kind of variables to fit the models and compare their results. The variables I used are shown in Table 2.

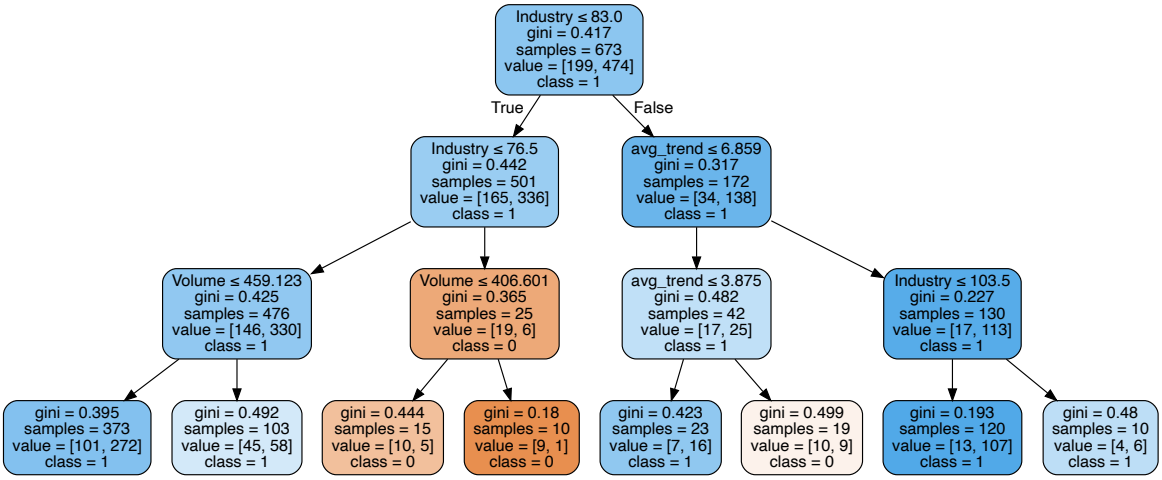
Table 2

Variable Category	Variable
Direct Variables	Average daily searching trend
	Industry
	Monthly share volume
Indirect Variables	Word count
	Positive%
	Negative%
	Uncertain%
	Weak Modal%
	Strong Modal%

A. Baseline Model – Decision Tree

Decision tree is a good baseline model for classification problems. It is relatively simple, using the entropy to decide the classification parameters and automatically adjust the data. In order to tune the hyperparameters, I did a random search for every model and used the optimal parameters to build the model and fit the data. Using all of the direct variables, I trained the following decision tree model (Figure 4):

Figure 4. Decision Tree



B. Random Forests

As a bagging approach, the random forest contains multiple trees, each of which uses a random subset of features to build a tree. It allows us to better explore the full set of possible predictors. For the sample with a large scale of features, random forests could have better prediction results. I also did a random search for every random forest model to tune the hyperparameters.

C. LSTM

Long-Short-Term Memory (LSTM) model improves the RNN model by adding a forget gate, which allows the long-term memory data (significant data that get from t-1) to go through the periods as well as the short-term memory (the input at t). This model allows us to train long sequential data. As we can regard a text as a sequence of words, we can apply it in training the text data. The basic formulas of LSTM are:

$$f_t = \sigma (W_f [h_{t-1}, x_t] + b_f)$$

$$i_t = \sigma (W_i [h_{t-1}, x_t] + b_i)$$

$$C'_t = \tanh (W_c [h_{t-1}, x_t] + b_c)$$

$$C_t = f_t C_{t-1} + i_t C'_t$$

$$o_t = \sigma (W_o [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t * \tanh (C_t)$$

In addition to using the text data to train the LSTM model and get the classification results (whether the IPO would be underpriced or not), I also used the similar method that Liu and Liu (2009)[6] used in their sales forecasting model- extracting the hidden layer results from the trained LSTM model, which can be considered as investor sentiments features and using those features combined with the existing variables to fit a random forest model.

3. Results

I used MSE as the measurement of the models, the MSE of each model is shown in Table 3. Among these models, using the direct variables and indirect variables to fit an LSTM-Random Forests model fit the best. However, it didn't improve a lot compared with other models.

Table 3

Model	MSE
Direct Variables	
Decision Tree	0.275556
Random Forest	0.273376
Direct Variables + Text Sentiment Variables	
Decision Tree	0.285450
Random Forest	0.251087
Pure Text	
LSTM	0.251812

[6] Liu, Y., & Liu, L. (2009). Sales Forecasting through Fuzzy Neural Networks. 2009 International Conference on Electronic Computer Technology. doi: 10.1109/icect.2009.65

Direct Variables + Indirect variables (Text Sentiment Variables + Variables extracted from LSTM hidden layer)	
LSTM + Random Forests	0.238289

This may due to the relatively less data I used while using the indirect variables to fit the model. As I only got 597 text data from the prospectuses, and I randomly split the training set and test set at the test size = 0.4. Thus, compared with only using direct variables to make the prediction, the data I used to train the LSTM-Random Forest model is not enough.

Besides, in the whole dataset (both the 2993 pieces of data with all of the direct variables and the 597 pieces of data that contains the text data), the number of underpriced IPO: number of non-underpriced IPOs = 3:1, which is unbalanced. Thus, the models would have relatively high accuracy while predicting the underpriced IPOs and low accuracy while predicting the non-underpriced ones. The result of using pure text to train the LSTM model shows this problem (Table 4).

Table 4

	precision	recall	f1-score
0	0.21	0.21	0.21
1	0.73	0.73	0.73
accuracy			0.60
macro avg	0.47	0.47	0.47
weighted avg	0.60	0.60	0.60

The prediction precision of underpriced IPOs is much higher than the non-underpriced IPOs. In order to solve this problem, I need to get more text data and rebalancing the sample.