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

Author: Yanwei Pan Adivisor: Dr. Richard Evans Date: Jun 1st, 2020

Contact Information
Email: panyw@uchicago.edu
Computational Social Science Program

INTRODUCTION

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. Many of them focused on explaining the cause of this anomaly and using linear regression to examine the issuer-based determinants of this phenomenon, whereas the most famous ones including the winner's curse (Rock, 1986) and the signaling model (Welch, 1989).

As behavioral finance came into our sight, some scholars tried to use it to study the underpricing, which gives me a lot of inspirations. In this paper, I would like to:

1. Stand on investors' side and use investor sentiments to predict whether the IPO would be underpriced or not based on the behavioral finance theories;

2. Combine text analysis approach to analyze the investor sentiments;

3. Use a random Forests-LSTM ensembled model to make the prediction.

DATA

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.

IPO detailed data (3486)- including issue dates, issuer names, symbols (tickers), leading underwriters, offer prices and their first day closed prices of IPOs in the US since 2000

IPO prospectus (597)- including the text of S-1 filings and Form 424 from the Electronic Data Gathering, Analysis, and Retrieval system (EDGAR)

Daily trend & Company industry- scraped from Google Trend and Yahoo! Finance

Table 1 Data Summary Statistics

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%

EDA & VARIABLES

Exploratory Data Analysis

These three figures demonstrate the differences between the underpriced IPO and the non-underpriced IPO.

As Fig 1 shows, the two groups of IPOs are different in the distribution of the number of words in their prospectuses and the average daily trend. Thus, word count and average daily trends can be good classifier.

Fig 2 shows that the underpriced ratio in different industries is significantly different. There are some industries having a high proportion of underpriced IPOs, e.g. Software-Application, Banks-Regional, etc. However, industries like Asset Management, REIT-Mortgage have a relatively low proportion of underpriced IPOs. Therefore, industry might be a good classifier.

Fig 3 demonstrates that there is a sharp decrease in about 2008 (during the financial crisis) when the number of IPOs and the proportion of underpriced IPOs reached an extremely low rate. This justifies that there might be a correlation between whether the IPO would be underpriced and the share volume in the market.

Figure 1. Boxplot of Number of Words & Google Trend

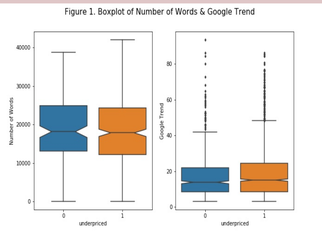


Figure 2. Underpriced IPO Industry

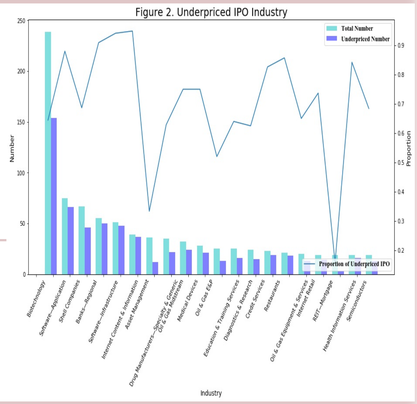
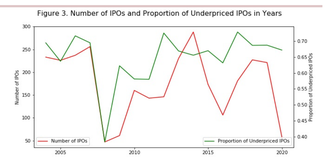


Figure 3. Number of IPOs and Proportion of Underpriced IPOs in Years



Variables

Below is the variables I selected after EDA (Table 2)

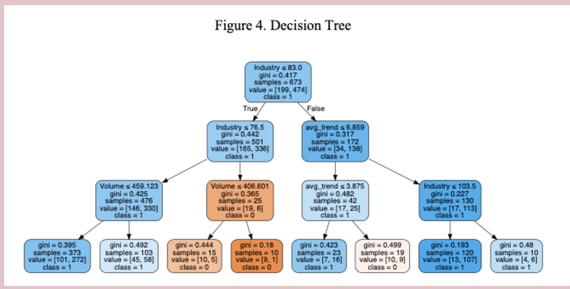
Table 2 Variable Description

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

MODELS

A. Baseline Model – Decision Tree

Figure 4. Decision Tree



B. Random Forests

C. LSTM

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f)$$
$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$$
$$C'_t = \tanh(W_c[h_{t-1}, x_t] + b_c)$$
$$C_t = f_t C_{t-1} + i_t C'_t$$
$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$$
$$h_t = o_t * \tanh(C_t)$$

RESULTS

Table 3 Model Prediction Summary

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

Table 4 LSTM Model Prediction Summary

	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

CONCLUSTION & FUTURE WORK

Among these models, using the direct variables and indirect variables to fit an LSTM-Random Forests model fit the best. Thus, we could predict the IPO whether it would be underpriced or not better using this new model. However, it didn't improve a lot compared with other models and 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.