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

Yanwei Pan

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, using investor sentiments, which related to behavioral finance, to predict whether the IPO would be underpriced or not based on the behavioral finance theories. This paper used four kinds of models to make the prediction: decision tree, random forest, LSTM and an ensembled model - Random Forests-LSTM model. After comparing the MSE of these models, we can find that the LSTM model has the best performance on prediction. The new ensembled model perform better than decision tree and random forest.

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

1. Introduction

IPO underpricing ("New issues Puzzle") has become one of the most famous market anomalies since Ibbotson (1975) documented in their research. Ibbotson used the data of new issue companies in the US from 1960 to 1969 and built up a time-series model to test the dependency of new issue premia, which is the first empirical analysis that justifies the IPO underpricing phenomenon in the US capital market. Ibbotson's research shows that the average underpricing rate is 11.4% among the sample data, which triggered researchers to explore the factors that cause this "puzzle".

Much work has been done in this exploration. As an IPO is associated with three groups of people: issuers, underwriters (usually are the investment banks), investors (including institutional investors and individual investors), most of the research that relates to IPO underpricing focus on these three perspectives. The most common one is focus on the underwriters' perspectives. That is to say, the majority of researchers considered the underpricing was caused by the low pricing. Logue firstly took this perspective. He built a linear regression model to explore the relationship between new issue performance (underpricing level relative to the market index) and competing issues variable, market ebullience variable, etc., as he considered the investment banker would underprice a new issue due to the consideration of minimizing his own cost, risks and gaining favors from issuers. Then, Baron utilized principal-agent theory to explain the IPO underpricing in 1982, which built a more generalized theory model in this perspective. His model indicated that, as the underwriters had more advantages in gathering information and resources in the stock markets, and they had more experience in issuing shares, they would underprice the new issues in order to ensure the success of issuing. As a result, the underpricing rate would be positively related to the risks of the new issue. Johnson and Miller (1988) specifically took Logue's consideration of investment banker's prestige to explore the impact of investment banker prestige on the IPO underpricing. They used two methodologies to measure the prestige level of each investment bank in their sample with a size of 196, including binary measurement and a four-point ranking scale. The empirical result shows that, the underpricing level is positively related to the prestige of investment banker, using the OLS regression. In 1989, Welch proposed to use a signaling model to explain the underpricing, which is one of the most well-known theory in explaining the "New Issue Puzzle". He assumed that there are two kinds of firms in the stock market, including low-quality firms and high-quality firms. As the low-quality firms tended to imitate the high-quality firms in their issuing

process, they would invest in imitation expenses. However, it was possible that their imitations would be discovered by the high-quality firms between offerings. In order to give sufficient signals to let the investors identify the high-quality firms, the high-quality firms would underprice at the IPO. Welch also used data to support his assumption. He used 1028 IPO firms in the 1977-1982 period as his sample, comparing the IPO data with the SO data and found that the empirical result was consistent with the implications of signaling model. Hanley (1993) also used signaling model to explore the IPO underpricing phenomenon. She considered the rate of final offer price to the range of anticipated offer prices that the firms disclosed in the preliminary prospectus as an efficient signal, which could affect the initial returns. She used the IPO data of 1,430 firms from January 1983 to September 1987, which were compiled from Investment Dealer's Digest Corporate Database, and built a regression model to examine the relationship. The result shows that the prices only partially adjust to new information, and the underwriters would like to increase the underpricing rate to compensate the ones who reveal the real information. Those researches only focused on the game between the issuers and underwriters, but not consider the investors. Among the research that stand on the underwriters' point of view, which considers that the underpricing was generated in the pricing process, a few of scholars also take the investors into account. The most famous theory in this perspective is known as the "Winner's curse", which was proposed by Rock in 1986. He built up an adverse selection model and assumed that there were two kinds of investors in the stock market, one is the informed investors who have more information of stocks, the other is the uninformed investors who were lack of information. If the stocks were overpriced, the uninformed investors, who played important roles in new issue process, wouldn't actively invest in the stocks. Therefore, in order to ensure the success of IPO, the underwriters would underprice the new issues. As Rock started to consider the impact of investors in the pricing process, Welch followed his step. In 1992, Welch established a herd behavior model to explain the underpricing. In the aspect of social psychology, individual behavior would be greatly affected by the collective behavior. Welch's theory model demonstrates that during the new issue process, the potential investors would pay attention to other investors' behavior, which would then affect their investment behaviors. In order to attract those potential investors to get involved and trigger the herd effect, the underwriters would underprice the new issues. Those researches, which focus on the underwriters' perspectives, give me lots of inspiration. As they only

consider the pricing process, it is possible that the underpricing is generated from the trading process.

A relatively novel perspective that the researchers take in this area is standing on the side of investors, which is the point of view I would like to focus on. As behavioral finance came into our sight, more and more researchers utilize relevant theories to explore the underpricing phenomenon. Jaggia and Thosar (2004) referred to the DHS theory (Daniel et al, 1998) to build up an ordered logit regression model and analyzed the high-tech IPO underpricing. Their sample included all IPOs from January 1, 1998, through October 30, 1999, in the specific sectors. They also used the Day 1 open price for each firm, which were collected from yahoo finance, as I will use in my own research. Their result shows that investors' overconfidence and biased self-attribution contributed to the underpricing in high-tech companies. However, they didn't provide much evidence that this model could apply to other sectors and how to generalize this model. In 2003, Ljungqvist, Nanda, and Singh did a simulation, which shows how could the sentiment investors behavior influence the IPO's first-day return. They modeled a firm that is going public in a "hot" IPO market and two types of investors: sentiment investors and rational investors, which then justify the underpricing phenomenon was triggered by those sentiment investors.

In addition to the theory models provided by the above research, some scholars focus on using OLS to explore the factors that could explain and predict the underpricing. Tian (2011) used some explicit variables to find the determinants of Chinese extreme IPO returns. She utilized the data of Chinese IPO to examine the factors that are significant in the regression model and found that although the asymmetric information about the quality of firms would cause the IPO underpricing, the effects of financial regulations played an important role in this phenomenon. Similar to Tian, Butler, Keefe, and Kieschnick did a research about robust determinants of IPO underpricing in the US IPO market in 2014. They examined the variables (more than 40) from previous studies and found that half of these variables were significant in the regression model. They gathered all the IPO data and the stock price data from 1981 to 2007. Their results demonstrated that the total share volume of the specific month would affect IPO issuing.

In recent years, many researchers begin to pay more attention to the impact of the IPO prospectus. They regard it as a significant factor that could influence investors' sentiment, which could indirectly cause the underpricing. Loughran and McDonald (2013) studied the relationship

between the first day returns of the IPO and the sentiments of the prospectus. They used the word lists they established in their previous research in 2011 to classify each prospectus in his sample (1,887 IPOs in the US with an offer price higher than $5 per share during 1997-2010) into uncertain, weak modal, negative, etc. And the result shows that the IPOs with uncertain text in their prospectus have higher first-day returns, which means as the uncertainty of the text increased, the underpricing is more severe. Their research gives me strong evidence that the prospectus could be a factor that causes the underpricing phenomenon and there could be a significant correlation between these two aspects. Then, Ly and Nguyen furtherly studied this relationship in 2020. They used different models, including the OLS regression model, random forest, decision tree, naïve Bayes, etc., to examine the impact of prospectus sentiments on IPO performance. They used similar sample data as Loughran and McDonald, which were both from the Electronic Data Gathering, Analysis, and Retrieval system. However, they used a wider sample, including all the IPOs in the US from 1975 – 2019. In addition, they not only considered the sentiments of the prospectus as the independent variables but also took the complexity of the text, count of characters, and other factors into account. Their result shows that the models they trained can predict the first-day price with an accuracy of up to 9.6% higher than chance. These two pieces of research only focus on the sentiments of the text in the prospectus but didn't explore the contents of the prospectus.

In order to explore deeply into the contents of the prospectuses, it is reasonable to use a deep learning model - the LSTM model in the prediction. This is a brand-new attempt in studying the causes of the IPO underpricing, basically, no researcher has used this model in this area. Inspired by Liu and Liu's (2009) research in sales forecasting, used an LSTM model to train the users' behavioral data, which led to a great prediction. Therefore, I would like to use LSTM-Random Forest model to explore the relationship between investor sentiments and IPO underpricing. I would take out the variables I get from the hidden layer of the LSTM model to make the prediction. Then, combining those variables with other market variables to train a random forest model to catch both general market sentiment patterns and specific sentiment patterns towards each issuer, which could be useful in making the classification prediction.

2. Data

2.1. 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 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 842 IPOs' prospectuses, which requires improvement in the subsequent research. The last dataset is the Loughran and McDonald (2013) word lists[3]. 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) 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.

2.2. 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,

[1] https://www.iposcoop.com/scoop-track-record-from-2000-to-present/
[2] https://github.com/javedqadruddin/EDGAR
[3] https://sraf.nd.edu/textual-analysis/resources

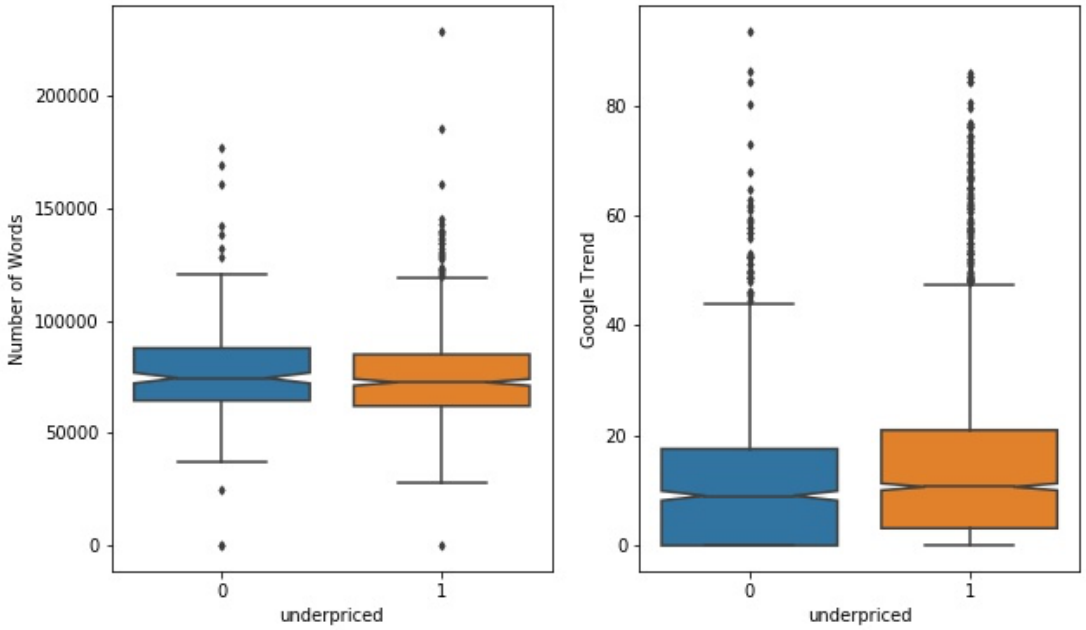
reflecting that most of the IPOs in this sample are underpriced. In the 'Risk Factors' part in the prospectuses, the sentiment of texts is more likely to be negative and uncertain, which corresponds to our common sense.

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	13.867183	15.275235	0.000000	9.515625	19.976562
	Monthly share volume	413.7920	79.0970	366.8631	401.5312	445.0541
Text Variables	Word count	78186.831354	41760.909835	62698.750000	73424.500000	87557.750000
	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.

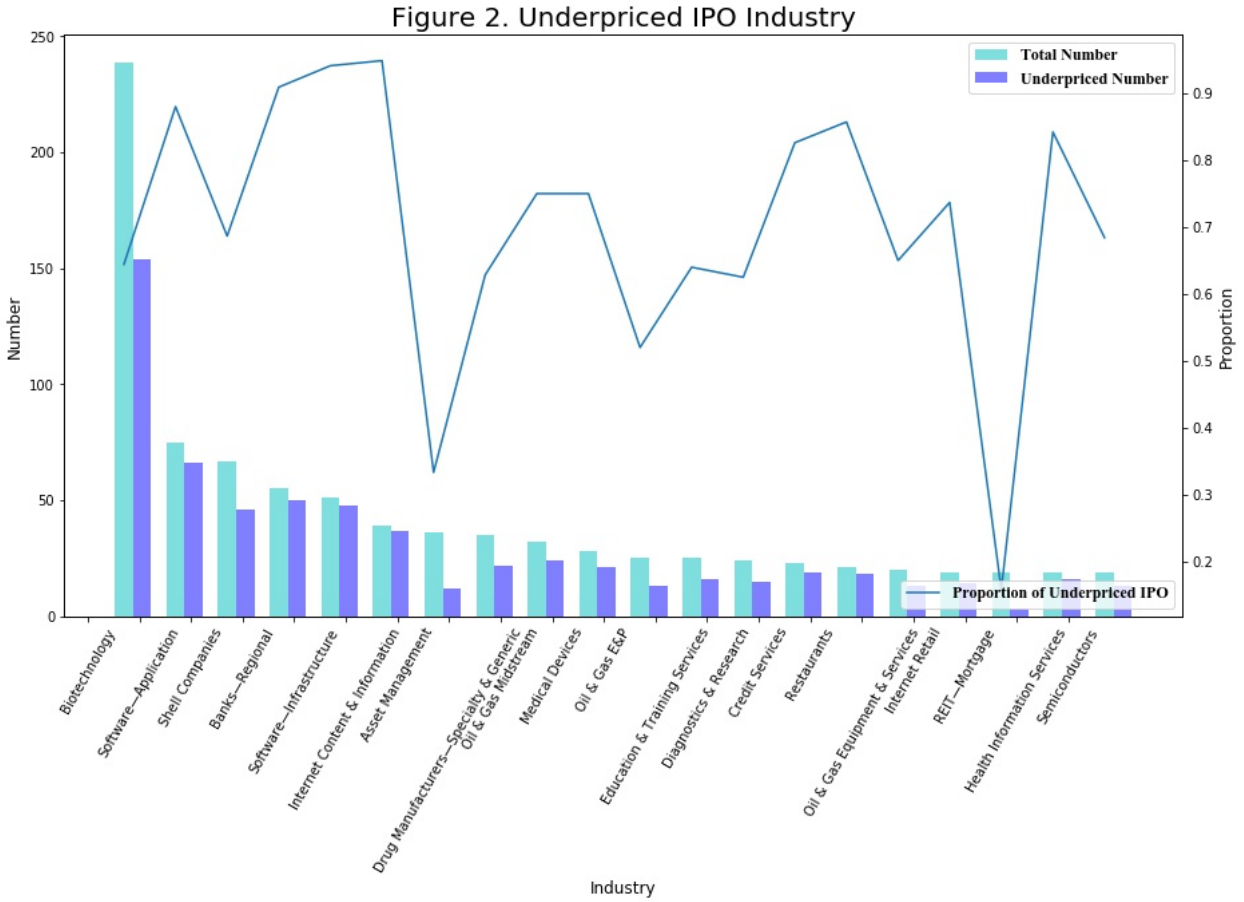
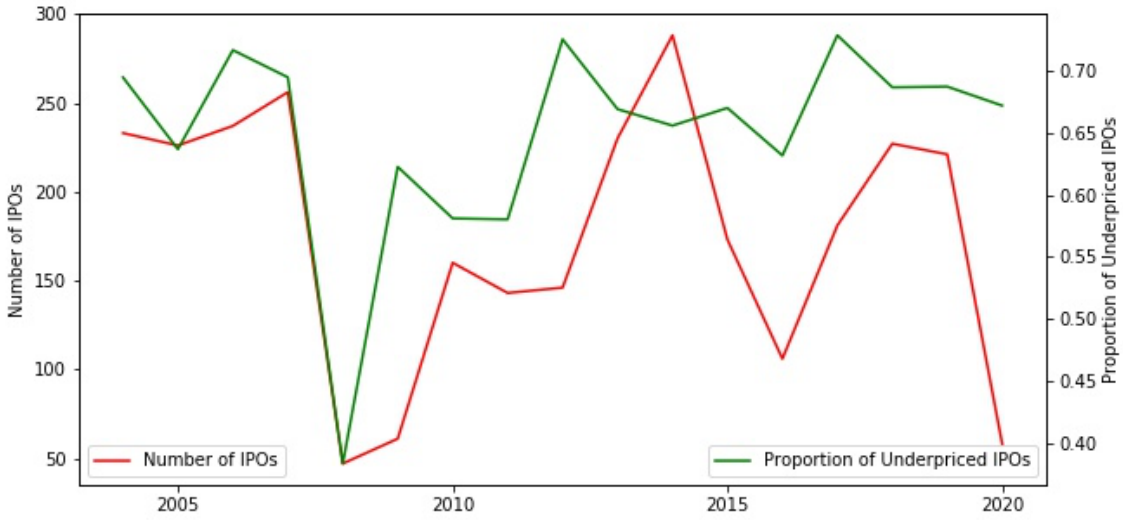


Figure 3 shows the change in the number of IPOs and the proportion of underpriced IPOs from 2004 to 2020 (the data in 2020 only includes the IPOs issued before May 15th, 2020). We can identify that there is a sharp decrease in about 2008 when the number of IPOs and the proportion of underpriced IPOs 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



3. 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.

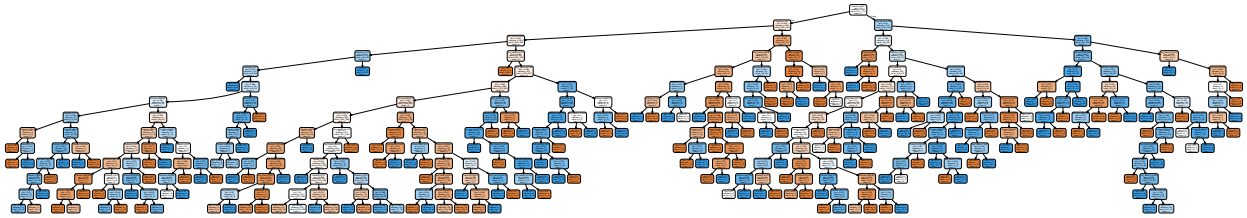
Table 2

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:

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

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

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

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

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

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

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) 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.

4. 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 better than the pure decision tree and random forest model. However, it didn't improve a lot compared with other models. And the new ensembled model doesn't perform better than LSTM model.

Table 3

Model	MSE
Direct Variables	
Decision Tree	0.369963
Random Forest	0.395604
Direct Variables + Text Sentiment Variables	
Decision Tree	0.389221
Random Forest	0.353293
Pure Text	
LSTM	0.284192
Direct Variables + Indirect variables (Text Sentiment Variables + Variables extracted from LSTM hidden layer)	
LSTM + Random Forests	0.359281

This may due to the relatively less data I used while using the indirect variables to fit the model. As I only got 842 text data from the prospectuses, and I randomly split the training set and test set at the test size = 0.2. 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 842 pieces of data that contains the text data), the number of underpriced IPO: number of non-

underpriced IPOs = 3:1, which is unbalanced. I used the SMOTE algorithm to balance the data in the model associated to the direct variables, but considering the data size of text, I didn't rebalance the indirect variables dataset, which may generate this result. 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	0.31	0.36	0.33
1	0.77	0.74	0.75
accuracy			0.64
macro avg	0.54	0.55	0.54
weighted avg	0.66	0.64	0.65

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 text sample.

5. Conclusion

The results of models show that combining the sentiment variables generated from the text of prospectus with the direct market variables can improve the accuracy of prediction while using decision tree and random forest model. This demonstrates the effectiveness of taking the text of prospectus into consideration in predicting whether an IPO will be underpriced or not and justified the correlation between the underpricing and the prospectus.

This relationship has been furtherly justified by using the LSTM model to train the text data and make the prediction. As we get a much lower MSE using this model, there is a close relationship between the text of prospectus and IPO underpricing. The content of the prospectus and the sentiment inclination would affect the issue process. This is reasonable because the IPO prospectus contains all the detailed information of the issuer, including its financial circumstance, risk factors, future strategy, etc., which is the most significant document that the investors concern about. Thus,

it would affect investors' valuation and expectations towards the issuer company, which can be summarized as the investor sentiments and then affect the first-day performance of the IPO.

While combining the LSTM model with random forests, we expect to get a better result over the other models. However, the result shows that this ensembled model only outperformed the decision tree and random forest but underperformed the LSTM model. This may due to the limitations of sample size and imbalanced issue in the sample.

In order to improve the robustness and effectiveness of the models, I need to gather more text data, balanced the sample, and try different LSTM models to predict. In addition, I can try to use more direct variables, like the underwriter ranking, to make the prediction. More data collection work needs to be done in the future.

References:

- Ibbotson, R. G., & Jaffe, J. F. (2012, April 30). "HOT ISSUE" MARKETS. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1975.tb01019.x
- Logue, D. E. (1973). On the Pricing of Unseasoned Equity Issues: 1965-1969. The Journal of Financial and Quantitative Analysis, 8(1), 91. doi: 10.2307/2329751
- Baron, D. P. (1982). A Model of the Demand for Investment Banking Advising and Distribution Services for New Issues. The Journal of Finance, 37(4), 955–976. doi: 10.1111/j.1540-6261.1982.tb03591.x
- Welch, I. (1989). Seasoned Offerings, Imitation Costs, and the Underpricing of Initial Public Offerings. The Journal of Finance, 44(2), 421–449. doi: 10.1111/j.1540-6261.1989.tb05064.x
- Hanley, K. W. (1993). The underpricing of initial public offerings and the partial adjustment phenomenon. Journal of Financial Economics, 34(2), 231–250. doi: 10.1016/0304-405x(93)90019-8
- Johnson, J. M., & Miller, R. E. (1988). Investment Banker Prestige and the Underpricing of Initial Public Offerings. Financial Management, 17(2), 19. doi: 10.2307/3665523
- Rock, K. (1986). Why new issues are underpriced. Journal of Financial Economics, 15(1-2), 187–212. doi: 10.1016/0304-405x(86)90054-1
- Welch, I. (1992). Sequential Sales, Learning, and Cascades. The Journal of Finance, 47(2), 695–732. doi: 10.1111/j.1540-6261.1992.tb04406.x
- Jaggia, S., & Thosar, S. (2004). The medium-term aftermarket in high-tech IPOs: Patterns and implications. Journal of Banking & Finance, 28(5), 931–950. doi: 10.1016/s0378-4266(03)00040-2
- Ljungqvist, A., Nanda, V. K., & Singh, R. (2003). Hot Markets, Investor Sentiment, and IPO Pricing. SSRN Electronic Journal. doi: 10.2139/ssrn.282293
- Tian, L. (2011). Regulatory underpricing: Determinants of Chinese extreme IPO returns. Journal of Empirical Finance, 18(1), 78–90. doi: 10.1016/j.jempfin.2010.10.004
- 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

- 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
- Ly, T. H., & Nguyen, K. (2020). Do Words Matter: Predicting IPO Performance from Prospectus Sentiment. 2020 IEEE 14th International Conference on Semantic Computing (ICSC). doi: 10.1109/icsc.2020.00061
- Liu, Y., & Liu, L. (2009). Sales Forecasting through Fuzzy Neural Networks. 2009 International Conference on Electronic Computer Technology. doi: 10.1109/icect.2009.65