IPO Underpricing and Prospectus Readability:
A Machine Learning Approach

Christophor Sai Kit Tsui
The Open University of Hong Kong

Kevin Chi Keung Li
The Open University of Hong Kong

IPO prospectus is the crucial document available to investors, allowing investors to understand the company and the IPO. IPO underpricing occurs when the closing price of the initial public offer (IPO) is higher than the offering price on its first trading day. If investors know whether the IPO is likely to be underpriced, they can earn a significant return by subscribing to those underpriced IPOs and selling the shares on the first trading day. In this study, the relationships between the readability of the IPO prospectus and IPO underpricing of firms listed in the Hong Kong Stock Exchange are analyzed using the gradient boost decision tree approach. This study shows that the readability scores of the chapters in the IPO prospectus are relevant to identify underpriced IPOs. Additionally, several indicators are more crucial to identify underpriced IPOs.

Keywords: IPO, underpricing, machine learning

INTRODUCTION

In 2019, Hong Kong was the No. 1 in the initial public offer (IPO) market globally, and about USD40 billion IPO equity funds were raised from the IPO market in Hong Kong (KPMG, 2020). The Hong Kong stock market is an attractive place for capital raising because the Hong Kong stock market listing requirements are more straightforward and have fewer regulations than other markets. Hong Kong is close to China, where the companies can easily attract investors in China. Thus, many leading companies such as Budweiser Brewing Co APAC, Prada have chosen their IPOs in the Hong Kong stock market.

Underpricing is a condition that the offering price of the shares in IPO is less than the true value of the shares. Singh (2019, p. 5) finds that the mean underpricing of IPO in Hong Kong was 14.46%, with a standard deviation of 0.44, and he also finds that IPO underpricing can be close to 100%.

IPO prospectus is the crucial document available to investors, allowing investors to understand the company and the IPO. Thus, the readability of the IPO prospectus will affect the motivation of investors to invest in the companies. If investors know whether the IPO is likely to be underpriced, they can earn a significant return by subscribing to those underpriced IPOs and selling the shares on the first trading day.

The objective of this study is to explore the IPO underpricing on the Hong Kong Stock Exchange using the IPO prospectus and attempts to model the relationship between the readability of the contents of the IPO prospectus and underpricing.

The organization of this paper is like this. Firstly, we will look into the theoretical background of IPO underpricing. Next, we will introduce the research methods and then, follow by the discussion of our results. Finally, a conclusion will be provided, and future research direction will also be provided.

LITERATURE REVIEW

Reasons for Becoming a Listed Company

Firms try to get a listing status due to numerous reasons, and here are just a few. The IPO market offers an excellent opportunity to firms to obtain additional funding for their expansion. For example, in 2014, the IPO of Alibaba Group Holding Limited raised USD21.77 billion from its IPO in the New York Stock Exchange, and it became the largest IPO in the global stock market (Weidner, 2014). Maksimovic and Pichler (2001) argue that the listing status can give customers, creditors, and suppliers confidence. The transfer of control of the company can also be seen in IPO as Black and Gilson (1998) argue that entrepreneurs can obtain controls of the firm from the hands of venture capitalists at the IPOs.

IPO Underpricing

Underpricing is a condition that the offering price of the shares in IPO is less than the actual value of the shares. Marcato, Milcheva, and Zheng (2018) point out that Hong Kong has a higher average IPO underpricing than in the United States. There are a lot of underpricing examples in the international stock markets. For example, the IPO of Alibaba Group Holding Limited in 2014 was underpriced by 38% on the first trading day (Weidner, 2014).

Prior studies had looked into the causes of underpricing. Rock (1986) argues that some underpricing is used to compensate uninformed investors who are reluctant to participate in IPO because of information asymmetry. In China, accounting conservatism and IPO underpricing are more pronounced when high information asymmetry (Lin & Tian, 2012). Marcato et al. (2018) argue that financial market integration is negatively related to IPO underpricing because the overall financial services are improved. Megginson and Weiss (1991) suggest a significant negative relationship between the firm's age and IPO underpricing since old-aged firms are more familiar with the investors. The need to compensate for information asymmetry is reduced.

The characteristics of IPO also affect underpricing. Fung, Gul, and Radhakrishnan (2014) suggest that when investment banks want to obtain business in the IPO market, the investment banks will offer high offer prices to the issuer, which will cause less underpricing. Carter, Dark, and Singh (1998) argue that underpricing is uncommon when reputable underwriters manage IPO because they want to protect their reputations. Interestingly, Elston and Yang (2010) find that venture capital has no effect on IPO underpricing.

The IPO prospectus is the most critical document to investors because it provides the key information of the company prospects, including the risk factors, use of proceeds, creditworthiness, etc. Therefore contents of the prospectus can affect underpricing. Park and Patel (2015) argue that IPO underpricing is low when the prospectus contains more information. Their view is supported by Nam, Arthurs, Nielsen, Mousa, and Liu (2009) where they have identified a negative relationship between information disclosure and IPO underpricing. Leone, Rock, and Willenborg (2007) suggest that when IPO issuers provide extensive disclosures on their intended use of proceeds, they will lower IPO underpricing because that will reduce the uncertainty of the IPO investors. Arthurs, Hoskisson, Busenitz, and Johnson (2008) have established positive relation between risk factors and IPO underpricing.

Readability Measurement

Readability is the measurement by which a reader can understand a written text. However, Loughran and McDonald (2014, p. 1643) point out that *'meant by "readability" is difficult to define precisely sure has evolved predominantly in the process of grade-leveling books, insurance contracts, and the understandability of instructions in military applications'.* This limitation can be overcome by using objective measurements of text difficulty using readability formulae (Saini, 2014).

This study uses Automated Readability Index (ARI) to measure the readability of the IPO prospectus. This index is designed for real-time monitoring of readability on electric typewriters (Senter & Smith, 1967). ARI is derived from ratios based on the number of letters per word representing word difficulty and the number of words per sentence describing sentence difficulty. This index produces an approximate representation of the US grade level needed to comprehend the document. ARI is more suitable to use computer programs to calculate because computers are good at counting characters. Other readability indexes such as Flesch–Kincaid readability index require that identifying the total number of syllables in the document is not easy for computer programs and can easily result in incorrect data for analysis. This index is commonly used in readability studies (see Saini, 2014).

Automated Readability Index formula:

$$4.71 \times (\text{no. of characters/ no. of words}) + 0.5 \times (\text{no. of words/ no. of sentences}) - 21.43$$

Table 1 identifies the ARI scores with corresponding US student grade levels.

TABLE 1
ARI SCORES WITH CORRESPONDING US STUDENT GRADE LEVEL

Score	Grade Level
1	Kindergarten
2	1st/2nd Grade
3	3rd Grade
4	4th Grade
5	5th Grade
6	6th Grade
7	7th Grade
8	8th Grade
9	9th Grade
10	10th Grade
11	11th Grade
12	12th Grade
13	Colleague Student
14+	Professor

Prospectus Readability and IPO Underpricing

Loughran and McDonald (2013) find that IPOs with difficult-to-read prospectus have higher first-day returns. In another study, Shette and Kuntluru (2018) find that in India, the first-year earnings performance of IPO firms with challenging to read prospectus is higher than that of firms with easy to read prospectus. These studies suggest that the readability of IPO has some relationship with underpricing. However, these studies used data outside Hong Kong, and the relationship of IPO underpricing and readability of IPO prospectus in Hong Kong listed companies are unexplored. This study tries to fill this research gap.

Structure of IPO Prospectus in Hong Kong

All companies listed in the Hong Kong Stock Exchange must make available a prospectus that complies with both the Companies (Winding Up and Miscellaneous Provisions) Ordinance and the Listing Rules of the Hong Kong Stock Exchange (Hong Kong Stock Exchange, 2015). The prospectus must be approved by the Hong Kong Stock Exchange before IPO.

A typical IPO prospectus of the Hong Kong listed company contains the following chapters(China Everbright Greentech Limited, 2017):

- Expected Timetable

- Summary
- Definitions
- Glossary of Technical Terms
- Forward-looking Statements
- Risk Factors
- Waivers from Strict Compliance with the Listing Rules and Exemption from Strict Compliance with the Companies (Winding Up and Miscellaneous Provisions) Ordinance
- Information About This Prospectus and the Offering
- Directors and Parties Involved in the IPO
- Corporate Information
- Industry Overview
- Regulatory Overview
- History and Corporate Structure
- Business
- Relationship with Controlling Shareholders
- Directors and Senior Management
- Substantial Shareholders
- Share Capital
- Financial Information
- Future Plans and Use of Proceeds
- Cornerstone Investors
- Underwriting
- Structure of the Offering
- How to Apply for Public Offer Shares
- Accountants' Report
- Unaudited Pro Forma Financial Information
- Unaudited Preliminary Financial Information
- Property Valuation Report
- Summary of the Constitution of the Company and Relevant Company Law
- Statutory and General Information
- Documents Delivered to the Registrar of Companies and Available for Inspection

Firms can provide other voluntary information in their prospectus, and firms can omit sections that do not apply to them.

RESEARCH METHODS

Research Objective

The main objective of this study is to assess the readability of the IPO prospectus of companies listed in the Hong Kong Stock Exchange from 2010 to 2017. Our second objective is to use the readability of the chapters in the IPO prospectus to predict the possibility of underpricing of the IPO on the first trading day.

Description of Observed Data

Our samples come from the IPO prospectus of firms listed in the Mainboard and the Growth Enterprise Market (GEM) of the Hong Kong Stock Exchange from 2010 to 2017. The prospectus is collected from the Hong Kong Stock Exchange website or the respective companies' websites. All prospectus are in Portable Document Format (pdf). We develop computer programs using Python 3.7.4 to convert the pdf files into plain text files, and the plain text files become the source data for this study. Since companies may have their prospectus chapters using different names or sometimes rearranging the contents into different

chapters, the researchers regroup the chapters into new chapters and calculate the ARI scores of each new chapter and the entire prospectus.

The descriptive statistics of the new chapter are shown in Table 2.

TABLE 2
DESCRIPTIVE STATISTICS OF READABILITY SCORES

Chapters	Min	Max	Mean	Standard Deviation
1. Accountants' report	16.74	43.43	23.5	3.17
2. Business	0.00	48.75	20.52	2.60
3. Connected transactions	0.00	54.92	17.98	12.01
4. Future plans and use of proceeds	10.02	116.58	28.42	9.95
5. History, organization and corporate structure	11.90	78.61	21.00	4.68
6. Industry	0.00	29.34	19.95	2.69
7. Profit forecast	14.67	61.60	7.58	13.77
8. Financial information	0.00	57.06	21.82	5.58
9. Underwriting	0.00	160.40	75.28	21.03
10. Property valuation report	0.00	25.72	6.55	7.17
11. Regulatory overview	18.19	58.91	27.40	3.49
12. Risk factors	16.25	28.47	19.98	1.53
13. Statutory and general information	10.95	106.22	31.00	6.26
14. Summary	16.03	36.61	22.55	2.89
15. Waivers from strict compliance with the listing rules	0.00	76.96	19.33	14.53
16. Overall	16.38	27.40	22.64	1.56

It is possible that firms did not provide the relevant information, such as waivers from listing rules in the prospectus. Therefore, these firms are still included in this study.

As shown in Table 2, the IPO prospectus of listed companies in Hong Kong generally have low readability. By referencing the ARI, the user of the IPO prospectus has to be at the 'professor' level of English to understand the contents. This is undesirable for investors to understand the firm before making their investment decision.

IPO Underpricing

The researchers collected the offering price and first-day closing price data of IPOs from 2010 to 2017 using the Bloomberg terminal. To calculate the IPO underpricing, the following formula is used:

$$\text{IPO underpricing} = \frac{\text{final IPO offering price} - \text{first-day closing price}}{\text{final IPO offering price}}$$

TABLE 3
DESCRIPTIVE STATISTICS OF IPO UNDERPRICING

	Min	Max	Mean	Standard Deviation
IPO underpricing	-91%	1,042%	-76%	4.48

For some firms, the first-day closing prices were less than their final offering price, leading to negative IPO underpricing.

Boosted Tree Regression

With the rapid development of machine learning, many new methods have been developed to improve the accuracy of classification problems. Gradient Boosting Decision Tree (GBDT) is a robust and accurate method in machine learning. It can cover different types of classifications and offer better performance than traditional classification methods (Friedman, 2001). The learning process in GBDT trains the decision trees using the results of the previous trees. In each round of processing, a new tree is created from the residuals of the last tree. The output of GBDT is the accumulation of the classification results from all the trees. There is a growing trend in using GBDT in the study of stock market behaviour. For example, GBDT has been used to detect insider trading in the stock market in China (Deng, Wang, Wang, & Sun, 2019).

In this study, the Gradient Boosting Decision Trees (GBDT) is employed to classify underpriced IPOs from non-underpriced IPOs using the ARI of the chapters in the IPO prospectus. Following Deng et al. (2019) approach, this study uses a Differential Evolution algorithm (DE) for parameter optimization. Storn and Price (1997) introduce DE which is a parallel direct search method that uses differences of random pairs of object vectors during the mutation process. Chandrasekar and Ramana (2012) argue that DE can produce better optimization results than other methods.

The researchers perform the following procedures to carry out GBDT in this study.

Step 1: Data Collection

Researchers collect IPO prospectus (in pdf format) of firms listed in the Hong Kong Stock Exchange from 2010 to 2018 from the Hong Kong Stock Exchange website or the listed firm's official websites. Researchers collect the final offering price and the closing price of the IPO's first trading day and determine the underpricing.

Step 2: Train the Identification Model

Researchers use DE to optimize the initial parameter of GBDT. GBDT is used to train the underpricing model using the first fifteen parameters in Table 2. This study applies the DE approach of Dang et al. (2019).

Stage 1 Initialization. The population of this study (P) of generation (G) contains N_p solution vectors that represent the potential solution of the optimization problem:

$$P_x^{(G)} = \left(X_i^{(G)} \right) = (x_{i,j}^{(G)})$$

where $i=0,1,\ldots, N_p\text{-}1$;
        $j=0,1\ldots D\text{-}1$;
        $G=0,1\ldots,g_{\max}$

Initialize the DE population for the starting point for an optimum search:

$$P^{(0)} = x_{i,j}^{(0)} = x_{j,L} + rand_j[0,1] \cdot (x_{j,U} - x_{j,L})$$

where $i=0,1,\ldots,N_p\text{-}1$;
        $j=0,1,\ldots D\text{-}1$;
        $rand_j[0,1]$ is a random number generated uniformly between 0 and 1

Stage 2 Determine fitness. For each sample in the population, a fitness value is calculated.

Stage 3 Mutation. A trial population of candidate vectors is generated from the DE recombination scheme.

Stage 4 Crossover. Trial vectors from the outputs of DE mutation are created by uniform crossover operation with a mutant vector:

$$u_{i,j}^{(G)} = \begin{cases} v_{i,j}^{(G)} & if \ \left(rand_{i,j}^{(G)} \le C_r \ or \ j = j_{rand}\right) \\ x_{i,j}^{(G)} & otherwise \end{cases}$$

where $C_r \in [0,1]$ is the crossover rate to control the mutant vector. A given uniform crossover compares C_r to $rand_{i,j}^{(G)}$ which is a uniform random number between 0 and 1. $V_i^{(G)}$ will be assigned to the trial element when the number is less than or equal to C_r. Otherwise, $x_i^{(G-1)}$ will be assigned to the trial element.

Stage 5 Selection. In this stage, solutions are selected from stage 4. If the trial vector $u_i^{(G)}$ has objective function value smaller than or equal to that of the target vector $x_i^{(G)}$, the trial vector becomes the target vector in the next generation. Otherwise, the target vector is kept in the population.

Stage 6 Termination. Stage 3 to 5 are repeated until a predetermined generation number is achieved or a termination condition is met.

The learning procedure of GBDT is that it iteratively constructs decision trees to create a more accurate estimate of the response variables. In each generation, a decision tree is trained from the residuals of the previous tree. The final result is the accumulated results from all trees. The optimal model of GBDT is:

$$F * (x) = F_0(x) + v * \sum_{t=1}^{m} p_t * h_t(x)$$

where $F(x)$ is a linear combination of individual decision trees
  v is the shrinkage parameter that controls learning rate (0<v<1)
  $h_t(x)$ is the tree trained in the tth iteration
  p_t is the weight of $h_t(x)$

For each sample, $F(x_i)$ is the classification of the ith decision tree, and L(y_i,F(x_i)) is the loss between them. The accumulated loss from L(yi,F(xi)) in the optimal model should be minimal.

Step 3: Performance Evaluation

Five-fold cross-validation is used to evaluate the GBDT result. This validation method is suitable to evaluate machine learning models with limited samples (Brownlee, 2018). Under this validation method, the data set is shuffled and divided into five equal-size groups (Liang, Luo, Zhao, & Wu, 2020). A single subsample is picked from a group as a validation set, and the other samples in that group are used for the training set. This procedure repeats until all groups have been picked for validation. The average accuracy of these five validation sets determines the average prediction accuracy of the machine learning model.

RESULTS

The five-fold cross-validation shows that the accuracy of our model is 65.56% which means our model is of high predictive power. GBDT result indicates that the ARI of the following chapters of the IPO prospectus are the essential factors for the identification of underpriced IPO in the Hong Kong stock market:
- Future plans and use of proceeds
- Waivers from strict compliance with the listing rules
- Financial information
- Underwriting
- History, organization and corporate structure
- Statutory and general information
- Connected transactions

These chapters are essential chapters that allow the investors to understand the IPO firms fully. If the ARI of the chapters in the prospectus is high (difficult to read and causes information asymmetric), rational investors are not willing to invest in the IPO. To compensate for information asymmetry, the offering price has to be lower than the true value of the shares and underpricing will occur. This study supports the argument of Rock (1986).

CONCLUSIONS AND FUTURE RESEARCH DIRECTION

In this study, data of 739 IPO prospectus of firms listed in the Hong Kong Stock Exchange, their final offering price and the closing price of the first trading day from 2010 to 2017 are collected as samples. A machine learning model that makes use of GBDT is created, which can successfully identify IPO underpricing from the ARIs of fifteen chapters from the IPO prospectus. From the results of GBDT, seven ARIs show higher importance than other ARIs. This suggests that they are dominant factors for the identification of underpriced IPOs. The high accuracy level of our GBDT model indicates that our model is relevant to identify underpriced IPOs from prospectus readability.

Machine learning has opened the door for academic research. This research can be extended by including financial data from the IPO prospectus into the model. This study can also be applied to other fields, such as the prospectus of debentures.

ACKNOWLEDGMENT

REFERENCES

Arthurs, J., Hoskisson, R., Busenitz, L., & Johnson, R. (2008). Managerial agents watching other agents: Multiple agency conflicts regarding underpricing in IPO firms. *Academy of Management Journal*, *51*(2), 277–294. doi:10.5465/AMJ.2008.31767256

Black, B.S., & Gilson, R.J. (1998). Venture capital and the structure of capital markets: Banks versus stock markets. *Journal of Financial Economics*, *47*(3), 243–277. doi:10.1016/S0304-405X(97)00045-7

Brownlee, J. (2018). *A Gentle Introduction to k-fold Cross-Validation*. Retrieved from https://machinelearningmastery.com/k-fold-cross-validation/

Carter, R.B., Dark, F.H., & Singh, A.K. (1998). Underwriter reputation, initial returns, and the long-run performance of IPO stocks. (includes appendix). *Journal of Finance*, *53*(1), 285. doi:10.1111/0022-1082.104624

Chandrasekar, K., & Ramana, N.V. (2012). Performance Comparison of GA, DE, PSO and SA Approaches in Enhancement of Total Transfer Capability using FACTS Devices. *Journal of Electrical Engineering & Technology*, *7*(4), 493–500. doi:10.5370/JEET.2012.7.4.493

China Everbright Greentech Limited. (2017). *Global Offering*.

Deng, S., Wang, C., Wang, M., & Sun, Z. (2019). A gradient boosting decision tree approach for insider trading identification: An empirical model evaluation of China stock market. *Applied Soft Computing Journal*, *83*. doi:10.1016/j.asoc.2019.105652

Elston, J.A., & Yang, J.J. (2010). Venture capital, ownership structure, accounting standards and IPO underpricing: Evidence from Germany. *Journal of Economics and Business*, *62*(6), 517–536. doi:10.1016/j.jeconbus.2010.08.003

Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist.*, *29*(5), 1189–1232. doi:10.1214/aos/1013203451

Fung, S., Gul, F., & Radhakrishnan, S. (2014). Investment Banks' Entry into New IPO Markets and IPO Underpricing. *Management Science*, *60*(5), 1297–1316. doi:10.1287/mnsc.2013.1817

Hong Kong Stock Exchange. (2015). *Prospectus*. Retrieved from https://en-rules.hkex.com.hk/sites/ default/files/net_file_store/new_rulebooks/h/k/HKEX4476_2435_VER10.pdf

KPMG. (2020). *Mainland China and Hong Kong 2020 Q1 review: IPOs and other market trends*.

Leone, A.J., Rock, S., & Willenborg, M. (2007). Disclosure of Intended Use of Proceeds and Underpricing in Initial Public Offerings. *Journal of Accounting Research*, *45*(1), 111–153. doi:10.1111/j.1475-679X.2007.00229.x

Liang, W., Luo, S., Zhao, G., & Wu, H. (2020). Predicting Hard Rock Pillar Stability Using GBDT, XGBoost, and LightGBM Algorithms. *Mathematics*, *8*(5). doi:10.3390/math8050765

Lin, Z.J., & Tian, Z. (2012). Accounting conservatism and IPO underpricing: China evidence. *Journal of International Accounting, Auditing and Taxation*, *21*(2), 127–144. doi:10.1016/j.intaccaudtax.2012.07.003

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

Loughran, T., & McDonald, B. (2014). Measuring Readability in Financial Disclosures. *Journal of Finance*, *69*(4), 1643–1671. doi:10.1111/jofi.12162

Maksimovic, V., & Pichler, P. (2001). Technological Innovation and Initial Public Offerings. *The Review of Financial Studies*, *14*(2), 459–494. doi:10.1093/rfs/14.2.459

Marcato, G., Milcheva, S., & Zheng, C. (2018). Market integration, country institutions and IPO underpricing. *Journal of Corporate Finance*, *53*, 87–105. doi:10.1016/j.jcorpfin.2018.10.002

Megginson, W.L., & Weiss, K.A. (1991). Venture Capitalist Certification in Initial Public Offerings. *Journal of Finance*, *46*(3), 879–903. doi:10.1111/j.1540-6261.1991.tb03770.x

Nam, D-i., Arthurs, J.D., Nielsen, M.L., Mousa, F.T., & Liu, K. (2009). Information Disclosure and IPO Valuation: What Kinds of Information Matter and is More Information Always Better? *Entrepreneurship & Finance eJournal*.

Park, H.D., & Patel, P.C. (2015). How Does Ambiguity Influence IPO Underpricing? The Role of the Signalling Environment. *Journal of Management Studies*, *52*(6), 796–818. doi:10.1111/joms.12132

Rock, K. (1986). Why new issues are underpriced. *Journal of Financial Economics*, *15*(1), 187–212. doi:https://doi.org/10.1016/0304-405X(86)90054-1

Saini, J. (2014). Estimation of Comprehension Ease of Policy Guides of Matrimonial Websites Using Gunning Fog, Coleman-Liau and Automated Readability Indices. *IUP Journal of Information Technology*, *10*(4), 19–33.

Senter, R.J., & Smith, E.A. (1967). *Automated Readability Index*. Wright-Patterson Air Force Base, p. iii, ARML-TR-6620.

Shette, R., & Kuntluru, S. (2018). Readability of Initial Public Offering Prospectus and Earnings Performance. *International Journal of Economics and Financial Issues*, *8*(3), 68–74.

Singh, V. (2019). *Modelling the First-Day Returns of an IPO - IPO Underpricing in Hong Kong*.

Storn, R., & Price, K. (1997). Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces. *Journal of Global Optimization*, *11*(4), 341–359. doi:10.1023/A:1008202821328

Weidner, D. (2014). *Opinion: This is where the Alibaba IPO went wrong*. Retrieved from https://www.marketwatch.com/story/this-is-where-the-alibaba-ipo-went-wrong-2014-09-23