DOI:10.18311/sdmimd/2023/33253

Prediction of IPO Subscription – A Logistic Regression Model

Ellur Anand[1*] **and Ganes Pandya**[2]
[1]Assistant Professor, Department of Business Analytics, Jagdish Sheth School of Management,
Bengaluru – 560100, Karnataka, India
[2]Associate Professor, Department of Business Analytics, Jagdish Sheth School of Management,
Bengaluru – 560100, Karnataka, India

Abstract

The main objective of this research paper is to apply logistic regression to estimate IPO subscription status in terms of oversubscription or under subscription. For this purpose, we used SMOTE (Synthetic Minority Over-sampling Technique) to generate minority class cases to rectify class imbalance problems and classification model logistic regression function to further classify the cases into majority class and minority class. KNIME (Konstanz Information Miner) and R Studio were used, as Integrated Development Environments (IDE), to develop the model. The results were quite encouraging with more than 90% accuracy levels for both training and testing datasets. The model was tested with different train-to-test ratios. The model and the results of the study can be used by firms and individuals involved in capital markets to predict the subscription status of a public offering. Further, there is ample scope to improvise the model by using different sets of variables and by applying different machine learning algorithms.

Keywords: Financial Analytics, IPO Subscription, Logistic Regression, Predictive Analytics, SMOTE

1. Introduction

IPO (Initial Public Offering) is a route used by most private companies planning to go public. Further, this is also a way to test out the credibility of a company issuing an IPO among the public, investors, and all stakeholders. IPO provides an investment opportunity to individual investors as well as institutional investors. Recent IPOs in India have created enough enthusiasm for new-age companies and startups to go for the public issue to establish and assess the financial markets. However, IPOs face challenges due to the issue of under-subscription. Further, predicting the subscription status of the IPO help investors and companies take appropriate investment decision like price and volume of investing. Hence, this research paper attempts to develop a prediction model based on the information available to investors about the firm in the advertisement of an IPO issue, particularly in print media. The subscription status of an IPO would result either in oversubscription or under-subscription. Oversubscription of an IPO refers to a situation where the demand for shares of that company by investors is more than the number of shares issued and advertised in print media. Undersubscription is exactly the opposite, a situation where demand for shares is less than the shares issued advertised by the company in print media. There are several research works which have used a different methodology to predict stock market subscription i.e., text analytics and investor sentiments Bi (2022), neural network Zhao (2021), or other machine learning techniques Fathali *et al*., (2022). This research paper has used Logistic regression to predict the categorical variable of whether an IPO will be over-subscribed or under-subscribed in this research paper. The logistic regression model is a supervised learning algorithm and classification model to predict the class of a dichotomous dependent variable.

This research paper attempts to predict the IPO subscription status using the dichotomous categorical variable ‘Subscription status’ as the dependent variable

*Email: ellur.anand@ifim.edu.in

in the logistic regression model. Further, several variables can be extracted from an advertisement of an IPO like Issue Size, Market Lot Size in several shares, Market Lot in Value (Rs.), Face Value, Minimum Price Band, Maximum Price Band, Qualified Institutional Buyers (QIB) Portion, Retail Portion, and Non-institutional Portion. The research paper has considered all these variables as independent variables to predict the dependent variable 'Subscription status'.

The developed model will be validated using upcoming IPO advertisements in print media as a test dataset. However, the number of IPOs that were undersubscribed was less than the number of IPOs that were oversubscribed leading to 'Class Imbalance'. The majority class in the dataset is 'Oversubscription' and the minority class is 'Under subscription'. When a minority class is less in number (less than or equal to 5% of total cases), then a situation of 'Class Imbalance' is created. Yet, there are several ways of balancing the dataset and managing the class imbalance. For instance, in this research work we use, Synthetic Minority Over-sampling Technique (SMOTE) to solve the 'Class Imbalance' problem. SMOTE generates an optimum number of minority class cases to balance the dataset. By addressing the class imbalance accuracy of the predictable model shall be improved. Improved predictability of subscription status can help individual investors and even small investors who would be investing in the stock market through the IPO route as beginners. Hence, in the process this research paper is trying to address the following key research questions:

RQ 1: What are the issue factors that significantly determine the IPO subscription status?

RQ 2: Does SMOTE work to rectify the class imbalance problem for IPO-related data?

RQ 3: What are the accuracy levels achieved in the diagnostics of the training dataset and testing dataset?

The research paper is divided into different sections. Section 2 covers the literature review briefly about IPO subscription and the usage of machine learning algorithms in developing classification models for stock market data. Section 3 explains the data collection methods and methodology adopted to execute and generate a logistic regression model for the IPO dataset. This provides metadata information about the features used in the model and analysis, with results discussed in detail. Further Section 4 explains the process of the development of a logistic regression model for the IPO dataset. Finally, Section 5 concludes by summarizing the research paper, mentioning the limitations of this research paper and the scope for future research.

2. Literature Review

The prediction of the status of the equity issues is highly uncertain. However, some mechanisms help analysts to manage and predict even in an uncertain environment. Krishnamurti & Kumar (2002) explored IPO subscriptions between 1992-94 of 386 IPOs. They have observed that Indian IPOs are underpriced, and they have attempted to identify the underlying causes for such behaviour. Arora & Singh (2020) have tried to explore various factors that affect the oversubscription of IPOs of SMEs in India. The authors have studied 403 IPOs from the SME sector which were issued between 2012 February to 2018 May listed on the Bombay Stock Exchange and National Stock Exchange. The study concludes that various factors like issue price, pricing mechanism, size of the firm, listing delay, under-pricing, and hot market impact the IPO subscription. Testing of the hypothesis has been carried out using ordinary least square regression and quantile regression. Liu *et al*., (2021) have carried out a similar study in China. They have used a regression model to understand how information asymmetry affects the under-pricing of IPO.

Media attention and tone in the media directly affect investor sentiment which in turn affects the under-pricing of IPO. Under-pricing or overpricing of an IPO has been studied extensively by different authors in different parts of the world, Wei & Marsidi (2019) and Xin-Er *et al*., (2020) have studied under-pricing of an IPO in the Malaysian stock market separately. Wei and Marsidi's study was conducted considering IPO under-pricing as the dependent variable and factors like issue price, offer size, age of the company and market capitalization as independent variables. Multiple regression was used to study the relationship between IPO under-pricing and the independent variables of 59 companies from 2012 to 2015. They have found that market capitalization has a positive relationship with the underpricing of an IPO. Mehmood *et al*., (2020) have studied the pricing mechanism for overpricing of IPOs in Pakistan. The authors have studied 85 IPOs in the period spanning 2000-2017 listed on Pakistan Stock Exchange. Ordinary least square, robust regression and quantile regression methods were deployed to test the hypothesis related to various factors impacting the IPO oversubscription. Liu *et al*., (2022) have studied the impact of the location of the companies on the under-pricing of IPOs in China. Authors have studied 1842 IPOs that were floated during the period 2005-2017. Authors argue that the companies which are located in rural areas or semi-urban areas will end up with higher under-pricing of an IPO as they will not get access to expert underwriters and support from investors. The location of the company is calculated as the distance from the company's headquarters to an urban centre. The authors have considered three urban centres for the study- Shangai, Beijing and Shenzhen. They have also developed propensity scores to rank the firms based on the impact the firms' location has on the under-pricing of an IPO.

2.1 Machine Learning and IPO Subscription

Machine learning models have supervised learning algorithms widely used for prediction. The supervised learning algorithms allow models to learn from the training dataset and validate the results with the testing dataset. Selvamuthu *et al*., (2019) have used machine learning algorithm neural networks to predict stock market data available in the form of tick data. The accuracy of 99.9% was observed using three different neural network algorithms namely, Levenberg-Marquardt, Scaled Conjugate Gradient and Bayesian Regularization methods. Since the collected dataset was having a case of 'Class Imbalance', the Synthetic Minority Oversampling Technique (SMOTE) was used to generate cases and balance the minority class and majority class. Chawla *et al*., (2002) were the first researchers to work on SMOTE to overcome the challenge of class imbalance due to relatively few cases of the minority class. In their research paper, the authors have explained the entire procedure of SMOTE. Gupta *et al*. (2022) carried out the sentiment analysis of media articles collected for a sample of 222 Indian IPOs between 2009-2018 and applied 'robust regression' for the collected data. Authors have also studied more than 2000 news articles to execute sentiment analysis. Singla (2021) has studied whether sentiment in the market and the ownership structure will affect IPO performance. The authors have used panel data and applied a systemic dynamic panel regression model to the dataset. The research mainly concentrated on construction sector IPOs in India.

Baba & Sevil (2020) have found that random forest provides better prediction over many machine learning algorithms like linear regression and that, as per the variable importance IPO proceeds and IPO Volume are the most important predictors. The results were like the ones this research paper found for the logistic regression model. The objective of the research paper is to identify the significant factors that affect the IPO subscription status and to develop a logistic regression model to predict the status of the IPO subscription in terms of 'oversubscription' or 'under subscription' of an IPO issue.

3. Data and Methodology

The analysis is conducted using the data of 245 IPOs listed companies in India between 1st January 2010 and 31st March 2022. The data has been collected using different secondary sources like Money control, Chittorgarh, IPO Ji, BSE SENSEX, and NSE websites. Moneycontrol.com has an 'IPO Historic Table' with the data for most of the parameters. Chittorgarh has various tabs to provide IPO information. The 'Subscription' tab provides data about the way different stakeholders have subscribed for.

It provides information regarding the percentage of shares bid by Qualified Institutional Buyers (QIBs), Non-Institutional Investors (NIIs), and Retail Individual Investors (RIIs). IPO ji, BSE SENSEX, and NSE websites provide similar information and were used to fill out missing data obtained by Moneycontrol.com and Chittorgarh.

The collected data was then scrutinized for relevant features for the study than performing feature analysis. The scrutiny was done by discussing with a few experts from the capital markets domain to understand how different numbers that appear in an IPO issue advert influence the listing price and subscription behaviour. The feature importance function was used in Python to check the importance of each of the features that were selected based on the discussion with the domain experts. The result of the feature importance function is provided below in Table 1.

Table 1. Feature importance values

Feature	Importance
Market Lot Value (Rs.)	0.1424
Non-institutional Portion (%)	0.1296
Retail Portion (%)	0.1043
QIB Portion (%)	0.0886
Issue Size (Cr.)	0.0386
Market Lot Size (Shares)	0.031
Floor Price	0.027
Cap Price	0.026

Source: Prepared by authors based on Python output

The importance index for Market Lot Value (Rs.) is the highest. Using the above features as independent variables and subscription type as the dependent variable, logistic regression was carried out to predict the class. The classes in the categorical dependent variable are “Oversubscribed” and “Undersubscribed”.

3.1 Metadata

- Subscription_type**:** Whether the subscription of IPO was oversubscribed or undersubscribed.

  Oversubscribed: When there is a huge demand for the new issue of a company than the number of shares that are made available to the public, then there is the situation of ‘Oversubscription’.

  Undersubscribed: When there is less demand for the new issue of a company than the number of shares that are made available to the public, then there is the situation of ‘Under subscription’.

- Face_value: It is the value at par or par value. It is a fixed price of an issue of a company. Usually, the face value of an issue will be Rs. 10, Rs. 5, or Rs. 100. The issue price is the sum of the face value of an issue plus the premium that the company charges on an issue.

- Issue Size_cr: It is the number of shares issued to the public multiplied by the price of a share.

- MinPrice_Band: A price band is the maximum and minimum limit of the share price which the company wants to sell to the public. The minimum Price Band is also known as the floor price of an issue.

- MaxPrice_Band: It is the maximum limit of a share price that the company is offering to the public in the form of an IPO. This is also referred to as the ceiling price of an issue.

- QIBPortion_Per: Qualified Institutional Buyers’ portion mentioned in percentage. QIBs are as defined by the Securities Exchange Board of India (SEBI). These are institutions that are eligible to invest in an IPO.

- RetailPortion_Per: Retail portion mentioned in percentage. The retail portion is the portion released for the public to buy shares of a company during the release of an IPO.

- NonInstiPortion_Per: Non-Institutional Investors’ Portion mentioned in percentage. These are investors who need not register with SEBI but are ready to invest more than 2 lakh rupees. Usually, this set comprises of HNIs, Societies, etc.

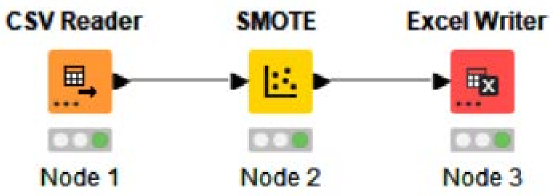
- MarketLotSize_Shares: This is the minimum order quantity an investor must buy if an investor wishes to participate in an IPO. This is the minimum order investor must bid for or in multiples thereafter.

- MarketLotValue_Rs: The Market lot value is the market lot size multiplied by the maximum band of an IPO issue. This is the lot value in Rupees which is held by the company before allotting shares in an apportioned manner if an IPO is oversubscribed.

4. Results and Analysis

The dataset collected comprised 245 IPOs listed companies in India between 1st January 2010 and 31st March 2022. The categorical dependent variable ‘Subscription_Type’ with two categories was having a case of ‘Class Imbalance’. The data comprised 225 IPOs which were ‘Over Subscribed’, the majority class, while only 20 IPOs were ‘Under Subscribed’, the minority class. The problem of class imbalance was rectified by using SMOTE – Synthetic Minority Over-Sampling Technique. The minority cases were less than 1% of the total number of cases in the dataset. SMOTE was carried out using KNIME software. A simple workflow shown in Figure 1 is used to generate an equal number of cases for minority and majority classes.

The new dataset generated by SMOTE consisted of 225 cases of the majority class and 225 cases of the minority class.



Source: Prepared by authors based on workflow in KNIME

Figure 1. SMOTE workflow.

Logistic Regression was used to develop a classification model, using R Software. The new balanced dataset after applying SMOTE was split into training and testing datasets using a 70:30 split ratio. The training dataset consisted of 316 cases and the testing dataset of 134 cases.

4.1 Logistic Regression

The result after applying logistic regression is shown in Table 2.

The glm() model in R provided the output with features Face_value, IssueSize_Cr, MarketLotSize_Shares, and MarktLotValue_Rs were all shown as insignificant as their p-value was greater than 0.05. Since these variables were considered after discussions with experts in this area, these features were not excluded from the logistic regression model. The importance () function in Python executed earlier had shown high importance index for these features. These features are equally significant or more significant than other features in influencing the listing price of the share as observed by experts.

In Table 2, for the Face value of a share, after adjusting for other features, the odds ratio is 1.023, with a 95% Confidence Interval (CI) being 0.973 and 1.086. The odds ratio is the ratio of successes to failures or the ratio of wins to losses. This implies odds of oversubscribing an issue increase by 2.3% for every unit change in the face value of the share. The odds ratio of all other

Table 2. Logistic regression output

Feature	Coefficient	Std. Error	p-value	Odds ratio	95% CI	
					2.5%	97.5%
Face_value	0.02	0.02	0.358	1.023	0.973	1.086
IssueSize_Cr	8.28E-05	2.046e-04	0.685662	1.000	0.999	1.0004
MinPrice_Band	-1.866e-01	4.110e-02	5.61e-06*	0.83	0.76	0.90
MaxPrice_Band	1.804e-01	4.038e-02	7.93e-06*	1.2	1.11	1.31
QIBPortion_Per	1.475	4.264e-01	0.000542*	4.371	2.112	11.26
RetailPortion_Per	2.988	8.059e-01	0.000210*	19.84	4.884	126.26
NonInstiPortion_Per	-2.231e-01	9.579e-02	0.019874*	0.8	0.797	1.105
MarketLotSize_Shares	2.105e-04	5.488e-03	0.969410	1.000	0.99	1.011
MarketLotValue_Rs	-3.100e-05	6.679e-05	0.642526	0.99	0.99	1.000

*Significance at 5% level

features can be inferred similarly. IssueSize_Cr, MarketLotSize_Shares, and MarketLotValue_Rs do not increase or decrease the odds of oversubscribing for a unit change in these features, as the odds ratio of all these three features is almost equal to one. MinPrice_Band and NonInstiPortion_Per features decrease the odds of oversubscription while MaxPrice_Band, QIBPortion_Per, and RetailPortion_Per increase the odds of oversubscription for a unit change in these features.

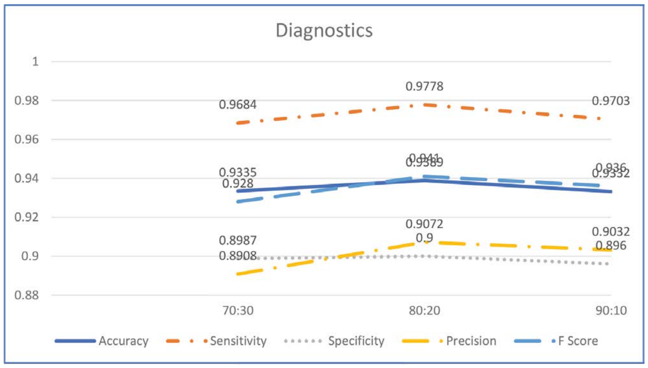
The diagnostic result for the logistic regression model executing the training dataset is shown in Table 3.

Figure 2 indicates that the 80:20 split ratio of training and testing data provides better diagnostics than other ratios. The diagnostic result for the logistic regression model executing the testing dataset with an 80:20 split ratio is provided below in Table 4.

Table 3. Diagnostics of the logistic regression model

Split Ratio Train: Test	Accuracy	Sensitivity	Specificity	Precision	F-measure
70:30	0.9335	0.9684	0.8987	0.8908	0.928
80:20	0.9389	0.9778	0.9000	0.9072	0.941
90:10	0.9332	0.9703	0.8960	0.9032	0.936

Source: Prepared by authors based on R Studio output



Source: Prepared by authors based on R Studio output in MS Excel

Figure 2. Diagnostics.

Table 4. Diagnostics for the test dataset

Accuracy	Sensitivity	Specificity	Precision	F-measure
0.9254	0.9701	0.8806	0.8904	0.929

Source: Prepared by authors based on R Studio

The results of the testing dataset are encouraging and are no different from the training dataset. The accuracy of the model has decreased by 1.44% only, for the testing dataset.

5. Conclusion

The main purpose of this research paper is to predict the subscription behaviour of IPOs. The subscription type that an IPO will be undersubscribed or oversubscribed was a dichotomous variable. A logistic regression[v] model was deployed as the dependent variable was binomial by nature. A logistic regression model was developed using the data from 245 IPOs. The results of the analysis to develop a logistic regression model found certain features significant – Maximum of the Price Band, Minimum of the Price Band, Percentage of QIB investors, Percentage of NIIs, and Percentage of RIIs. The accuracy levels for both training and testing datasets were greater than 90%. The model was verified with different split ratios of 70:30, 80:20, and 90:10 for training and testing datasets respectively. The results were best for the 80:20 split ratio than the other two. The limitation of this research paper is that the data collected is limited to the advertisement of IPO in print media and does not consider the information provided in Draft Red Herring Prospectus (DRHP) released by the company before the IPO issue. The model was tested using only one of the classification methods. A lot of scopes exist for future research as model accuracy could increase if the information provided in DRHP is processed instead of using the information provided only in issue advert in print media. Assessment of the performance of the stock post-listing day is another area of study regarding IPOs of new-age companies. There is tremendous scope and space to conduct variations to this model. Future research can try applying neural networks to predict the subscription type of any new issue to the public. Decision trees can also be applied to the dataset. Instead of predicting whether an issue will be oversubscribed or undersubscribed, the research paper can try to predict the number of times an issue could be oversubscribed or to predict the listing day opening or closing price of an issue.

6. Acknowledgments

The authors would sincerely thank the reviewer(s) for their valuable input and guidance at all stages. We sincerely thank the reviewer(s) for reviewing so meticulously that there was great learning concerning how a research paper must be written.

EndNotes

1 Konstanz Information Miner – It is called as KNIME in short. It is a workflow-based software to execute machine learning models. It has GUI interface where the nodes can be used in drag/drop format. It is an open-source and free software. It can also be easily linked to other platforms like Python and R programming software. It is easy to connect to social media websites like Twitter through API to collect data.

2 Random Forest – Random Forest is an advanced version of Decision Tree algorithm. Random Forest is highly flexible because of two reasons: a) it can handle both classification as well as regression models and b) it develops multiple decision trees and selects the one which provides the best result.

3 Python Programming – This is the most widely used free, open-source software for building machine learning models. There are packages to execute machine learning models and statistical functions.

4 R programming – This is widely used as statistical software and as a software to build machine learning models. R program too has packages like Python Programming language.

5 Logistic Regression Model – Logistic regression model is a classification model. It helps to predict the class of new cases based on the past data. Past data is used to develop the model using the training dataset. Later, the model is tested with the testing dataset.

7. References

Arora, N., & Singh, B. (2020). Determinants of oversubscription of SME IPOs in India: Evidence from quantile regression. Asia-Pacific Journal of Business Administration, 12(3/4), 349-370. https://doi.org/10.1108/APJBA-05-2020-0160

Baba, B., & Sevil, G. (2020). Predicting IPO initial returns using random forest. Borsa Istanbul Review, 20(1), 13-23. https://doi.org/10.1016/j.bir.2019.08.001

Bi, J. (2022). Stock market prediction based on financial news text mining and investor sentiment recognition. Mathematical Problems in Engineering, 2022, 1-9. https://doi.org/10.1155/2022/2427389

Chawla, N. v., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357. https://doi.org/10.1613/jair.953

Fathali, Z., Kodia, Z., & ben Said, L. (2022). Stock market prediction of NIFTY 50 index applying machine learning techniques. Applied Artificial Intelligence, 36(1). https://doi.org/10.1080/08839514.2022.2111134

Gupta, V., Singh, S., & Yadav, S. S. (2022). The impact of media sentiments on IPO underpricing. Journal of Asia Business Studies, 16(5), 786-801. https://doi.org/10.1108/JABS-10-2020-0404

Krishnamurti, C., & Kumar, P. (2002). The initial listing performance of Indian IPOs. Managerial Finance, 28(2), 39-51. https://doi.org/10.1108/03074350210767681

Liu, L., Neupane, S., & Zhang, L. (2022). Firm location effect on underwriting, subscription, and underpricing: Evidence from IPOs in China. Economic Modelling, 108, 105778. https://doi.org/10.1016/j.econmod.2022.105778

Liu, L., Zhang, Z., & Lyu, K. (2021). A study of IPO underpricing using regression model based on information asymmetry, media, and institution. Advances in Economics, Business and Management Research. https://doi.org/10.2991/aebmr.k.210917.051

Mehmood, W., Mohd-Rashid, R., & Ahmad, A. H. (2020). Impact of pricing mechanism on IPO oversubscription: Evidence from Pakistan stock exchange. Pacific Accounting Review, 32(2), 239-254. https://doi.org/10.1108/PAR-04-2019-0051

Selvamuthu, D., Kumar, V., & Mishra, A. (2019). Indian stock market prediction using artificial neural networks on tick data. Financial Innovation, 5(1), 16. https://doi.org/10.1186/s40854-019-0131-7

Singla, H. K. (2021). Do ownership structure and market sentiment affect the performance of IPOs in India in the short run? A dynamic panel data analysis. Journal of Financial Management of Property and Construction, 26(1), 1-22. https://doi.org/10.1108/JFMPC-10-2019-0077

Wei, F. J., & Marsidi, A. (2019). Determinants of Initial Public Offering (IPO) underpricing in malaysian stock market. International Journal of Academic Research in Business and Social Sciences, 9(11). https://doi.org/10.6007/IJARBSS/v9-i11/6657

Xin-Er, C., Sin Huei, N., Tze San, O., & Boon Heng, T. (2020). Underpinning theories of IPO underpricing. Evidence from Malaysia. International Journal of Asian Social Science, 10(10), 560-573. https://doi.org/10.18488/journal.1.2020.1010.560.573

Zhao, Y. (2021). A novel stock index intelligent prediction algorithm based on attention-guided deep neural network. Wireless Communications and Mobile Computing, 2021, 1-12. https://doi.org/10.1155/2021/6210627