A COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS FOR IPO UNDERPERFORMANCE PREDICTION

Pravinkumar M Sonsare[1*†], Ashtavinayak Pande[1], Sudhanshu Kumar[1], Akshay Kurve[1], Chinmay Shanbhag[1]

[1] *Shri Ramdeobaba College of Engineering and Management, Nagpur*

Abstract

Initial Public Offerings (IPOs) are a popular way for companies to raise capital and enter the public markets. However, many IPOs underperform and fail to meet the expectations of investors. In this research paper, we explore the use of different machine learning models, namely AdaBoost, Random Forest, Logistic Regression, ANN and SVM for predicting IPO underperformance. We collect and pre-process a dataset of IPOs from the past few years and use it to train. We evaluate the performance of each model. Our results show that Artificial Neural Network model is better suited for predicting IPO underperformance. Additionally, our analysis provides insights into the factors that contribute to underperformance and highlights the importance of certain features in predicting IPO performance. Our research provides valuable information for investors and financial analysts interested in predicting the performance of IPOs and mitigating the risks associated with IPO investments. We have tested machine learning models, namely AdaBoost, Random Forest, Logistic Regression, ANN and SVM. After Comparing the accuracy of all the models, we arrived at the conclusion that ANN model performed the best with an accuracy of 68.11%.

Keywords: Initial Public Offerings (IPOs), Machine learning, AdaBoost, Logistic regression, Support Vector Machines (SVM), Financial analysis, Supervised learning Classification

1 INTRODUCTION

Initial Public Offerings (IPOs) are an important way for companies to raise capital by offering their shares to the public. However, the success of IPOs is not guaranteed, as many newly issued stocks experience underperformance shortly after going public. This underperformance can

*Corresponding author.
†E-mail: sonsarep@rknec.edu*

have significant implications for investors, underwriters, and the broader economy.Traditional methods of analysing IPOs rely heavily on financial and accounting metrics. However, in recent years, machine learning algorithms have shown promise in predicting stock performance. By leveraging the power of these algorithms, it may be possible to develop more accurate models for predicting IPO underperformance.

Our work aims to investigate the use of all machine learning algorithms for predicting the performance of IPO. We will explore the potential of various machine learning techniques such as regression models, decision trees and neural networks in predicting IPO underperformance. We will also examine the role of different variables in predicting underperformance such as financial metrics, market trends, and company-specific factors. The prediction of IPO underperformance has largely relied on traditional financial and accounting metrics such as p/e (price to earning) ratio, earnings per share and market capitalization [4] [5]. These metrics can provide valuable insights. They may not capture all of the factors that influence IPO performance.

This article will begin with a review of relevant literature on IPO underperformance and machine learning in finance. We will then describe the data used in our analysis, including financial and market data for a sample of IPOs. We will then apply various machine learning algorithms to this data, evaluating their performance in predicting underperformance.

Overall, the findings of this research could have important implications for investors, underwriters and regulators. By developing more accurate models for predicting IPO underperformance. We may be able to improve the efficiency and stability of the IPO market.

2 DATA PREPARATION

The data used for predicting IPO performance typically includes a range of financial and non-financial variables related to the IPO issuer such as historical financial performance, company size, industry sector, market conditions, and underwriter reputation. We have splitted the pre-processed data in a ratio of 70:30 into training and testing sets. We have considered historical financial ratios such as price to earnings (P/E) ratio, earning per share (EPS) and return on equity (ROE). We also trained company-specific characteristics such as industry classification, management experience, and ownership structure. Market conditions at the time of the IPO, such as stock market performance and interest rates are also considered. Information related to the IPO underwriter, such as reputation and track record also played an important role.

The data can be sourced from a variety of public and private databases, such as financial statements, prospectuses, and company filings. In some cases, additional data may be gathered

through surveys or expert opinions. It is important to note that the quality and relevance of the data can have a significant impact on the accuracy and generalizability of the IPO prediction models. Careful data pre-processing, feature engineering, and validation procedures are necessary to ensure the models are robust and reliable.

3 PROPOSED METHODOLOGY

We pre-processed the data by first handling missing values. For numerical variables, we have replaced the missing values with the mean value of variable, and for categorical variables. We replaced the missing values with the mode of the variable. We then removed outliers using the interquartile range (IQR) method. The value that falls below Q1 -1.5IQR or above Q3+1.5 IQR were defined as outliers where Q1 and Q3 are the first and third quartiles, respectively. Outliers were replaced with the nearest non-outlier value.

We intend to utilize classification algorithms to predict IPO under-pricing and compare the techniques based on their predictive ability. We also aim to identify variables that are significantly better predictors for IPO under-pricing. Accordingly, we selected the following set of methods to analyse our data - logistic regression, decision trees and ensemble techniques like bagging, random forest, gradient boosting as shown in Figure 1.

4 PREDICTIVE MODELS

4.1 Gradient Boosting

It is a one of algorithm used in predicting stock performance. However, when working with imbalanced datasets, where one class is significantly underrepresented compared to the other. The algorithm may struggle to identify the minority class. One common technique used to address this issue is resampling. The dataset is balanced artificially where the minority class is oversampled or the majority class is under sampled [9].

In the context of predicting IPO underperformance, resampling the minority class may help to improve the accuracy of Gradient Boosting models. The model may be better able to identify patterns and relationships that are specific to underperforming stocks by artificially increasing the number of underperforming IPOs in the dataset. Table 1. summarizes the predictive performance of Gradient Boosting model.

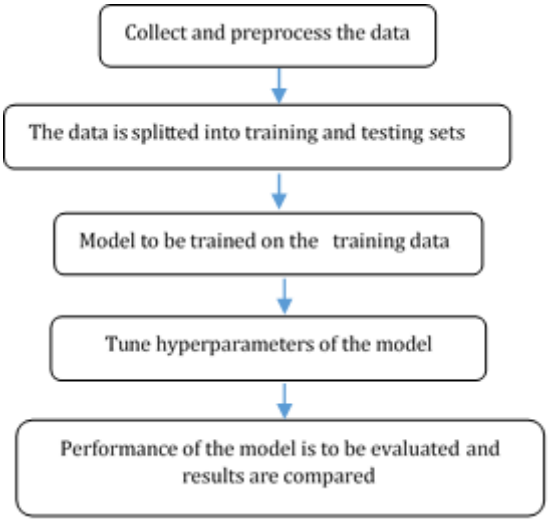

FIGURE 1

F lowchart that outlines the process

TABLE 1

Output of Gradient Boosting model

Actual successes	22	
Predicted successes	12	
Predicted gain	$64.64 for 12 IPOS	$5.39 per IPO
Actual gain:	$1075.66 for 22 IPOs	$48.89 per IPO
Naive gain:	$-507.29 for 69 IPOs	$-7.35 per IPO

4.2 Logistic Regression

It is another algorithm that can be used for predicting IPO performance. A relation between a set of input features and a binary target variable is modelled using Logistic Regression. A dataset of historical IPOs can be used to train a Logistic Regression model. The model is trained on the training set and evaluated on the basis of testing sets. Another important consideration when using Logistic Regression for IPO prediction is how to handle missing or incomplete data. This can be addressed using techniques such as imputation, which involves estimating missing values based on the available data [6].

Overall, it is a valuable algorithm for predicting IPO performance and when used appropriately in combination with other methods. It also helps to identify the important factors that influence IPO underperformance and overperformance.

It estimates the probability of the positive class using the logistic function:

$$p(y = 1|x) = \ 1 \ / \ [1 \ + exp(-(w.T * x)) \ ] \tag{1}$$

where x be the input feature vector and w be the weight vector

The weight vector is learned by minimizing the negative log-likelihood of the data, subject to a regularization term:

$$L(w) \ = \ -[sum(y * log(p) \ + \ (1-y) * log(1-p))] \ + \ lambda/2 \ * \ ||w||2 \tag{2}$$

where p be the be the predicted probability of the positive class, y be the true class label, and lambda is the regularization parameter. Table 2 summarizes the predictive performance of logistic regression model.

TABLE 2

Output of logistic regression model

Actual successes	22	
Predicted successes	20	
Predicted gain	$110.83 for 20 IPOS	$5.54 per IPO
Actual gain	$1075.66 for 22 IPOs	$48.89 per IPO
Naive gain	$-507.29 for 69 IPOs	$-7.35 per IPO

4.3 Random Forest

It is another popular algorithm that can be used for IPO prediction. It is a supervised algorithm that is trained on historical data to predict the performance of new IPOs [2] [3]. A dataset of historical IPOs can be used where the target variable is a binary indicator of whether the stock underperformed or overperformed to train a Random Forest model.

Another advantage of using Random Forest for IPO prediction is its ability to handle missing or incomplete data. Random Forest can work with data that has missing values without the need for imputation which can be a useful feature when working with real-world datasets [1].

Overall, Random Forest is a powerful machine learning algorithm for predicting IPO performance. By combining multiple decision trees and using appropriate input features and evaluation metrics, we may be able to develop more accurate and robust models for predicting IPO underperformance and overperformance. Table 3 summarizes the predictive performance of Random Forest model.

TABLE 3

Output of Random Forest model

Actual successes	22	
Predicted successes	4	
Predicted gain	$-212.93 for 4 IPOS	$-53.23 per IPO
Actual gain:	$1075.66 for 22 IPOs	$48.89 per IPO
Naive gain:	$-507.29 for 69 IPOs	$-7.35 per IPO

4.4 AdaBoost

The basic idea behind AdaBoost is to iteratively train a sequence of weak classifiers on the training data, where each classifier focuses on the most difficult examples that were misclassified by the previous classifiers. One of the advantages of using AdaBoost for IPO prediction is its ability to handle noisy and complex datasets [7]. AdaBoost can focus on the most important and informative examples, while ignoring irrelevant or misleading ones. However, AdaBoost can be sensitive to noisy or mislabelled data. If the training dataset contains a large number of mislabelled examples, the model's performance may be negatively affected. To address this issue, it is important to carefully pre-process the data and remove any outliers.

The prediction of the AdaBoost model is calculated as follows:

$$f(x) \; = \; sign[\sum(alpha\_t \; * \; h\_t(x))] \tag{3}$$

where weight of the t-th weak learner is alpha_t (decision stump), h_t(x) is the prediction of the t-th weak learner, and the sign function returns the value of -1 or +1 depending on the sign of the value which is denoted by sign.

The weight alpha_t is calculated as:

$$alpha\_t \; = \; 0.5 \; * \; ln[(1 - \varepsilon\_t)/\varepsilon\_t] \tag{4}$$

Where the weighted error is denoted by ε_t. Table 4. Summarizes the predictive performance of Ad boost model.

4.5 Support Vector Machine (SVMs)

SVMs work by looking at a bunch of different factors such as financial data, market trends, and company-specific information, and trying to find a pattern that separates IPOs that do well from those that don't.

TABLE 4

Output of AdaBoost model

Actual successes	22	
Predicted successes	10	
Predicted gain	$-402.65 for 10 IPOS	$-40.27 per IPO
Actual gain	$1075.66 for 22 IPOs	$48.89 per IPO
Naive gain	$-507.29 for 69 IPOs	$-7.35 per IPO

One of the advantages of SVMs is that they can handle complex data even when there are a lot of different factors to consider [8]. They do this by looking for the best possible boundary or hyperplane, that separates the different IPOs into two groups those that do well and those that don't.

Another advantage of SVMs is that they can handle situations where there are more IPOs that do poorly than those that do well. This is called class imbalance and SVMs have special techniques that can help them work better in these situations. However, SVMs can be computationally expensive which means they can take a long time to run or require a lot of computer power. However, careful parameter tuning and evaluation are necessary to ensure the model's performance and avoid overfitting or underfitting.

The decision boundary of the SVM is defined by a hyperplane:

$$w.T * x + b = 0 \tag{5}$$

Where w denotes the weight factor, b be the bias term and x is the input feature vector. The SVM tries to maximize the margin between the decision boundary and the closest training samples subject to a constraint that all samples are classified correctly. The optimization problem can be formulated as:

$$min\ 1/2 * ||w||\text{?}2 \tag{6}$$

Subject to:

$$y\_i * (w.T * x\_i + b) >= 1 \tag{7}$$

where y_i is the true class label of the i-th training sample. The solution of the optimization problem gives the weight vector w and the bias term b. Table 5 summarizes the predictive performance of SVM.

TABLE 5

Output of SVM

Actual successes	22	
Predicted successes	8	
Predicted gain	$263.33 for 4 IPOS	$32.92 per IPO
Actual gain:	$1075.66 for 22 IPOs	$48.89 per IPO
Naive gain:	$-507.29 for 69 IPOs	$-7.35 per IPO

4.6 Artificial Neural Networks (ANN)

It has been used in IPO prediction to identify the relationship between input factors and IPO under-pricing. In this approach, a feedforward neural network is trained to predict the under-pricing of IPOs using historical data on IPOs. The neural network is trained using a backpropagation algorithm that adjusts the weights of the connections between the neurons in the network to minimize the difference between the predicted and actual under-pricing values.

Studies have shown that ANN models can be effective in predicting IPO under-pricing. However, the performance of the model depends on the quality of the data on the basis of which the model is trained and the of input factors. Therefore, it is important to carefully select and pre-process the datal to ensure accurate predictions. Table 6. summarizes the predictive performance of ANN.

TABLE 6

Output of ANN

Actual successes	22	
Predicted successes	8	
Predicted gain	$263.33 for 4 IPOS	$32.92 per IPO
Actual gain	$1075.66 for 22 IPOs	$48.89 per IPO
Naive gain	$-507.29 for 69 IPOs	$-7.35 per IPO

5 RESULTS

We have applied various classification methods on a training dataset which was assigned 70% of the observations randomly using stratified sampling. On the remaining 30% sample, we tested our models. Table 7 summarizes the accuracy obtained on the application of models to the testing dataset.

TABLE 7

The accuracy obtained on the application of models

Gradient Boosting Classifier	59.42%
Logistic Regression	59.42%
Random Forest Classifier	65.21%
Ada boost Classifier	57.97%
SVM	65.33%
ANN	68.11%

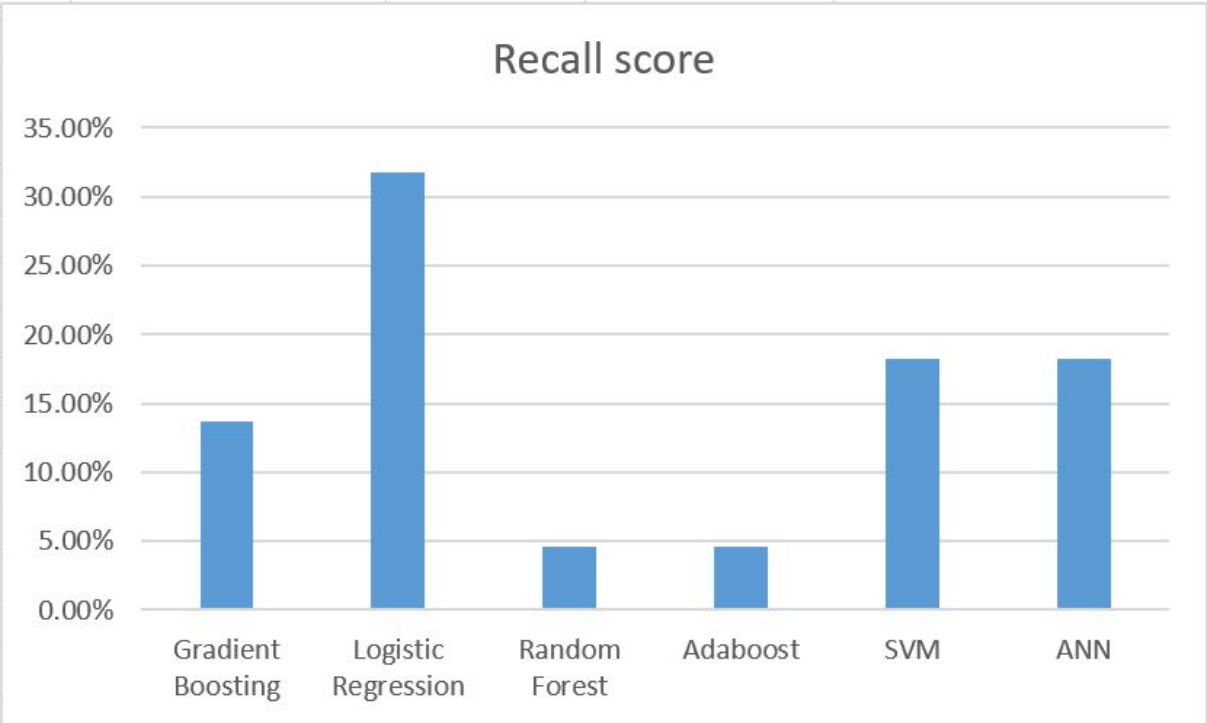

FIGURE 2

2. Recall score

Fig 2 shows the compare the performance of different models based on their recall scores. Recall score for each model on your test dataset is calculated and plot them against each other.

We have set the accuracy values on the y-axis and the model names on the x-axis. Fig 3 compare the accuracy of each model and determine which one performs best for our specific dataset and problem.

Fig 4. shows the f-beta which can be calculated by calculating the weighted harmonic mean of precision and recall which takes into account both false positives and false negatives precision which is the measures the proportion of true positives (i.e., correctly identified underperforming

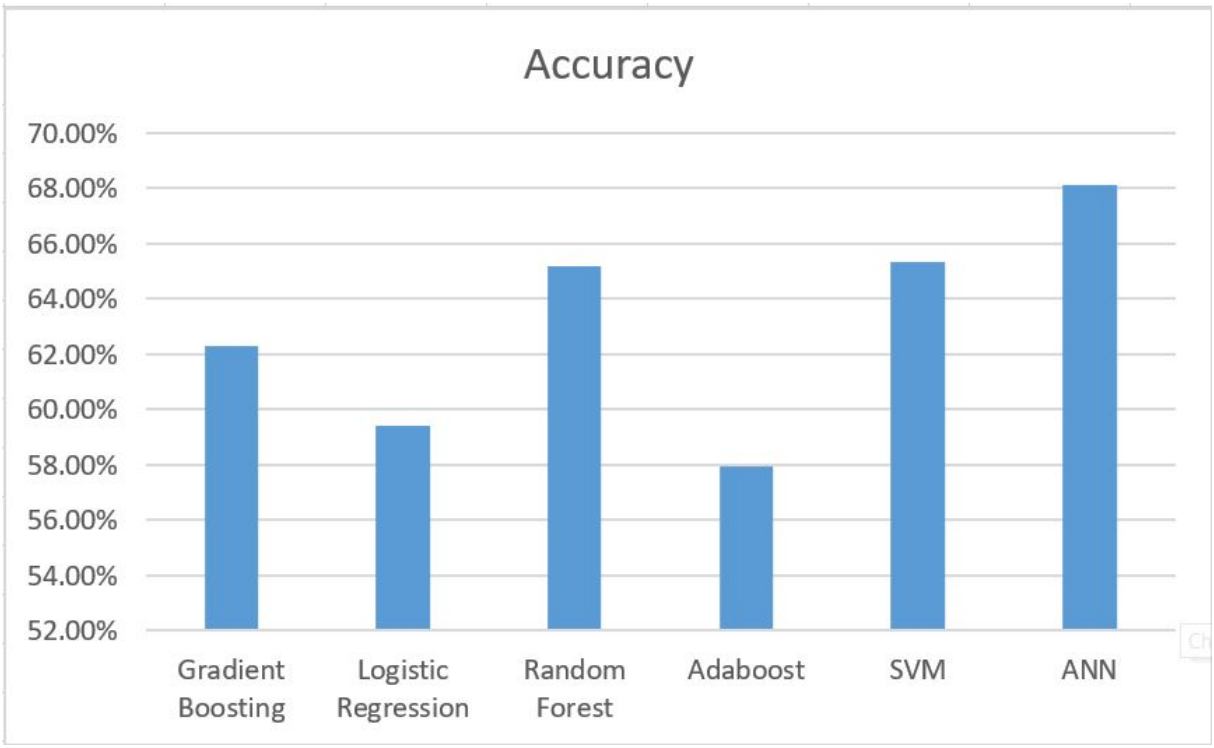

FIGURE 3

3. Accuracy obtained from different model

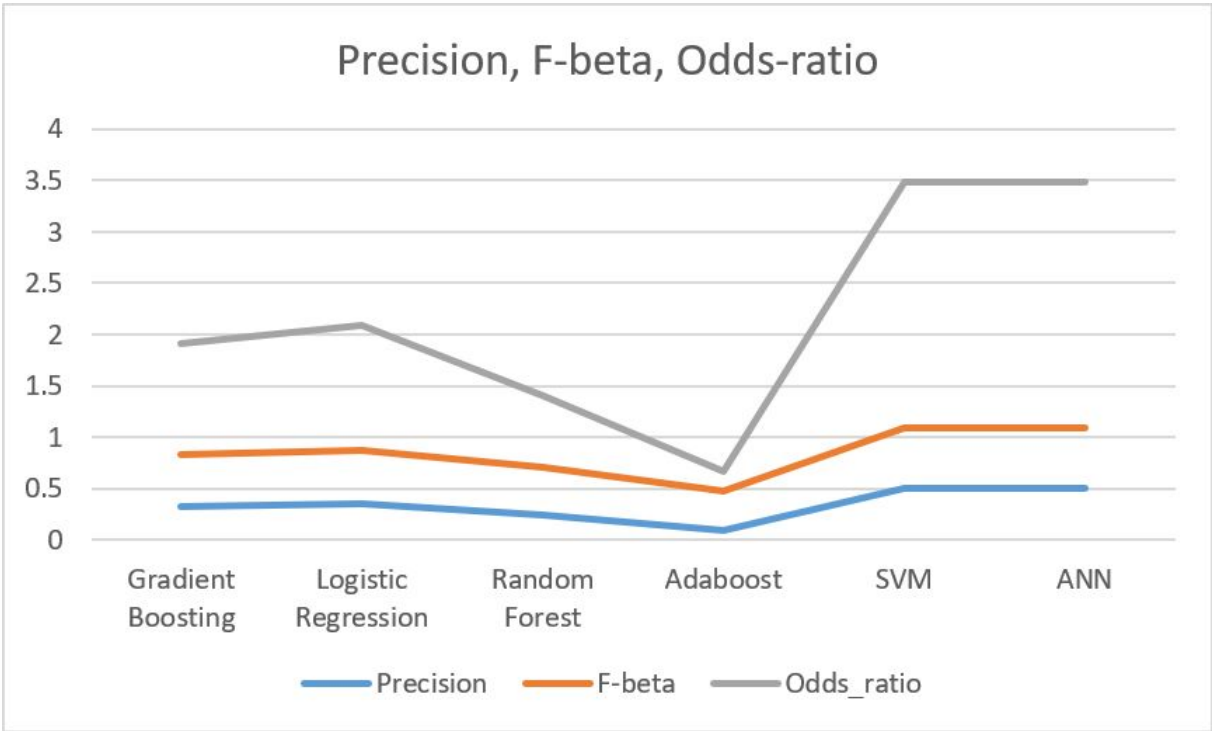

FIGURE 4

Precision, F-beta and Odds-ratio

IPOs) out of all the positives (i.e., all the IPOs identified as underperforming). Odds ratio can be used to measure the likelihood of an IPO underperforming compared to not underperforming.

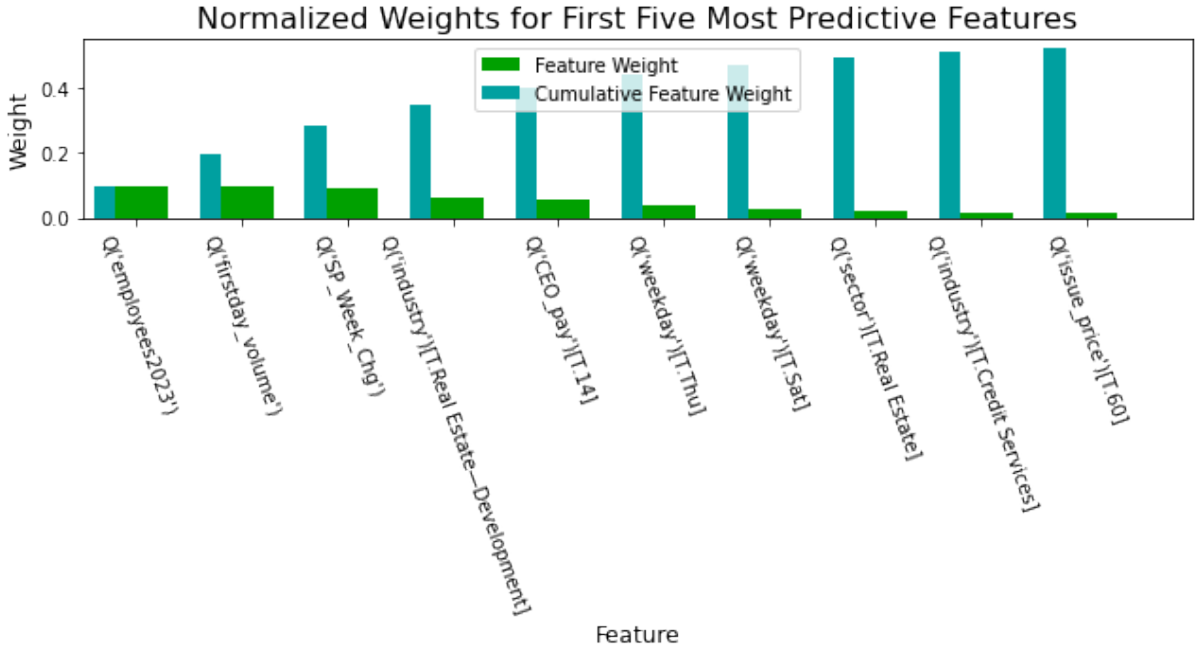

FIGURE 5

Normalized weights for first five most predictive features

Fig 5 shows the normalized weights indicate the relative importance of each feature in predicting IPO underperformance with feature 1 being the most important and feature 5 being the least important. Similarly Table 8. shows the comparison of the existing results and the results obtained by us.

TABLE 8

Comparison of obtained results with existing results

Model	Existing Results (Accuracy in %)	Obtained Results (Accuracy in %)
Logistic Regression	59	59.42
Random Forest	64	65.21
ANN	65	68.11

6 Conclusion

In conclusion our research paper explored the use of several machine learning algorithms including for predicting IPO underperformance. We found that each algorithm had its own

strengths and weaknesses in this task. Gradient Boosting and Random Forest showed promising results in handling the imbalanced nature of the data. The Logistic Regression provided insights into the importance of individual features in predicting IPO performance. AdaBoost and SVMs showed promising results in improving the classification accuracy and handling nonlinear data. Overall, our research highlights the potential of machine learning algorithms for predicting IPO underperformance and provides insights into the factors that drive IPO performance. However, a lot of research is needed to improvise the accuracy and interpretability of these models and to evaluate their performance in real-world investment scenarios. In summary, our research paper provides insights into the use of machine learning algorithms for IPO prediction and highlights the potential for these models to inform investment decisions. However, caution should be exercised in applying these models in practice as the accuracy and generalizability of the models depend on data quality and the specific context of investment decision.

References

[1] Jerome & Friedman and Jacqueline Meulman. Multiple additive regression trees with application in Epidemiology. *Statistics in medicine* , 22:1365–81, 2003.

[2] B Baba and G Sevil. Predicting IPO initial returns using random forest. *Borsa Istanbul Review* , 20(1):13–23, 2020.

[3] R Bansal and R Khanna. IPOs under-pricing and money “left on the table” in Indian market. *International Journal of Research in Management* , 2:106–120, 2012.

[4] S Chhabra, R Kiran, and A N Sah. Information asymmetry leads to under-pricing: Validation through SEM for Indian IPOs. *Program*, 51(2):116–131, 2017.

[5] S Chhabra, R Kiran, A N Sah, and V Sharma, 2005.

[6] V S Desai and R Bharati. A comparison of linear regression and neural network methods for predicting excess returns on large stocks. *Annals of Operations Research*, 78(0):127–163, 1998.

[7] Y Freund and R E Schapire. Experiments with a new boosting algorithm. *Machine Learning: Proceedings of the Thirteenth International Conference* , 13:148–156, 1996.

[8] Y Freund and R E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* , 55(1):119–139, 1997.

[9] Jerome Friedman. Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics,29:1189–1232, 2000.