*sustainability*

MDPI

*Article*

# Artificial Intelligence and Exploratory-Data-Analysis-Based Initial Public Offering Gain Prediction for Public Investors

**Manushi Munshi [1], Manan Patel [1], Fayez Alqahtani [2], Amr Tolba [3], Rajesh Gupta [4], Nilesh Kumar Jadav [1], Sudeep Tanwar [1,*], Bogdan-Constantin Neagu [5] and Alin Dragomir [5,*]**

1 Department of Computer Science and Engineering, Institute of Technology, Nirma University, Ahmedabad 382481, India
2 Software Engineering Department, College of Computer and Information Sciences, King Saud University, Riyadh 12372, Saudi Arabia
3 Computer Science Department, Community College, King Saud University, Riyadh 11437, Saudi Arabia
4 Department of Computer Engineering, U. V. Patel College of Engineering, Ganpat University, Mehsana 384012, India
5 Department of Power Engineering, Faculty of Electrical Engineering, "Gheorghe Asachi" Technical University of Iasi, 67 D. Mangeron Blvd., 700050 Iasi, Romania
* Correspondence: sudeep.tanwar@nirmauni.ac.in (S.T.); alin.dragomir@tuiasi.ro (A.D.)

**Abstract:** An initial public offering (IPO) refers to a process by which private corporations offer their shares in a public stock market for investment by public investors. This listing of private corporations in the stock market leads to the easy generation and exchange of capital between private corporations and public investors. Investing in a company's shares is accompanied by careful consideration and study of the company's public image, financial policies, and position in the financial market. The stock market is highly volatile and susceptible to changes in the political and socioeconomic environment. Therefore, the prediction of a company's IPO performance in the stock market is an important study area for researchers. However, there are several challenges in this path, such as the fragile nature of the stock market, the irregularity of data, and the influence of external factors on the IPO performance. Researchers over the years have proposed various artificial intelligence (AI)-based solutions for predicting IPO performance. However, they have some lacunae in terms of the inadequate data size, data irregularity, and lower prediction accuracy. Motivated by the aforementioned issues, we proposed an analytical model for predicting IPO gains or losses by incorporating regression-based AI models. We also performed a detailed exploratory data analysis (EDA) on a standard IPO dataset to identify useful inferences and trends. The XGBoost Regressor showed the maximum prediction accuracy for the current IPO gains, i.e., 91.95%.

**Keywords:** initial public offering (IPO); stock market; random forest; XGBoost Regressor; exploratory data analysis (EDA)



**Citation:** Munshi, M.; Patel, M.; Alqahtani, F.; Tolba, A.; Gupta, R.; Jadav, N.K.; Tanwar, S.; Neagu, B.-C.; Dragomir, A. Artificial Intelligence and Exploratory-Data-Analysis-Based Initial Public Offering Gain Prediction for Public Investors. *Sustainability* **2022**, *14*, 13406. https://doi.org/10.3390/su142013406

Academic Editors: Kamalakanta Muduli, Rakesh Raut, Balkrishna Eknath Narkhede and Himanshu Shee

Received: 21 September 2022
Accepted: 12 October 2022
Published: 18 October 2022



## 1. Introduction

The study of initial public offering (IPO) markets, their changing trends, and the stock market has been an essential arena of financial analysis over the years. An IPO refers to the mechanism by which private corporations generate capital by offering their shares to public investors while issuing a new stock [1]. It is considered one of the significant transitions in ownership of shares, as the existing private company can offer their shares in the public market to generate more capital. Additionally, IPO allotment is a quick and easy inflow of capital to finance the various ventures of the firm. Moreover, it improves the company's public image once it enters the global market. Publicly listed companies are bound to attract more investors and stakeholders; therefore, the profit generated with an IPO is shared equally among all of the stakeholders.

The profitability of a given IPO depends on the company's image, public sentiments

about the company's IPO, its valuation, its profit potential, and how well the IPO attracts the investment community. With the advent of technology and globalization, the IPO market has evolved drastically over the years. The period of 2020–2021 showed a significant increase in start-up firms to create more jobs, innovation, and long-run-capital growth. It has been observed from the literature that these firms have listed multiple IPO shares to generate substantial capital interest and come out in the public market to showcase their existence to hold higher valuations. The statistics show that 63 companies in India mobilized a colossal sum of 1.19 trillion rupees in 2021, which amounted to over four times the money raised in the previous year [2]. New-age technology-based start-ups have shaped the Indian market IPO trend by creating a profitable environment with low interest rates and robust retail participation.

The trend of IPO under-pricing, i.e., listing an IPO at a price below its actual stock price, plays a significant role in improving the prediction of IPO performance. However, this task has a few challenges due to the fragile and unstable nature of the stock market. The unpredictability of the stock market arises from the influences of external factors, such as the political conditions of a country, natural calamities, and exchange rate fluctuations, which pose further challenges in predicting the return of an IPO or its closing price. The data available for prediction are also vast and nonlinear, thereby adding to the challenges. Prediction of the performance of an IPO in the stock market has been of significant interest to the research community over the years [3,4]. The advancements in technology and the introduction of artificial intelligence (AI) algorithms have propelled the scientific community toward accurately predicting the performance of an IPO in the stock market. Ideally, prediction algorithms use linear models, such as linear and logistic regression; however, the irregularity of data makes them sensitive to outliers, making regression models less efficient.

Research has proved that ensemble algorithms such as Random Forest (RF) and XGBoost Regressor are more efficient and adept at dealing with significant problems that linear regression models cannot solve. For example, in [3], Baba et al. used RF to predict the initial returns of an IPO in Borsa, Istanbul. They performed a comparative study of AI methods and showed that the RF model outperformed regression models.. However, their study was limited to the listings in Borsa, Istanbul. Along with RF, researchers have also explored artificial neural networks (ANNs) as a probable algorithm for predicting stock prices and IPO gains. In [5], Vijh et al. performed a comparative analysis of ANN and RF models on stock market data. They evaluated the performance of their model using the mean absolute percentage and mean bias error. However, the data size was limited; thus, it did not provide intuitive inferences from the predictions.

The emerging field of deep learning (DL) in the AI domain has opened doors to a new world for exploration. In [6], Selvin et al. explored and compared different DL techniques, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and long short-term memory (LSTM), for the prediction of the stock prices of companies listed in the National Stock Exchange. The authors of [7] proposed an RNN model with a low complexity that worked on financial-type time-series data, such as stock market indices, to predict their future performance over time. They tested their model on the Bombay stock exchange data for stock price prediction. Further, in [8], Roman et al. presented a study on market trend prediction by training an RNN model on stock market data belonging to multiple countries' markets. The authors of [9,10] also showed promising results for DL-based feature engineering and price prediction algorithms. Transfer learning and federated learning are also being explored and implemented for stock market analysis and prediction. In [11], Nguygen et al. employed transfer learning, wherein a pre-trained LSTM model was applied to the target dataset and fine-tuned to enhance the performance in the prediction of IPOs. In [12], Lim et al. presented a comparative study between federated LSTM and traditional LSTM for stock price prediction. They concluded that the performance of traditional LSTM was better than that of federated LSTM regardless of parameter optimization and model fine-tuning.

IPO under-pricing is when the closing price of an IPO in the market at the end of the listing day is higher than the initial offer price [13]. This implies a gain for investors investing in that IPO and a loss for the company issuing that IPO. In [4], Agarwal et al. analyzed the under-pricing trends of historical IPO listings to study factors affecting under-pricing. From the viewpoint of the Indian IPO market, in [14], Krishnamurti et al. discussed the applicability of various reasons for the under-pricing of an IPO in the Indian market. Their study determined a vital feature affecting under-pricing—the time lag between the final allotment of the IPO and the listing of the IPO. This is because the time elapsed between these two periods is considered perilous by investors, who have started to require additional compensation.

The prediction and analysis of IPO performance are essential for facilitating more profitable investment decisions. AI domains such as ML and DL can be used to make significant headway in such research. Past research has shown much promise in this domain, but the solutions are constrained to a single data source and are affected by several challenges, such as the inability to handle outliers, the insufficient amount of data being used in such studies, and deceptive accuracy. The study presented in [3] used an RF Regressor to predict IPO gains in Istanbul. However, the dataset used in this study had IPO listings from 1998 to 2018, so it was not updated with the latest trends in the financial market. In addition, the model used can be improved with other AI models that better fit the dataset. Motivated by the scope of development provided by past research in this field, we performed an exhaustive analysis and predictive study of IPO performance in the Indian stock market. Useful inferences were derived by using exploratory data analysis (EDA) to better understand the market trends. We employed four regression models—the Decision Tree Regressor, K-Nearest Neighbors (KNN) Regressor, RF Regressor, and XGBoost Regressor. Further, the proposed architecture was evaluated by using evaluation parameters such as the mean squared error (MSE), mean absolute error (MAE), root mean squared error (RMSE), and accuracy. XGBoost Regressor outperformed all other models in the prediction of IPO gains.

### *1.1. Motivations*

Predicting the performance of an IPO in a stock market is an essential domain of study in financial analysis. Past research has proposed several AI-based ML and DL models for IPO gain predictions and stock market trend analysis [3–6,15]. However, the existing solutions have some lacunae, such as a lack of use of sufficient and reliable data on IPO listings, lower accuracy in prediction models, and sensitivity to outliers. Motivated by these, we conducted an in-depth analysis of IPO data and incorporated AI models, i.e., RF Regressor and XGBoost Regressor, in order to predict IPO performance in the stock market.

### *1.2. Contributions*

The major contributions of this paper are as follows.

- We present an in-depth analysis of IPO data from the Indian market over the last decade. Useful trends and inferences on the stock market were derived from the data.
- We adopted AI-based models, i.e., RF and XGBoost Regressors, to enhance the efficiency of the prediction models for IPO performance in the stock market. A comparative study between the two algorithms in terms of their predictions and the feature importance curves was conducted.
- Evaluation parameters such as the MSE, MAE, RMSE, and accuracy were used to evaluate the performance of the models and compare the predictions to the actual values given in the IPO dataset.

### *1.3. Novelty*

In today's age, where investors fear risks owing to the highly volatile nature of the stock market, investing in an IPO without prescience can prove to be disadvantageous. Thorough knowledge of past IPO listings and their profitability can go a long way toward

helping an investor make a decision. However, it is observed from the literature that researchers have not explored the potential of AI in IPO price prediction. Most of the work that is done is for stock price prediction using AI algorithms, not IPO price prediction. It is difficult for us to find recent and reputable research articles that support our ideas and facts; moreover, there is no recent standard dataset available that can be used for AI training to improve IPO price prediction. A few research articles that support our work were gathered and included in the above-mentioned section; however, their work was on an obsolete dataset with a very old feature space, which will not provide any intuitive information about today's IPO prices. Motivated by this, we propose an AI-based intelligent IPO price prediction architecture that improves IPO performance. The proposed architecture was trained on all past records of IPO listings on the Indian stock market. It then evaluates all major features that influence an IPO's performance in the stock market to finally give a prediction. The prediction accuracy of existing regression models, such as the Decision Tree Regressor and RF Regressor, can be further expanded and improved. Our proposed model, XGBoost Regressor, is intended to overcome these challenges, improve prediction accuracy, and reduce the loss function. This model can significantly benefit investors by providing a comprehensive overview of IPO performance in the market and market trends.

### *1.4. Organization*

The rest of this paper is organized as follows. Section 2 describes the problem's formulation and the system model. Section 3 presents the proposed architecture and includes a description of the data, preprocessing, analysis, and the proposed model. The results of the study are presented in Section 4. Finally, the paper is concluded in Section 5, which also includes the future scope of the study.

## 2. Problem Formulation and System Model

### *2.1. System Model*

Figure 1 shows a pictorial representation of the system model, indicating the step-by-step process of IPO data analysis and IPO performance prediction. First, we collected Indian IPO data for 2010–2022 from different Internet sources, which included companies' official blogs, competitive data science websites, and research articles. The dataset had essential IPO features, such as issue size, qualified institutional buyers (QIBs), high-net-worth individuals (HNIs), retail individual investors (RIIs), listing open, and listing close. First, preprocessing steps were utilized to normalize the dataset by using the Z-score normalization technique. This method deducts the mean from the data value and scales each data value to the unit variance. The normalized value $\zeta$ for a specific feature value $\chi$ is calculated in the following way [11].

$$\zeta = \frac{(\chi - m)}{s} \tag{1}$$

where $m$ is the mean of the sample value and $s$ is the standard deviation. Further, EDA was performed on the dataset to obtain results and inferences regarding the trends of the IPO market. Various EDA-based graphs are illustrated to support the results, including scatter plots, density plots, histograms, and a correlation heatmap. To conduct a more insightful analysis and obtain efficient prediction results, the IPO data listed over the last decade were split into three logical partitions based on the year in which the IPOs were listed in the market. Different regression models were trained for the entire dataset, including the three data sub-parts. Finally, a comparative study was performed based on the results obtained from the predictions.

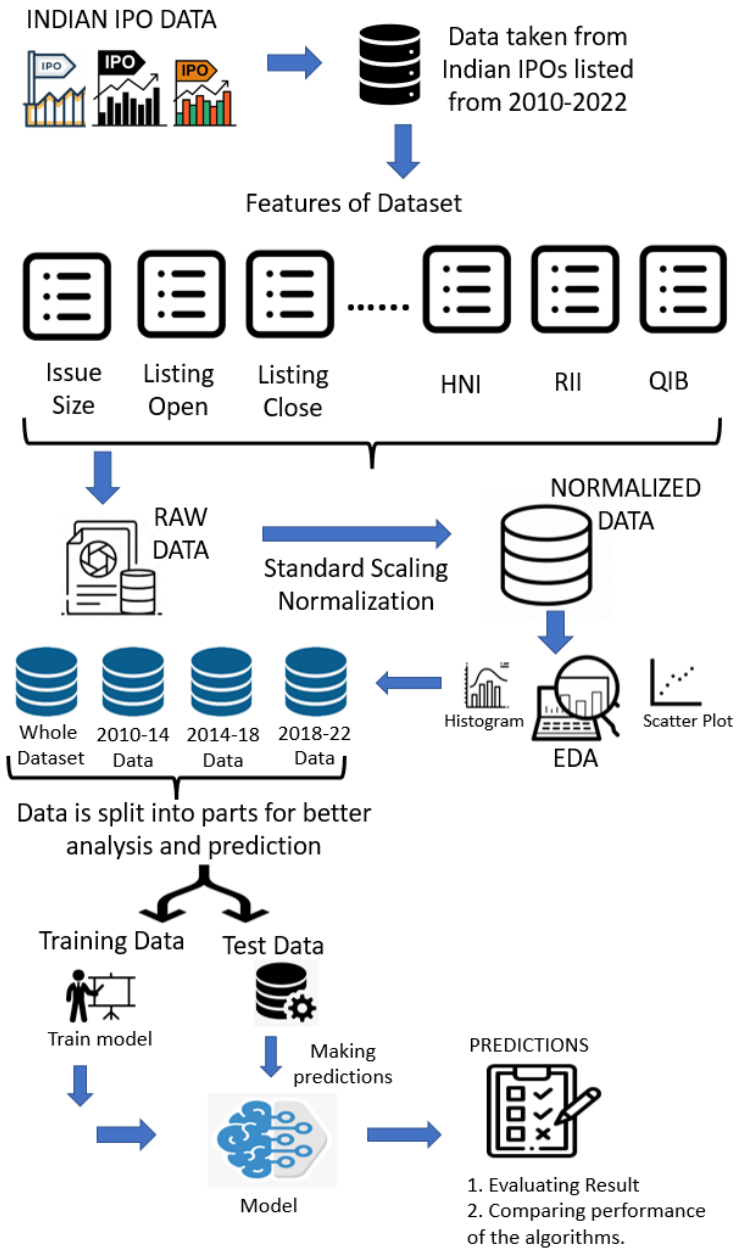


**Figure 1.** System model.

### *2.2. Problem Formulation*

In this study, we utilized two different datasets of IPO information from the Indian stock market to enhance the IPO prediction performance. Both datasets were then merged by analyzing the correlations between the feature spaces of each dataset. The IPO dataset contained over 500 rows of information on IPOs listed in the Indian stock market over the last decade, i.e., 2010–2022. In addition, there were 12 columns in the IPO dataset, each containing specific information on the concerned IPO, such as the date of listing, IPO name, market price, and IPO gains.

Let $(D)$ be the IPO dataset containing $(n)$ rows and $(m)$ columns.

$$D = \mathbb{R}^{n \times m} \tag{2}$$

$$D = \{\varsigma_1, \varsigma_2, \ldots, \varsigma_m\} \tag{3}$$

$$\text{each } \varsigma \in \text{unique columns of } D \tag{4}$$

where each ς represents the feature space (unique features) of the dataset $D$. Moreover, the number of rows $(n)$ is greater than 500, and the number of columns $(m)$ is 12. For that, several researchers in studies such as [3,4,14] proposed AI-based solutions, such as applying ML, DL, and genetic algorithms to enhance IPO prediction performance in a stock market. However, their solutions have not prevailed for several reasons, such as the small and concise data size, their country-dependent datasets, and trivial and deceptive accuracies.

$$D \xrightarrow{\text{small data size}} \mathbb{M} \tag{5}$$

$$D \xrightarrow{\text{inefficient feature space}} \mathbb{M} \tag{6}$$

$$\mathbb{M} \implies \text{trivial IPO prediction accuracies} \tag{7}$$

where $\mathbb{M}$ is the AI model used to predict the IPO gains. Therefore, the objective of this study is to improve performance in the prediction of current IPO gains by applying ML models.

$$\mathbb{O} = max\left(\underset{n \times m \in D}{\mathbb{M}_{pred}}\right) \tag{8}$$

where $\mathbb{O}$ is the objective function and $\mathbb{M}_{pred}$ is the AI-based model intended to improve the performance in IPO prediction.

## 3. Proposed Architecture

This section presents the proposed architecture for efficiently predicting the current IPO gains in the stock market. This section comprises a description of the dataset, data preprocessing, and proposed model. A detailed description of each component of the proposed architecture is given in the following.

### *3.1. Dataset Description*

In this section, we explain the insights of the dataset that was utilized in the proposed architecture. We identified two datasets from the Kaggle platform with a potential feature space to be studied and adopted by the AI models. The two datasets were—"All Indian IPO-Initial Public Offering" [16] and "IPO Data India 2010–2021" [17]. By analyzing the correlations between the feature spaces of both datasets, we fused the two datasets. The final dataset $(D')$ comprised over 500 Indian IPO listings in the stock market from 2010 to 2022. The columns or features contained in $(D')$ are as follows.

- Date: Date on which the IPO was listed in the market.
- IPO name: Name of the IPO.
- Issue size: Total number of shares issued by the company listing the IPO.
- QIB: The institutional investors known to have the means and expertise to evaluate the market and invest. These include banks, insurance companies, financial institutions, etc.
- HNI: The category of investors who invest in shares worth more than 2 lakh rupees in an IPO.
- RII: The category of investors who invest in shares worth less than 2 lakh rupees in an IPO.
- Issue price: The price at which the shares are sold by the company.
- Listing open: The opening price listed on the stock exchange as the market opens on the listing day.
- Listing close: The closing price listed on the stock exchange after the market closes on the listing day.
- Listing gains: The profit or loss percentage incurred by the difference in issue price and listing open price.
- Current market price (CMP): Current price of the IPO in the market.
- Current gains: The gains obtained with the IPO. If they are negative, this is a loss for the investors.

### *3.2. Data Preprocessing*

In this subsection, we describe the data preprocessing that we performed on the raw IPO dataset $(D')$, which was obtained as described in Section 3.1. Before applying the preprocessing steps, $(D')$ was modified to offer an intuitive analysis for IPO prediction. $(D')$ did not contain any features for the year in which an IPO came out; it only contained the IPO listing date. From the viewpoint of IPO gain prediction, the exact date of the IPO listings is not relevant; only the year in which it came out in the stock market is useful for analyzing differing stock market trends over the years. So, we added a new column named "time in years" to the IPO dataset $(D')$, which signified the exact year in which the IPO came out in the stock market. The size of the original dataset, $D'_{n\times m}$, then became $D'_{n\times m+1}$.

$$\text{time in years} \xrightarrow{\text{add column}} D' \tag{9}$$

$$D' \rightarrow \mathbb{R}^{n\times m} \xrightarrow{\text{modified dimension}} \mathbb{R}^{n\times m+1} \tag{10}$$

The dataset had many inconsistencies, such as missing and not-a-number (NaN) values, unstandardized data values, incompatible data types, and trivial data columns; these needed to be preprocessed before being sent for model training [18]. The missing and NaN values are filled/replaced with the central tendency measures, i.e., the mean value of a particular column. Further, the data values were standardized by using the Z-score normalization technique, which normalized each value by using the mean and standard deviation values. This helped in solving the range-scaling problem, where the value of a particular column $m_{12} \gg m_{13}$ or $m_{12} \ll m_{13}$, resulting in inaccuracies in model training. Particularly for $D'$, the subscription values of QIB, HNI, and RII had a significantly smaller range compared to features such as current gains and current market price, which could affect the training time and performance of current gain prediction models. Therefore, the normalization scaled the values of all of the columns, i.e., $m_{12} = m_{13}$. In addition, it eradicated the outliers from $D'$ to improve the IPO prediction performance.

Irrelevant and trivial features, such as the IPO name, CMP, and date columns, were dropped from $D'$. This was achieved by calculating the cumulative variance of all of the features. The higher the variance value, the higher the feature importance. The target variable, i.e., current gains, was the class label for the prediction models. Then, the features forming the set of independent variables, such as the issue size, QIB, HNI, RII, issue price, listing open, listing close, listing gains, and time in years, were employed to train the AI models.

### *3.3. Proposed Model*

Once the dataset was processed, it was forwarded to the AI models to perform EDA and provide the results of the prediction of the current IPO gains. EDA was used to get intuitive inferences from the dataset that helped in understanding the stock market trends and the preference for the AI model to be used in predicting the current IPO gain. To support the EDA, several matplotlib-based visual representations, such as histograms, density plots, heatmaps, and scatter plots, were used, which helped to get a better insight into the IPO trends and relations of the IPO data. First, a correlation heatmap was plotted with features of the IPO, such as issue size, issue price, HNI, RII, QIB, listing open, listing close, listing gains, current gain, and time in years, which helped us identify pairs with significant correlations. Further, a density plot was plotted for the target variable, i.e., current gains, to showcase the distribution of the profit and loss incurred by the investors. Scatter plots were generated between features, such as QIB, HNI, RII, and current gains, to discover further trends. The three subscription columns, i.e., QIB, HNI, and RII, were specifically chosen because they showed significant correlations with each other and with the target variable, i.e., current gains, in the correlation heatmap. The plot generated between QIB and QIB, HNI, and RII also showed promising trends and inferences, which

are described in the later sections.

We incorporated different AI models in order to predict IPO gains in the stock market.

$$\mathbb{M} \xrightarrow{\text{AI model}} D' \tag{11}$$

First, the entire preprocessed dataset was split into training and testing sets; the training dataset was forwarded to $\mathbb{M}$ and the testing dataset was used to validate the prediction results of $\mathbb{M}$. The split was performed with an 80–20 ratio for the dataset $D'$, where *x_train* and *y_train* were part of the training dataset, and *x_test* and *y_test* were part of the testing dataset. Then, the training dataset was applied to $\mathbb{M}$ to fit the data and predict the results.

$$x\_train, y\_train \xrightarrow{\text{train model}} m, \forall m \in \mathbb{M} \tag{12}$$

where $m$ represents the AI model. Once $\mathbb{M}$ was trained on $D'$, *x_test* and *y_test* were used to validate the prediction of the current IPO gain. Let $P$ be the set containing predictions, i.e., $p_1$, $p_2$, ..., $p_m$, from $\mathbb{M}$, which contains predictions from model 1, model 2, ..., model $m$, respectively.

$$P = \{p_1, p_2, \ldots, p_m\} \tag{13}$$

$$x\_test, y\_test \xrightarrow[\text{predictions}]{\text{validate}} p, \forall p \in P \tag{14}$$

The proposed architecture employed various AI models for the task of the prediction of IPO gains. The four regression models applied were the KNN Regressor, Decision Tree Regressor, RF Regressor, and XGBoost Regressor. The KNN Regressor is a supervised, non-parametric ML algorithm. It intuitively estimates the relationships between independent and dependent variables by getting an average of the observations present in the same neighborhood. Further, the Decision Tree Regressor is a supervised ML algorithm wherein the dataset is broken into smaller parts and a related decision tree develops incrementally in a parallel manner. A decision node consists of multiple branches wherein each branch represents values for the different attributes tested. The leaf node depicts the decision made on the numerical target variable. The RF Regressor randomly selects small samples from the training dataset with replacements. Then, a feature is selected that iteratively splits the node of the aforementioned samples (small decision trees). Each individual tress has its class prediction result; using majority voting, one can estimate the model's best prediction.

$$\text{RF} \xrightarrow{\text{applied}} D'_{train} \in x\_train, y\_train \tag{15}$$

$$x\_train, y\_train = \begin{bmatrix} v_{11} & v_{12} & v_{13} & \ldots & v_{1m} \\ v_{21} & v_{22} & v_{23} & \ldots & v_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ v_{n1} & v_{n2} & v_{n3} & \ldots & v_{nm} \end{bmatrix} \tag{16}$$

where $v_{ij}$ represents the data value of the training dataset ($D'_{train}$).

$$\text{RF} \xrightarrow[\text{select samples}]{\text{Randomly}} \{s_1, s_2, \ldots, s_k\} \tag{17}$$

where each sample $s_i$ looks like:

$$s_i = \begin{bmatrix} v_{11} & v_{12} & v_{13} & \ldots & v_{1h} \\ v_{21} & v_{22} & v_{23} & \ldots & v_{2h} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ v_{j1} & v_{j2} & v_{j3} & \ldots & v_{jh} \end{bmatrix} \tag{18}$$

where $j$ and $h$ represent the rows and columns of each randomly selected sample $s_i$.

$$\forall s_i \xrightarrow[\text{prediction}]{\text{make}} \{p_1, p_2, \ldots, p_k\} \tag{19}$$

$$p_i \in s_i$$

$$\text{RF} \xrightarrow[\text{voting}]{\text{majority}} \{p_1, p_2, \ldots, p_k\} \tag{20}$$

$$\text{RF} \xrightarrow{\text{selects the best prediction}} p_l \tag{21}$$

The results obtained were then improved by fine-tuning the model using the method of randomized grid search.

In addition to the RF Regressor, the XGBoost Regressor algorithm was used to analyze and predict the current IPO gain. Here, we focused on the RF Regressor and used it as an example to showcase the working of the XGBoost Regressor. This is because both RF and XGBoost are tree-based AI algorithms and show minor accuracy and error differences. The only difference between them is the way in which the algorithms train on a dataset. The XGBoost algorithm tries to accurately predict the value of a target variable by combining the results of simple and weaker models. Model fitting is done by applying any loss function and the optimal gradient descent algorithm. The loss function is then minimized with each iteration in the algorithm. It first randomly selects a sample from $D'_{train}$. Each sample is individually trained using the objective function of XGBoost, wherein the residual of the first sample is inserted into the second sample to improve the results of the first iteration. Table 1 shows the hyperparameters used in the XGBoost algorithm. However, certain implementation constraints were encountered when training the model for the XGBoost Regressor. Firstly, the model had a long computation time and high complexity, so we fine-tuned the model to get optimized values of the hyperparameters. Secondly, XGBoost has limitations while handling large amounts of sparse data, such as the IPO dataset, so we split the dataset into logical parts according to the year of the IPO listing and separately applied the model to all the sub-parts to improve the model efficiency. A detailed explanation of each step is given in the following.

First, the XGBoost model ($M_1$) takes the training dataset $D'train$ and predicts the best data samples ($S_d$) from the entire $D'train$.

$$M_1 \xrightarrow{\text{trains}} D'train \xrightarrow{\text{Output (in 1}st\text{ iteration)}} S_d, W_d \tag{22}$$

$$M_1 \in \text{XGBoost} \in \mathbb{M} \tag{23}$$

where $S_d$ and $W_d$ are strong and weak data samples from the training of the XGBoost Regressor algorithm. Then, another model $M_2$ is prepared to minimize the errors of the weak data samples ($W_d$) from the first model $M_1$ in order to obtain more fine-tuned and optimal data samples ($S'_d$)

$$M_2 \xrightarrow{\text{trains}} W_d \xrightarrow{\text{Output (in 2}nd\text{ iteration)}} S'_d \tag{24}$$

This process is an iterative process until a better accuracy and a minimum error are achieved by the XGBoost model.

$$M_q \xrightarrow{\text{trains}} W_d^q \xrightarrow{\text{Output (in }qth\text{ iteration)}} S_d^q \tag{25}$$

Here, $M_q$ is the final XGBoost Regressor model that optimizes the accuracy and minimizes the error of the $qth$ weak data sample $W_d^q$ of $D'_{train}$ to get strong optimal data samples $S_d^q$ with high accuracy and minimal error rates.

**Table 1.** Optimal values in the XGBoost Regressor grid search.

| Parameter | Description | Optimal Value |
|---|---|---|
| n_estimators | Number of decision trees | 1000 |
| learning_rate | Rate set to reduce overfitting | 0.05 |
| max_depth | Maximum depth of tree | 6 |
| num_parallel_tree | No. of trees formed in each iteration | 1 |

## 4. Result Analysis

In this section, we present and analyze the results obtained from the EDA performed on the dataset. In addition to the EDA results, this section showcases the predictions derived after applying the regression models. A detailed explanation of each result is given in the following.

### *4.1. EDA Results*

This subsection presents the plots obtained from the data analysis. It includes a density plot of the current gains, a scatter plot showing a comparison of IPO subscriptions of different types of investors, and a correlation map of important features in the given dataset. Figure 2 shows the density plot of the current gain feature of the dataset. The current gain density plot aids us in getting an overview of the relative imbalance in the dataset with regard to the target variable of current gains. From the analysis of our plot, we inferred that a higher proportion of IPO listings had negative gains, i.e., losses, than those with profits that reaped benefits for their investors. This plot behaves as a skewed Gaussian plot. It is clear from the graph that a higher number of IPO listings had incurred losses as compared to those that gave profits. This shows that investing in IPOs has been a risky trend in recent years. On the other hand, careful analysis and study of a company along with market sentiments could help one in investing wisely. This greatly improves the chances of gaining profits from these early investments. The plot shows that some investors earned massive profit percentages (200 and above).

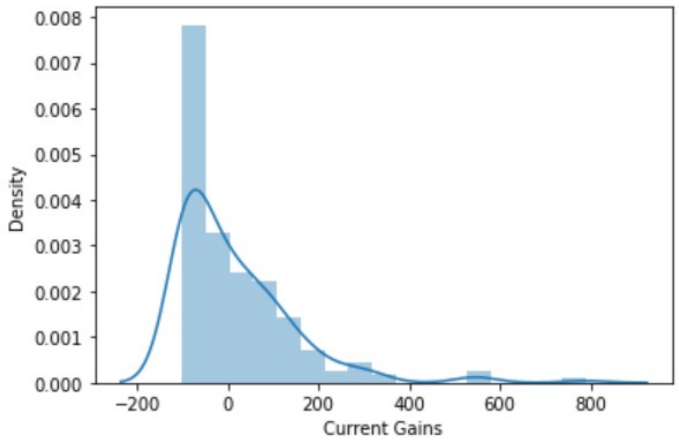


**Figure 2.** Density plot of the current gain feature of the dataset.

Figure 3 shows a scatter plot of QIB over-subscriptions vs. QIB, HNI, and RII over-subscriptions. The purpose of this plot is to show the correlation between the investing patterns of different levels of investors in the IPO. It depicts how small-scale investors are affected by the investing trends of large-scale investors. The QIB, RII, and HNI counts greater than 150 were considered outliers and were removed from the analysis to get more insights from the result. A significant trend inferred from the graph is that the HNIs largely tended to over-subscribe in comparison with the QIBs. The RII count was observed to lie below 20.

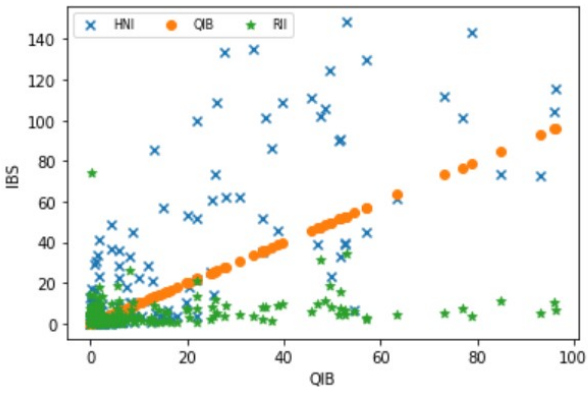


**Figure 3.** Scatter plot of QIB over-subscriptions vs. QIB, HNI, and RII over-subscriptions.

Figure 4 is a heat map of the correlations between the features of the dataset. The correlation heat map helps in understanding the potential relationships between IPO features and how closely the features are related. Significant correlations were observed between HNIs and QIBs (0.76), RIIs and HNIs (0.62), and RIIs and QIBs (0.43). HNIs and QIBs have higher stakes in companies than small-scale investors do. So, they tend to conduct a deeper analysis before investing in an IPO. They also consider the market's investment patterns, other competing investors, and the market sentiment. Hence, they have a higher correlation. RIIs are small-scale investors; hence, they are influenced by the investing patterns of large institutions and HNIs. So, the correlation between the former and latter is significant.

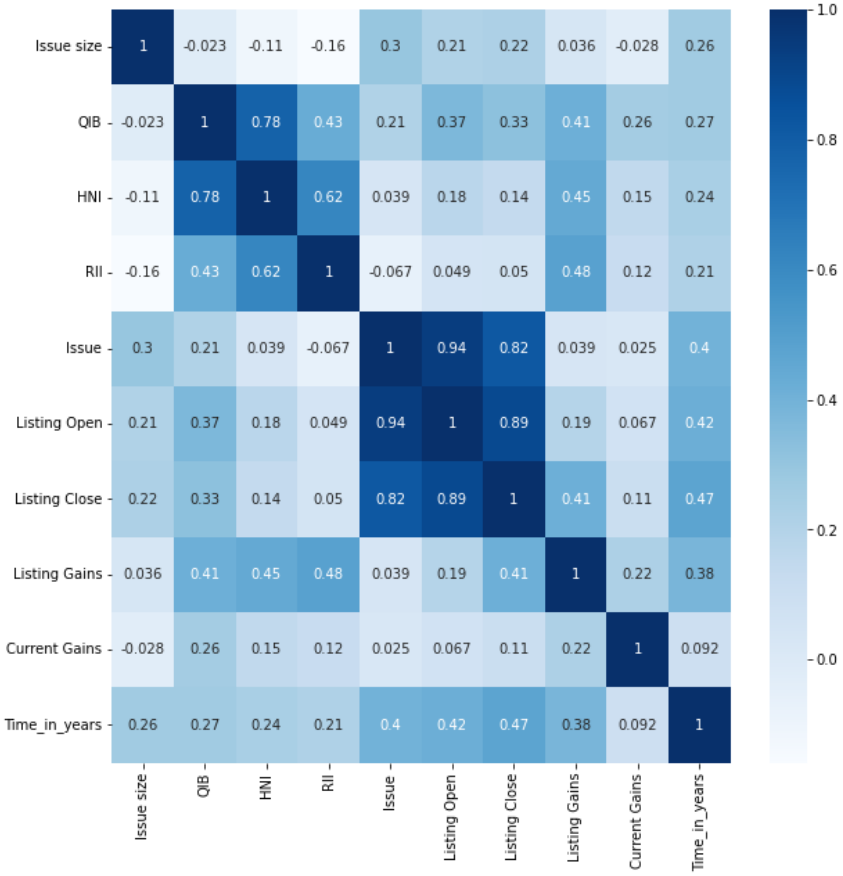


**Figure 4.** Heat map of the correlations between dataset features.

### *4.2. Prediction Results*

#### 4.2.1. RF and XGBoost Regressors Applied to the Dataset (2010–2022 Time Period)

Figures 5a and 6a are the feature importance graphs of the RF Regressor and XGBoost Regressor algorithms. XGBoost outperformed the RF algorithm because RF requires hyperparameters to optimize the results, but XGBoost focuses on functional and feature space. XGBoost had the top four features allocated in the order of issue size, QIBs, HNIs, and RIIs, while RF had the order of HNIs, QIBs, issue size, and RIIs. Moreover, RIIs were not given significant importance in RF, thereby affecting the accuracy of the algorithm.

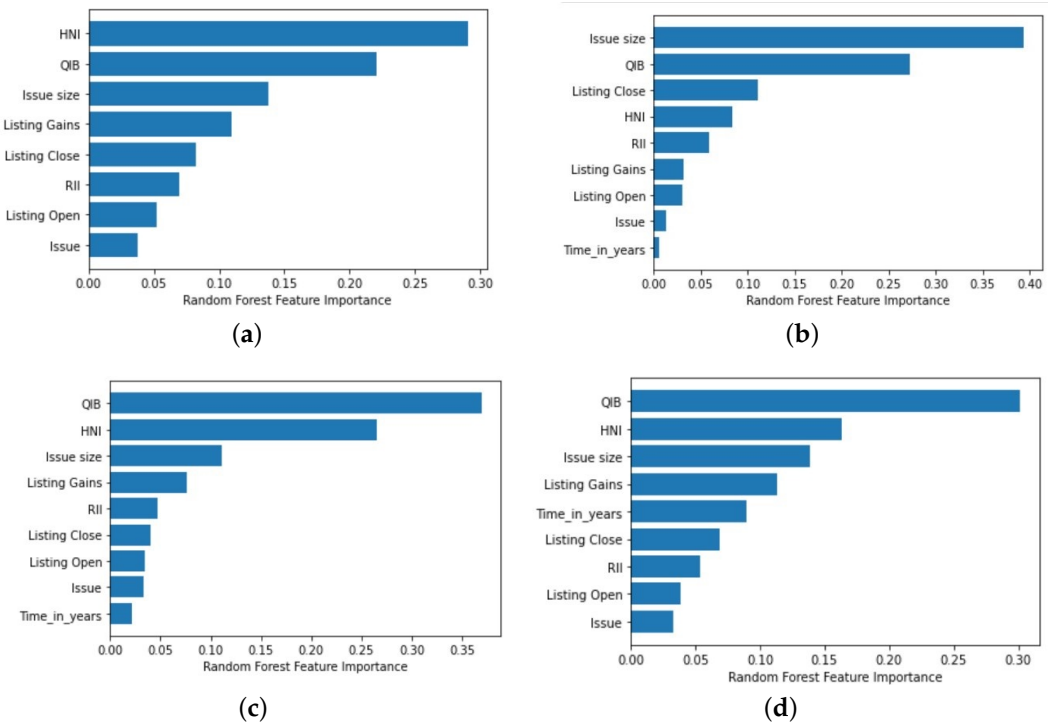


**Figure 5.** Feature importance plot for the RF Regressor for different year slots—(**a**) 2010–2022, (**b**) 2010–2014, (**c**) 2014–2018, and (**d**) 2018–2022.

In order to perform a more organized analysis and study the changing market scenarios in India over the last decade, we segregated the data into three parts depending on the year in which a particular IPO was released.

#### 4.2.2. RF and XGBoost Regressors Applied to the IPO Data from 2010 to 2014

Figures 5b and 6b show the feature importance plots for the RF and XGBoost Regressors when applied to IPO data from 2010 to 2014. From the graph, we can see that the "issue size" feature was an important feature that played a significant role in predicting the current gains. Furthermore, it determined whether or not profitable results were present for the investor. Listing close was a more significant feature for RF than for XGBoost, as the closing price of an IPO on a listing day indicates the initial performance of the IPO in the market and affects the initial sentiments of the public towards the IPO. Substantial profit on a listing day helps boost the confidence of people in the future performance of an IPO. In contrast, losses on a listing day lead to negative sentiments towards an IPO. For the IPO data from 2010 to 2014, XGBoost had a better prediction accuracy than that of the RF Regressor.

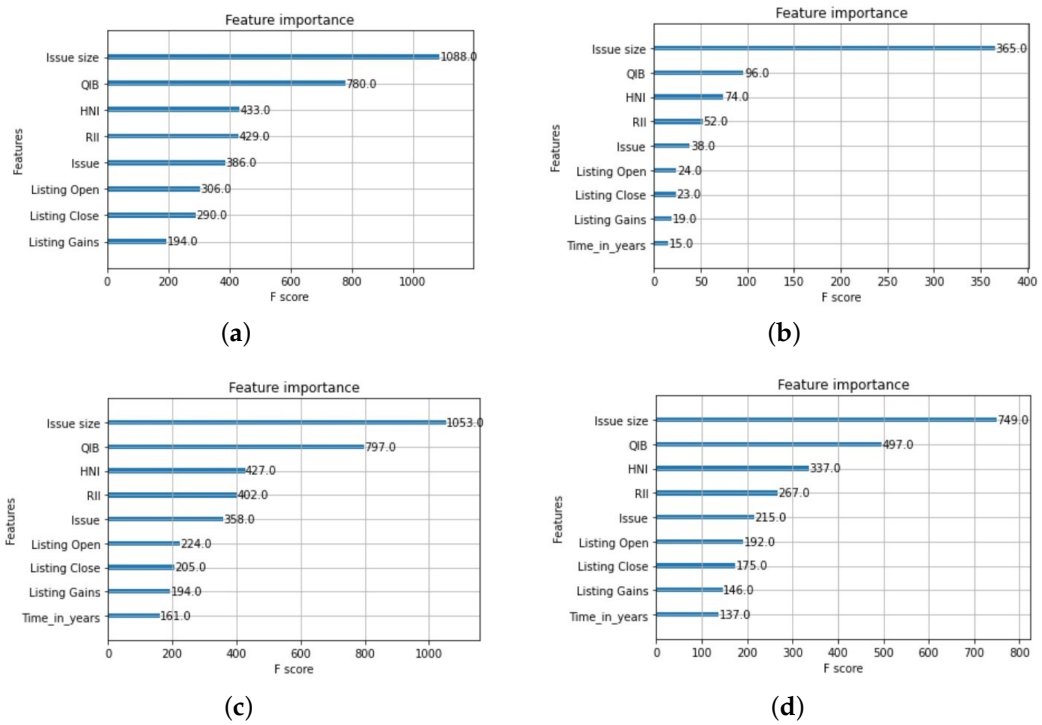


**Figure 6.** Feature importance plot for the XGBoost Regressor for different year slots—(**a**) 2010–2022, (**b**) 2010–2014, (**c**) 2014–2018, and (**d**) 2018–2022.

#### 4.2.3. RF and XGBoost Regressors Applied to the IPO Data from 2014 to 2018

Figures 5c and 6c illustrate the feature importance curves of the RF and XGBoost Regressors, respectively, for the IPO data from the years 2014–2018. The curves were observed to differ from the feature importance plots of the previous time period, i.e., 2010–2014. This change in trend can be attributed to the emergence of a new government in India, which led to a shift in financial policies. As a result, the RF Regressor had greater feature importance for the QIB and HNI features in this time period. In contrast, the XGBoost Regressor maintained a similar order of feature importance, i.e., issue size, QIBs, HNIs, and RIIs.

#### 4.2.4. RF and XGBoost Regressors Applied to the IPO Data from 2018 to 2022

Figures 5d and 6d display the feature importance plots for the IPO data of the 2018–2022 period. The COVID-19 pandemic impacted the financial market for around half of this period. The feature importance of RIIs in the RF algorithm was observed to decrease significantly due to the pandemic's impact on small-scale retail investors' investing patterns. Moreover, issue size, QIBs, and HNIs remained important features for both algorithms for this time period.

### *4.3. Performance Analysis of the XGBoost Regressor*

To evaluate the performance of the proposed architecture, we used different evaluation parameters, such as the MAE, MSE, RMSE, and accuracy, to analyze and compare the performance of the models. The MAE is the average of all the absolute errors between the actual values and predictions. Further, the MSE is the average of the squared error; the RMSE is the square root of the MSE and mainly defines the standard deviation of the prediction errors. The accuracy of the model is a measure of how precise the model is at predicting values. We can infer from Table 2 that the XGBoost Regressor had the best MAE, MSE, and RMSE values, i.e., 12%, 2%, and 15%, respectively, for the year of 2010–2014. The accuracy of prediction was also increased by almost 5% by the XGBoost Regressor; hence, it outperformed the other AI models, as shown in Table 3.

**Table 2.** Evaluation parameters of the XGBoost algorithm.

| Time Period | MAE | MSE | RMSE | Accuracy |
|---|---|---|---|---|
| 2010–2022 | 0.29 | 0.27 | 0.52 | 84.51% |
| 2010–2014 | 0.12 | 0.02 | 0.15 | 91.95% |
| 2014–2018 | 0.23 | 0.15 | 0.39 | 87.99% |
| 2018–2022 | 0.26 | 0.15 | 0.39 | 87.10% |

**Table 3.** Comparison of the evaluation parameters of the four regression models.

| Model | MAE | MSE | RMSE | Accuracy |
|---|---|---|---|---|
| KNN Regressor | 0.50 | 0.83 | 0.91 | 52.40% |
| Decision Tree Regressor | 0.31 | 0.41 | 0.64 | 76.25% |
| RF Regressor | 0.37 | 0.34 | 0.58 | 80.25% |
| XGBoost Regressor | 0.29 | 0.27 | 0.52 | 84.51% |

The XGBoost Regressor outperformed the other AI models in the prediction of IPO gains. The proposed architecture was compared to the model used in [3], wherein the authors employed the RF Regressor for prediction. The RF Regressor has certain limitations that are overcome by the proposed XGBoost Regressor XGBoost algorithm. XGBoost has a greater focus on the functional space while reducing the model cost. A slight change in the hyperparameters of the RF Regressor significantly affects all trees in the forest, as they are applied to every tree in the beginning and, hence, affect its predictions. Contrary to this, the hyperparameters of XGBoost are applied to only one tree in the beginning and are dynamically adjusted with each iteration. In addition, the XGBoost Regressor works better than the RF Regressor when we have to work with an unbalanced dataset, such as the IPO dataset in our case. The proposed XGBoost Regressor model thus provides better prediction results than those of other models.

Figure 7 depicts the error functions of the RMSE, MSE, and MAE of the KNN Regressor, Decision Tree Regressor, RF Regressor, and XGBoost Regressor. We can clearly see that the XGBoost Regressor was able to considerably lessen the loss functions and, thus, perform significantly better on outliers of the IPO dataset. It gave out an RMSE value of 0.52, MSE value of 0.29, and MAE value of 0.27, which were well below the RMSE, MSE, and MAE values of the existing models. Figure 8, on the other hand, depicts a comparison of the prediction accuracies of the four regression models employed to predict IPO gains in the form of a bar graph. Once again, it can be seen that models such as the KNN Regressor barely crossed the 50% mark in terms of prediction accuracy, while tree-based AI algorithms, such as the Decision Tree Regressor and RF Regressor, gave out 76.25% and 80.25% accuracies. The XGBoost Regressor was able to outperform them by giving an accuracy of 84.51%. Table 3 shows the values of the performance metrics of all of the regression models. From Table 3, we can infer the superior performance of the XGBoost Regressor over that of the other regressor models owing to its lower error values and higher accuracy.

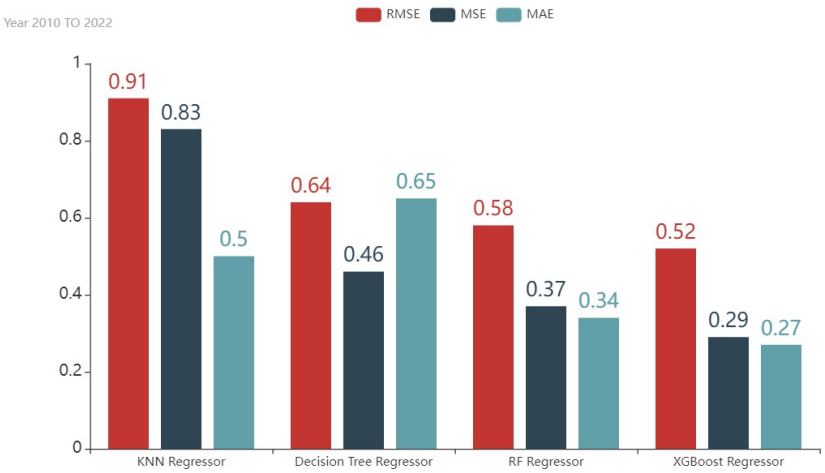


**Figure 7.** Error comparison of the different AI models.

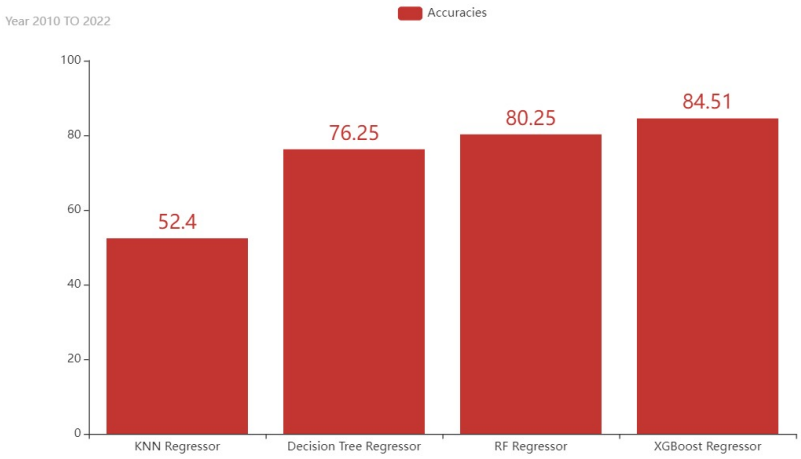


**Figure 8.** Accuracy comparison of the different AI models.

## 5. Conclusions and Future Plan

The prediction of IPO performance in the stock market comes with a set of challenges, such as the fragility of the stock market, irregularity in data, and external socioeconomic factors affecting the IPO market. Motivated by these challenges, we presented a comparative study of four regression models for predicting IPO performance in a market. The four regression models were the KNN Regressor, Decision Tree Regressor, RF Regressor, and XGBoost Regressor. We also presented an analysis of IPO data, providing essential inferences that allow a better understanding of IPO trends in the current financial market. For that, two standard datasets were identified and then merged into a single dataset by calculating their correlations. Then, the single dataset was preprocessed by using several data preprocessing steps. Then, critical conceptions were carried out using EDA and data visualization, such as correlations, current gains, and feature importance. Further, the regression models were applied to the standard dataset to predict the IPO performance. Finally, the performance of the proposed architecture was evaluated by using various evaluation metrics, such as the MAE, MSE, RMSE, and accuracy. The results show that the XGBoost Regressor outperformed the other regression models in terms of accuracy, RMSE, MSE, and MAE. The results show that the maximum accuracy obtained was 91.95% by the XGBoost Regressor.

The area of IPO performance prediction provides a vast scope for future research. Advanced AI models based on DL, federated learning, and transfer learning can be implemented to obtain better prediction results. More features can be incorporated into IPO datasets to train models, such as sentiments in a market about a particular IPO. Additionally,

a comparative study can be performed on the IPO performances of different countries, and the impact of one country's market trends on another country's trends can also be studied.

**Author Contributions:** Conceptualization: R.G., N.K.J., F.A., B.-C.N. and S.T.; writing—original draft preparation: M.M., M.P., R.G., B.-C.N. and N.K.J.; methodology: R.G., A.T., F.A., A.D. and S.T.; writing—review and editing: S.T. and N.K.J.; Software: M.M., M.P., R.G. and N.K.J.; Visualization: S.T., A.D., A.T. and N.K.J.; Investigation: N.K.J., M.M., M.P. and S.T.; Supervision: S.T., F.A. and A.T.; Validation: S.T., A.T. and A.D. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was funded by the Researcher Supporting Project (No. RSP2022R509) of King Saud University, Riyadh, Saudi Arabia.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Acknowledgments:** The authors would like to extend their gratitude to King Saud University for funding this research through Researchers Supporting Project No. (RSP2022R509) King Saud University, Riyadh, Saudi Arabia and the authors are thankful to the Gheorghe Asachi Technical University of Iasi for their valuable support.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. IPO Def. Available online: https://www.investopedia.com/terms/i/ipo.asp (accessed on 28 June 2022).
2. IPO Trends. Available online: https://www.ey.com/en_in/ipo/india-ipo-trends-report (accessed on 17 August 2022).
3. Baba, B.; Sevil, G. Predicting IPO initial returns using random forest. *Borsa Istanb. Rev.* **2020**, *20*, 13–23. [CrossRef]
4. Agrawal, R.; Sjmsom, U. Predicting IPO Underperformance Using Machine Learning. In Proccedings of the 51st Annual Conference of The Decision Sciences Institute, San Francisco, CA, USA, 21–23 November 2021; pp. 1–11.
5. Vijh, M.; Chandola, D.; Tikkiwal, V.A.; Kumar, A. Stock Closing Price Prediction using Machine Learning Techniques. *Procedia Comput. Sci.* **2020**, *167*, 599–606. [CrossRef]
6. Selvin, S.; Ravi, V.; Gopalakrishnan, E.; Menon, V.; Kp, S. Stock price prediction using LSTM, RNN and CNN-sliding window model. In Proceedings of the 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Udupi, India, 13–16 September 2017; pp. 1643–1647.
7. Rout, A.; Dash, P.; Dash, R.; Bisoi, R. Forecasting Financial Time Series Using A Low Complexity Recurrent Neural Network and Evolutionary Learning Approach. *J. King Saud Univ.-Comput. Inf. Sci.* **2015**, *29*, 536–552. [CrossRef]
8. Roman, J.; Jameel, A. Backpropagation and recurrent neural networks in financial analysis of multiple stock market returns. In Proceedings of the HICSS-29: 29th Hawaii International Conference on System Sciences, Wailea, HI, USA, 3–6 January 1996; Volume 2, pp. 454–460.
9. Long, W.; Lu, Z.; Cui, L. Deep learning-based feature engineering for stock price movement prediction. *Knowl.-Based Syst.* **2019**, *164*, 163–173. [CrossRef]
10. Kavinnilaa, J.; Hemalatha, E.; Jacob, M.S.; Dhanalakshmi, R. Stock Price Prediction Based on LSTM Deep Learning Model. In Proceedings of the 2021 International Conference on System, Computation, Automation and Networking (ICSCAN), Puducherry, India, 30–31 July 2021; pp. 1–4.
11. Nguyen, T.T.; Yoon, S. A Novel Approach to Short-Term Stock Price Movement Prediction using Transfer Learning. *Appl. Sci.* **2019**, *9*, 4745. [CrossRef]
12. Federated Learning. Available online: https://hdl.handle.net/10356/153212 (accessed on 17 July 2022).
13. IPO Underpricing. Available online: https://www.investopedia.com/terms/u/underpricing.asp (accessed on 23 November 2020).
14. Krishnamurti, C.; Kumar, P. The initial listing performance of Indian IPOs. *Manag. Financ.* **2002**, *28*, 39–51. [CrossRef]
15. Luque, C.; Quintana, D.; Isasi, P. Predicting IPO Underpricing with Genetic Algorithms. *Int. J. Artif. Intell.* **2012**, *8*, 133–146.
16. Data 1. Available online: https://www.kaggle.com/datasets/soumyadipghorai/all-ipo-stocks-of-moneycontrol (accessed on 18 July 2022).
17. Data 2. Available online: https://www.kaggle.com/datasets/aimack/ipo-data-india2021 (accessed on 10 June 2022).
18. Patel, N.P.; Parekh, R.; Thakkar, N.; Gupta, R.; Tanwar, S.; Sharma, G.; Davidson, I.E.; Sharma, R. Fusion in Cryptocurrency Price Prediction: A Decade Survey on Recent Advancements, Architecture, and Potential Future Directions. *IEEE Access* **2022**, *10*, 34511–34538. [CrossRef]