2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM) | 979-8-3503-4779-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICONSTEM56934.2023.10142428

Flight Ticket Prediction Using Gradient Boosting Regressor Compared With Linear Regression

N.Sri Sai Venkata Subba Rao, S.John Justin Thangaraj* and V Sheeja Kumari

Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu. India. Pincode: 602105.

E-mail : srisaivenkatasubbarao18@saveetha.com, johnjustinthangarajs.sse@saveetha.com sheejakumari.sse@saveetha.com*

Abstract- Aim: The purpose of this project is to predict airfare for ticket bookings using the Gradient Boosting. Regression device learning is the set of rules as opposed to a brand new linear regression. Materials and Methods: New Linear Regression Algorithm (with a sample size of ten) and Gradient Boosted Regression (with a sample size of ten). These algorithms are calculated on this picture using a total of 20 examples for the algorithm, and there are two firms that are used to calculate them. The size of the sample was determined to be 10, and it was compared with a group using a G Power value of 80%. Results: Values achieved in terms of accuracy are decided by Gradient Boosting regression (82.5%), as opposed to new Linear Regression (62.5%).8%. This is because Gradient Boosting regression is more accurate than new linear regression. In a test with one tail, the statistically significant difference between the new linear regression algorithm and the Gradient Boosting Regressor was found to be 0.00. This result was obtained With in Significance level of p 0.05. Conclusion: After going through all of the methods, it has been determined that the airfare forecast is more accurate than the brand new linear regression. This was revealed after going through all of the procedures.

Keywords: Reliability, Machine Learning, Gradient Boosting Regression, Novel Linear Regression, Flight Ticket Prediction, Flight Fare, and Flight Fare Prediction.

I.INTRODUCTION

Air travel is currently one of the fastest modes of transportation, but it is also one of the most expensive ones. [Case in point:] [Case in point:] [Case in point It is projected that the level of accuracy of value forecasts produced by machine learning algorithms would continue to rise (National Research Council et al. 2003). Machine learning and the algorithms that drive it are responsible for a considerable amount of the work that goes into making predictions possible. This is because the algorithms look at the past history of relevant airfare as well as patterns in shopping charts (Zhao et al. 2021), which enables them to deliver results that are both accurate and efficient (Yu 2021). As a result, in order to complete this work, we will be deploying a system that utilises machine learning to make forecasts regarding airfare. These algorithms are able to give results that are precise and efficient because they learn from the historical data and trends of prior match prices. This article makes use of two distinct approaches, notably Gradient Boosting Regression and Comparison with New Linear Regression, in order to discover which of these algorithms is the most accurate in predicting future flight expenses (Lok 2018).

Throughout the course of the past five years, from 2017 to 2021, a large amount of investment has been made, with Google Scholar publishing over 200 articles and IEEE publishing nearly 600 articles on future airfare projections. An empirical methodology is used to conduct an overview of the Gradient Boosting Regressor approach and the development of a novel linear regression algorithm (Purey and Patidar 2018). Both of these are evaluated in terms of their excellent performance. The new model generates results that are more accurate than the one that came before it. The experiment was carried out in its entirety, and the suggested algorithmic approach brought about an improvement in accuracy that is comparable to the improvement brought about by the earlier model thanks to the upgrade (Tavana, Nedjah, and Alhajj 2020).

In an effort to find a solution to this problem, we are going to compare the efficacy of the Gradient Boosted Regression method with that of the recently developed linear regression.

In this part of our survey, we are going to talk about

the advantages and disadvantages of the two different machine learning algorithms that were covered in the previous section (Deepak., John Justin Thangaraj and Rajesh Khanna 2020). After applying gradient boosting regression to a data set, observing the data with the new linear regression, and finally plotting the results of the comparison, one can make a comparison of the two methods. This can be done by following these steps: first, applying gradient boosting regression to the data set; then, observing the data with the new linear regression; and finally We are now in possession of a method that provides a more precise estimate of the future prices of aeroplane tickets.

II. Materials And Methods

The Machine Learning Laboratory at the Saveetha Institute of Technology in Chennai, which also houses the Saveetha Institute of Medical and Engineering Sciences, is where the research is being conducted at the moment. The sample size was figured out with the assistance of the G Power software and by contrasting the performance of two controllers in an environment of supervised learning. This work uses a total of 20 sets, with each sample size consisting of 10 sets for both the Gradient Boosting and Novel Linear Regression techniques. The total number of sets is referred to as "sets chosen." The pre-test T-power values are computed with the assistance of the G Power 3.1 Software (G power settings: Statistical Test difference Between two Independent means, = 0.05, power = 0.80, and two algorithms (Gradient Boosting Regressor and a new linear regression algorithm) implemented using technical analysis software) [G power settings: statistical test difference between two independent means, = 0.05, power = 0.80, and two algorithms (Gradient Boosting Regressor and a new linear regression A permit from an ethics committee is not required for this study because it did not employ any materials derived from humans or animals.

Pseudocode for Gradient Boosting Regression

```
from Sklearn.ensemble import Random
GradientBoostingRegressor
from sklearn.metrics import precision_score
gradient = GradientBoostingRegressor()
gradient.fit(X train, y train)
gradient predict = gradient.predict(X test score)
gradient .
score(X train, y train)
print('accuracy score in whole:', score)
print('accuracy score in percent:',
round(score*100,2))
```

Gradient Boosting Machines, more commonly referred to as GBMs, are computer programmes that produce final predictions by merging the findings of many decision trees (Rafique et al. 2020). It is imperative that you do not forget that the Gradient Boosting Machine employs decision trees for all of its underperforming students. With the assistance of gradient-boosting techniques, it is possible to predict both categorical and continuous aspects of the targets (Harrington 2012). When the variable is being utilised in the role of a regressor, the cost function is the mean squared error (MSE), but when it is being utilised in the role of a classifier, the cost function is the logarithmic loss.

Pseudocode for Novel Linear Regression

```
From Sklearn.linear model import Linear Regression
Linear = Linear Regression()
linear. Fit(X train, y train)
Linear predict = linear. Predict(X test)
score = linear. Score(X train, y train)
print('accuracy_score total:',
print('accuracy_score процент:',
round(score*100,2))
```

Equations that characterise the relation among independent and dependent variables are being returned by modern linear regression models (Chandler, Pachter, and Mears, 1993). A regression equation with one dependent and one independent variable can be written in its most basic form as follows:

$$Y=M*X+C \ldots \quad (1)$$

If Y is the estimated dependent variable, then m is the regression coefficient. (also known as the slope), x is the independent variable, and c is a constant. The slope is often referred to as the slope of the line. To put it another way, m, x, and care are the inputs, while y is the output of that combination. The latest linear regression method makes an effort to forecast future values and trends.

$$j(\theta) = i/2m \, {}^{m}\Sigma i=1 \, (h_{\theta}(x^{(i)})-y^{(i)})^2 \ldots \quad (2)$$

There are several names for this function, including "Error Squared Function" and "Root Mean Squared Error." As a result of the fact that the quadratic function's derivative term cancels out one half of the

terms, we now get the mean, which simplifies the computations for gradient descent. The processor is an Intel Core I5 and there are 8 gigabytes of RAM in this hardware combination. The operating system was 64 bits, and the hard disc was 1 terabyte. Windows 10 was employed as the computer's Operating System, and Jupyter Notebook, which incorporates Python as a Programming language, was utilised in the process of implementation. The information was obtained through Kaggle, which is a free online community for anyone interested in data science and machine learning. This information was gathered from the following website: https://www.kaggle.com/nikhilmittal/flight-fare-prediction-mh.

The 11 variables in the data set and a description of each variable.

1. Airline: The name of the airline used for travel.
2. Date_of_Journey: The date of the trip.
3. Source: The starting point of the flight.
4. Destination: The final destination of the flight.
5. Itinerary: A standard format used by airlines that contains information about where the journey begins and ends.
6. Dep_Time: Flight departure time from the starting location.
7. Arrival Time: The time the flight arrives at its destination.
8. Duration: Flight time (hours / minutes).
9. Total_Stops: The total number of times the flight stopped before landing at its destination.
10. Additional_Info: It displays additional information about the flight.
11. Price: The cost of the flight.

Statistical Analysis

With the assistance of the SPSS Statistics software programme, one is able to carry out either an interactive or batch examination of the statistical significance of the data. With the assistance of the SPSS software package, it is now possible to conduct an analysis of the statistical significance of approaches that are founded on linear regression and gradient regression. The ticket price, the route, and the date are Independent variables, while number of flights and the flight numbers themselves are the independent variables. In light of the results of the experiment, group statistics and independent sample tests were carried out, and graphs depicting the two parameters that were the subject of the inquiry were developed.

III. RESULTS

Table 1 details the outcomes of a simulation that was run by simultaneously executing Proposed Gradient Boosting Regressor method and existing Novel linear regression system in a Jupyter notebook with a sample size of 500. The simulation was carried out with the goal of determining which of the two systems would produce the best results. The simulation was run at a few different times during the day. In Table 1, the Gradient Boosting Regressor algorithm achieved an accuracy of 82.5% on average, but the new linear regression technique could only achieve an accuracy of 62.8% overall.

Table 1. Comparison of Linear Regression and Gradient Boosting Regression using N=10 samples of the dataset with best accuracy of 82.52% and 62.81% in sample 1 (when N=1) using dataset size= 9650 and 70% training and 30% testing data.

Sample (N)	Dataset Size	Gradient Boosting Regressor Accuracy in %	Novel linear Regression Accuracy in %
1	9650	82.52	62.81
2	8500	82.49	62.58
3	7900	82.05	62.24
4	7000	81.68	62.05
5	6500	81.52	61.85
6	5500	81.25	61.68
7	4000	81.02	61.25
8	3500	80.96	58.96
9	1500	80.65	57.02
10	1000	80.32	56.85

Table 2: In Table 2, you can see the statistical analysis of the GBR and the new linear regression algorithm. For 20 different sample data sets, the mean accuracy, standard deviation, and standard error are calculated.

Group		N	Mean	Std. Deviation	Std. Error Mean
Accuracy	GBR	10	80.9340	1.21607	.38456
	LR	10	61.0760	1.32783	.41990

The new linear regression approach has a standard deviation of 1.32783, but the mean gradient boosting regression method presented in Table 2 has a standard deviation of 1.32783. Table 2 displays the mean gradient boosting regression method. The mean gradient boosting regression technique has a standard deviation of 1.21607, and it is comparable to the recently created algorithm for linear regression in terms of both mean and precision. The results indicate that the Gradient Boosting Regressor algorithm possesses a higher degree of accuracy (82.5%), as compared to the new Linear Regressor technique, which only possesses a degree of accuracy of (62.5%). 8%.

Table 3 displays the mean, standard deviation, and standard error for every data that was collected. With the use of independent Student's t-tests, we were able to calculate the mean values for the data collected from each of the research groups. In comparison, the new linear regression method has an impact size of 0.050 whereas the Gradient Boosting Regressor methodology has a value of 0.826. Both of these algorithms are significantly different as a result of this divergence.

Table 3. T-tests on independent samples establish significance and standard error. Wet basis P < 0.05.

Accuracy	Levene's test for equality of variances		T-test of Equality of Means					95% of the confidence interval of the Difference	
	F	Sig.	t	df	Sig (2-tailed)	Mean Difference	Std.Error Difference	Lower	Upper
Equal variance assumed	.050	.826	34.876	18	.000	19.85800	.56938	18.66177	21.05423
Equal variances			34.876	17.863	.000	19.85800	.56938	18.66111	21.05489

A comparison chart is included in figure 1, and it demonstrates that the Gradient Boosting technique is superior to the new Linear Regression algorithm. As a result, we are able to draw the conclusion that Gradient Boosting Regressor performs significantly better than the new Linear Regressor. The graph that was produced as a result is displayed in Figure 1 below. The images can be found at the bottom of the article.

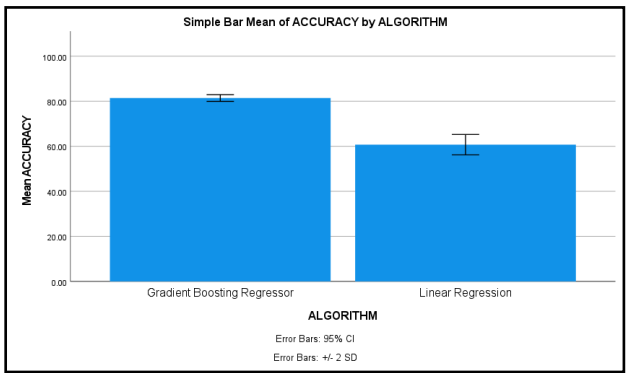

Fig. 1. A comparison is presented between the Gradient Boosting Regressor and the Novel linear Regression approach with regard to mean and accuracy. Both of these methods are regression techniques. The Gradient Boosting Regressor has a greater mean accuracy than the Novel Linear Regression, which has a lower mean accuracy than the Gradient Boosting Regressor. The comparison of Innovative Linear Regression and Gradient Boosting Regression is shown along the X-axis, and the mean accuracy of detection compared to two standard deviations is shown along the Y-axis.

IV. Discussion

The previously established models have had new regressors that are based on linear regression and gradient boosting added to them in in order to boost efficiency of the models and provide more accurate predictions regarding flight costs. However, based on the findings that we have obtained, we have come to the conclusion that Gradient Boosting Regressor is a more effective and accurate method of prediction on huge data sets when compared to the new linear regression. This conclusion was reached as a result of the fact that we came to the conclusion that the Gradient Boosting Regressor was more effective and accurate.

A recent research paper made a recommendation that the Gradient Boosting Regressor method be

used to improve ticket prediction. This could be beneficial to the implementation of the technique. Even when applied to enormous volumes of data, models that use gradient boosting regressors display low error rates. [Case in point:] We developed a method called Gradient Boosting Regressor that analyses the daily movement of tickets purchased from a number of different airlines in order to predict the price of a ticket for a specific company. This method can be used to anticipate the price of a ticket for any company. In addition, the most up-to-date linear regression techniques are not optimal for the task of improving the accuracy of ticket prediction.

In the prior discussion, only a few papers are guaranteed to provide better performance than the proposed Gradient Boosting Regressor and a new linear regression algorithm (Ataman and Kahraman 2021) to improve the accuracy of flight ticket prediction. These two pieces of research were conducted by Ataman and Kahraman. On the other hand, both the newly developed algorithm for linear regression and the proposed Gradient Boosting Regressor have the same objective: to enhance the precision of aircraft ticket prediction. In addition, the most recent price projections do not take into consideration any additional fees, which is something that has received a lot of attention as of late due to the fact that it has come to the forefront of people's minds.

Consequently, in order to increase accuracy of aircraft ticket prediction, proposed Gradient Boosting Regressor technique as well as a novel linear regression strategy can be utilised (Panwar et al. 2021).

Although the ability of airline forecasting to anticipate future prices is limited, these skills are contingent on significant margins of future pricing, which makes it possible for stronger price forecasting in the future. The algorithms that are the driving force behind deep learning are able to take into consideration forecasts of the future.

V. Conclusion

The main objective of the study is to identify how accurate ticket estimates actually are, and this will be accomplished by analysing the data. In this study piece, the GradientBoosting regression model and the new linear regression are both examined, as well as compared to one another. The results indicate that the accuracy of the GradientBoosting regression is 82.5%, which is a substantial improvement over the accuracy of the new linear regression, which is 62.8%.

Declarations

Conflict of Interest
No conflict of interests in this manuscript.

Author Contributions
The author NSSVS was involved in the process of collecting data, analysing data, and producing the publication. The conceptualization of the study, the validation of the data, and the critical assessment of the text were all contributed by the author JJT.

Acknowledgement
The author would like to use this opportunity to extend their profound gratitude to the Saveetha School of Engineering as well as the Saveetha Institute of Medical and Technological Sciences for providing the required infrastructure to successfully carry out this work.

Funding
We would like to extend our gratitude to the following organisations for providing the financial support that made it possible for us to finish the study.
1. Qbec Infosol, Chennai.
2. Saveetha University
3. Saveetha Institute of Medical And Technical Sciences
4. Saveetha School of Engineering

References

[1] Ataman, Görkem, and Serpil Kahraman. 2021. “STOCK MARKET PREDICTION IN BRICS COUNTRIES USING LINEAR REGRESSION AND ARTIFICIAL NEURAL NETWORK HYBRID MODELS.” The Singapore Economic Review. https://doi.org/10.1142/s0217590821500521.

[2] Chandler, P. R., M. Pachter, and M. Mears. 1993. “Constrained Linear Regression for Flight Control System Failure Identification.” 1993 American Control Conference. https://doi.org/10.23919/acc.1993.4793484.

[3] Deepak., S. John Justin Thangaraj, and M. Rajesh Khanna. 2020. “An Improved Early Detection Method of Autism Spectrum Anarchy Using Euclidean Method.” In 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC). IEEE. https://doi.org/10.1109/i-smac49090.2020.9243361.

[4] Harrington, Peter. 2012. Machine Learning in Action. Simon and Schuster.

[5] Lok, Johnny Ch. 2018. Prediction Factors Influence Airline Fuel

Price Changing Reasons.

[6] National Research Council, Transportation Research Board, Studies and Information Services, Division on Engineering and Physical Sciences, Aeronautics and Space Engineering Board, and Committee on Aeronautics Research and Technology for Vision 2050. 2003. Securing the Future of U.S. Air Transportation: A System in Peril. National Academies Press.

[7] Panwar, Bhawna, Gaurav Dhuriya, Prashant Johri, Sudeept Singh Yadav, and Nitin Gaur. 2021. "Stock Market Prediction Using Linear Regression and SVM." 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE). https://doi.org/10.1109/icacite51222.2021.9404733.

[8] Purey, Prateek, and Anil Patidar. 2018. "Stock Market Close Price Prediction Using Neural Network and Regression Analysis." International Journal of Computer Sciences and Engineering. https://doi.org/10.26438/ijcse/v6i8.266271.

[9] Rafique, Muhammad, Aleem Dad Khan Tareen, Adil Aslim Mir, Malik Sajjad Ahmed Nadeem, Khawaja M. Asim, and Kimberlee Jane Kearfott. 2020. "Delegated Regressor, A Robust Approach for Automated Anomaly Detection in the Soil Radon Time Series Data." Scientific Reports 10 (1): 3004.

[10] Tavana, Madjid, Nadia Nedjah, and Reda Alhajj. 2020. Emerging Trends in Intelligent and Interactive Systems and Applications: Proceedings of the 5th International Conference on Intelligent, Interactive Systems and Applications (IISA2020). Springer.

[11] Yu, Jiangni. 2021. "A New Way of Airline Traffic Prediction Based on GCN-LSTM." Frontiers in Neurorobotics 15 (December): 661037.

[12] Zhao, Zhichao, Jinguo You, Guoyu Gan, Xiaowu Li, and Jiaman Ding. 2021. "Civil Airline Fare Prediction with a Multi-Attribute Dual-Stage Attention Mechanism." Applied Intelligence(Dordrecht, Netherlands), August, 1–16.