FLIGHT TICKET PREDICTION USING XGBREGRESSION COMPARED WITH KNEIGHBOUR REGRESSION ALGORITHM

N. Sri Sai Venkata Subba Rao[1], S. John Justin Thangaraj[2*]

Abstract

Aim: To predict flight fare for ticket booking using machine learning algorithm XGBRegression compared with KNeighbour Regression

Materials and Methods: The XGB Regression (N=10) and KNeighbour Regression algorithm (N=10) these two algorithms are calculated by using two groups and a total of 20 samples taken for both algorithm and accuracy in this work. The sample size was measured as 10 per group using a G Power value of 80%.

Results and Discussion: The Values obtained in terms of Accuracy are Identified by XGB Regression (87.6%) over KNeighbour Regression (49.1%). Statistical significance difference between XGBRegression and KNeighbour Regression Algorithm was found to be 0.00 in the 2-tailed test ($p<0.05$).

Discussion and Conclusion: After all the Procedures the Prediction of Flight fare using the novel XGBRegression appears to be more accurate when compared to KNeighbour Regression.

Keywords: Machine Learning, Novel XGBRegression, KNeighbour Regression, Accuracy, Flight ticket prediction, Flight fare.

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

[2*]Project Guide, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu. India. Pincode: 602105.

1. Introduction

These days aviation routes are perhaps the quickest method of transport however the costs are not predictable (Lok 2018) machine learning calculations have a higher pace of foreseeing precise ticket fare, Here the machine learning and its calculation assume a significant part in predicting the flight ticket fare since the flight ticket majorly depends on the date and time of the travel (Turner, Griffin, and Holland 2000). So far my concern, machine learning calculations are utilized to anticipate the future costs of flight tickets (Turner, Griffin, and Holland 2000; Batra, Roy, and Panda 2020). These calculations concentrate on the previous history of the ticket fare and their patterns and get productive and precise outcomes (Boruah et al. 2019). The application of this research includes the prediction of stock markets to determine the future value of the company stock and supporting the successful prediction of price in the products trade exchange.

In the last 5 years, more than 60 papers have been published on IEEE Xplore and google scholar on flight ticket prediction. A comparative way of detecting the flight ticket can be helpful to many people (Keretna, Hossny, and Creighton 2013). In this paper analysis of the K-neighbours algorithm and XGB Regression in high-performance efficiency has been made using an experimental approach (Abulaish and Fazil 2021). In this work, no human and animal samples were used so no ethical approval is required (Boruah et al. 2019). On applying Novel XGBRegression to the dataset followed by performing observations using KNeighbour Regression and the results were plotted on a graph market-rate techniques are compared based on the result. Finally Getting the best algorithm for predicting the future prices of the flight fare (Abulaish and Fazil 2021).

Our institution is keen on working on latest research trends and has extensive knowledge and research experience which resulted in quality publications (Rinesh et al. 2022; Sundararaman et al. 2022; Mohanavel et al. 2022; Ram et al. 2022; Dinesh Kumar et al. 2022; Vijayalakshmi et al. 2022; Sudhan et al. 2022; Kumar et al. 2022; Sathish et al. 2022; Mahesh et al. 2022; Yaashikaa et al. 2022). The accuracy of existing research is not properly existing in the system. The existence of the experiment is total and the improvement of accuracy of a proposed algorithm system compared the existing model by improving (Tavana, Nedjah, and Alhajj 2020). To overcome this problem, a novel XGBRegression to improve the accuracy of the prediction is implemented and compared with KNeighbour Regression. Now by the above two Machine learning algorithms, we have taken have their own advantages and disadvantages in the current survey (Deepak., John Justin Thangaraj, and Rajesh Khanna 2020). So now in this article, two algorithms are used: XGBRegression and compare it with KNeighbour Regression to find out which among these algorithms is best to predict the future prices of flight prices.

2. Materials and Methods

The research work is carried out in the Machine Learning Laboratory at Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai. The sample size has been calculated using the G Power software by comparing both of the controllers in supervised learning. Each sample size is 10 sets for Novel XGBRegression and KNeighbour Regression total of 20 sets is selected for this work. The pre-test power value is calculated using G Power 3.1 software g power setting parameters: statistical test difference between two independent means, α=5.586, power=0.80, Two algorithms (XGBRegression and KNeighbour Regression) are implemented using Technical Analysis software.

XGB Regression

```
from xgboost import XGBRegression
from sklearn.metrics import mean_absolute_error, mean_squared_error
xgb = XGBRegression()
xgb.fit(X_train, y_train)
xgb_predict = xgb.predict(X_test)
score= xgb.score(X_train, y_train)
print('accuracy_score overall :', score)
print('accuracy_score percent :', round(score*100,2))
```

It is an XGB Regression algorithm that uses decision trees as its “weak” predictors. Beyond that, its implementation was specifically engineered for optimal performance and speed. The objective function contains a loss function and a regularization term. It tells about the difference between actual values and predicted values, i.e how far the model results are from the real values. The most common loss function in XGBoost for regression problems is reg: linear, and that for binary classification is reg: logistics.

Ensemble learning involves training and combining individual models (known as base learners) to get a single prediction, and XGBoost is one of the ensemble learning methods. XGBoost expects to have the base learners which are uniformly bad at the remainder so that when all the predictions are combined, bad predictions cancel out and a better one sums up to form final good predictions.

KNeighbour Regression

```
from sklearn.neighbors import KNeighborsRegression
knn = KNeighborsRegression()
knn.fit(X_train, y_train)
knn_predict = knn.predict(X_test)
score = knn.score(X_train, y_train)
print('accuracy_score overall :', score)
print('accuracy_score percent :', round(score*100,2))
```

K nearest neighbors is a simple algorithm that stores all available cases and predicts the numerical target based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition already since the beginning of the 1970s as a non-parametric technique. Algorithm A simple implementation of KNN regression is to calculate the average of the numerical target of the K nearest neighbors. Another approach uses an inverse distance weighted average of the K nearest neighbors. KNN regression uses the same distance functions as KNN classification.
Data was used from Kaggle which is a freely available platform for data scientists and machine learning enthusiasts. Source: https://www.kaggle.com/nikhilmittal/flight-fare-prediction-mh

The 11 variables in the dataset and description of each variable.

1. Airline: Name of the airline used for traveling.
2. Date_of_Journey: Date at which a person traveled.
3. Source: Starting location of the flight.
4. Destination: Ending location of the flight.
5. Route: This contains information on starting and ending location of the journey in the standard format used by airlines.
6. Dep_Time: Departure time of flight from starting location.
7. Arrival_Time: Arrival time of flight at the destination. Duration:
8. Duration of flight in hours/minutes.
9. Total_Stops: Number of total stops the flight took before landing at the destination.
10. Additional_Info: Shown any additional information about a flight.
11. Price: Predicted flight fare.

Statistical Analysis

SPSS Statistics is a software package used for interactive, or batched, statistical analysis. Now software is used for statistical analysis of XGB Regression and KNeighbour Regression Algorithms. The independent variable is flight name, flight number and the dependent variable is flight price, route, and date. The independent test analyses calculate the accuracy of the flight ticket fare prediction for both Methods.

3. Results

Table 1 shows the simulation result of the proposed algorithm XGB Regression algorithm and the existing system KNeighbour Regression were run at different times in the Jupyter notebook with a sample size of 500. From the table, it was observed that the mean accuracy of the XGB Regression algorithm was 87.61% and the K-Neighbour Regression algorithm was 49.17%. from the Mean, Standard Deviation, and Standard Error Mean were calculated by taking an independent variable T-test among the study groups. The XGB Regression algorithm produces a significant difference from the KNeighbour Regression algorithm with a value of 0.030 and effect size=5.586.
Table 2 represents the Mean of the XGB Regression algorithm which is better compared with the KNeighbour Regression algorithm with a standard deviation of 1.39634 and .71881 respectively. From the XGB Regression algorithm and KNeighbour Regression algorithm in terms of mean and accuracy. The mean results, the novel XGBRegression algorithm (87.61%) gives better accuracy than the KNeighbour Regression algorithm (49.17%).
Fig1 gives the comparison chart of the XGBRegression algorithm that is better than the KNeighbour Regression. It is, therefore, conclusive that XGBRegression performs better than KNeighbour Regression. The resultant plots are shown below in the figure. The figure has been placed at the end of the paper.

4. Discussion

KNeighbour Regression and XGBRegression both were Implemented to Predict the Flight Ticket and to Improve the Accuracy of the existing Model. But from the obtained results in our paper (Zhang 2020), it is concluded that the XGB Regression is more efficient and accurate in prediction compared with KNeighbour Regression for the larger datasets (Zhang 2020; Surrey Flying Services Limited 1937).
In the recent survey, the Proposed XGB Regression Algorithm is a Promising option for Flight ticket prediction (Ataman and Kahraman 2021). XGB Regression-based models have fewer error levels than the Larger data. Proposed XGB Regression Algorithm for predicting Flight fares of selected companies by comparing the daily Flight tickets movement in various airlines. Further, the neighbor Regression algorithm is not suitable for Improving the Accuracy of Flight Ticket Prediction (Panwar et

al. 2021). From the above discussion, only a few articles ensure that they provide better performance than the proposed XGBRegression and KNeighbour Regression algorithm (William Groves and Maria Gini, 2019) for improving the accuracy of Flight ticket prediction. Also, the present price prediction requires no additional cost and therefore received intense attention in recent years. So that the proposed XGBRegression algorithm and KNeighbour Regression Algorithm can be used to Improve the Accuracy of Flight ticket prediction (Panwar et al. 2021).

Flight ticket prediction has limited price prediction ability based on future price significant profit. But there are so many other factors that affect the market value of the product that can not be forecasted precisely which makes the system a little difficult in price prediction. Deep learning algorithms can address future predictions by considering the dynamic factors that influence the market price of the trades and products.

5. Conclusion

The main aim of the study is to measure the accuracy of flight ticket prediction. In this research paper a XGBRegression model with the KNeighbour Regression. The results obtained show that the XGBRegression has found 87.61% of accuracy on the flight ticket prediction than the 49.17% of the K-neighbor Regression.

Declarations

Conflicts of Interest

No conflict of interest in this manuscript.

Author Contributions

Author SSV was Involved in data collection, data analysis, and manuscript writing. Author JVT was involved in the conceptualization, data validation, and critical review of the manuscript.

Acknowledgement

The author would like to express their sincere gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences for providing the necessary infrastructure to carry out this work successfully.

Funding

We thank the following organizations for providing financial support that enabled us to complete the study.

1. Qbec Infosol, Chennai.
2. Saveetha University
3. Saveetha Institute of Medical And Technical Sciences
4. Saveetha School of Engineering

6. References

Abulaish, Muhammad, and Mohd Fazil. 2021. "A Machine Learning Approach for Socialbot Targets Detection on Twitter." *Journal of Intelligent & Fuzzy Systems*. https://doi.org/10.3233/jifs-200682.

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.

Batra, Usha, Nihar Ranjan Roy, and Brajendra Panda. 2020. *Data Science and Analytics: 5th International Conference on Recent Developments in Science, Engineering and Technology, REDSET 2019, Gurugram, India, November 15–16, 2019, Revised Selected Papers, Part I*. Springer Nature.

Boruah, Abhijit, Kamal Baruah, Biman Das, Manash Jyoti Das, and Niranjan Borpatra Gohain. 2019. "A Bayesian Approach for Flight Fare Prediction Based on Kalman Filter." *Advances in Intelligent Systems and Computing*. https://doi.org/10.1007/978-981-13-0224-4_18.

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.

Dinesh Kumar, M., V. Godvin Sharmila, Gopalakrishnan Kumar, Jeong-Hoon Park, Siham Yousuf Al-Qaradawi, and J. Rajesh Banu. 2022. "Surfactant Induced Microwave Disintegration for Enhanced Biohydrogen Production from Macroalgae Biomass: Thermodynamics and Energetics." *Bioresource Technology* 350 (April): 126904.

Keretna, Sara, Ahmad Hossny, and Doug Creighton. 2013. "Recognising User Identity in Twitter Social Networks via Text Mining." *2013 IEEE International Conference on Systems, Man, and Cybernetics*. https://doi.org/10.1109/smc.2013.525.

Kumar, J. Aravind, J. Aravind Kumar, S. Sathish, T. Krithiga, T. R. Praveenkumar, S. Lokesh, D. Prabu, A. Annam Renita, P. Prakash, and M. Rajasimman. 2022. "A Comprehensive Review on Bio-Hydrogen Production from Brewery Industrial Wastewater and Its Treatment Methodologies." *Fuel*. https://doi.org/10.1016/j.fuel.2022.123594.

Lok, Johnny Ch. 2018. *Prediction Factors Influence Airline Fuel Price Changing Reasons*.

Mahesh, Narayanan, Srinivasan Balakumar, Uthaman Danya, Shanmugasundaram Shyamalagowri, Palanisamy Suresh Babu, Jeyaseelan Aravind, Murugesan Kamaraj, and Muthusamy Govarthanan. 2022. "A Review on Mitigation of Emerging Contaminants in an Aqueous Environment Using Microbial Bio-Machines as Sustainable Tools: Progress and Limitations." *Journal of Water Process Engineering*. https://doi.org/10.1016/j.jwpe.2022.102712.

Mohanavel, Vinayagam, K. Ravi Kumar, T. Sathish, Palanivel Velmurugan, Alagar Karthick, M. Ravichandran, Saleh Alfarraj, Hesham S. Almoallim, Shanmugam Sureshkumar, and J. Isaac JoshuaRamesh Lalvani. 2022. "Investigation on Inorganic Salts K2TiF6 and KBF4 to Develop Nanoparticles Based TiB2 Reinforcement Aluminium Composites." *Bioinorganic Chemistry and Applications* 2022 (January): 8559402.

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.

Ram, G. Dinesh, G. Dinesh Ram, S. Praveen Kumar, T. Yuvaraj, Thanikanti Sudhakar Babu, and Karthik Balasubramanian. 2022. "Simulation and Investigation of MEMS Bilayer Solar Energy Harvester for Smart Wireless Sensor Applications." *Sustainable Energy Technologies and Assessments*. https://doi.org/10.1016/j.seta.2022.102102.

Rinesh, S., K. Maheswari, B. Arthi, P. Sherubha, A. Vijay, S. Sridhar, T. Rajendran, and Yosef Asrat Waji. 2022. "Investigations on Brain Tumor Classification Using Hybrid Machine Learning Algorithms." *Journal of Healthcare Engineering* 2022 (February): 2761847.

Sathish, T., V. Mohanavel, M. Arunkumar, K. Rajan, Manzoore Elahi M. Soudagar, M. A. Mujtaba, Saleh H. Salmen, Sami Al Obaid, H. Fayaz, and S. Sivakumar. 2022. "Utilization of Azadirachta Indica Biodiesel, Ethanol and Diesel Blends for Diesel Engine Applications with Engine Emission Profile." *Fuel*. https://doi.org/10.1016/j.fuel.2022.123798.

Sudhan, M. B., M. Sinthuja, S. Pravinth Raja, J. Amutharaj, G. Charlyn Pushpa Latha, S. Sheeba Rachel, T. Anitha, T. Rajendran, and Yosef Asrat Waji. 2022. "Segmentation and Classification of Glaucoma Using U-Net with Deep Learning Model." *Journal of Healthcare Engineering* 2022 (February): 1601354.

Sundararaman, Sathish, J. Aravind Kumar, Prabu Deivasigamani, and Yuvarajan Devarajan. 2022. "Emerging Pharma Residue Contaminants: Occurrence, Monitoring, Risk and Fate Assessment – A Challenge to Water Resource Management." *Science of The Total Environment*. https://doi.org/10.1016/j.scitotenv.2022.153897.

Surrey Flying Services Limited. 1937. *Flight Ticket*.

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.

Turner, M., M. J. Griffin, and I. Holland. 2000. "Airsickness and Aircraft Motion during Short-Haul Flights." *Aviation, Space, and Environmental Medicine* 71 (12): 1181–89.

Vijayalakshmi, V. J., Prakash Arumugam, A. Ananthi Christy, and R. Brindha. 2022. "Simultaneous Allocation of EV Charging Stations and Renewable Energy Sources: An Elite RERNN-m2MPA Approach." *International Journal of Energy Research*. https://doi.org/10.1002/er.7780.

Yaashikaa, P. R., P. Senthil Kumar, S. Jeevanantham, and R. Saravanan. 2022. "A Review on Bioremediation Approach for Heavy Metal Detoxification and Accumulation in Plants." *Environmental Pollution* 301 (May): 119035.

Zhang, Mu. 2020. *A Natural Language Flight Ticket Searching System*.

Tables and Figures

Table 1. Comparison between XGBRegression and KNeighbour Regression with N=10 samples of the dataset with highest accuracy of respectively 82.52% and 62.81% in sample 1 (when N=1) using the dataset size= 9650 and the 70% of training and 30% of testing data

Sample (N)	Dataset Size	XGBRegression Accuracy in %	KNeighbour RegressionAccuracy 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. Statistical analysis of XGB and KNN Algorithm. Mean accuracy, Standard deviation, and standard error values are obtained for 20 sample data sets.

Group		N	Mean	Std. Deviation	Std. Error Mean
Accuracy	XGB	10	85.8530	1.39634	.44156
	KNN	10	48.1340	.71881	.22731

Table 3. An Independent sample T-test is performed for the two groups for significance and standard error determination. P>0.05 for wet basis.

	Levene's Test for Equality of Variances	T-test of Equality of Means					95% of the confidence interval of the Difference
				Sig (2-	Mean	Std.Error	

	F	Sig.	t	df	tailed)	Difference	Difference	Lower	Upper
Accuracy									
Equal Variance Assumed	5.586	.030	75.949	18	.000	37.7190	.49664	36.6756	38.76239
Equal Variance Not Assumed			75.949	13.457	.000	37.7190	.49664	36.679	38.78822

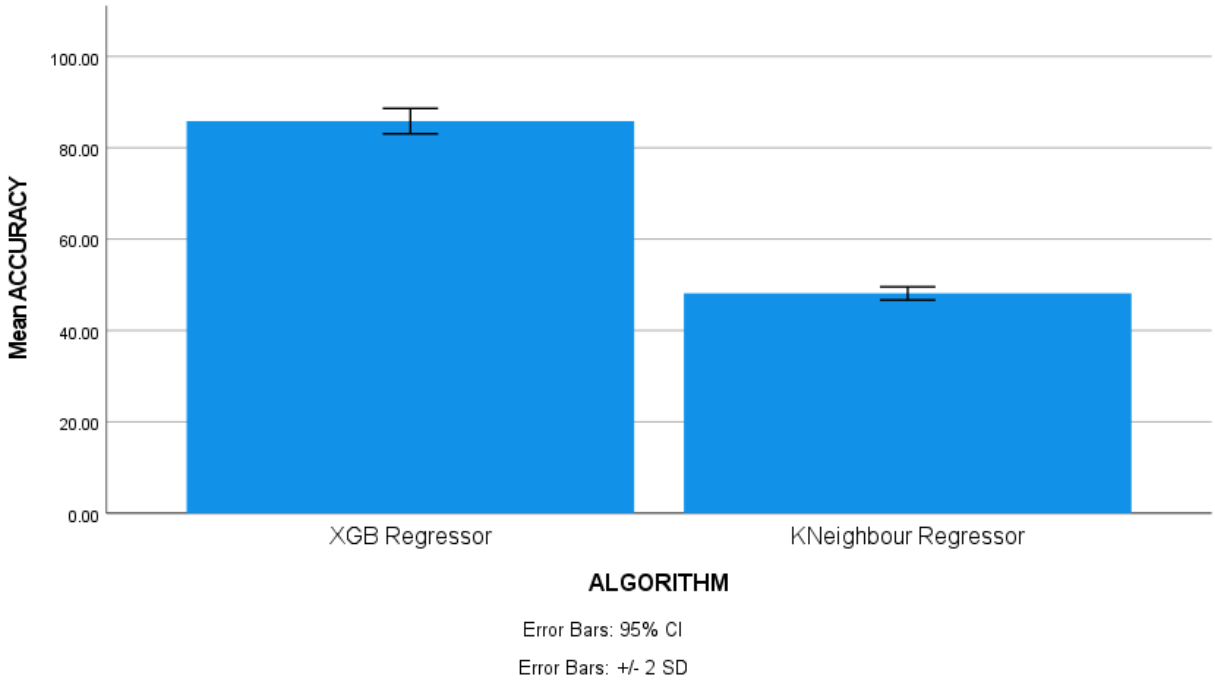

Fig. 1. Comparison of XGB Regression and KNeighbour Regression algorithm in terms of mean and accuracy. The mean accuracy of the XGB Regression is better than KNeighbour. X-axis: XGB Regression vs KNeighbour Regression, Y-axis: Mean accuracy of detection ±2SD.