IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.53

Volume 4, Issue 5, March 2024

An Ensemble Learning Method to Predict Airline Ticket Price Using Machine Learning

V. Ch. Jwala[1], K. Jahnavi[2], K. Mukthamukhi[3], K. Durga Madhavi[4], K. Jaya Lakshmi[5]

Assistant Professor, Department of Information Technology[1]

Students, Department of Information Technology[2,3,4,]

S.R.K.R. Engineering College, Bhimavaram, Andhra Pradesh, India

jwalavegesna@gmail.com, jahnavikasanneni0209@gmail.com

mukthamukhi23@gmail.com, kondurudurgamadhavi@gmail.com, jayalakshmikolli31@gmail.com

Abstract*: In recent years airline travel has become a common mode of transportation. People often prefer travelling through airlines to reach destination as early as possible. But the ticket booking for airlines is somewhat a hectic process because the price value changes day to day based on the availability of flights. So far so many Machine Learning algorithms were used to improve the prediction of ticket prices. This paper presents a comparison of two ensemble machine learning algorithms namely Random Forest and XGBoost to detect the airplane ticket price. Detecting airline price using XGBoost given more accuracy when compared to existing system.*

Keywords: Machine learning, Airline ticket, Random Forest, XGBoost.

I. INTRODUCTION

Airlines are used by various people for travelling purpose through one location to another location within less amount of time. It is considered as most sophisticated mode of transportation regardless of its ticket price. But booking ticket for an airline is difficult because of several reasons and the most common reason is change of the price day to day.

The ticket prices of an airline will always depend on change of the availability of seats, month of bookings, class of travelling and so on. For suppose when there is a huge demand for ticket booking of a particular airline then the ticket cost will also increase and at the same time if the number of seats available in the plane are booked but not filled then also there will be change in the price. So, there are many reasons that leads to the change of the Airline price.

Passengers travelling through Airlines regularly should be aware of the time at which the ticket price is low and should plan their trip or vacation accordingly. But to know the prices for each and every airline company one has to visit their website or corresponding application and should get their details. Hence, there are some Machine Learning [1] models that can be used to predict the airline ticket price based upon different parameters.

II. LITERATURE SURVEY

To predict the airline ticket price so far so many machine learning algorithms were used and all the work done till now has improved the efficiency of the problem. Not only the Machine Learning, Deep Learning [2] and Artificial Intelligence (AI) [3] can also be used.

Naresh Alapati et al. [4] in their paper has already discussed about the prediction of airline ticket price with an ensemble learning algorithm to improve the accuracy of the prediction.

Ankita Panigrahi et al. [5] tried to predict the flight fare price using machine learning algorithms which gives you the best time to purchase the ticket for travelling.

Archana Dirgule et al. [6], also provided an ensemble framework to predict the airline ticket price with the help of random forest.

III. MACHINE LEARNING MODELS

A. Definition: Machine learning is a part of Artificial Intelligence that is used to train a system by inputting some data and in response the trained model will be given as output. This model can further be used to predict the newly added

DOI: 10.48175/IJARSCT-16664

data value. There are variety of machine learning models available to predict the data. All these models are used to improve the accuracy of the system.

B. Classification of models: All the machine learning models or algorithms are mainly classified into three types. They are

- Supervised learning
- Unsupervised learning
- Reinforcement learning

a) Supervised learning: The data provided as input to this learning model is a labeled data which means a structured data. The problems classification[7]and regression [8] comes under supervised learning.

b) Unsupervised learning: Input data for this type of learning is an unlabeled data which means an unstructured data. Clustering[9] can be considered as a best example of unsupervised learning.

c) Reinforcement learning: It is a trial-and-error based learning that is used to get the accurate results for a given problem.

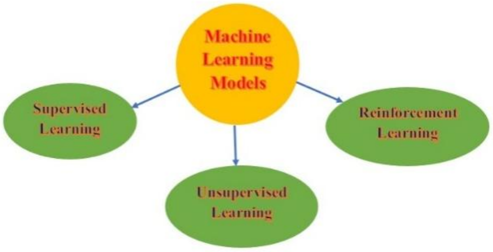

Fig 1: ML Models

IV. ENSEMBLE MODELS

Along with the above learning techniques there are some special models available in machine learning called Ensemble models[10]. These models are used widely to take a decision for solving a problem. This model works by taking into consideration of various parameters and other models. These models not only increase the accuracy but has an ability to recover the uncertain data.

The popular ensemble techniques available are Stacking [11], Blending [12], Bagging and Boosting. For predicting the price of an Airline ticket, both bagging and boosting techniques can be used.

A. Bagging: In bagging to get the final result, all the results of previous models were added. The input for the bagging technique is the dataset taken in the form of different subsets and all these subsets are evaluated with the help of bootstrapping [13] to increase the efficiency.

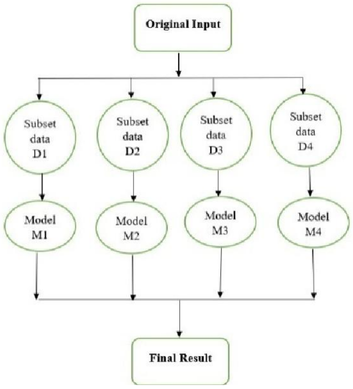

Fig 2: Bagging process

In order to implement the bagging technique on a dataset, there are some algorithms available and the most popular algorithm is Random Forest[14]. Random forests use decision trees [15] to analyse the dataset and the input for this algorithm is chosen randomly. A decision tree is constructed for each and every subset of the data and the final

DOI: 10.48175/IJARSCT-16664

prediction is obtained by doing average of all the results obtained by these decision trees. Since the bootstrapping technique is used to build the decision trees, the trees obtained are more complex and less flexible.

B. Boosting: To overcome the problem encountered in bagging, there is another ensemble technique available known as Boosting.

In boosting, initially a model is created for a subset of original data and a model is built. This model is further used to predict the complete dataset. The errors occurred while predicting the dataset are again fed along with new subset which produces another model. This produced model tries to correct the results of the previous model that were predicted wrongly. The process is continued until the entire training dataset is predicted correctly or maximum models were built. In this way the boosting algorithm combines all the wrongly predicted data to improve the total efficiency of the system.

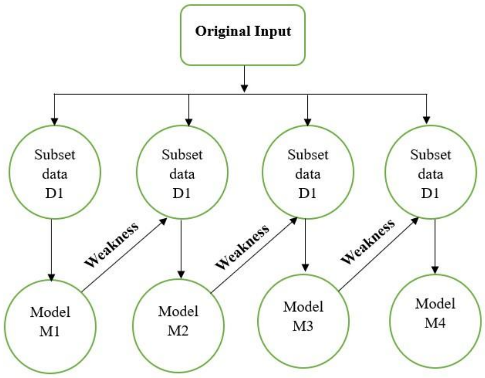

Fig 3: Boosting process

Both the bagging and boosting techniques were used to improve the accuracy of the system, but boosting gives more accuracy when compared to bagging.

To implement boosting technique there are many algorithms but here XGBoost algorithm is used.

V. METHODOLOGY

XGBoost stands for *Extreme Gradient Boosting* is one of the algorithms that comes under gradient boosting technique. The errors present in the current model are fed into the next model to correct the errors. During training phase to reduce the loss an optimization technique called Gradient Descent [16] is used.

XGBoost is an implementation of gradient boosting where every independent variable present in the dataset is assigned with some weights and decision trees will be constructed. The error occurred in one decision tree is fed into another decision tree to train them more precisely. All these errors were ensembled together to build a strong model.

The following is the working methodology for the XGBoost algorithm.

1. Initializing the model: The algorithm starts by setting an initial predictionfor each data point. That prediction is usually the average value of the target variable.
2. Calculating the residuals: The model then calculates the residualsor errors which is the difference between initial prediction and the actual target values.
3. Building a decision tree: Decision trees are used to predict the errors. The decision tree is built by recursively splitting the data based on different features and creates rules to make predictions.
4. Updating the prediction: To get new predictions, the trees are updated time to time which becomes the new target for the new decision tree.
5. Repeating the process: To get the accurate predictions, Steps 2 to 4 must be repeated for a limited number of iterations. Each iteration focuses on improving the model's predictions and reducing the overall error.
6. Combining the predictions: After the construction of all decision trees, to obtain the final prediction XGBoost combines all the predictions of decision trees. This is done by summing up the predictions from all the trees.

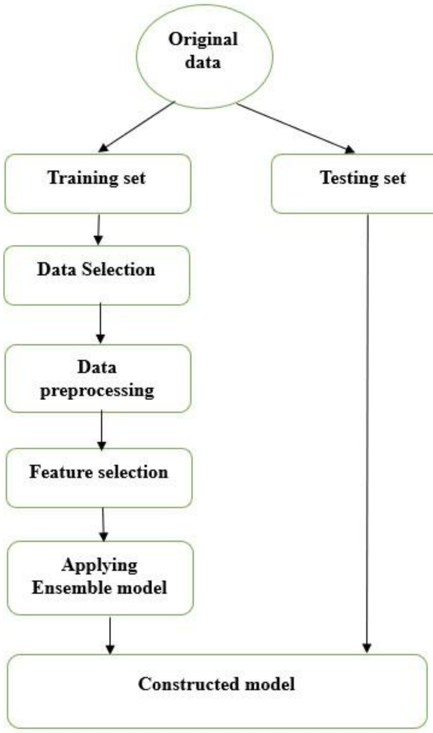

Fig 4: Proposed System Architecture

The dataset is considered from Kaggle which contains a total of 10002 records, each variable is separated by a comma delimiter providing details like name of the airline, date of journey, source, destination, route, departure time, arrival time, duration, and total stops. This data is divided into two parts namely training set with a percentage of 80 and remaining 20 percent as test set.

The training set is first pre processed to eliminate the noisy and unwanted data and then the features are extracted to apply the ensemble method for constructing a model. This model is then used for testing the remaining 20% data. If the accuracy observed is good then the model is applied to predict the new data.

VI. RESULTS AND DISCUSSION

The results obtained after using XGBoost algorithm on airline ticket price dataset were quite good when compared to the results occurred using random forest. The results can be represented in the form of accuracy for both the algorithms and the following graph shows the comparison of ticket price prediction for both the algorithms.

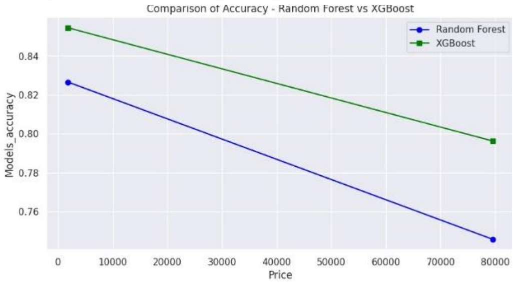

Fig 5. Graph representing the accuracy of two ML models

In the above graph along the X-axis the various prices were denoted and along Y-axis the accuracy levels were used. For the same dataset if the two models' random forest and XGBoost were used then the latter one gave the more accuracy with a percentage of 85 when compared to the first algorithm which gave only 82 percent.

VII. CONCLUSION

In conclusion, the paper provides prediction of optimal flight ticket price for customers who are willing to preplan their journey in airline. By using ensemble machine learning models and historical data, the paper demonstrates the accuracy of the prediction.

REFERENCES

[1]. Machine Learning. Retrieved from https://en.wikipedia.org/wiki/Machine_learning

[2]. Deep Learning. Retrieved from https://www.ibm.com/topics/deep-learning

[3]. Artificial Intelligence. Retrieved from https://www.britannica.com/technology/artificial-intelligence

[4]. Naresh Alapati et al., “Prediction of Flight-fare using machine learning”, Conference paper in Research Gate.

[5]. Ankita Panigrahi et al., “Flight Price Prediction Using Machine Learning”, Proceedings of the Workshop on Advances in Computational Intelligence, its Concepts & Applications, Vol. 3283, pp. 172-178, 2022.

[6]. Archana Dirgule et al., “Flight Fare Prediction using Random Forest Algorithm”, International Journal of Advanced Research in Science, Communication and Technology, Issue 3, Vol 2, pp 659-662, May 2022.

[7]. Classification. Retrieved from https://www.datacamp.com/blog/classification-machine-learning

[8]. Regression. Retrieved from https://builtin.com/data-science/regression-machine-learning

[9]. Clustering. Retrieved from https://www.javatpoint.com/clustering-in-machine-learning

[10]. Ensemble models. Retrieved from https://www.toptal.com/machine-learning/ensemble-methods-machine-learning

[11]. Stacking. Retrieved from https://www.scaler.com/topics/machine-learning/stacking-in-machine-learning/

[12]. Blending. Retrieved from https://www.scaler.com/topics/machine-learning/blending-in-machine-learning/

[13]. Bootstrapping. Retrieved from https://www.mastersindatascience.org/learning/machine-learning-algorithms/bootstrapping/

[14]. Random Forest. Retrieved from https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/

[15]. Decision trees. Retrieved from https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/

[16]. Gradient Descent. Retrieved from https://www.javatpoint.com/gradient-descent-in-machine-learning

DOI: 10.48175/IJARSCT-16664