International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 6 June 2021, pp: 1123-1127 www.ijaem.net ISSN: 2395-5252

Airfare Price Prediction using various Machine Learning Models

Ankita Harkude#1, Sarika Namade#2, Saloni Gholapl#3
1,2,3#Department of Information Technology and Computer Science and Technology,
Usha Mittal Institute of Technology, SNDT Women's University,
Juhu-Tara Road, Sir Vitthaldas Vidyavihar, Santacruz(W), Mumbai 400049

Submitted: 01-06-2021 Revised: 14-06-2021 Accepted: 16-06-2021

ABSTRACT—Nowadays air travelling is getting popular in our country and hence people are seeking to get the lowest price possible. While the airlines are using various strategies and methods to predict flight prices in a smart fashion. They keep their profit maximized and keep the revenue high. Due to this it becomes difficult for the customer to buy a ticket at the minimum cost as the price changes dynamically. This paper aims at implementing the machine learning regression methods to predict the prices at the desired time so the customer can decide a proper airline according to their budget.

Index Terms—Machine Learning Algorithm, Predictor, Airfare, Random forest, Naive Bayes.

I. INTRODUCTION

Any person who has booked a flight ticket very well knows how the price changes dynamically with the time, season, holidays, etc. That is why the customers try to book the tickets prior to the departure date to avoid the high prices at peak time. This is the reason why many techniques are being made using AI and Machine learning models to predict the price in the given time so the customer has a clear Idea beforehand. In this paper, we have tried to predict the price of the flight of several airlines from the year 2018 to 2020 and of several cities by making a flask web app.

II. LITERATURE SURVEY

Nowadays, buying a ticket at minimum price is not easy for the customer. To give the price of an air ticket many techniques are used. Machine learning models and Artificial intelligence are the most useful techniques. Using 75.3 percent precision the PLSR(Partial Least Square Regression) model is connected to get least cost of aircraft ticket buying to acquire greatest presentation from Utilizing AI models. William Groves and Maria Gini[2] took the Partial Least Square Regression(PLSR) for developing a model for predicting the best booking time for airline tickets. The data collected was from travel journey websites from 22 February 2011 to 23 June 2011. Further to this, additional data were also collected and used to check the comparisons of the performances of the final model.

Janssen [1] built up an expectation model utilizing the Linear Quantile Blended Regression strategy for San Francisco to New York course with existing every day airfares given by www.infare.com. The model has utilized two highlights which includes the number of days left until the departure date and whether the flight date is at the end of the week or weekday. The model predicts airfare very well before the days that are a long way from the departure date, anyway for a considerable length of time close to the takeoff date, the expectation isn't compelling.

Wohlfarth [3] proposed a airline ticket buying time enhancement model that is dependent on an extraordinary pre preparing step which is known as information mining systems (arrangement and bunching) and macked point processors and measurable investigation strategy. This system was mainly proposed to change the heterogeneous value arrangement information into added value arrangement direction that can be bolstered to unsupervised grouping calculation. The value direction is collected or grouped into gathering dependent on comparative estimating conduct. Advancement model gauge the value chain designs. It is a tree based order calculation used to choose the best coordinating group and then later on comparing the advancement model.
Before buying a flight ticket this model gives the most extreme number of days.

The ideal buying time dependent on nonparametric isotonic relapse methods for particular carriers,timeframe and course is recommended by DominguezMenchero [5]. For the expectation two sorts of the variables are

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 6 June 2021, pp: 1123-1127 www.ijaem.net ISSN: 2395-5252

considered. One is the date of procurement and another is the passage. This model is very helpful to buy flight tickets. Before the buying ticket of any flight this model gives the customer the most correct price.

III. DATA COLLECTION

We have tried to make a web spider that extracts data from a website and stores it in a csv file. Different sources from API's to customer travel sites are accessible for information scratching.

A. Data Collection

Data collection is performed using two data sets i.e. Data Train and Test Data. To train the data combination of categorical and numerical data is used. The test data is also the same as a data train except the price column. In data collection, feature engineering is the most important step as it includes cleaning the unwanted data, creating specific columns. So only the following features are considered as all are not required-

- Date of journey
- Time of Departure
- Place of Departure
- Time of Arrival
- Place of Destination/Arrival
- Airway company
- Total Stops
- Price

B. Cleaning and preparing data

Data needs to be cleaned and prepared according to the model requirements. After this is done it is further analysed and distribution is performed. The unnecessary data is removed like duplicates and null values. Several statistical methods are used to prepare and clean the data. For example, departure time and arrival time is split into hour and minute and converted into integer.

C. Analyzing data

After data cleaning and preparation machine learning mod-els are applied and some features are calculated on the basis of existing features. These include Random Forest, Logistic Regression, Gradient Boosting and combination of these mod-els to increase the accuracy. Also, time plays an important role. So, the time can be divided up as: morning, evening and night.

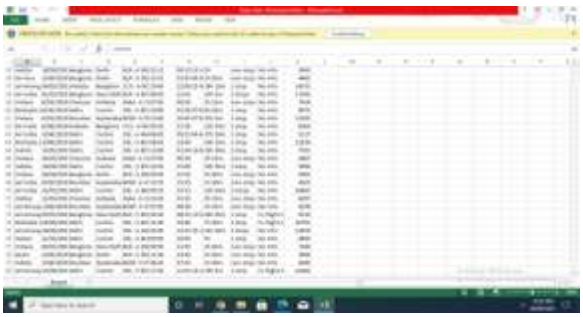

Fig.1.Dataset

IV. MACHINE LEARNING PERFORMANCE MODELS

Now,we have to predict the prices of the flight ticket for that we introduced many algorithms in machine learning which are: Support Vector Machine (SVM) , Linear regression, K Nearest neighbours , Decision tree,MultilayerPerceptron,Gradient Boosting and Random Forest Algorithm. For implementing these models we use python library scikit learn. To verify the performance of these models different parameters are considered which are R-square, MAE and MSE.

A. Linear Regression

Simple linear regression analysis is used to determine the correlation between two continuous variables. Predictor variable is the one of the two variables to which value we have to find. Linear regression does not give a deterministic relation but gives a statistical relationship between two variables. It gives prediction error which is minimum from best fit line of given data.To understand the linear regression two major factors used which is Gradient descent and cost function. Equation for linear regression is : y(pred) = b0+b1 x (1) We have to choose the value of b0 and b1 so the error value became as small as possible. It shows the difference between actual and square of predicted value. Mean square error(MSE) is used to deal with negative values. Here b1 means bias and b0 is used for giving the positive or negative relationship between x and y. R-squared, MAE and

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 6 June 2021, pp: 1123-1127 www.ijaem.net ISSN: 2395-5252

MSE these terms are used to measure the accuracy of regression problems.

B. Decision Tree

To make comparative same time persistent this tree count small subsets from isolating the collected information. The final result shows that tree with decision centres like the leaf centres.It contains two branches at any rate. We have to think about the root as an informational index. Then we have to discretize the model before structuring it. For the decision of tree computation information Gain and Gini index is essential. And it is defined as a change amount in entropy. Basic squared conditions for regression tree: Y is predicted value and having maximum number of expected value. The training example on leaf nodes assigned to stop slow and overfitting the model.

C. K-Nearest Neighbours(KNN)

K-nearest neighbour algorithm is the type of supervised ML classification algorithm that can also be used as a regression. The k-nearest algorithm is one of the most used ML algorithms due to its simplicity. In k-nearest neighbour regression analysis, the output is mean of its k nearest neighbours. Like SVM this is also a non-parametric method. Considering few values, results are computed to achieve the best value. It assigns a new data point to the class. It is non-parametric because it does not take any assumption. KNN keeps all training data since they are needed during the testing phase. K- entries in the data set are picked by the model that are close to the new data point.

D. Random Forest

Forest creates large models from aggregating the base model. To produce better predictive models it ensembles the less predictive model. In order to obtain the highly uncorrelated decision trees the features are sampled and then passed to trees without any replacement. It required less correlation between trees to select the best split. Aggregated uncorrelated trees are the main concept which makes it different from decision trees. It maintains accuracy and handles missing values for the missing data. It is basically a bagging technique. Hyperparameters as a decision tree or bagging classifier are nearly the same as the random forest. Over fitting is an error which occurs when a function is closely fit with a limited set of data points that is the reason why Random Forests do not over-fit.

V. EXPERIMENTAL RESULTS

Output of the model is plotted for the selected test dataset across the test dataset. Comparative study of original values and predicted results are shown by Graphs. The predicted values of the fare to purchase the fight ticket at the right time given by the analysis of results which is obtained from the algorithm such as Decision Tree, Random Forest, KNN, Linear Regression. Below table gives values of R-square, MAE, MSE. The given graph is plotted between the fare of the flight versus the days left until departure. The red color line shows the predicted value of flight tickets whereas the blue color line denotes the actual value of the flight ticket. Fig.3 shows a plot between Days remaining for the departure versus Actual and predicted values evaluated by Random Forest Algorithm. Compared to other algorithms for a given dataset, the Decision Tree algorithm has more accuracy. In the regression analysis it gives the highest R-square value with maximum accuracy.

A. Algorithm Evaluation

ML Algorithms	R-squared	MAE	MSE
Decision Tree	0.67	0.13	0.21
Random Forest	0.68	0.13	0.21
K-NN	0.65	0.13	0.22
Linear Regression	0.40	0.19	0.29

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 6 June 2021, pp: 1123-1127 www.ijaem.net ISSN: 2395-5252

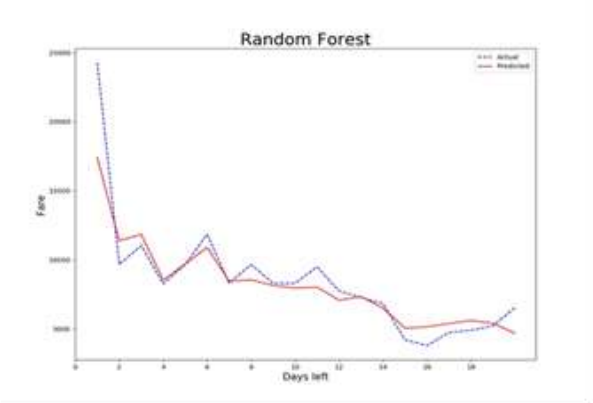

Fig. 2. Graphical result for random forest

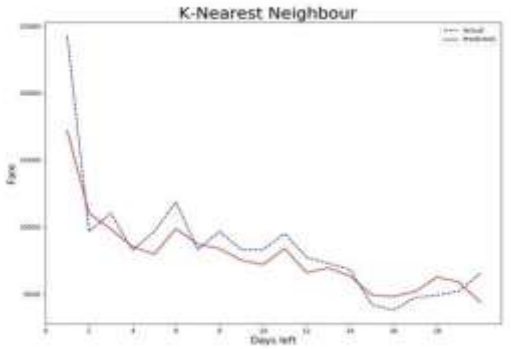

Fig. 3. Graphical result for K-Nearest Neighbour

VI. ACKNOWLEDGMENT

Foremost, We would like to express our sincere gratitude to Professor Dr. Sanjay Shitole and Dr.Sanjay Pawar Head of Department Information Technology and Computer Science and Technology respectively and our guide, Dr.Sanjay Pawar Sir for his valuable guidance during the Major project phase. My sincere gratitude to Dr. Sanjay Pawar, Principal (Usha Mittal Institute of Technology)for his valuable encouragement. We would like to give special thanks to our parents and friends for their valuable time and support.

VII. CONCLUSION AND FUTURE SCOPE

We gathered airfare data from the web and showed that it is feasible to predict prices for flights based on historical fare data. The experimental results show that ML models are a satisfactory tool for predicting airfare prices. Other important factors in airfare prediction are the data collection and feature selection from which we drew some useful conclusions.

BIBLIOGRAPHY

[1]. T. Janssen, —A linear quantile mixed regression model for prediction of airline ticket prices,‖ Bachelor Thesis, Radboud University, 2014

[2]. William Groves and Maria Gini Department of Computer Science and Engineering University of Minnesota, USA groves,gini@cs.umn.edu

[3]. Wohlfarth, T. Clemencon, S.Roueff, —A Data mining approach to travel price forecasting‖, 10 th international conference on machine learning Honolulu 2011

[4]. Viet Hoang Vu, Quang Tran Minh and Phu H. Phung,"An Airfare Prediction Model for Developing Markets ", IEEE paper 2018

[5]. Dominguez-Menchero, J.Santo, Riviera, ‖optimal purchase timing in airline markets‖, 2014

[6]. P. Malighetti, S. Paleari and R. Redondi, "Pricing strategies of low-cost airlines: The Ryanair case study," Journal of Air Transport Management, vol. 15, no. 4, pp. 195-203, 2009

[7]. Supriya Rajankar and Neha Sakharkar, —A Survey on Flight Pricing Prediction using MachineLearning International Journal of

Engineering Research and Technology, vol 8, issue 6, June 2019
[8]. M. Papadakis, "Predicting Airfare Prices," 2014
[9]. R. Ren, Y. Yang and S. Yuan, "Prediction of airline ticket price," Technical Report, Stanford University, 2015
[10]. Qiqi Ren,"When to book:Predicting Flight Pricing", Stanford University