NAME OF THE PROJECT

FLIGHT PRICE PROJECT USING MACHINE LEARNING

Submitted by:

VARSHA V. SHINDE

ACKNOWLEDGMENT

We first would like to thank god for giving us the opportunity to be here and make this project. Secondly, we thank our families for their unwavering support and great sacrifices in order to help us reach this moment. We would also like to thank the faculty of Mohd Kashif sir and FLIP ROBO for their work and efforts to give us the best education possible in order to reach our full potential, especially our SME Mr.Mohd Kashif for his guidance and assistance in not only this but Throughout our entire journey

INTRODUCTION

Perfect time for purchasing plane ticket by the passenger's view is difficult since passengers get very less information of future business price rates. Different models figure out future business price on plane and categorise the best time to obtain flight ticket.

Airlines use different strategies of pricing for their tickets, later taking the decision on price because order shows higher value for the approximation models. The causes behind the difficult system are each Plane has limited number of seats to be filed, so airlines must regulate demand. Suppose when demand is expected to increase capacity, the airline may increase prices, to decrease the rate at which seats fill.

Also, seating arrangements in flight which is not occupied shows the loss of the amount invested for the business airline companies and making them purchase the ticket to fill the seats for any price this would be the best idea to get profit in loss too.

Passengers should be compatible with the airline companies to get adjusted for the increase and decrease of the price. Passengers or customers should make their own planning to get the best offers available on different airlines and travel through less price.

Planes ticket prices changes as time passes, pulling out the elements which creates the difference.

Reporting the correlated and models which is used to price the flight tickets. Then, using that information, building the model which helps passengers to make pull out the ticket to buy and predicting air ticket prices which progresses in the future. Duration, Arrival time, Price, Source, Destination and much more these are the attribute used for flight price prediction.

ANALYTICAL PROBLEM FRAMING

DATA COLLECTION

The important a part of the project is data collection. Data on different websites is gathered with unique attribute to provide the best accuracy. The data is collected from website YATRA.com and the models are implemented using python.

The python-script helps to easily pre-process the data and forecast the output. The duplicate values are avoided in the pre-processing step. This dataset is more concentrated on calculating the plane price value. The dataset contains the data with attributes such as

- Journey_Date
- Departure
- Designation
- Arrival
- Airline
- Duration
- Source
- Price

CLEANING AND PREPARING DATA

The gathered data must be cleaned and pre-processed and after improving the data, it is read to run on the algorithms. The duplicate values are removed, data is arranged with numerical values by pre-processing and by this model building and selecting the features becomes easier. Pre-processing plays the vital role for the whole dataset.

ANALYZING DATA

Constructing of the data is the huge task, by knowing the various patterns of data visualization and later using the required machine learning models. Also, from the current attribute the new small elements can be acquired. If it is on holiday, festival day or a weekday or weekend, plane date plays main role. Travelling during weekends is costlier than the planes on weekdays and time is considered in classes as: Morning, afternoon, evening and night, so time plays important role. Travelling days is computed with plane date and the date on which data is collected.

MODEL/S DEVELOPMENT AND EVALUATION

Proposed Model through Regression Analysis the visualization and forecasting are performed for the presented model. Blending of technologies, processing is called conceptually the Intelligence that is cloud computing, machine learning and virtualization etc.

ML is in trend to build our skills and it is one of the highest growth fields in computer science and health care informatics. As the time passes by the algorithm should be learnt is the main goal in Machine.

Also, used for predicting algorithm that makes the communication with agent and makes easier for learning. In this paper, random forest and decision tree algorithms is to find solutions for flight price problems in machine learning tasks and a hybrid method is formulated from Chi square, anova and Correlation tests is performed.

The data collecting is performed followed by data pre-processing. Before data modelling is done, data must be split into train and test dataset to ignore the data leakage.

Based on the various attributes in the dataset for example departure and arrival features play the important role for predicting the price. Running the random forest and decision tree models, grouping the maximum price of airlines. Next performing the feature engineering and calculating the accuracy

Flight price forecasting using various algorithms in machine learning. The algorithms for forecasting purpose are: Support Vector Machine, Linear regression, K-Nearest neighbours, Multilayer Perception, Gradient Boosting and Random Forest Algorithm, Decision tree. Traversing the python library and parameters like R-square, MAE and MSE area unit to verify the production of those models.

LINEAR REGRESSION

Variable quantity of that price is to be found for this we are employing statistical regression analysis such as correlation between 2 continuous variables, from the 2 variables. The equation for statistical regression is: y (pred) = b0+b1 $*$ x (1) the two major factors to grasp statistical regression is gradient descent and price operate area unit. It gives the simplest match line to the given data that the forecast error is minimum and provides the applied mathematics relationship not the settled relationship between 2 variables. The sq. of expected and actual price distinction gives the error. To alter the negative values, the mean sq. error is taken (MSE). Value of the coefficients b1 and b0 area unit chosen in order that the error value is as little as doable. Choosing a random data from a dataset with replacement is called Bootstrap aggregating. By gradient boosting and random forest strategies achieves the greater accuracy.

DECISION TREE

It is used to make any decision and have multiple branches which are the Decision Node and Leaf Node. Decision Tree used for both classification and Regression problems, but mostly it is preferred for solving Classification problems which is a supervised learning technique. Decision tree has the two nodes, represents the features of a dataset, each leaf node shows the outcome is internal nodes and branches shows the decision rules. Based on features of the given dataset the test must be performed. For getting all the possible solutions to a problem on the given condition's visualization is done.

RANDOM FOREST

One of the popular machine learning algorithms which belongs to the supervised learning technique is Random Forest. Process of combining multiple classifiers to solve complicated problem and increasing the performance of the model it is based on the concept of learning. To avoid the over fitting problem demand the greater number of trees in the forest tends to greater accuracy. Random forest is used for both classification and regression problems are the huge advantage. The resultant accuracy the random forest gives are 70% as shown in the result and analysis graph and table.

CONCLUSION

Evaluating the algorithmic rule, a dataset is collected, pre-processed, performed data modelling and studied a value difference for the number of restricted days by the passengers for travelling. Machine learning algorithms with square measure for forecasting the accurate fare of airlines and it gives accurate value of plane price ticket at limited and highest value.

Information is collected from Kaggle websites that sell the flight tickets therefore restricting data which are often accessed. The result obtained by the random forest and decision tree algorithm has better accuracy, but best accuracy is predicted by decision tree algorithm as shown is the above analysis. Accuracy of the model is also forecasted by the R-squared value.

In Upcoming days when huge amount of information is accessed as in detailed information in the dataset, the expected results in future are highly correct. For further research anyone desire to expand upon it ought to request different sources of historical data or be a lot of organized in collection knowledge manually over amount of your time to boot, a lot of different combination of plane are going to be traversed.

There is whole possibility that planes differ their execution ideas consisting characteristics of the plane. At last, it is curious to match our model accuracy with that of the business models accuracy offered nowadays.