

FLIGHT\_PRICE\_PREDICTION

Submitted by:

Harshika Mahesh

**ACKNOWLEDGMENT**

Optimal timing for airline ticket purchasing from the consumer’s perspective is challenging principally

because buyers have insufficient information for reasoning about future price movements. In this project we majorly targeted to uncover underlying trends of flight prices in India using historical data

and also to suggest the best time to buy a flight ticket.

For this project, we have collected data from 18 routes across India while the data of 4 routes were extensively used for the analysis due to the sheer volume of data collected over 4 months resulting in

5.28 lakh data points each across the Mumbai-Delhi and Delhi-Mumbai route and 1.05 lakh data points each across the Delhi-Guwahati and Guwahati-Delhi route.

The project implements the validations or contradictions towards myths regarding the airline industry, a comparison study among various models in predicting the optimal time to buy the flight ticket and the amount that can be saved if done so. A customized model which included a combination of ensemble and statistical models have been implemented with a best accuracy of above 90% for a few routes, mostly from Tier 2 to metro cities.

These models have led to significant savings and produced average positive savings on each transaction.Remarkably, the trends of the prices are highly sensitive to the route, month of departure, day of departure, time of departure, whether the day of departure is a holiday and airline carrier. Highly competitive routes like most business routes (tier 1 to tier 1 cities like Mumbai-Delhi) had a non-decreasing trend where prices increased as days to departure decreased, however other routes (tier 1 to tier 2 cities like Delhi - Guwahati) had a specific time frame where the prices are minimum. Moreover, the data also uncovered two basic categories of airline carriers operating in India – the economical group and the luxurious group, and in most cases, the minimum priced flight was a member of the economical group. The data also validated the fact that, there are certain time-periods of the day where the prices are expected to be maximum. With a high probability (about 20-25%) that a person has to wait to buy a ticket, the scope of the project can be extensively extended across the various routes to make significant savings on the purchase of flight prices across the Indian Domestic Airline market.

Background & Objective

Anyone who has booked a flight ticket knows how unexpectedly the prices vary. Airlines use using sophisticated quasi-academic tactics known as "revenue management" or "yield management". The cheapest available ticket for a given date gets more or less expensive over time. This usually happens as an attempt to maximize revenue based on - 1. Time of purchase patterns (making sure last-minute purchases are expensive) 2. Keeping the flight as full as they want it (raising prices on a flight which is filling up in order to reduce sales and hold back inventory for those expensive last-minute expensive purchases) So, if we could inform the travellers with the optimal time to buy their flight tickets based on the historic data and also show them various trends in the airline industry we could help them save money on their travels. This would be a practical implementation of a data analysis, statistics and machine learning techniques to solve a daily problem faced by travellers.

The objectives of the project can broadly be laid down by the following questions -

1. Flight Trends

Do airfares change frequently? Do they move in small increments or in large jumps? Do they

tend to go up or down over time?

2. Best Time To Buy

What is the best time to buy so that the consumer can save the most by taking the least risk? So

should a passenger wait to buy his ticket, or should he buy as early as possible?

3. Verifying Myths

Does price increase as we get near to departure date? Is Indigo cheaper than Jet Airways? Are

morning flights expensive?

Main body of text

❏Automated Script to Collect Historical Data

For any prediction/classification problem, we need historical data to work with. In this project,

past flight prices for each route needs to be collected on a daily basis. Manually collecting data

daily is not efficient and thus a python script was run on a remote server which collected prices

daily at specific time.

❏Cleaning & Preparing Data

After we have the data, we need to clean & prepare the data according to the model's

requirements. In any machine learning problem, this is the step that is the most important and

the most time consuming. We used various statistical techniques & logics and implemented

them using built-in R packages

❏Analysing & Building Models

Data preparation is followed by analysing the data, uncovering hidden trends and then applying

various predictive & classification models on the training set. These included Random Forest,

Logistic Regression, Gradient Boosting and combination of these models to increase the

accuracy. Further statistical models and trend analyzer model have been built to increase the accuracy of the ML algorithms for this task.

❏Merging Models & Accuracy Calculation

Having built various models, we have to test the models on our testing set and calculate the savings or loss done on each query put by the user. A statistic of the over Savings, Loss and the mean saving per transaction are the measures used to calculate the Accuracy of the model implemented.

Method

Figure 1 : Overview of the model

❏Data Collection

Since the APIs by Indian companies like Goibibo returned data in a complex format resulting in a lot of time to clean the data before analysing, therefore we decided to build a web spider that extracts the 4

required values from a website and stores it as a CSV file. We decided to scrape travel service providers

website using a manual spider made in Python. Further we also developed a Python script to run the API

provided by Google flights which is more reliable, but it allows only 50 queries each day.

Such scrapping returns numerous variables for each flight returned and we had to decide the

parameters that might be needed for the flight prediction algorithm. Not all are required and thus we

selected the following -

1. Origin City

2. Destination City

3. Departure Date

4. Departure Time

5. Arrival Time

6. Total Fare

7. Airway Carrier

8. Duration

9. Class Type - Economy/Business

10. Flight Number

11. Hopping - Boolean

12. Taken Date - date on which this data was collected

❏Data Cleaning

The data was further processed based on the parameters mentioned below and cleaned based on

appropriate considerations -

1. Days to Departure

2. Day of Departure

3. Duration

4. Hopping

5. Holiday

6. Outliers

Further, the data was analysed and tests on the distribution were performed. Conclusions of the tests

revealed that our data followed Log-Normal distribution and the same has been positively confirmed

through statistical methods.

Based on previous history, the trend in the flight prices were modelled and the same was used to

provide the user with an approximation of the number of days to wait from the current day, and if at all

he waits, the amount he can say on the ticket.

In order to predict if the customer has to wait or not, we used a combination of statistical models and

machine learning models. The statistical model provided with a probability corresponding to each airline

having the least cost while the machine learning model further went ahead to predict the specific

conditions taking into account the days to departure and the day of departure.

The machine learning algorithms implemented started off with basic Regression models and were

extended to Decision Trees followed by Random Forests and Gradient Boosting methods. Later we

developed an algorithm which had a combination of Rule based learning, Ensemble models and

Statistical models to increase the accuracy.

Based on the prediction made by the model and the estimated time to wait, we calculated the savings

we could achieve and the losses we incurred based on the predictions.

❏Data Preparation

Data preparation was a critical part, as we had multiple airlines on a specific day and we had to predict

the future prices for all those airlines, or the airline which would have the lowest fare.

1. We do the predictions for each flight id. The problem with this is that, if there is a change in

flight id by the airline (which happens frequently) or there is an introduction or a new flight for a

specific route then our analysis would fail.

2. We group the flight ids according to the airline and the time of departure and do the analysis on

each group. For this we need to combine the prices of the airlines lying in that group such that

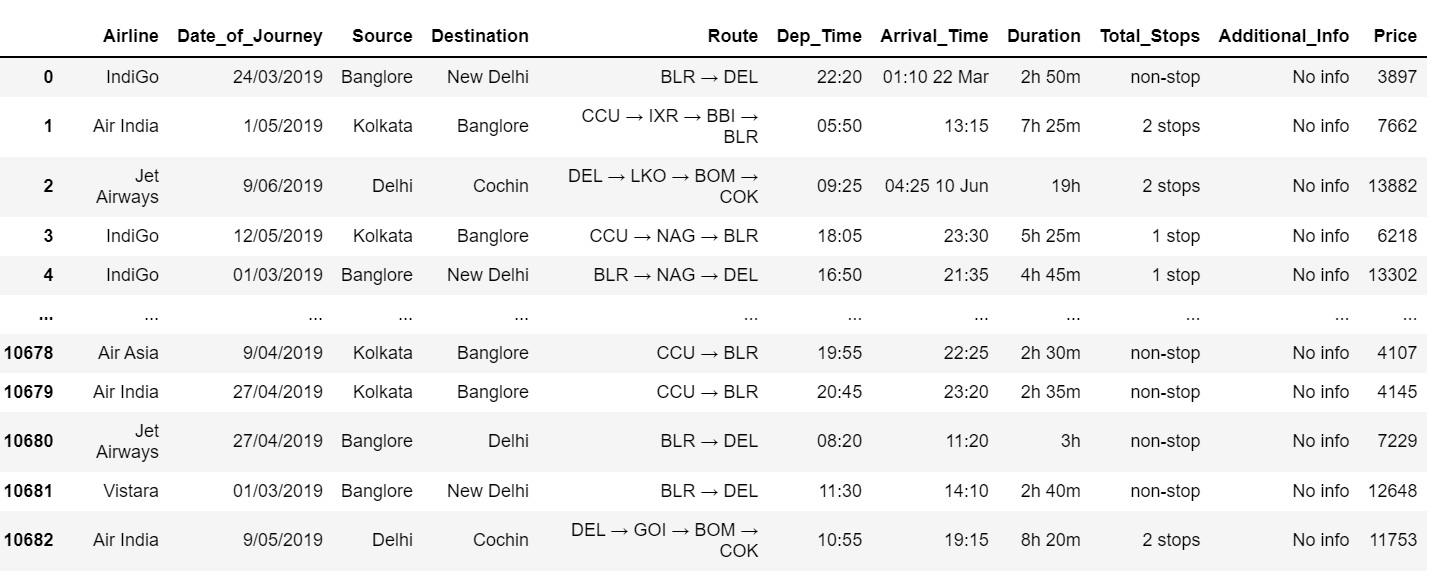
the basic trend in captured.

Moving ahead with the second option, we created the group according to the airlines and the departure

time-slot created earlier (Morning, Evening, Night) and calculated the combined flight prices for each

group, day of departure and depart day. Since these three are the most influencing factors which

determine the flight prices. Also, we calculated the average number of flights that operated in a particular group, since competition could also play a role in determining the fare



MACHINE LEARNING ALGORITHM

To develop the model for the flight price prediction, many conventional machine learning algorithms are evaluated. They are as follows: Linear regression, Decision tree[8], Random Forest Algorithm[9], K-Nearest neighbors[7], Multilayer Perceptron[10], Support Vector Machine (SVM) [11]and Gradient Boosting. All these models are implemented in the scikit learn. To evaluate the performance of this model, certain parameters are considered. They are as follows: R-squared value, Mean Absolute Error (MAE) and Mean Squared Error (MSE). The formulas for these three parameters are as follows:

(1)

(2)

(3)

Linear Regression

Regression is a method of modeling a target value based on predictors that are independent. It is mostly based on the number of independent variables and the relationship between independent and dependent variables. linear regression is a type of analysis where the number of independent variables is one and the relationship between the dependent and independent variables vary linearly. The important concept to understand linear regressions are cost function and Gradient decent.

y(pred) = b0+b1 x (4)

Decision tree

The Decision tree calculation separates the informational collection into small subsets, at a similar same time it creates gradually. The last outcomes are the tree with the decision nodes, whats more, the leaf nodes. A decision hub may have at least two branches. In the beginning, consider the entire informational collection as root. Highlight esteems are wanted to be downright. On the off chance that the qualities are constant then they are discretized before structure the model. Based on characteristic qualities records are dispersed recursively. There are two primary characteristics in the decision tree calculation. One is Information Gain and another is the Gini index. Information Gain is the proportion of Change in entropy. Higher the entropy more the instructive substance, where the entropy is a proportion of vulnerability of arbitrary variable. Gini Index is a component that measures how frequently an arbitrarily picked

component would be mistakenly distinguished. It implies a characteristic with a lower Gini index ought to be liked.

Random Forest

It is a supervised learning algorithm. The benefit of the random forest is, it very well may be utilized for both characterization and relapse issue which structure most of current machine learning framework. Random forest forms numerous decision trees, whats more, adds them together to get an increasingly exact and stable expectation. Random Forest has nearly the equivalent parameters as a decision tree or a stowing classifier model. It is very simple to discover the significance of each element on the expectation when contrasted with others in this calculation.

The regular component in these techniques is, for the kth tree, a random vector theta k is produced, autonomous of the past random vectors theta 1, … , theta k-1 however with the equivalent distribution,while a tree is developed utilizing the preparation set and bringing about a classifier. x is an information vector. For a period, in stowing the random vector is created as the includes in N boxes where N is the number of models in the preparation set of information. In random split, choice includes various autonomous random whole numbers between 1 to K. The dimensionality and nature of theata rely upon its utilization in the development of a tree. After countless trees are created, they select the most famous class. These methodology are called as random forests.[6]

K-Nearest Neighbours

In regression techniques, the output obtained is an average value of its k nearest neighbors. It is a non-parametric method like SVM. Using some values, results are evaluated and the best performance value is obtained.

Multilayer perceptron (MLP)

It is the class of feedforward artificial neural networks. It includes the input layer, output layer and the number of the hidden layers. The hidden layer gives the depth of the neural network. The setup includes 1 hidden layer, the number

TABLE I ALGORITHMEVALUATION

|  |  |  |  |
| --- | --- | --- | --- |
| ML algorithms | R-squared | MAE | MSE |
| Random forest | 0.67 | 0.08 | 0.04 |
| Multilayer Perceptron | 0.65 | 0.09 | 0.04 |
| Gradient Boosting | 0.47 | 0.13 | 0.06 |
| Decision tree | 0.45 | 0.09 | 0.06 |
| K-nearest neighbour | 0.38 | 0.14 | 0.07 |
| SVM | 0.19 | 0.15 | 0.08 |
| AdaBoost | -0.12 | 0.21 | 0.11 |

of neurons starts from 100 to 2000 with different intervals depending upon the required condition. To fire each neuron it requires activation energy. The logistic sigmoid function is used as an activation function.

Gradient boosting

It is an additive regression model by fitting simple function to current pseudo residuals sequentially by least-squares at each iteration. It uses the Decision tree as a basic estimator in sci-kit implementation. Starting from 10 to 1000 with the interval of 10 boosting stages are used with maximum numbers.

The loss function is an important parameter in the gradient boosting. It can be calculated with options: least squares regression, least absolute deviation, and quantile regression.

The most important feature in the flight pricing prediction is the the day is a holiday or not, the day is weekday or weekend and the difference between the days.

Support Vector Machine ( SVM )

In the proposed paper Support Vector Machine used as regression analysis that relays on kernel function considered as non parametric technique. The following kernels are used: Linear, Polynomial, Radial Basis Function[10].

As per the previous studies Random forest and the gradient boosting gives the maximum accuracy[7]. The values of R square, MAE and MSE are given in the table:

PREDICTORS

After evaluating the performance of the all machine learning models , further improvements are made using a correct predictor model for the best result. Two separated train models are developed by applying the trained datasets. Also the appropriate weights are assigned to them to get a better predictor model[14].

Stacked Prediction Model

The performance of machine learning model are ealuated. The Random forest and Multilayer Perceptron, these two models have better results compared to other models. Some weights are applied on the prediction results of these two models to get better prediction results. This is called as stacked prediction model as defined as follows:

Naive Bayes Model: Naive Bayes model simply assigns the equal weights to the results of both the models. There are three techniques used for the prediction. They are as follows:

Bernoulli: It is good for making the prediction from binary features.

Gaussian: It is good for making prediction from normally distributed features.

Multinomial: It is good for when the features (categorical or continuous) describe discrete frequency count (e.g. word counts)

(5)

The result for the naive method is better than random forest model and multilayer perceptron model.

Exhausted search method: This method finds the optimal values for theta by running an exhausting search over a domain of finite interval from -n to n at the difference of a perticular step size.

COCLUSION

In the proposed paper the overall survey for the dynamic price changes in the flight tickets is presented. this gives the information about the highs and lows in the airfares acording to

the days, weekend and time of the day that is morning, evening and night. also the machine learning models in the computational intelligence feild that are evaluated before on different datasets are studied. their accuracy and performances are evaluated and compared in order to get better result. For the prediction of the ticket prices perfectly differnt prediction models are tested for the better prediction accuracy. As the pricing models of the company are developed in order to maximize the revenue management. So to get result with maximum accuracy regression analysis is used. From the studies , the feature that influences the prices of the ticket are to be considered. In future the details about number of availble seats can improve the performance of the model.

.