

FLIGHT PRICE PREDICTION PROJECT

Submitted by:

SANJAY SIVAKUMAR

**ACKNOWLEDGMENT**

1. Manolis Papadakis. Predicting Airfare Prices.

2. Groves and Gini, 2011. A Regression Model For Predicting Optimal Purchase Timing For Airline Tickets.

3. Modeling of United States Airline Fares – Using the Official Airline Guide (OAG) and Airline Origin and Destination Survey (DB1B), Krishna Rama-Murthy, 2006.

4.B. S. Everitt: The Cambridge Dictionary of Statistics, Cambridge University Press, Cambridge (3rd edition, 2006). ISBN 0-521-69027-7.

5. Bishop: Pattern Recognition and Machine Learning, Springer, ISBN 0-387-31073-8.

6. E. Bachis and C. A. Piga. Low-cost airlines and online price dispersion. International Journal of Industrial Organization, In Press, Corrected Proof, 2011.

7. P. P. Belobaba. Airline yield management. an overview of seat inventory control. Transportation Science, 21(2):63, 1987.

**INTRODUCTION**

* Business Problem Framing

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.

* Conceptual Background of the Domain Problem

Anyone who has booked a flight ticket knows how unexpectedly the prices vary. The cheapest available ticket on a given flight gets more and 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)

* Review of Literature

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?

* Motivation for the Problem Undertaken

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.

**Analytical Problem Framing**

* Mathematical/ Analytical Modelling of the Problem

The basic target for this problem is to provide the predicted flight prices, the client wants some predictions that could help them in

Predicting the constantly varying flight prices. Thus we use regression algorithms for this prediction.

* Data Sources and their formats

The dataset consists of 9 columns out of which 8 input and one target variable. This dataset is scrapped from the yatra website while searching for different flight prices.

Table

Description automatically generated

* Data Pre-processing Done

1.Firstly,as our dataset is taken directly from the website it requires much tuning process to be ready for the analysis stage.

2.Then we can check for any null values present in the dataset.

Initailly there is no null values in the dataset.

3.Then Encoding the objective data.

* Data Inputs- Logic- Output Relationships

Currently we have is as much as 9 columns in our dataset out of which the column “price” is our target/output column.Thus we have 8 columns as input to predict our target column.

* Hardware and Software Requirements and Tools Used

For this project I have used python as programming language and Jupyter notebook as its tool.The libraries and packages I have used in this project are,

Text

Description automatically generated

1.pandas- for dataframing the dataset

2.numpy- for any analytical functions

3.seaborn and matplotlib-for visualization process

4.train\_test\_split- for splitting data into training and testing

5.And I have imported regression algorithms such as Random Forest Regressor,KNeighbors Regressor etc… for model building.

6.And finally imported metrics for regression problem for proper model selection.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

Methods used during this project are,

1.Z-Score method- for removing skewness/outliers

2.Box-Cox transformation- for removing skewness/outliers

3.Correlation method- for analyzing the correlation of input columns with our target column

4.VIF method- to check for any multicollinearity present in our input columns

5.Standard Scaler method- to standardize our input column data.

* Testing of Identified Approaches (Algorithms)

1.Linear regression

2.Decision Tree Regressor

3.Random Forest Regressor

4.KNeighbors Regressor

* Run and Evaluate selected models

Text, letter

Description automatically generated

* Key Metrics for success in solving problem under consideration

1.r2 score- to know the accuracy of the model

2.mean squared error-error predicting metrics

* Visualizations

Chart, box and whisker chart

Description automatically generated

-Jet airways has the highest price range

* Interpretation of the Results

From Data preprocessing I have found out there is one date column which needs to be encoded.Then while using histogram and box plot I found that there are more outliers and I tried to reduce it.Then I go for model building where I used various algorithms and found out the best fit algorithm for my model.

**CONCLUSION**

* Key Findings and Conclusions of the Study

Initially I have founded that there is much skewness present in the dataset.So I used various methods to reduce it.Also found that there is no multicolliniearity present in the columns.Then I standardize the dataset for model building and I used as much as 4 algorithms to build my model.Later I found that the algorithm Random Forest Regressor is the best suit for my model by analyzing using the cross-val score and its desired metrics.

* Learning Outcomes of the Study in respect of Data Science

By this project I have learned about various parameters that affect the constantly changing flight prices.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.

* Limitations of this work and Scope for Future Work

After my final model I got a accuracy which is not that great in this scenario and this accuracy can be further increased by adding more columns and using advanced feature Engineering to this dataset.