

A project report submitted in partial fulfilment of the requirement for
the degree of

Bachelor of
Technology in
Computer Science Engineering

On

Decision AI Framework for Airline Price Prediction

Submitted by

Y. Mokshith ramendra(18bcs112)

K Gireesh Kumar(18bcs035)

Godina Pranav(18bcs028)

Mareedu sai rama Krishna (18bcs050)

Under the guidance of

Dr. Uma.S

Professor in Computer Science & Engineering

IIIT Dharwad



INDIAN INSTITUTE OF INFORMATION TECHNOLOGY
DHARWAD

Certificate

This is to certify that the work contained in the project report titled “Decision AI Framework for Airline Price Prediction” by Y. Mokshith ramendra(18bcs112), K Gireesh Kumar(18bcs035), Godina Pranav(18bcs028), Mareedu sai rama Krishna (18bcs050) was completed during the VII semester - IV Year as a Minor Project under the guidance of Dr.Uma.S

Signature of Supervisor

Dr.Uma.S

Professor in Computer Science & Engineering

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Y. Mokshith ramendra(18bcs112)

K Gireesh Kumar(18bcs035)

Godina Pranav(18bcs028)

Mareedu sai rama Krishna (18bcs050)

Approval Sheet

This project report entitled Decision AI Framework for Airline Price Prediction by Y. Mokshith ramendra(18bcs112), K Gireesh Kumar(18bcs035), Godina Pranav(18bcs028), Mareedu sai rama Krishna (18bcs050) of Indian Institute of Information Technology, Dharwad is approved for the degree of Bachelor of Technology in Computer Science and Engineering.

Supervisor

Uma.S

Professor

Computer Science & Engineering,

IIIT Dharwad

Head of Department

Dr. Uma S

Computer Science & Engineering,

IIIT Dharwad

Table of contents

Chapter-1 : Introduction

Chapter-2 : Abstract

Chapter-3 : Literature work

Chapter-4 : Problem statement

Chapter-5 : Working model

5.1: Introduction to model

5.2: Model diagram

5.3: Requirements

5.4: Steps involved in training

Chapter-6 : Results

Chapter-7 : Future scope

Chapter-8 : Conclusion

Chapter-9 : References

Chapter-1: Introduction

In today's fast-growing world transportation is the key for development of a company or nation. In order to meet its(world) requirement people has started preferring fastest way of transport i.e., Airlines. As number of people choosing airlines for transport has increased in order to meet the demand number of routes wee increased so does the prices.

In airlines, ticket prices vary rapidly depending on various factors i.e., sometimes they are high and sometimes low, which makes difficult for passengers to choose when they have to purchase ticket and end paying more than actual cost of that travel.

So, we tried to develop an AI based Airline Price prediction model, so that a passenger can know when to purchase ticket.

Chapter-2: Abstract

There are various projects done on airline price prediction but many failed to use multiple factors that effect the actual price of a ticket. So, in our project we tried to include as many as useful features to include in training of our model and achieved a better percentage of success.

Chapter-3: Literature Review

In the modern era air travel is considered to be fastest mode of transportation. However, the major factor effecting is prices. Especially in case of flight tickets the price won't be constant as in bus fare/train fare it varies abnormally. The main reason for it is that airline uses a technique called dynamic pricing with factors such as demand for a airline ticket over a specific route or type of seat and cat what price customers are willing to pay with a goal of selling maximum number of tickets at the same time making maximum profit possible for its airline company. Where as the customer willing to pay lowest price available for ticket. Because of these factors in same category of tickets one passenger pay less while other for same ticket end up paying higher.

Although previously many projects are done towards development of model for prediction but many skipped few features in training model, we are working on a model random forest regression tree using various useful features to develop an efficient model/application so that we can predict the price of an airline ticket.

Chapter-4: Problem Statement

Transportation plays an important role in today's economy out of which the fastest way of transport is through airlines. One of the major problems when it comes to airlines is price as they are considered to be the highest among all forms of transportation.

There are two kinds of travelling passengers namely leisure travellers, who book their travel well ahead of the travelling date and secondly business travellers who make their booking in a short time where prices vary in both cases.

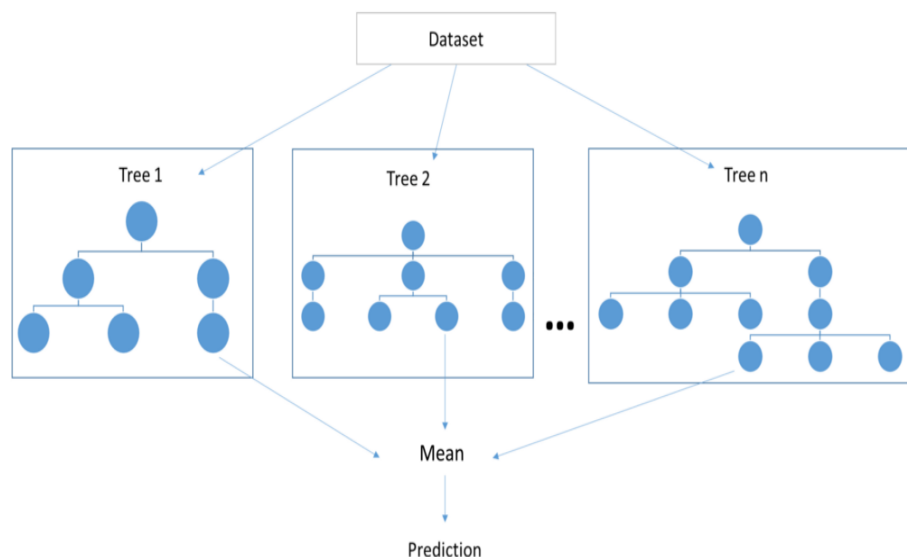
In order to make their travel economically affordable we need to make a model that could predict the price of an airline ticket.

Chapter-5: Working model

5.1: Introduction to model

The model used is “Random Forest regression tree”. It is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning is a technique that combines predictions from various machine learning algorithms to make a more accurate model. In this scenario those machine learning algorithms are decision tree. So, we can say that random forest regression tree forms an output based on various outputs from multiple decision trees.

5.2: Model Diagram



5.3: Requirements

- **Software Used:** Jupyter or google colaboratory
- Data set for training

5.4: Steps involved in training

Step-1: Data Pre-processing

Since the dataset we have consists of factors such as

- Airline name
- Date of journey
- Source and destination
- Duration
- Route
- Departure and arrival times
- Total stops and additional info

All above mentioned are in string format and cannot be understood by our model so using Exploratory Data Analysis (EDA), one-hot encoding and Categorical data we convert them into int format so that our model could understand and train the data set.

Image of the dataset at beginning of pre-processing:

```
data_train.head()
```

	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price
0	IndiGo	24/03/2019	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2h 50m	non-stop	No info	3897
1	Air India	1/05/2019	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7h 25m	2 stops	No info	7662
2	Jet Airways	9/06/2019	Delhi	Cochin	DEL → LKO → BOM → COK	09:25	04:25 10 Jun	19h	2 stops	No info	13882
3	IndiGo	12/05/2019	Kolkata	Banglore	CCU → NAG → BLR	18:05	23:30	5h 25m	1 stop	No info	6218
4	IndiGo	01/03/2019	Banglore	New Delhi	BLR → NAG → DEL	16:50	21:35	4h 45m	1 stop	No info	13302

```
data_train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10683 entries, 0 to 10682
Data columns (total 11 columns):
#   Column              Non-Null Count  Dtype
---  ---
0   Airline              10683 non-null  object
1   Date_of_Journey      10683 non-null  object
2   Source               10683 non-null  object
3   Destination          10683 non-null  object
4   Route                10682 non-null  object
5   Dep_Time             10683 non-null  object
6   Arrival_Time         10683 non-null  object
7   Duration             10683 non-null  object
8   Total_Stops          10682 non-null  object
9   Additional_Info      10683 non-null  object
10  Price                10683 non-null  int64
dtypes: int64(1), object(10)
```

Image of the dataset at after pre-processing:

mini_project_2 Last Checkpoint: 10 hours ago (autosaved)

```
In [28]: data_train.head()
```

station	Route	Total_Stops	Additional_Info	Price	journey_day	journey_month	Dep_hour	Dep_min	Arrival_hour	Arrival_min	Duration_hours	Duration_mins
New Delhi	BLR → DEL	non-stop	No info	3897	24	3	22	20	1	10	2	50
Bangalore	CCU → DXR → BBI → BLR	2 stops	No info	7862	1	5	5	50	13	15	7	25
Cochin	DEL → LKO → BOM → COK	2 stops	No info	13882	9	6	9	25	4	25	19	0
Bangalore	CCU → NAG → BLR	1 stop	No info	6218	12	5	18	5	23	30	5	25
New Delhi	BLR → NAG → DEL	1 stop	No info	13302	1	3	16	50	21	35	4	45

Handling Categorical data

```
In [29]: data_train["Airline"].value_counts()
```

Step-2: Removing extra columns

In the previous step we formed new columns in our required format for model training so we will be removing previous column from dataset through “drop” feature (data_train.drop()).

mini_project_2 Last Checkpoint: 10 hours ago (autosaved)

```
##onehotEncoding on Airline
Airline = data_train[["Airline"]]
Airline = pd.get_dummies(Airline, drop_first=True)
Airline.head()
```

	Airline_Air India	Airline_GoAir	Airline_IndiGo	Airline_Jet Airways	Airline_Jet Airways Business	Airline_Multiple carriers	Airline_Multiple carriers Premium economy	Airline_SpiceJet	Airline_Trujet	Airline_Vistara	Airline_Vistara Premium economy
0	0	0	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	1	0	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0	0	0	0
4	0	0	1	0	0	0	0	0	0	0	0

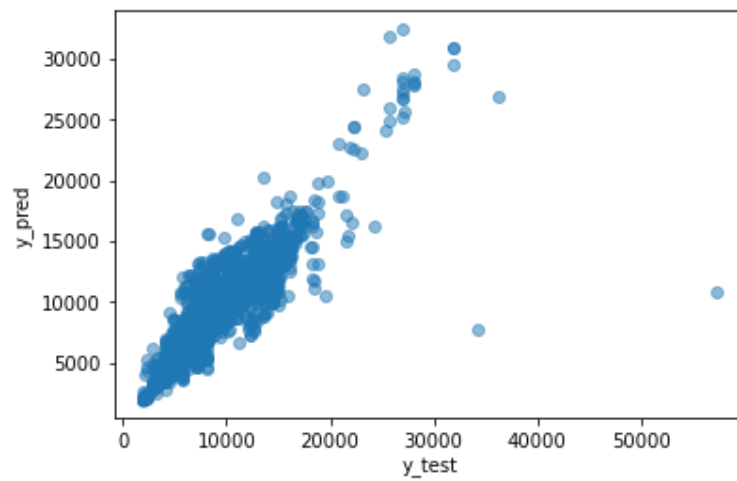
```
##Source
data_train["Source"].value_counts()
```

Step-3: Training the model

With the pre-processed data set we will train the dataset using Random Forest Regression tree model.

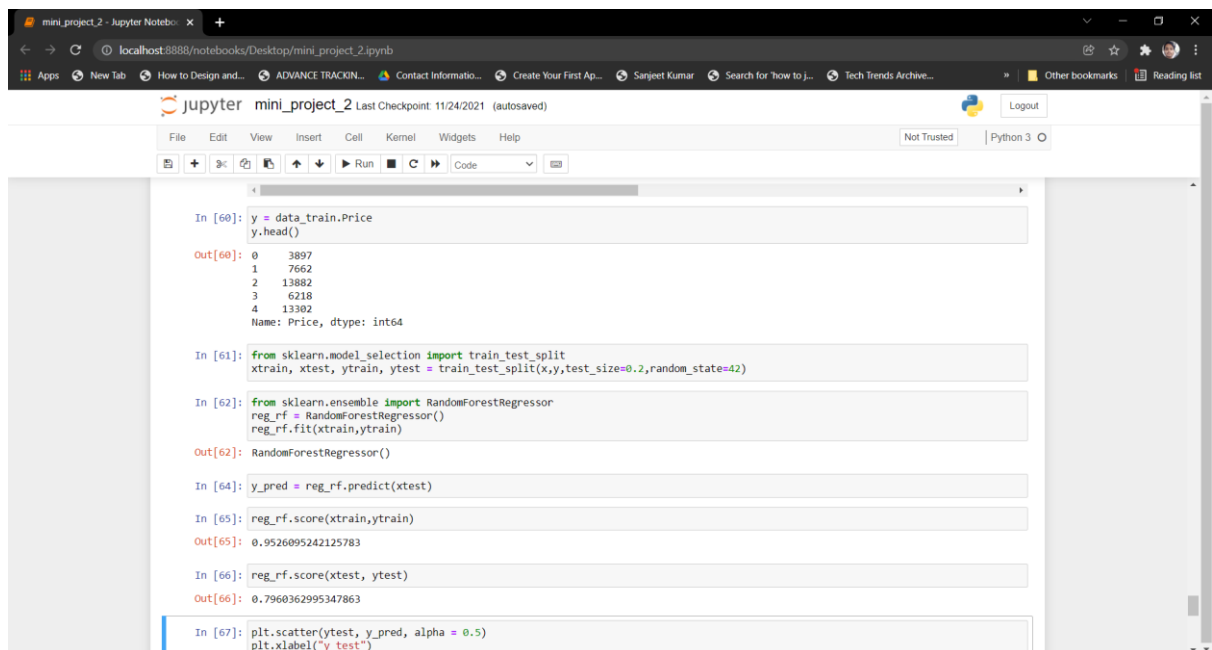
Step-4: After training the model we test our model with our test data set pre-processed in same way as training dataset to check our accuracy.

Step-5: We plot the graph based on test predictions and results as (y_test, y_pred).



Chapter-6: Results

We have trained our dataset using random forest regression tree and tested with test dataset and we have achieved an training accuracy of 96.526% and a test accuracy/prediction accuracy of 79.603



```
In [60]: y = data_train.Price
y.head()

Out[60]: 0    3897
         1    7662
         2   13882
         3   6218
         4   13302
         Name: Price, dtype: int64

In [61]: from sklearn.model_selection import train_test_split
xtrain, xtest, ytrain, ytest = train_test_split(x,y,test_size=0.2,random_state=42)

In [62]: from sklearn.ensemble import RandomForestRegressor
reg_rf = RandomForestRegressor()
reg_rf.fit(xtrain,ytrain)

Out[62]: RandomForestRegressor()

In [64]: y_pred = reg_rf.predict(xtest)

In [65]: reg_rf.score(xtrain,ytrain)

Out[65]: 0.9526095242125783

In [66]: reg_rf.score(xtest, ytest)

Out[66]: 0.7960362995347863

In [67]: plt.scatter(ytest, y_pred, alpha = 0.5)
plt.xlabel("y_test")
```

Chapter-7: Future scope

As a fast growing economy, changes happens very quickly in now-days so does the pricing techniques followed by various airline industries. So we are looking forward for further development of our model towards greater accuracy also for developing a web based application for easier access to all to find a better price for airline ticket.

Chapter-8: Conclusion

The aim of this project is to develop a working model that will be able to predict the price of an airline ticket price and to discuss about various factors that are used in deciding the price of an airline ticket by airline industry and factors that can be used to predict the price.

Chapter-9: References

1. https://www.kaggle.com/vinayshaw/airfare-price-prediction?select=Data_Train.xlsx
2. <https://www.analyticsvidhya.com/blog/2021/06/flight-price-prediction-a-regression-analysis-using-lazy-prediction/>
3. <https://medium.com/analytics-vidhya/regression-flight-price-prediction-6771fc4d1fb3>
4. <https://www.geeksforgeeks.org/random-forest-regression-in-python/>
5. <https://towardsdatascience.com/machine-learning-basics-random-forest-regression-be3e1e3bb91a>
6. <https://simpleflying.com/how-airline-ticket-pricing-works/>