A
Project Report
On
"Airline Fare Prediction"

Prepared by

MITI NAYAK (17DCS034)
DHRUMIL PATEL(17DCS040)
JAHNAVI SHAH (17DCS058)

Under the guidance of

Assistant Prof. Phenil Buch

A Report Submitted to

Charotar University of Science and Technology

for Partial Fulfillment of the Requirements for the

7th Semester Data Science And Analytics Project (CS442)

Submitted at

(CSE)

DEPSTAR

At: Changa, Dist: Anand – 388421

Oct 2020

CHARUSAT
CHAROTAR UNIVERSITY OF SCIENCE AND TECHNOLOGY

CERTIFICATE

This is to certify that the report entitled “**Airline Fare Prediction**” is a bonafied work carried out by **Ms. Miti Nayak (17DCS034)** under the guidance and supervision of **Assistant Prof. Phenil Buch** for the subject CS442 Data Science And Analytics (CSE) of 7th Semester of Bachelor of Technology in **DEPSTAR** at Faculty of Technology & Engineering – CHARUSAT, Gujarat.

To the best of my knowledge and belief, this work embodies the work of candidate himself, has duly been completed, and fulfills the requirement of the ordinance relating to the B.Tech. Degree of the University and is up to the standard in respect of content, presentation and language for being referred to the examiner.

Phenil Buch Assistant Professor Branch Name(CE/CSE/IT) DEPSTAR, Changa, Gujarat.	Prof. Parth Goel HOD Branch Name(CE/CSE/IT) DEPSTAR, Changa, Gujarat

Dr. Amit Ganatra
Principal, DEPSTAR
Dean, FTE
CHARUSAT, Changa, Gujarat.

Devang Patel Institute of Advance Technology And Research At: Changa, Ta. Petlad, Dist. Anand, PIN: 388 421. Gujarat

CERTIFICATE

This is to certify that the report entitled "**Airline Fare Prediction**" is a bonafied work carried out by **Ms. Jahnavi Shah (17DCS058)** under the guidance and supervision of **Assistant Prof. Phenil Buch** for the subject CS442 Data Science And Analytics (CSE) of 7th Semester of Bachelor of Technology in **DEPSTAR** at Faculty of Technology & Engineering – CHARUSAT, Gujarat.

To the best of my knowledge and belief, this work embodies the work of candidate himself, has duly been completed, and fulfills the requirement of the ordinance relating to the B.Tech. Degree of the University and is up to the standard in respect of content, presentation and language for being referred to the examiner.

Phenil Buch
Assistant Professor
Branch Name(CE/CSE/IT)
DEPSTAR, Changa, Gujarat.

Prof. Parth Goel
HOD
Branch Name(CE/CSE/IT)
DEPSTAR, Changa, Gujarat

Dr. Amit Ganatra
Principal, DEPSTAR
Dean, FTE
CHARUSAT, Changa, Gujarat.

Devang Patel Institute of Advance Technology And Research At: Changa, Ta. Petlad, Dist. Anand, PIN: 388 421. Gujarat

CERTIFICATE

This is to certify that the report entitled "**Airline Fare Prediction**" is a bonafied work carried out by **Mr. Dhrumil Patel (17DCS040)** under the guidance and supervision of **Assistant Prof. Phenil Buch** for the subject CS442 Data Science And Analytics (CSE) of 7th Semester of Bachelor of Technology in **DEPSTAR** at Faculty of Technology & Engineering – CHARUSAT, Gujarat.

To the best of my knowledge and belief, this work embodies the work of candidate himself, has duly been completed, and fulfills the requirement of the ordinance relating to the B.Tech. Degree of the University and is up to the standard in respect of content, presentation and language for being referred to the examiner.

Phenil Buch
Assistant Professor
Branch Name(CE/CSE/IT)
DEPSTAR, Changa, Gujarat.

Prof. Parth Goel
HOD
Branch Name(CE/CSE/IT)
DEPSTAR, Changa, Gujarat

Dr. Amit Ganatra
Principal, DEPSTAR
Dean, FTE
CHARUSAT, Changa, Gujarat.

Devang Patel Institute of Advance Technology And Research At: Changa, Ta. Petlad, Dist. Anand, PIN: 388 421. Gujarat

ACKNOWLEDGEMENT

We have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. I would like to extend my sincere thanks to all of them.

I am highly indebted to Asst Prof. Phenil Buch for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project.

I would like to express my gratitude towards member of DEPSTAR for their kind co-operation and encouragement which help me in completion of this project.

ABSTRACT

The airline implements dynamic pricing for the flight ticket. According to the survey, flight ticket prices change during the morning and evening time of the day. Also, it changes with the holidays or festival season. There are several different factors on which the price of the flight ticket depends. The seller has information about all the factors, but buyers are able to access limited information only which is not enough to predict the airfare prices.

Considering the features such as departure time, the number of days left for departure and time of the day it will give the best time to buy the ticket. The purpose of the paper is to study the factors which influence the fluctuations in the airfare prices and how they are related to the change in the prices. Then using this information, build a system that can help buyers whether to buy a ticket or not.

LIST OF FIGURES

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1PROJECT DEFINITON

Flight ticket prices can be something hard to guess, today we might see a price, check out the price of the same flight tomorrow, it will be a different story. We might have often heard travelers saying that flight ticket prices are so unpredictable. Almost all airline companies base their ticket price on demand estimation models and implement various dynamic pricing strategies in order to regulate seats demand and maximize their revenue. These corporations are said to use some proprietary software to evaluate ticket price per seat on a given day for a particular flight but the algorithms used are guarded with commercial secrets. These companies usually tie up with various online ticket sale channels (yatra.com, makemytrip.com, paytm.com) which maintains real time data on ticket price and constantly updates this price per seat over time. These channels are usually available over the internet where the traveler can book the ticket conveniently paying some convenience charges. This constant updating of prices results in high fluctuation which often confuses consumers as to when book their flight tickets to get best of the deals. This project deal with prediction of the best prices for the customers as they are the most affected due to the fluctuation in ticket price. So in this project we are using various machine learning model analyzing them and finding the most suitable one. Then the prediction for the given dataset is carried out.

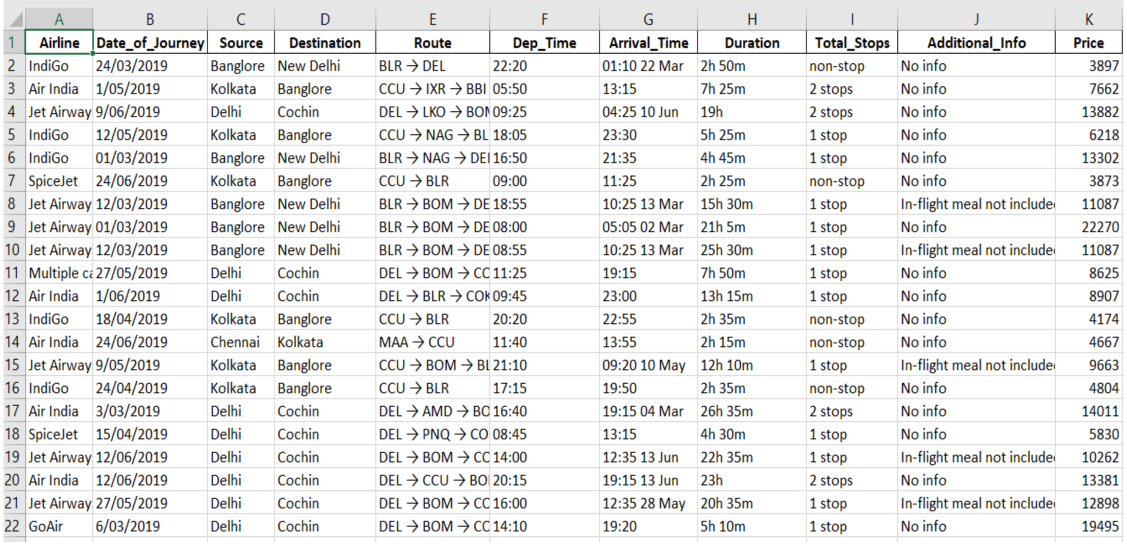
CHAPTER 2

DESCRIPTION

2.1 Data Collection

Data Collection is one of the most important aspect of this project. There are various sources of airfare data on the Web, which we could use to train our models. A multitude of consumer travel sites supply fare information for multiple routes, times, and airlines. We are using the Excel dataset which have all the attributes required for the correct prediction of the value.

Now an important aspect is to decide the parameters that might be needed for the flight prediction algorithm.



	A	B	C	D	E	F	G	H	I	J	K
1	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price
2	IndiGo	24/03/2019	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2h 50m	non-stop	No info	3897
3	Air India	1/05/2019	Kolkata	Banglore	CCU → IXR → BBI	05:50	13:15	7h 25m	2 stops	No info	7662
4	Jet Airway	9/06/2019	Delhi	Cochin	DEL → LKO → BOM	09:25	04:25 10 Jun	19h	2 stops	No info	13882
5	IndiGo	12/05/2019	Kolkata	Banglore	CCU → NAG → BL	18:05	23:30	5h 25m	1 stop	No info	6218
6	IndiGo	01/03/2019	Banglore	New Delhi	BLR → NAG → DEL	16:50	21:35	4h 45m	1 stop	No info	13302
7	SpiceJet	24/06/2019	Kolkata	Banglore	CCU → BLR	09:00	11:25	2h 25m	non-stop	No info	3873
8	Jet Airway	12/03/2019	Banglore	New Delhi	BLR → BOM → DE	18:55	10:25 13 Mar	15h 30m	1 stop	In-flight meal not include	11087
9	Jet Airway	01/03/2019	Banglore	New Delhi	BLR → BOM → DE	08:00	05:05 02 Mar	21h 5m	1 stop	No info	22270
10	Jet Airway	12/03/2019	Banglore	New Delhi	BLR → BOM → DE	08:55	10:25 13 Mar	25h 30m	1 stop	In-flight meal not include	11087
11	Multiple ca	27/05/2019	Delhi	Cochin	DEL → BOM → CC	11:25	19:15	7h 50m	1 stop	No info	8625
12	Air India	1/06/2019	Delhi	Cochin	DEL → BLR → COK	09:45	23:00	13h 15m	1 stop	No info	8907
13	IndiGo	18/04/2019	Kolkata	Banglore	CCU → BLR	20:20	22:55	2h 35m	non-stop	No info	4174
14	Air India	24/06/2019	Chennai	Kolkata	MAA → CCU	11:40	13:55	2h 15m	non-stop	No info	4667
15	Jet Airway	9/05/2019	Kolkata	Banglore	CCU → BOM → BL	21:10	09:20 10 May	12h 10m	1 stop	In-flight meal not include	9663
16	IndiGo	24/04/2019	Kolkata	Banglore	CCU → BLR	17:15	19:50	2h 35m	non-stop	No info	4804
17	Air India	3/03/2019	Delhi	Cochin	DEL → AMD → BO	16:40	19:15 04 Mar	26h 35m	2 stops	No info	14011
18	SpiceJet	15/04/2019	Delhi	Cochin	DEL → PNQ → CO	08:45	13:15	4h 30m	1 stop	No info	5830
19	Jet Airway	12/06/2019	Delhi	Cochin	DEL → BOM → CC	14:00	12:35 13 Jun	22h 35m	1 stop	In-flight meal not include	10262
20	Air India	12/06/2019	Delhi	Cochin	DEL → CCU → BOI	20:15	19:15 13 Jun	23h	2 stops	No info	13381
21	Jet Airway	27/05/2019	Delhi	Cochin	DEL → BOM → CC	16:00	12:35 28 May	20h 35m	1 stop	In-flight meal not include	12898
22	GoAir	6/03/2019	Delhi	Cochin	DEL → BOM → CC	14:10	19:20	5h 10m	1 stop	No info	19495

2.2 Data Preparation

All the collected data needed a lot of work so after the collection of data, it is needed to be clean and prepare according to the model requirements. All the unnecessary data is removed like duplicates and null values. In all machine learning this technology, this is the most important and time consuming step. Various statistical techniques and logic built in python are used to clean and prepare the data.

- **Append**

  Appending of the data set is done to work together with both train and test at a same time and don't have to make changes separately. After we apply the transformation then we can separate them again into test and train

- **Feature Engineering**

  Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data. Feature engineering turn inputs into things the algorithm can understand.

- **Converting Categorical into Integer values**

  Many machine learning tools will only accept numbers as input. This may be a problem if you want to use such tool but your data includes categorical features. To convert categorical text data into model-understandable numerical data, we use the Label Encoder class. So all we have to do, to label encode a column is import the Label Encoder class from the sklearn library, fit and transform the column of the data, and then replace the existing text data with the new encoded data.

- **Missing Value Validation**

  Missing data treatment is very important to avoid biased results. Generally, missing data in training data set can reduce the power of the model which can lead to wrong classification/ prediction.

- **Split into Test Set and Train Set**

  The data we use is usually split into training data and test data. The training set contains a known output and the model learns on this data in order to be generalized to other data later on. We have the test dataset (or subset) in order to test our model's prediction on this subset.

2.3 Our Model

To develop the model for the flight price prediction, many conventional machine learning algorithms are evaluated.

Algorithm	RMS
Linear Regression	3238.316
Ridge Regression	3238.153
Lasso Regression	3238.3169
Light GBM	1395.095

- **Comparative Analysis**

  From the above different Regression Technique we can see Light GBM is performing really good in regards to all .Finally we will use this to predict our test data.

CHAPTER 3

SYSTEM REQUIRNMENT STUDY

3.1 User Characteristics

User should know how to run a program .The user should know how to respond to the code.

3.2 Hardware Requirements

- Processor: Intel dual core or above
- Processor Speed:1.0GHZ or above
    - RAM: 2 GB RAM or above
    - Hard Disk: 10 GB hard disk or above

3.3 Software Requirements

The computer should have Windows operating system. Software used is Google Collaboratory.

CHAPTER 4

SYSTEM ANALYSIS

4.1 Functional and Non-Functional Requirements

- **Functional Requirements**–
- Customer should be able to search flights for a specific date for one-way trips.
- Customer should be able to search flights for specific dates for round trips.
- Customer should be able to search flights for multiple destinations.
- Customer should be able to manually enter the names of departure and arrival cities.
- Customer should be able to sort the list of possible flights by price.
- Customer should be able to sort the list of possible flights by flight duration.
- System should allow a customer to specify only departure date for one-way trips.
- System should allow a customer to specify both departure and arrival dates for round trips.
- System should provide the list of possible flights matching criterion of user inputs.

Other Non-functional Requirements:

Performance Requirements

- The system shall accommodate high number of items and users without any fault.
- Responses to view information shall take no longer than 5 seconds to appear on the screen.

Safety Requirements

- System use shall not cause any harm to human users.

Security Requirements

- System will use secured database.
- Normal users can just read information but they cannot edit or modify anything expect their personal and some other information.
- System will have different types of users and every user has access constraints.

Error handling

- System shall handle expected and non expected errors in ways that prevent loss in information and long downtime period.

CHAPTER 5

DIAGRAMS

Uses Case Diagram

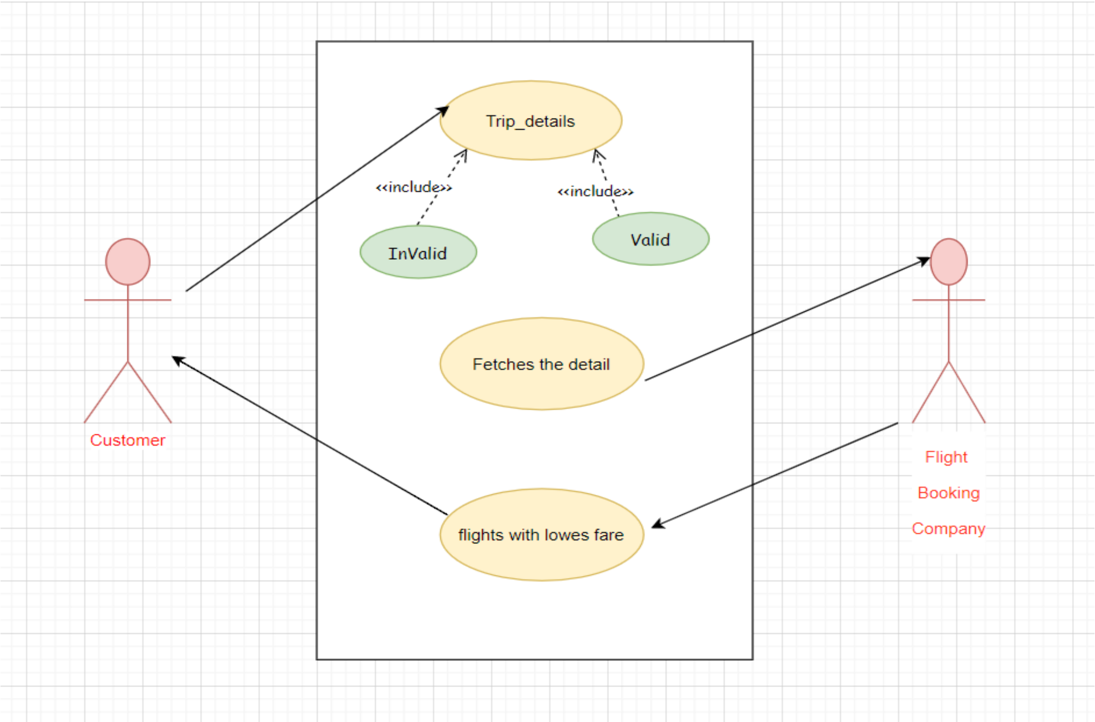


Fig 5.1 Use case diagram

DFD (Data Flow Data)

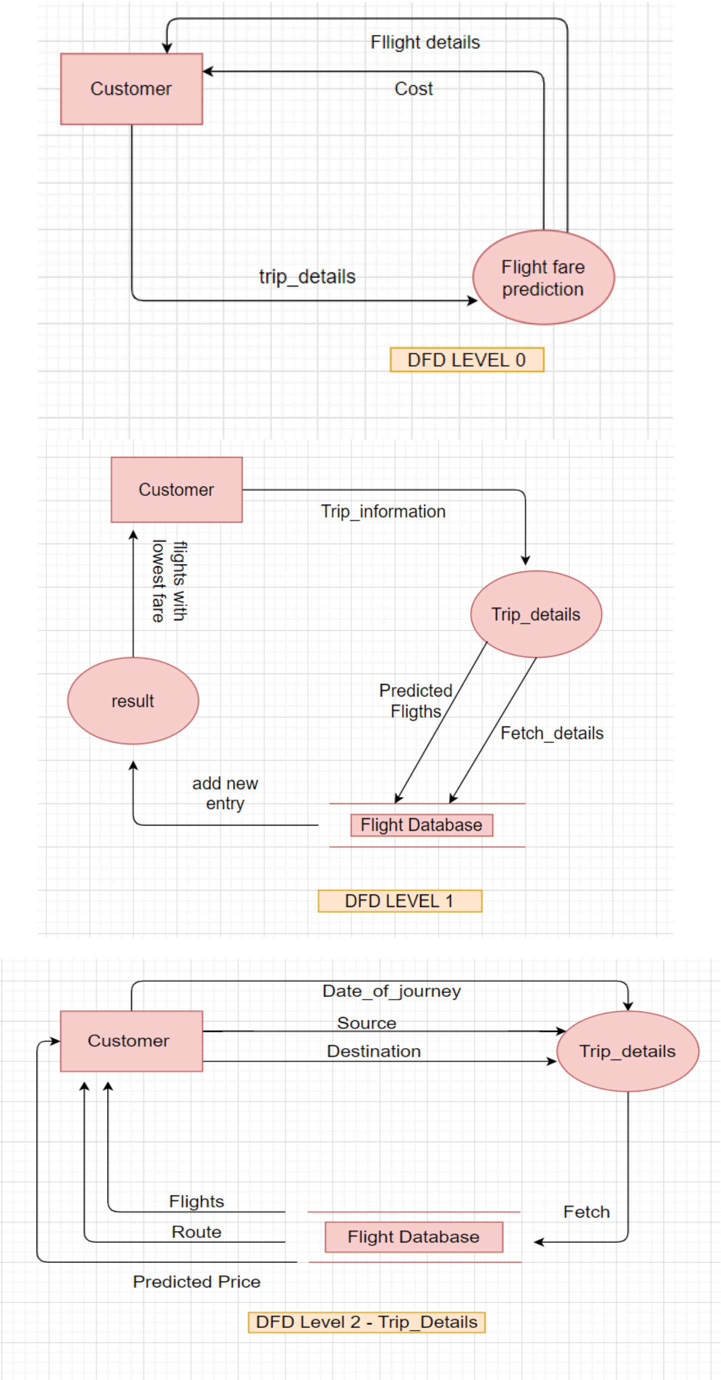


Fig 5.2 Data Flow diagram

Sequence diagram

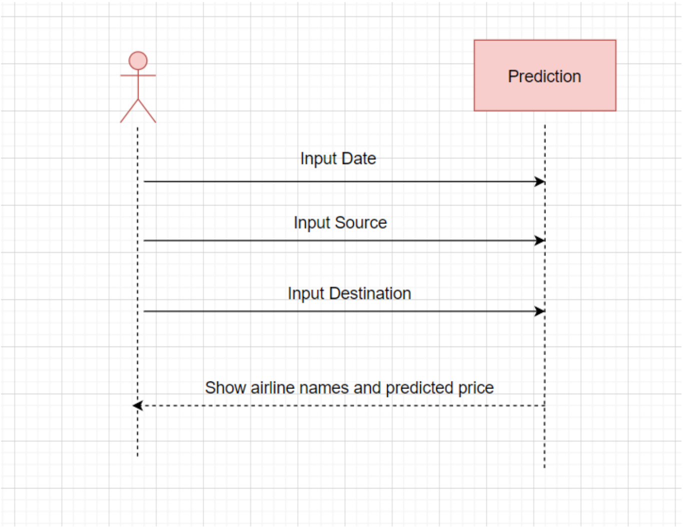
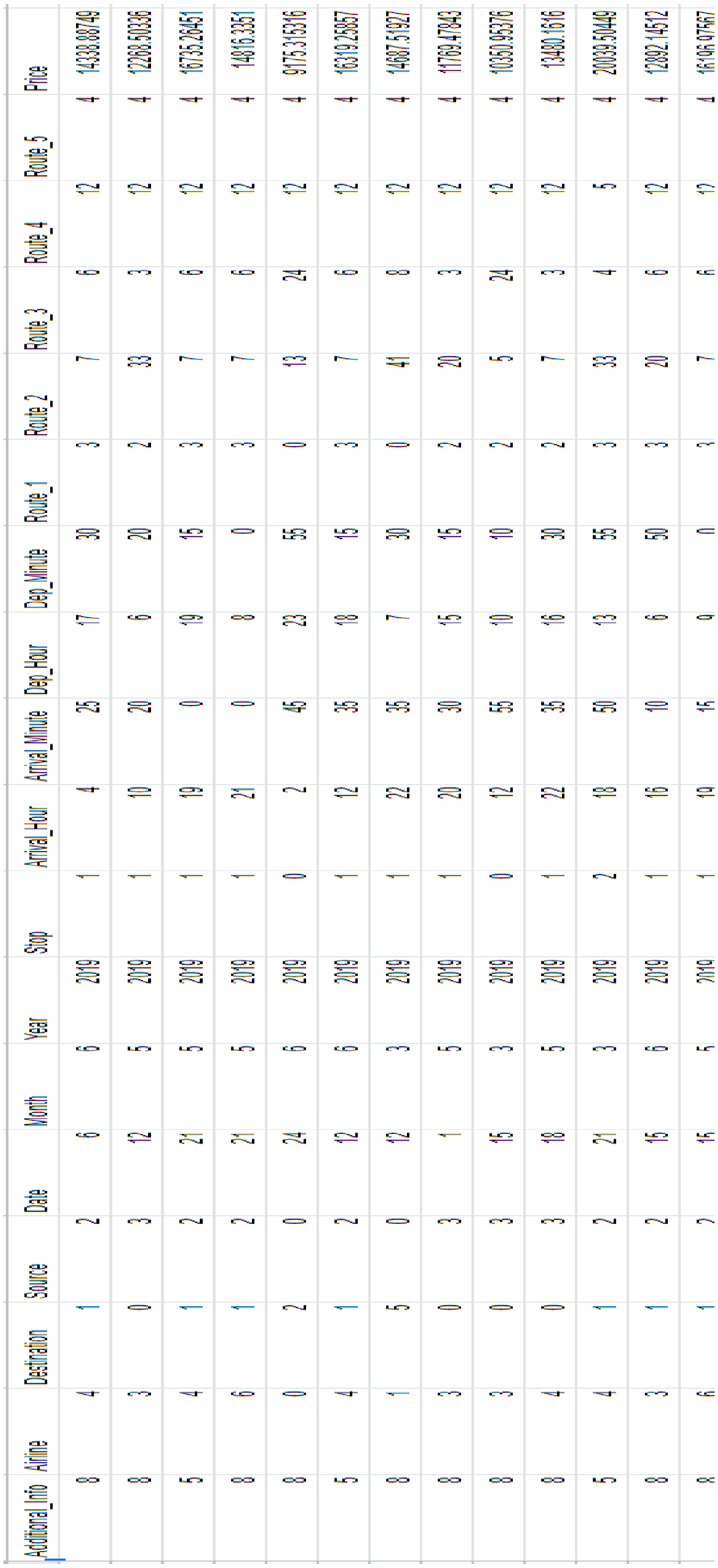


Fig 5.3 Sequence diagram

CHAPTER 6

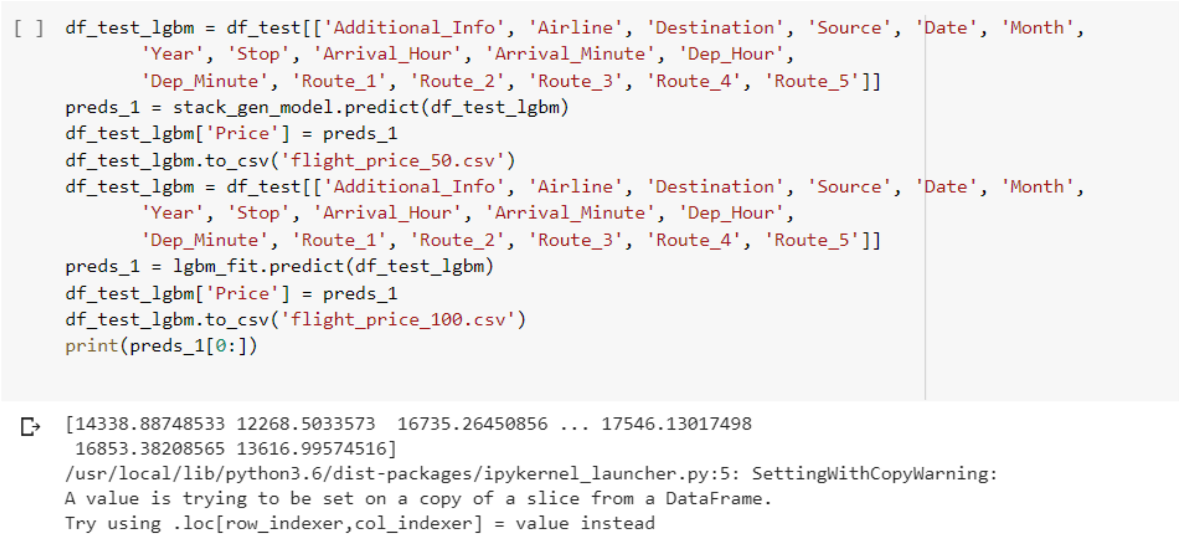
SCREENSHOTS

6.1 Screen shots for the project:



Additional_Info	Airline	Destination	Source	Date	Month	Year	Stop	Arrival_Hour	Arrival_Minute	Dep_Hour	Dep_Minute	Route_1	Route_2	Route_3	Route_4	Route_5	Price
8	4	1	2	6	6	2019	1	4	25	17	30	3	7	6	12	4	14338.88749
8	3	0	3	12	5	2019	1	10	20	6	20	2	33	3	12	4	12268.50336
5	4	1	2	21	5	2019	1	19	0	19	15	3	7	6	12	4	16735.26451
8	6	1	2	21	5	2019	1	21	0	8	0	3	7	6	12	4	14816.3351
8	0	2	0	24	6	2019	0	2	45	23	55	0	13	24	12	4	9175.315316
5	4	1	2	12	6	2019	1	12	35	18	15	3	7	6	12	4	16319.25857
8	1	5	0	12	3	2019	1	22	35	7	30	0	41	8	12	4	14687.51927
8	3	0	3	1	5	2019	1	20	30	15	15	2	20	3	12	4	11769.47843
8	3	0	3	15	3	2019	0	12	55	10	10	2	5	24	12	4	10350.95376
8	4	0	3	18	5	2019	1	22	35	16	30	2	7	3	12	4	13480.1616
5	4	1	2	21	3	2019	2	18	50	13	55	3	33	4	5	4	20039.50449
8	3	1	2	15	6	2019	1	16	10	6	50	3	20	6	12	4	12892.14512
8	6	1	2	15	5	2019	1	19	15	9	0	3	7	6	12	4	16196.97567

Fig 6.1 Predicted price



```
[ ] df_test_lgbm = df_test[['Additional_Info', 'Airline', 'Destination', 'Source', 'Date', 'Month',
           'Year', 'Stop', 'Arrival_Hour', 'Arrival_Minute', 'Dep_Hour',
           'Dep_Minute', 'Route_1', 'Route_2', 'Route_3', 'Route_4', 'Route_5']]
    preds_1 = stack_gen_model.predict(df_test_lgbm)
    df_test_lgbm['Price'] = preds_1
    df_test_lgbm.to_csv('flight_price_50.csv')
    df_test_lgbm = df_test[['Additional_Info', 'Airline', 'Destination', 'Source', 'Date', 'Month',
           'Year', 'Stop', 'Arrival_Hour', 'Arrival_Minute', 'Dep_Hour',
           'Dep_Minute', 'Route_1', 'Route_2', 'Route_3', 'Route_4', 'Route_5']]
    preds_1 = lgbm_fit.predict(df_test_lgbm)
    df_test_lgbm['Price'] = preds_1
    df_test_lgbm.to_csv('flight_price_100.csv')
    print(preds_1[0:])
```

```
[14338.88748533 12268.5033573  16735.26450856 ... 17546.13017498
 16853.38208565 13616.99574516]
/usr/local/lib/python3.6/dist-packages/ipykernel_launcher.py:5: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

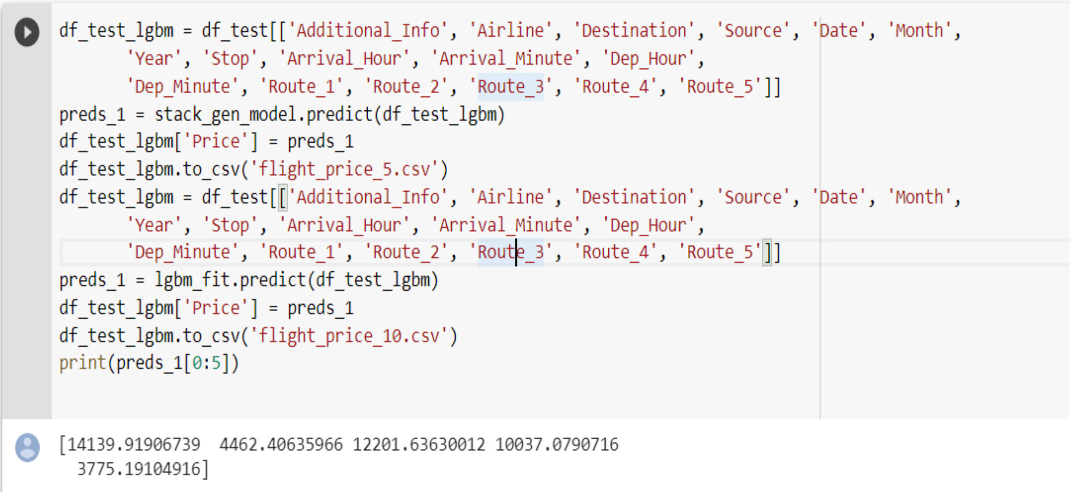
Fig 6.2 final project outcome

CHAPTER 7

PRACTICAL OUTCOME

7.1 PRACTICAL OUTCOME

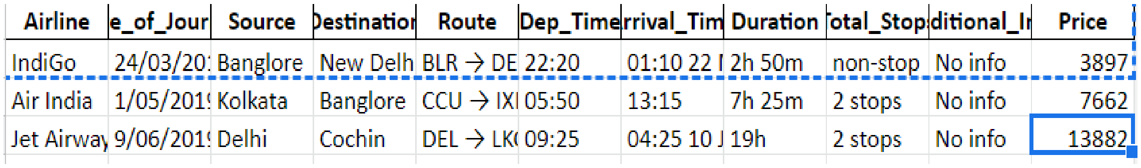
- Since the rsme of **LIGHT GBM (1395.153)** machine learning algorithm is lowest of the the different models implemented i.e. Linear regression (3238.316),Ridge Regression (3238.153),Lasso Regression(3238.169).
- Thus the airline fare is predicted using LIGHT GBM model.



```
df_test_lgbm = df_test[['Additional_Info', 'Airline', 'Destination', 'Source', 'Date', 'Month',
        'Year', 'Stop', 'Arrival_Hour', 'Arrival_Minute', 'Dep_Hour',
        'Dep_Minute', 'Route_1', 'Route_2', 'Route_3', 'Route_4', 'Route_5']]
preds_1 = stack_gen_model.predict(df_test_lgbm)
df_test_lgbm['Price'] = preds_1
df_test_lgbm.to_csv('flight_price_5.csv')
df_test_lgbm = df_test[['Additional_Info', 'Airline', 'Destination', 'Source', 'Date', 'Month',
        'Year', 'Stop', 'Arrival_Hour', 'Arrival_Minute', 'Dep_Hour',
        'Dep_Minute', 'Route_1', 'Route_2', 'Route_3', 'Route_4', 'Route_5']]
preds_1 = lgbm_fit.predict(df_test_lgbm)
df_test_lgbm['Price'] = preds_1
df_test_lgbm.to_csv('flight_price_10.csv')
print(preds_1[0:5])
```

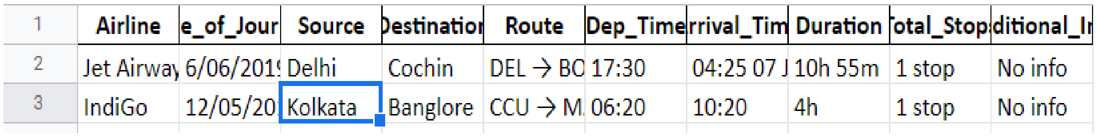
```
[14139.91906739  4462.40635966 12201.63630012 10037.0790716
  3775.19104916]
```

Training set



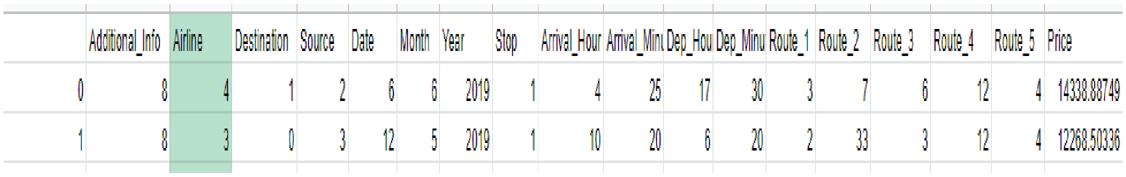
Airline	e_of_Jour	Source	Destinatio	Route	Dep_Time	rrival_Tim	Duration	otal_Stop	ditional_Ir	Price
IndiGo	24/03/20:	Banglore	New Delh	BLR → DE	22:20	01:10 22 I	2h 50m	non-stop	No info	3897
Air India	1/05/201!	Kolkata	Banglore	CCU → IXI	05:50	13:15	7h 25m	2 stops	No info	7662
Jet Airway	9/06/201!	Delhi	Cochin	DEL → LK(	09:25	04:25 10 J	19h	2 stops	No info	13882

Test set



1	Airline	e_of_Jour	Source	Destinatio	Route	Dep_Time	rrival_Tim	Duration	otal_Stop	ditional_Ir
2	Jet Airway	6/06/201!	Delhi	Cochin	DEL → BO	17:30	04:25 07 J	10h 55m	1 stop	No info
3	IndiGo	12/05/20:	Kolkata	Banglore	CCU → M.	06:20	10:20	4h	1 stop	No info

Prediction



	Additional_Info	Airline	Destination	Source	Date	Month	Year	Stop	Arrival_Hour	Arrival_Min	Dep_Hou	Dep_Minu	Route_1	Route_2	Route_3	Route_4	Route_5	Price
0	8	4	1	2	6	6	2019	1	4	25	17	30	3	7	6	12	4	14338.88749
1	8	3	0	3	12	5	2019	1	10	20	6	20	2	33	3	12	4	12268.50336

CHAPTER 8

LIMITATIONS AND FUTURE ENHANCEMENT

8.1 Limitations

- This project predicts airlines fare of the dates given in the dataset only.
- Limited availability of the data.

8.2 Future Enhancement

- Live data can be used by using web scrapping to improve usability of the project.
- UI/UX design can be created for better user experience.

CHAPTER 9

REFERENCES

9. 1 References:

- https://www.semanticscholar.org/paper/Airfare-prices-prediction-using-machine-learning-Tziridis-Kalampokas/124250a5ff813e30d9305c26db8896c2278dca8d
- https://youtu.be/jxKg65AimSI
- https://youtu.be/72hlr-E7KA0
- https://analyticsindiamag.com/flight-ticket-price-prediction-hackathon-use-these-resources-to-crack-our-machinehack-data-science-challenge/
- https://ieeexplore.ieee.org/document/8081365