FLIGHT PRICE PREDICTION

SUBMITTED BY JUHI MISHRA

ACKNOWLEDGEMENT

To complete this project I had taken help from few of the sources listed below:
1. Viman, Tripdeal, EaseMyTrip, HappyFare these were the few websites from where scrapped data of flight price for model building.
2. StackOverflow, Kaggle were the few websites from where taken help for coding and visualisation part.
3. PDF document provided by Flip Robi helped a lot in understanding the problem statement and guided how to start and what are the steps to follow to complete this project.

INTRODUCTION

· Business Problem Framing :

Here I do have to predict the flight price which may vary depending upon various factors. As we all know that flight price fluctuates a lot. Price depends upon how early the ticket been taken and for which time period. Like if any special occasion is there flight price even after 2 months will be higher compare to other days. And if tickets been taken 2 or 3 days back or on the same day price will be too high. Most of the time we see that same day ticket price will be too high, similarly 10 days prior or before that compare to same day price will be low. And if someone taken almost a month back , flight ticket will be at cheapest cost. But as per the business point of view if ticket been taken on an urgent or few days back , then the profit will be good , as for same flight where one passenger taken at say for an example rs.2000, and on an urgency one taken at 10000 so profit on 10000 will be more compare to 2000. Considering this situation most of the aviation house they hold some seats for an emergency purpose so that last minute rate will be high and they will earn good profit . Few special occasion state wise also price of flight tickets generally higher . Even weekend tickets too are comparatively higher . So depending upon consideration where they can earn good profit and will get customer easily target most and keep the price high intentionally.

· Conceptual Bankground of the

Domain Problem :

In order to get the prediction on flight I.e., when the price will be high and when low depending upon situation an aviation house needs a data scientist who can help them using there domain knowledge in prediction of flight tickets when is the possibility of the price to be high and low. As any of the aviation house will have profit when the tickets been taken at last minute as the cost will be almost 3 times higher compare to tickets been taken almost a month a back. For this reason they hold few of the seats for last minute sale. Depending upon prior's dataset will analyse and can able to predict what will be the future price so that they can plan accordingly there business strategy and can earn good revenue and profit.

· Review of Literature :

Here in this dataset where I as a data scientist need to predict the flight price depending upon which a business house can make their proper strategy and planning to have good revenue and profit. Flight price depends upon various factors. Like compare to weekdays weekends price will be high. Any occasion tickets will be high depending upon state. Last minutes ticket will always be too high comparatively. Though now a days to attract more customers and grab more and more even middle class family too flight price declined a lot. So that they can have more number of customers. If we go 7 or 8 years back where any of the flight whose cost was starting from 5k now they reduced it to almost a half which is even 2k or less than that depending upon distance between city. But there are few aviation house who still maintaining higher price like vistara which we seen everywhere there cost were too high comparatively. Here we can consider that this type of aviation doesn't focus on number of customers I.e., quantity they mainly focus on business class people who invest more for their comforts. But there are few aviation house like Go Air, Indigo who mainly focus on economical class which anyone can afford now a days who travel via train in 3rd AC or 2nd AC. As there is not much difference in price between two. Even flight takes less time compare to train so these type of customers who are looking for cheaper flight they do travel a lot. And this not only for middle class person but even for those who are looking for pocket friendly flight they do prefer these type of flights only. Here by lowering

the price aviation house do focus on more number of customers. Now the problem comes where if all the customers will have ticket at cheapest cost then these aviation house will not able to earn good profit. For this reason they use to hold few seats for last minute and higher there cost so that they can maintain both which means more number of customers and revenue too. Even now a days we seen that for selecting seats too flight charges. These are the few new things where flight do try to increase there revenue. Depending upon customer targeted by particular aviation house there revenue and profit varies. Like vistara where they charge from Kolkata to Hyderabad approx 10k indigo do have 3k so the difference between two are really huge. Clearly we can say that vistara main target is only business class customer, while indigo target all type of customer. Price also depends on timing and no of stoppage do they have. Like if there's a early morning flight they will charge less even late night flight charges are less. Similarly we seen that no of stop of those flights are 1 or more than 1 there charges are less because of more time taken by them.

- Motivation for the Problem Undertaken :

This dataset is all about flight price prediction where as a data scientist I do have job to predict the price using different algorithm. When I started doing this project very first that motivated me is there are lot many aviation house. But when we go for ticket booking excluding few most of the flights do have same price trend . But those few why do they charge more and what type of customer do they target these were the few questions which raised in my mind. As we know that flight saves travelling time a lot. And most of the flight do take almost same time for same distance if they don't have stoppage in between. Then why there's difference in price. As few of the aviation house only targets business class people so even for same distance they charge more. And there are few flights who targets almost all sort of customers , so according to their guidelines they set their price. In what situation do price change and when the price is high in which time and how much is the variance, these were the few questions which arised in my mind, and motivated me to take this project . By doing this project I came across when the price become high and when cheap. As a data scientist we help them in making decision for What strategy do these type of business wants to have to improve there business and earn more revenue.

ANALYTICAL PROBLEM FRAMING

· Mathematical / Analytical Modeling of the Problem :

As there were no continuous column on which basically we perform mathematical functions like using describe method we generally analyse statistical prospect. Also removing skewness and outliers generally we perform for those columns which are having mathematical calculation means only for continuous columns which are not discrete . Here basically we were having almost all the columns in object form which means string datatype so no need to remove skewness or outliers for those. Rest for data column it was in date datatype and date columns are generally considered as discrete columns which means there are fixed aspect no fluctuation between two data.

· Data Sources and their Formats :

When I get this project I was not having data to build model. So first step which I was suppose to do was to scrap data from relevant websites. Scrapped required columns from few of the websites which were Viman, EaseMyTrip,TripDeal and HappyFare. After scrapping all the columns using selenium save them to excel format. Then compiled all the data into one excel sheet and loaded the same using pandas in python for model building and flight price prediction which was my main motive. Below is the screenshot for data loaded in python:

```
df = pd.read_excel('/Users/juhimishra/Downloads/flight scrapping - excel sheet.xlsx')
df.head()
```

	Filght	Date	Arrival	Departure	Duration	Total Stop	Source	Destination	Price	Site
0	Indigo Air	2021-12-08	07:00	11:20	04h 20m	1 Stop	Kolkata	Hyderabad	5338	viman
1	Indigo Air	2021-12-08	07:00	17:00	10h 00m	1 Stop	Kolkata	Hyderabad	5467	viman
2	Indigo Air	2021-12-08	14:50	22:45	07h 55m	1 Stop	Kolkata	Hyderabad	5646	viman
3	Indigo Air	2021-12-08	11:05	17:20	06h 15m	1 Stop	Kolkata	Hyderabad	5817	viman
4	Air Asia	2021-12-08	04:40	06:45	02h 05m	Non Stop	Kolkata	Hyderabad	6096	viman

· Data Preprocessing Done :

To have good model very first thing as a data scientist we need to analyse our model properly. Means what are the steps we need to follow. One very important step is to have proper data cleaning. If data cleaning not done in proper way then it will affect our model performance. Here in this dataset there were no null values which I checked using info and isna method. So no need to do anything for missing values. Next using info method checked what datatype do this dataset have , accordingly will perform. Here we were having only one numerical column which was target column means price, And there were one date column which was in date time format. Rest all were in object datatype. As we know that our machine algorithm doesn't understand string datatype so to convert the same to numerical datatype we need to apply encoding technique. Here I used one hot encoder as my columns were less so chances were less to have huge number of columns which is possible if I do have lot many variables in one particular column as one hot encoder generally increases the number of columns. Here Data, Arrival, Departure,Duration, these columns using date time method converted to numerical columns. So only for source, destination and flight I used one hot encoder . Due to which my number of columns from 9 increased to 28. For total stop used replace method as the data inside that were less. After data cleaning and preprocessing done checked relation among features using correlation method. As if there is high correlation among feature then are chances of overfitting as both the features may share same relation which our model can learn twice and predict accordingly. So to avoid such situation generally we use few techniques and remove such columns which will show high correlation among them just to avoid overfitting. Here there were no such columns. So not removed any of the column except site column which was not

required for our prediction purpose. Removed date, Arrival, Departure columns as already extracted data from there.

· Data Inputs - Logic - Output Relationship :

Depending upon label generally our target depends. Means feature column always considered as independent variable and target column is dependent variable. Here our target column is price which depends upon different features like flight, date, arrival, departure, duration, total stop, source and destination. As we observed that Different flight do have different price. Few of the flight do have same range but there are few which are having higher price like vistara which is having almost 3 times higher price compare to indigo or go air. We can consider that vistara generally targets high class customers who doesn't go for economical instead they prefer comfort more. While indigo, Air Asia or Go Air they target all type of customers. Price also varies on when the ticket been taken means last moment ticket price will always be high. Arrival and departure matters when customer o prefer any time. As few customers may opt for noon or late morning flight which generally cost high . Coming to duration generally customers prefer flight because of less travelling time and if the duration will be more then price will be low for such flights. No coming to source and destination , distance is also one of the important factor on which price depends a lot. For less distance price will be lower while for longer distance price will be high. These are the few factors on which flight price depends a lot.

· State the Set of Assumptions (if any) Related to the Problem Under Consideration :

As this dataset was related to flight price prediction so first I thought was what type of dataset is his means regression or classification based problem. But when I gone through the PDF provided by Flip Robo came across that we need to predict the price of flight which is always considered as continuous column and continuous columns are considered to be regression problem. Based on which I decided which model to run for this problem statement. Few protocols which we

generally follow is removing null values, skewness , outliers etc, here no such operations were required as there were no null values so no treatment required for the same. Moreover skewness and outliers generally we perform for continuous columns means its a kind of mathematical operation but here none of the column were continuous column. It's not mandatory that all the columns which are in numerical form they are continuous, Generally mathematical operation we perform for those where predicting between two continuous value is difficult means the data is continue in nature . While there are few numerical columns which are considered to be discrete column where we can assume what next can come as non other possibility between will take place. Also few model which I assumed will perform best for this dataset , but they were not turned so good as per expectation, so tried few other algorithms so that will get best model and prediction will be more accurate.

· Hardware and Software Requirements and Tools Used :

To build a good model we should have a cleaned dataset and if not then need to work on same so that will get best possible outcome. Data Processing and data cleaning are one of the vital job role of data scientist. In this problem . Hardware tools which being used while coding were :

1. CPU - This is one of the important part while performing machine learning task because most of the computation will be done in learning environment which is most likely done on CPU.
2. Power Supply - This is also one of the important hardware while performing any machine learning task as if power supply will not be there , will not able to perform the task.

3. RAM - While performing coding on machine learning minimum 8 GB ram is required to perform the task.

Coming to software part the important libraries and packages required was for machine learning algorithm :

1. Pandas - One of the important library which helps in importing the data to perform data manipulation and analysis.
2. Numpy - This library is used to perform all the mathematical

operations required during coding

3. Seaborn and Matplotlib - Used for visualisation analysis. As we know that to understand any data in a more proper way and for better presentation, visualisation is one of the important technique used.

Model/s Development and Evaluation

· Identification of Possible Problem-Solving Approaches :

For knowing the data, means what type of data is and what approaches need to be followed we have to analyse the data thoroughly. This dataset is about price prediction of flight. Here our main task was to know the trend by observing previous dataset and train our model on the same so that on the basis of understanding our model can predict. Very first approach we can do is cleaning the data properly. And cleaning data means analysing null values if any, what datatype each column do have. Statistical analysis, EDA etc. using all these we can have a proper analysis of our data before model building. In this dataset there were no missing values so no need treat them, next using info method analysed what are the datatype , where most of the column were in string datatype , according to column variance changed them to numerical form like date column converted them to data datatype and using the same separated year, month and date columns respectively , similarly for arrival, departure, duration did the same using hour and min. Next for few object columns like flight, source and destination used one hot

encoder to convert them to numerical form as the no of columns were less so chances to have more number of columns were low, As id we use one hot encoder no of columns get increased which turns our data to be huge . Next for total stop used replace method. After feature engineering done next observed if any continuous column is there or not for which we need to remove skewness and outliers. But there were no such columns. Next checked whether any feature is correlated to each other or not. As if there is high correlation among any feature there is a chance of overfitting , which may affect our model performance. But here none of the columns were highly correlated to each other. So no need to remove any of the columns. Just dropped Date, Arrival, Departure, Duration columns as already extracted data from these columns separately so these columns were not required anymore. After doing all preprocessing, Data Cleaning, EDA spliced feature and target column separately then scaled all the features so that all the features come to same unit. Then stated model building after finding best random state.

· Testing of Identified Approaches (Algorithms) :

For this problem statement used the below mentioned algorithms for training and
testing purpose :
1. Linear Regression
2. LassoCV
3. RidgeCV
2. Decision Tree Regressior
3. Random Forest Regressior
For Hyper parameter Tuning used Grid Search CV on Random Forest Regressor as this is one of best performing model among all.

Different metrics used to measure the score are:
1. R2Score
2. MAE
3. MSE
4. RMSE

· Run and Evaluate Selected Models :

For knowing the data, means what type of data is and what approaches need to be followed we have to analyse the data thoroughly. This dataset is about price prediction of flight. Here our main task was to know the trend by observing previous dataset and train our model on the same so that on the basis of understanding our model can predict. Very first approach we can do is cleaning the data properly. And cleaning data means analysing null . Before selecting any model first found best Radom State using one of the model which is Decision Tree Regressor. We find best random state so that will train our model on that particular score. Below is the snapshot for the same :

```
maxscore = 0
maxrs = 0

for i in range(1,1000):
    x_train,x_test,y_train,y_test = train_test_split(x_scaler,y,test_size = 0.30,random_state = i)
    dt = DecisionTreeRegressor()
    dt.fit(x_train,y_train)
    pred = dt.predict(x_test)
    rsc = r2_score(y_test,pred)
    if rsc>maxscore:
        maxscore=rsc
        maxrs=i
print("Best r2 score is:",maxscore,"On Random state: ",maxrs)
```

```
Best r2 score is: 0.7054411380757111 On Random state:  841
```

Here best R2 score we got as 70% approx on 841 best random state . Further will train Our model on the same.

Training model using x and y as variable for feature and target respectively on best random state as 841 which we get above.

```
x_train,x_test,y_train,y_test=train_test_split(x_scaler,y,test_size=0.30,random_state=i)
```

Models which I selected for this problem statement are : 1) Decision Tree Regression 2) Linear Regression 3) Lasso 4) Ridge 5) Random Forest Regressor. Here we need to predict flight price which is continuous in nature so considering this assumed that this problem statement is regression based problem statement and selected models accordingly. Further for metrics selected R2 score as this metrics we use for regression based problem statement. Few more metrics I used here to check error : MAE, MSE, RMSE. And for hyperparameter tuning selected GridSearchCV to improve the score if possible. Snapshot for the same mentioned below :

```
from sklearn.tree import DecisionTreeRegressor
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import r2_score
from sklearn.model_selection import GridSearchCV,RandomizedSearchCV
from sklearn.model_selection import KFold, cross_val_score
from sklearn.linear_model import Ridge,Lasso,LassoCV,RidgeCV
```

Models selected run below to check which model performance is best for this problem statement :

1) **Decision Tree Regression** - One of the supervised machine learning algorithm where the data is continuously split according to certain parameters . Score which the model understood and given were really not upto the mark. Score mentioned was really less which was just 38% and CV score was in negative. Which here I considered that my model was not good performer in Decision Tree Regression algorithm. Below is the snapshot for the same :

```
dt = DecisionTreeRegressor()
dt.fit(x_train,y_train)
pred = dt.predict(x_test)
print(r2_score(y_test,pred))
```

```
0.3965781256150046
```

```
print(cross_val_score(dt,x_scaler,y,cv=5).mean())
```

```
-1.2878378354342135
```

2) **Linear Regression** - This is also one type of supervised machine learning algorithm which performs the task to predict a dependent variable value based on a given independent variable. This technique find the linear relationship between input and output. The algorithm metrics score was better then Decision Tree Regression which is 51%, but CV for this too were in negative which is -7% . To check whether the model is overfitting or not used lasso and ridge , if the score between two having difference or not. Below is the screenshot for the same.

```
lr = LinearRegression()
lr.fit(x_train,y_train)
pred = lr.predict(x_test)
print(r2_score(y_test,pred))
```

```
0.511319031049276
```

```
print(cross_val_score(lr,x_scaler,y,cv=5).mean())
```

```
-6.994825671269636e+25
```

Checking whether model is overfitting or not using lasso and ridge

```
lassocv = LassoCV(alphas = None, max_iter=1000, normalize = True)
lassocv.fit(x_train,y_train)

alpha = lassocv.alpha_

lasso_reg = Lasso(alpha)
lasso_reg.fit(x_train,y_train)

lasso_reg.score(x_test,y_test)
```

```
0.5114178870519259
```

```
ridgecv = RidgeCV(alphas =(0.1,1.0,10.0),normalize = True)
ridgecv.fit(x_train,y_train)

alpha = ridgecv.alpha_

ridge_reg = Ridge(alpha)
ridge_reg.fit(x_train,y_train)

ridge_reg.score(x_test,y_test)
```

```
0.511456560223522
```

As for both lasso and ridge we get same score as 51% approx which is similar to Linear regression score. So here on the basis of score considering that our model is not an overfitting model.

3) **Random Forest Regression** - This is also one kind of supervised machine learning algorithm which builds multiple decision trees and merges them together to get a more accurate and stable prediction. Compare to other two algorithms this model performed best whose r2 score was approx 74% but CV score for the same was in negative which is -43% . Snapshot for the same mentioned below:

```
rf = RandomForestRegressor()
rf.fit(x_train,y_train)
pred = rf.predict(x_test)
print(r2_score(y_test,pred))
```

```
0.7427301595269199
```

```
print(cross_val_score(rf,x_scaler,y,cv=5).mean())
```

```
-0.438807349095784 56
```

As got all the algorithm CV score in minus so checked other metrics score to check how much error is there using MAE, MSE, RMSE. Snapshot for the same given below :

```
from sklearn import metrics
print('MAE:', metrics.mean_absolute_error(y_test, pred))
print('MSE:', metrics.mean_squared_error(y_test, pred))
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, pred)))
```

```
MAE: 1420.0240128503542
MSE: 4816139.47915921
RMSE: 2194.570454362131
```

Here MAE score get as 1420 means this much is the error in this particular dataset. MSE score is too high . Here MSE score means Mean Squared Error , Next coming to RMSE score which means Root mean squared error that got as 2194. So almost error score is really high.

Next did hyperparameter tuning using GridSearchCV to improve the model score if possible :

GRIDSEARCHCV = GridSearchCV is a library function that is a member of sklearn's model_selection package. It helps to loop through predefined hyperparameters and fit your estimator (model) on your training set. So, in the end, you can select the best parameters from the listed hyperparameters. After getting best parameters tuned the same with best performing model which was Random Forest Regression algorithm. And the final score which I got was 61%. So instead of getting improved score my score get diminished. Need to improve more by including more parameters or some other hyperparameter tune model. Below mentioned the screenshot for the same:

```
# RandomForestRegressor
param = {'n_estimators':[30,40,50,60],
         'criterion':['mse','mae'],
         'max_depth':[2,3,4,5,6],
         'max_features':['auto','sqrt','log2']}
```

```
GC = GridSearchCV(rf,param,cv=5)
```

```
GC.fit(x_train,y_train)
```

```
GridSearchCV(cv=5, estimator=RandomForestRegressor(),
             param_grid={'criterion': ['mse', 'mae'],
                         'max_depth': [2, 3, 4, 5, 6],
                         'max_features': ['auto', 'sqrt', 'log2'],
                         'n_estimators': [30, 40, 50, 60]})
```

```
GC.best_params_
```

```
{'criterion': 'mse',
 'max_depth': 6,
 'max_features': 'auto',
 'n_estimators': 30}
```

```
final_rfc = RandomForestRegressor(criterion = 'mse',max_depth = 6, max_features = 'auto',n_estimators = 30)
final_rfc.fit(x_train,y_train)
pred = final_rfc.predict(x_test)
```

```
metrics.r2_score(y_test,pred)
```

```
0.6126589926094568
```

· Key Metrics for Success in Solving Problem Under Consideration :

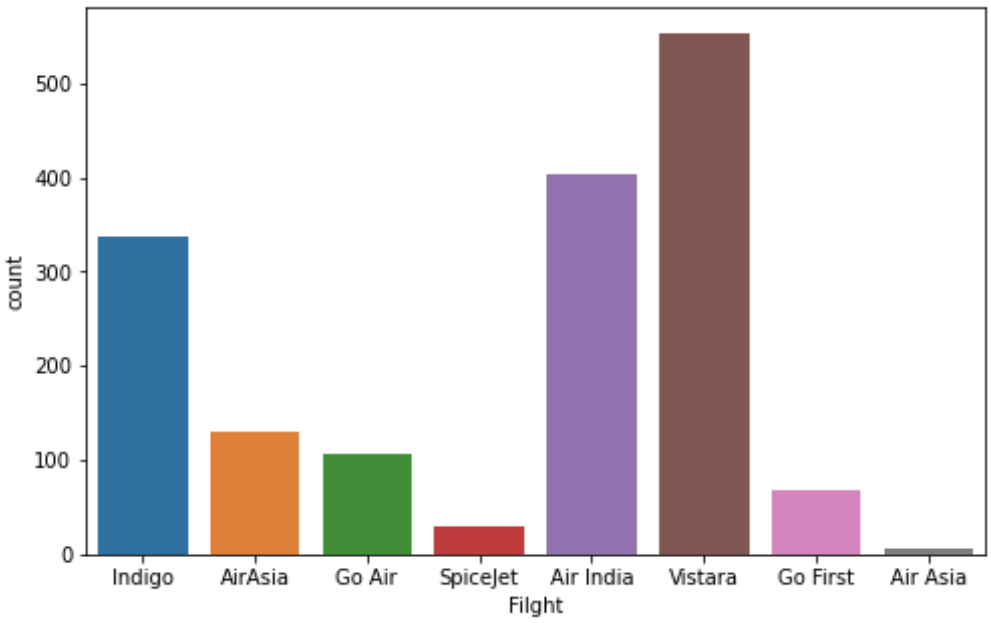
To solve any problem statement first we need to know what type of problem is? Means as we know that we are working on supervised machine learning but in that also there are two different types which are - Regression problem, Classification problem. Regression problem generally we use when our target column is in continuous nature while classification problem generally we use when the target column is having classified or binary based problem . This problem statement where I need to predict flight price considered as regression problem as the target column is continuous in nature. After identifying what problem statement is then analysed the data after loading it in python using pandas. After loading checked the shape of data. Then to have good model we need to do proper data cleaning and preprocessing method. Best model performance depends how good the data cleaning done. Next important step is to decide which model would perform best for a particular problem statement. These are the few factors which are important to get success in best model performance.

· Visualisations :

Visualisation is one of the important aspect to analyse any data. Visualisation helps in understanding the data more deeper and clearer , if anyone who doesn't able to understand by text they can even understand by seeing the graph and plot . Here for getting visualisation analysis import seaborn and matplotlib library. Here compared feature with target and tried to get insight which feature do play more important role. Below are the few plotted graphs image with more description.

```
plt.figure(figsize=(8,5))
sns.countplot('Filght',data=df)
```

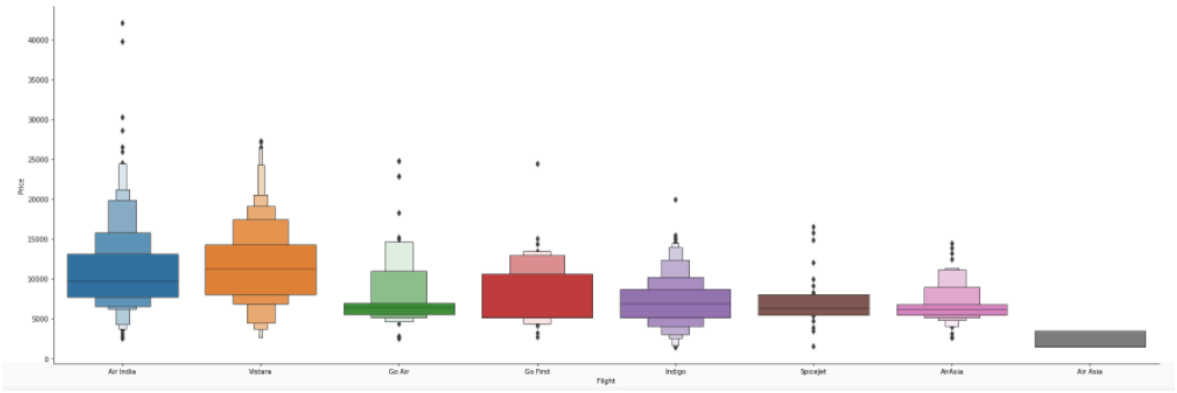
```
<AxesSubplot:xlabel='Filght', ylabel='count'>
```



Above using countplot analysed which flight are more in number. Vistara followed by Air India which is ruined by government is more in number compare to all other flights and Air Asia is the which is having very less flights available . In economical flights we can see that indigo is more in number. Even Air Asia too are considered as economical flight only as the price these flights are cheaper depending upon time and distance.

```
plt.figure(figsize=(3,2))
sns.catplot(x='Filght',y='Price',data=df.sort_values('Price',ascending=False),kind='boxen',height = 8, aspect = 3)
plt.show()
```

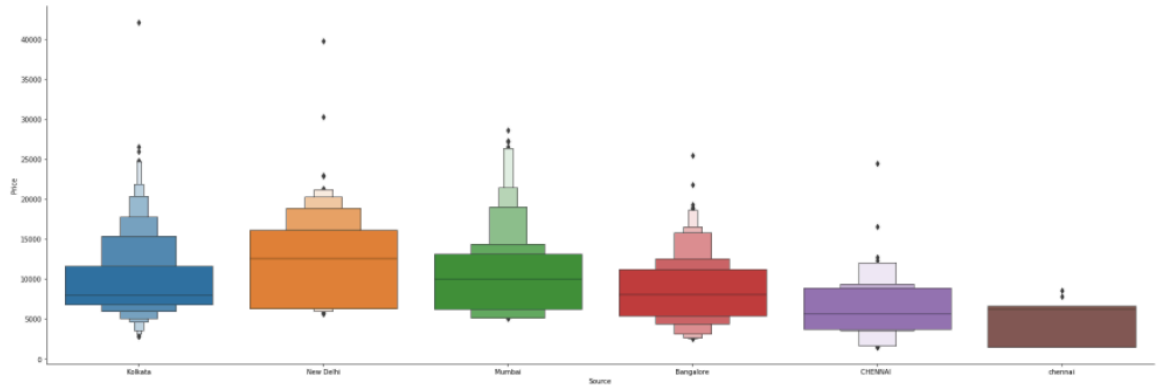
```
<Figure size 216x144 with 0 Axes>
```



Here using catplot trying to get the insight of which is cheaper and which flight do have higher price. Vistara do have higher price compare to all other flights while Air India too shows higher rate only . By plot we can see that there are minor difference between vistara and Air India. Here we can say that these two flights generally targets business class customers. While rest all are having economical price. They target all type of customers especially Indigo and Go Air do have cheaper price depending upon distance and time.

```
plt.figure(figsize=(4,3))
sns.catplot(x='Source',y='Price',data=df.sort_values('Price',ascending=False),kind='boxen',aspect=3,height=8)
plt.show()
```

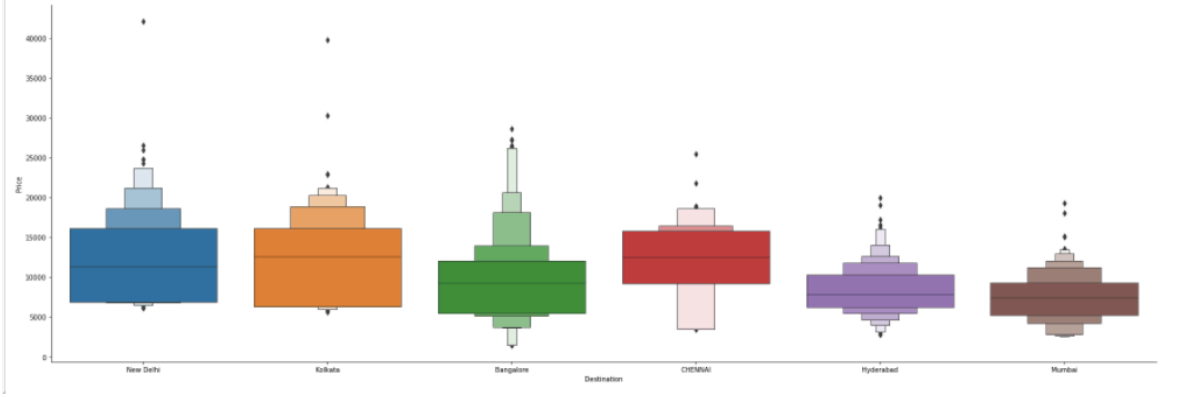
```
<Figure size 288x216 with 0 Axes>
```



Source means from where the customer takes flight. Here New Delhi we can see are having higher price compare to any other city . As I am analysing few of the metropolitan cities like Kolkata, New Delhi, Mumbai, Bangalore and Chennai. Among all New Delhi having higher price for any of the flight and lowest is for Chennai.

```
plt.figure(figsize=(4,3))
sns.catplot(x='Destination',y='Price',data=df.sort_values('Price',ascending=False),kind='boxen',aspect=3,height=8)
plt.show()
```

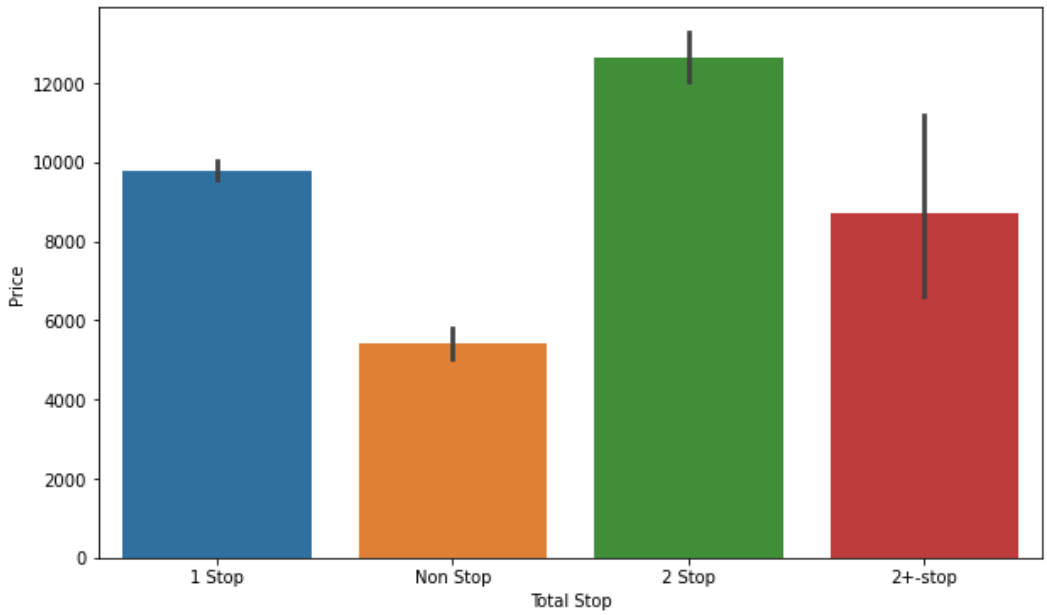
```
<Figure size 288x216 with 0 Axes>
```



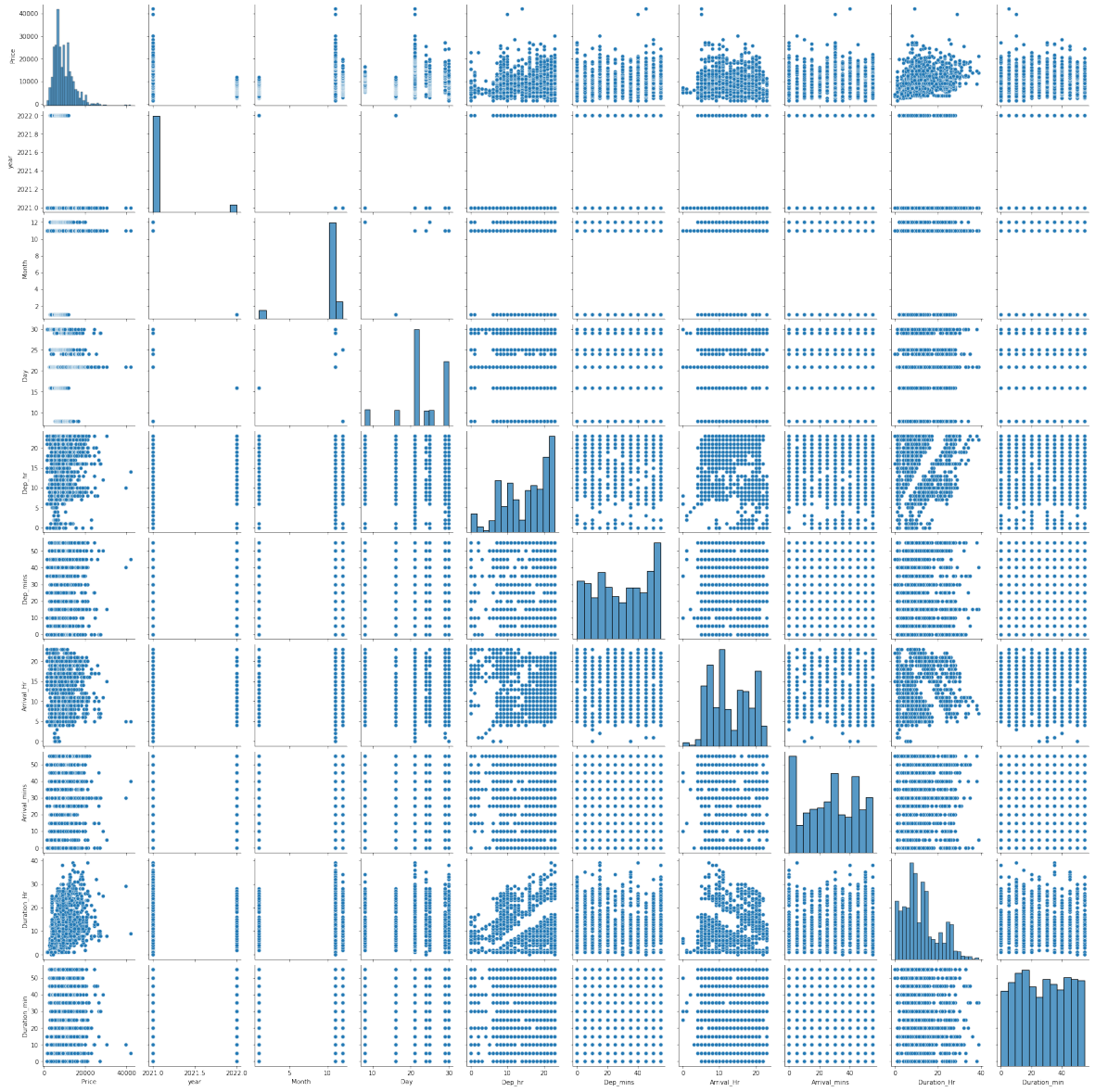
Destination here means where customer wants to go means final destination. Again for New Delhi we can see having higher price likewise in source we observed. Even Kolkata too have higher flight rate compare to any other city.

```
plt.figure(figsize=(10,6))
sns.barplot(x='Total Stop',y= 'Price',data=df)
```

```
<AxesSubplot:xlabel='Total Stop', ylabel='Price'>
```



For more stoppage price of flight is high. Likewise above using barplot we can see that flight which is having 2 stop is having higher price followed by 2+ stop.. And the cheapest we can see is for non stop flights as these flights covers less distance . And for less distance it's obvious that price will be less as distance coverage is less and fuel consumption even are too less.



With the help of pair plot we can see the relationship among features and target too.Here we get insight of date, arrival, departure and duration . Here date shows no relationship with price.Whereas departure hour and duration hour do have some relationship among them. Coming to arrival and departre none have direct relation with price. But price and duration relation we can see is showing towards upward which means tere are some relation between them.

· Interpretation of the Result :

Using visualisation got the insight of data. Like above we seen that Vistara Flights were more in number compare to all other flights. And

least number of flights were SpiceJet because of less customers we can assume. Price wise we observed that Vistara is one of the costly flight . Here we can assume that these flights generally targets Business class customers while the cheapest flight is SpiceJet above using countplot we observed that even in number SpiceJet is less . Else rest all flights considered to be economical as those flights price is neither too high nor too less. Depending upon time, distance and duration their price been set. Even these flights generally hold few seats for last minute sale so that they can earn profit by selling them on higher rate. Once gone through the insight of data did preprocessing .This dataset is a regression based problem statement because it's target column is continuous in nature. There is no missing value in this dataset. Few of the columns were in object datatype so accordingly need to convert them to numerical form as our algorithm doesn't understand string datatype. Using one hot encoder converted object datatype to integer datatype. Next checked the correlation among any feature to avoid overfitting . There were very less correlated relation shown among any of the features so not removed any of the column. Not performed skewness and outliers removal as there were no continuous data and generally we skewness and outliers we perform for mathematical operation. After data cleaning and preprocessing done splitter feature and target into x and y respectively for model building. Next found best random state. After that used 5 algorithms which were : Decision Tree Regression, Random Forest Regression, Linear Regression, Lasso, Ridge. Among all Random Forest Regression were the best performing model. Using few metrics which were MAE, MSE, RMSE checked the error in model building. Here error were too high. Even CV score for any of model got in negative. Did Hyperparameter tuning using GridSearchCV so that will able to improve score if possible. But score not improved . Random Search R2score were 74% while after tuning the model got only 63%. For improving score and to get best possible we can use some other model too if they can perform well on this datatype. As a data scientist need to explore more and more to have best performing model.

Conclusions

· Key Findings and Conclusions of the Study :

This dataset is all about flight price prediction where as a data scientist I need to predict the price of flight depending upon previous price dataset. Flight price depends upon various factors like destination i.e.,source and final destination, arrival and departure time, duration, total stop. If the total stop increases flight price go high as the distance increases . Even arrival time matters a lot means usually we observed that early morning and late night flights are more cheaper than usual day and evening flights. Also to earn more profits generally aviation house do hold few of the seats for last minute sale. As last minute tickets are generally sold at very high price where any of the flight will have good profit. So reserving seat is one of the technique which most of the aviation house do to earn more profit and increase their revenue. Here in this dataset after loading it to python checked the shape of data . Where total no of rows were 1632 and 9 columns. After getting data started data cleaning and preprocessing. Data cleaning is one of the important aspect for best model building. First started with whether there is any missing values or not . There were no missing value in this dataset. Next checked the type of data. There were most of the columns which were in string datatype. So to convert them to numerical form used encoding method. Encoding method generally converts object datatype to numerical form. Here I used one hot encoder to convert few of the object columns like flight, source and destination. Rest date, arrival, departure and duration column using date time method converted them to numerical form. And for total stop used replace method. Once feature engineering done checked the correlation among feature to avoid overfitting. But there were very less correlation among any feature. Not dropped any of the feature. As the number of feature be more model performance will be much better. Then splitting feature and target into x and y respectively for model building. Scales all the feature using standard scaler to bring them to same unit. Next find best random state based on which trained the model. After finding best random state selected few of the algorithms

for model building . These are - Decision Tree Regression, Random Forest Regression, Linear Regression, Lasso and Ridge. Among all Random Forest Regression were the best performing model with R2 score of 72%. But CV score was in negative for all the models. So checked using other metrics like MAE,MSE,RMSE error score. And the error was too high. Next to improve the model score used GridSearchCV for hyperparameter tuning. There also score not improved. Instead of increase score get reduced. This may be possible because of I had to add more parameters so that model could have trained themselves more better or could have performed more algorithms based on which would have done hyperparameter tuning.

· Learning Outcomes of the Study in Respect of Data Science :

Outcomes on the basis of visualisation was that came across which flight are more in operation means more in number where came across that vistara were more in number compare to all other flights followed by Air India. Less number of flights in operation was SpiceJet. This we can say because of less customer they do have or are in more loss so they reduced the number of flights. Then compared each flight with price. Where get insight that vistara though covering same distance on same time from same destination having more price compare to all, almost 3 times more expensive comparatively. While spice jet was the cheapest flight compare to all. There were few flights which were considered as economical because they were maintaining cheaper to expensive flight tickets. Depending upon distance, stoppage and timing these flights were deciding their rate. Coming to data cleaning part which is one of the important aspect in model building here only need to do feature engineering means converting all object datatype to integer or float datatype. As there were no null values so no need to do anything for that. No continuous columns so no need to check skewness or outliers. After converting all the columns to numerical form checked correlation among features. There were no such correlation among any feature so less chance of overfitting. Next chosen algorithms which would work better on this model. Among all Random Forest Regression was having best R2 score but cv score was in negative. So we should have chosen some other algorithm assuming that would have performed much better than Random Forest Regression as this algorithm score was 72% and to improve the same used hyperparameter tuning using GrdSearchCV ,

But instead score get improved it reduced. This may be because of more parameter to be added so chances of score improvement would have increased.

· Limitations of this Work and Scope for Future Work :

Outcomes on the basis of visualisation was that came across which flight are more in operation means more in number where came across that vistara were more in number compare to all other flights followed by Air India. Less. In this dataset I was having very less things to do like no require of missing data treatment as there were no null values. Then Statistical inferences or observation were not required much as no continuous column was there except price which was our target column. And if there were no continuous column then no need to remove skewness and outliers as these are mathematical operation and mainly mathematical operation be performed on continuous column. Few more data cleaning part and preprocessing not able to do due to less columns , even exploring visualisation was also less. These were the few limitations which I faced with this dataset. Scope is flight price is one of requirement which decides when getting more revenue and when the price could be increased so that they can earn more and more revenue and profit. Also as a data scientist getting chance to work on a new type of project where will get experience on how to analyse these types of dataset and what are best possible algorithms could be applied.