NAME OF THE PROJECT

“Flight Price Prediction”

Submitted by:

Ashish Kumar Samal

ACKNOWLEDGMENT

It is a matter of great pleasure to express my profound feeling of reference to all the people who helped and supported me during the project. I would like to convey my sincere thanks to FLIPROBO TECHNOLOGIES AND DATATRAINED EDUCATION for constantly helping me with the valuable inputs during the project duration. Their inspiring guidance and everlasting enthusiasm have been valuable assets during the tenure of my project.

ASHISH KUMAR SAMAL

INTRODUCTION

- Business Problem Framing

  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) So, you have to work on a project where you collect data of flight fares with other features and work to make a model to predict fares of flights.

- Motivation for the Problem Undertaken

  To find out the following analysis about the airfares.

Do airfares change frequently? Do they move in small increments or in large jumps? Do they tend to go up or down over time? What is the best time to buy so that the consumer can save the most by taking the least risk? Does price increase as we get near to departure date? Is Indigo cheaper than Jet Airways? Are morning flights expensive?

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  We used different regression models to predict the flight price from the given features.

- Data Sources and their formats

At first hand we scraped the data of flights from different websites (yatra.com, skyscanner.com, official websites of airlines, etc). The number of columns for data doesn't have limit, it's up to you and your creativity. Generally, these columns areairline name, date of journey, source, destination, route, departure time, arrival time, duration, total stops and the target variable price.

- Data Preprocessing Done
  At first we had to clean the columns and then converted all the categorical features into numerical for further analysis and model building.Then we removed the skewness from the dataset.The data collected has no null values.

- Data Inputs- Logic- Output Relationships
  After the data is cleansed.We split the data for training and testing.And found out that it is performing best with the random forest regressor as compared to other regression models.

- Hardware and Software Requirements and Tools Used:
  Libraries like pandas,numpy.
  Punctuations were used for data cleaning.
  For visualization we used matlplotlib and seaborn.
  And Scikit Learn is used for model building and then model is hypertuned.

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

-Data Cleansing

-Visualization and EDA

-Train Test Split

-Model Building and evaluation

-Cross Validation

-Hyperparameter tuning of best model

-Saving the best model

- Testing of Identified Approaches (Algorithms)
  Random Forest Regressor
  Linear Regression
  Gradient Descent Regressor
  KNeighborsRegressor

- Run and Evaluate selected models

```
from sklearn.model_selection import GridSearchCV
parameter={'min_samples_leaf': [2,5,7,10],'min_samples_split': [2,5,7,10],
     'n_estimators':[100,200,300],
          'max_depth':[5,10,15,20,30]}
RCV=GridSearchCV(RandomForestRegressor(),parameter,cv=5,scoring='accuracy',verbose=2,n_jobs=-1)
RCV.fit(x_train,y_train)

RCV.best_params_
```

```
Fitting 5 folds for each of 240 candidates, totalling 1200 fits

{'max_depth': 5,
 'min_samples_leaf': 2,
 'min_samples_split': 2,
 'n_estimators': 100}
```

```
final_mod=RandomForestRegressor(min_samples_split=2,max_depth=5,min_samples_leaf= 2, n_estimators=100)
final_mod.fit(x_train,y_train)
pred=final_mod.predict(x_test)
acc=r2_score(y_test,pred)
print(acc)
```

```
0.7323237782498229
```

- Key Metrics for success in solving problem under consideration

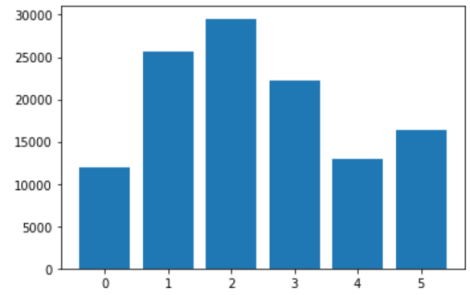
```
from sklearn.model_selection import cross_val_score


def rmse_cv(model, x,y):
    rmse =cross_val_score(model,X,Y, cv=5)
    return(rmse)


models = [LinearRegression(),
             RandomForestRegressor(),
             DecisionTreeRegressor(),
             GradientBoostingRegressor()]



names = ['LR','RF','DTR','GBR']

for model,name in zip(models,names):
    score = rmse_cv(model,X,Y)
    print("{}    : {:.6f}, {:4f}".format(name,score.mean(),score.std()))
```

```
LR    : -0.947390, 0.825490
RF    : 0.353811, 0.461564
DTR    : 0.088920, 0.558125
GBR    : 0.469224, 0.357570
```

- Visualizations

```
1 plt.bar(df['Month'],df['Price'])
```

```
<BarContainer object of 4169 artists>
```

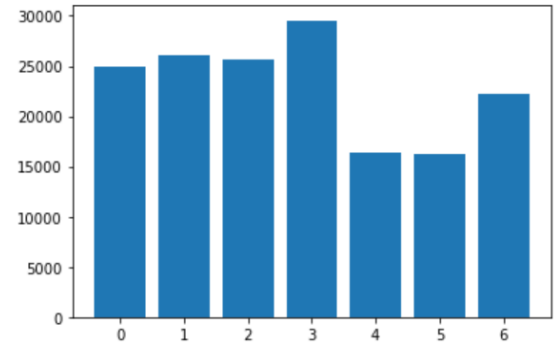


Apr 0 Feb 1 Jan 2 Jun 3 Mar 4 May 5

--The data was collected in the month of january which shows a rise in the price for month January and february which infers that there is a rise in price when the departure date is nearby. --The flight price is high if the departure date is nearby i.e.Jan and Feb(as the data collected in month of Jan) and seems to go down gradually in the month of Mar,Apr and grow afterwards in June,May.

```
1 plt.bar(df['Day'],df['Price'])
```

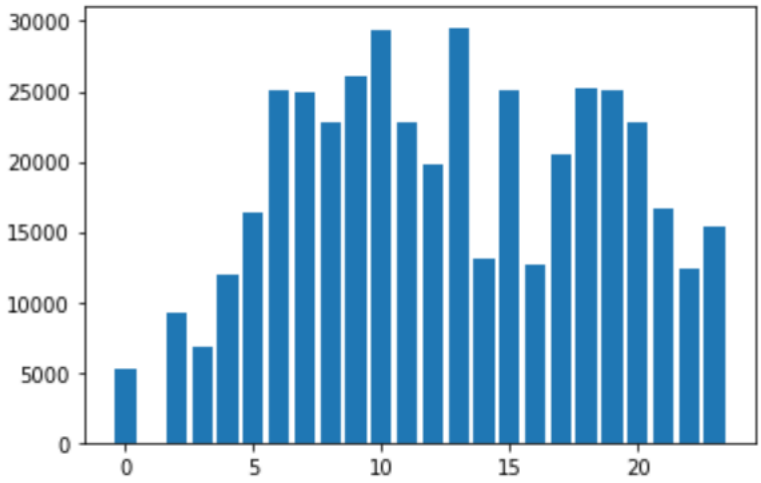
```
<BarContainer object of 4169 artists>
```



Tuesday,Thursday,Wednesday has less flight price. Friday,Saturday,Sunday,Monday has high flight prices as compared to other days.

```
1 plt.bar(df['Departure_Time'],df['Price'])
```

```
<BarContainer object of 4169 artists>
```



The above plot shows that the early morning flights are cheaper.

- Interpretation of the Results

  After trying different models we found out Random Forest Regressor is performing well and hence hypertuned the model to find the accuracy of model as 0.73.

CONCLUSION

- Key Findings and Conclusions of the Study

  The model is performing with an accuracy score of 0.73.

  Air Asia and Indigo flights seem to be generally cheaper as compared to Air India and Vistara. Tuesday,Thursday,Wednesday has less flight price. Friday,Saturday,Sunday,Monday has high flight prices as compared to other days.

  The data was collected in the month of January therefore it shows a rise in the price for month January and february which infers that there is a rise in price when the departure date is nearby.

  The flight price is high if the departure date is nearby i.e.Jan and Feb(as the data collected in month of Jan) and seems to go down gradually in the month of Mar,Apr and grow afterwards in June,May.

  Early morning flights are cheaper than flights at other time.

  Hence, the best time to book a flight is before one or two month and look for days between Tuesday to thursday so that the consumer can save the most by taking the least risk.

- Learning Outcomes of the Study in respect of Data Science

  Use of different regression models and find out the Required insights to help booking the ticket with less risk and cheaper rate.

- Limitations of this work and Scope for Future Work

The only limitation of the work which can get better with adding more data to the dataset which will help to increase the accuracy of the model for better prediction.