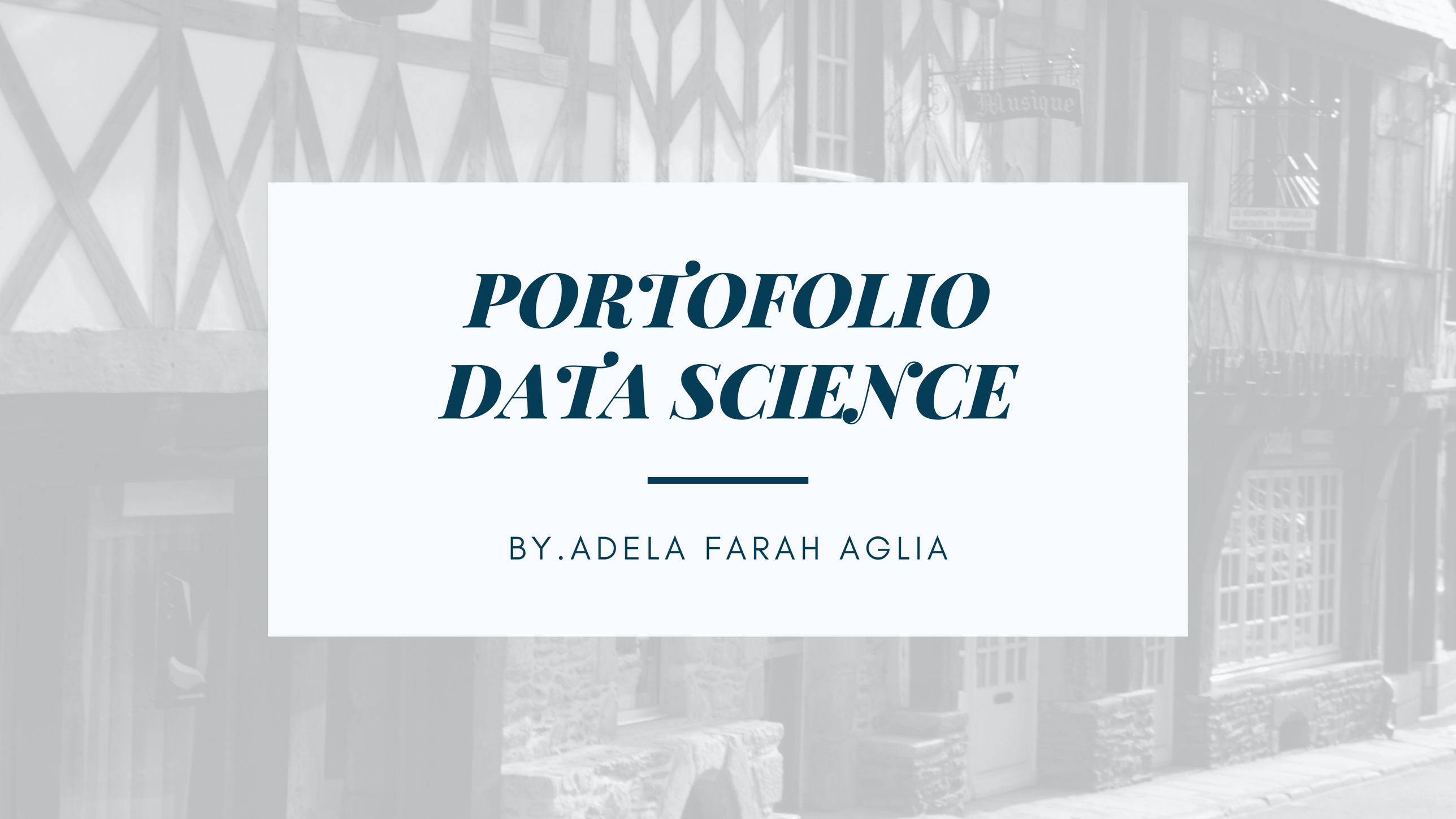


PORTOFOLIO DATA SCIENCE

BY.ADELA FARAH AGLIA

ADELA FARAH AGLIA

EDUCATION BACKGROUND

Bachelor's of Information systems from Sriwijaya Univeristy (2016-2020)

COURSE

Dbimbing,id

- DATA SCIENCE BOOTCAMP (JUL '21 - OKT '21)

KOMINFO

- DIGITAL TALENT SCHOLARSHIP - CLOUD FONDATION WITH AWS (JUN '21 - JUL '21)
- DIGITAL TALENT SCHOLARSHIP - DATA SCIENCE (SEP '20 - OCT '20)

Reach me out at :

: linkedin.com/in/AdelaFarahAglia

: adelafarah13.medium.com/

PORTOFOLIO
DATA SCIENCE

FLIGHT FARE PREDICTION

Tabel of Content

INTRODUCTION

Nowadays, Airlines is one of the preferred transportation options because it's effective in term of time.
But, Flight Fare often **fluctuate** due to several conditions.

At this project, the author wants to find the best machine learning model to predict flight fare which hopefuly can be help Airlines predict what prices they can manage based on certain circumstances.

DATA COLLECTION

DESCRIBING DATA

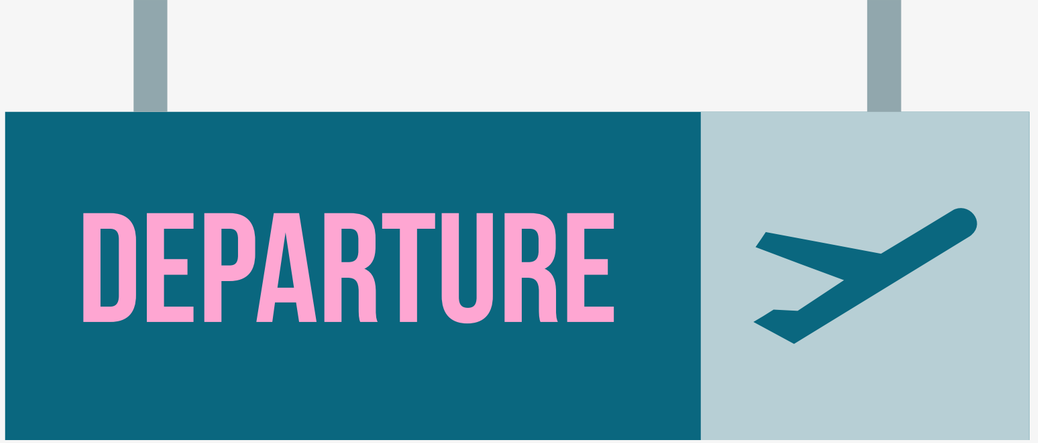
Data Understanding

EXPLORING DATA

Data Numerical

EXPLORING DATA

Data Categorical

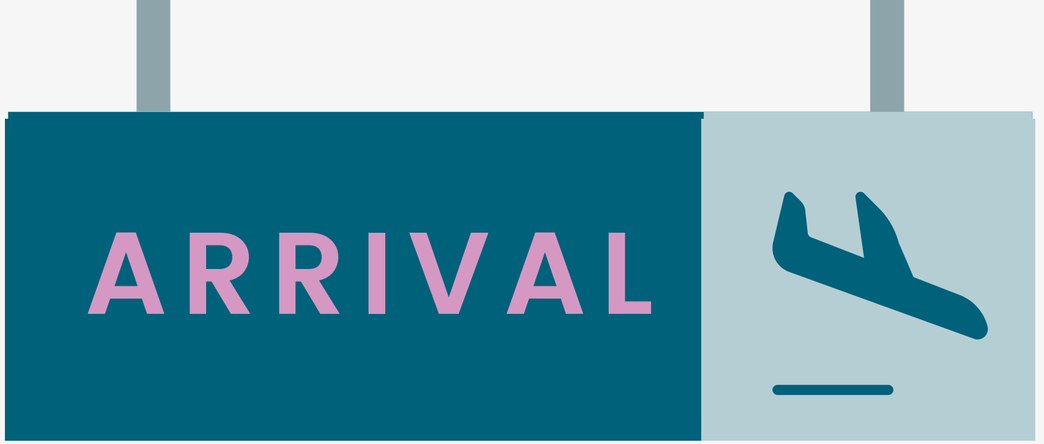


DATA COLLECTION

The Data Airlines is taken from Kaggle Platfrom. Consists of 10.683 row and 11 columns.

Source: https://www.kaggle.com/absin7/airlines-fare-prediction

DESCRIBING DATA



Kolom	Deskripsi
Airline	Nama Maskapai
Date_of_Journey	Tanggal Keberangkatan
Source	Lokasi Keberangkatan
Destination	Lokasi Kedatangan/tujuan
Route	Informasi tentang lokasi awal dan akhir perjalanan
Dep_Time	Waktu keberangkatan penerbangan dari lokasi awal
Arrival_Time	Waktu kedatangan penerbangan di tempat tujuan
Durasi	Durasi perjalanan dalam jam/menit
Total_stop	Jumlah total stop penerbangan sebelum mendarat di lokasi tujuan
Additional_info	Informasi tambahan tentang penerbangan
Price	Harga penerbangan

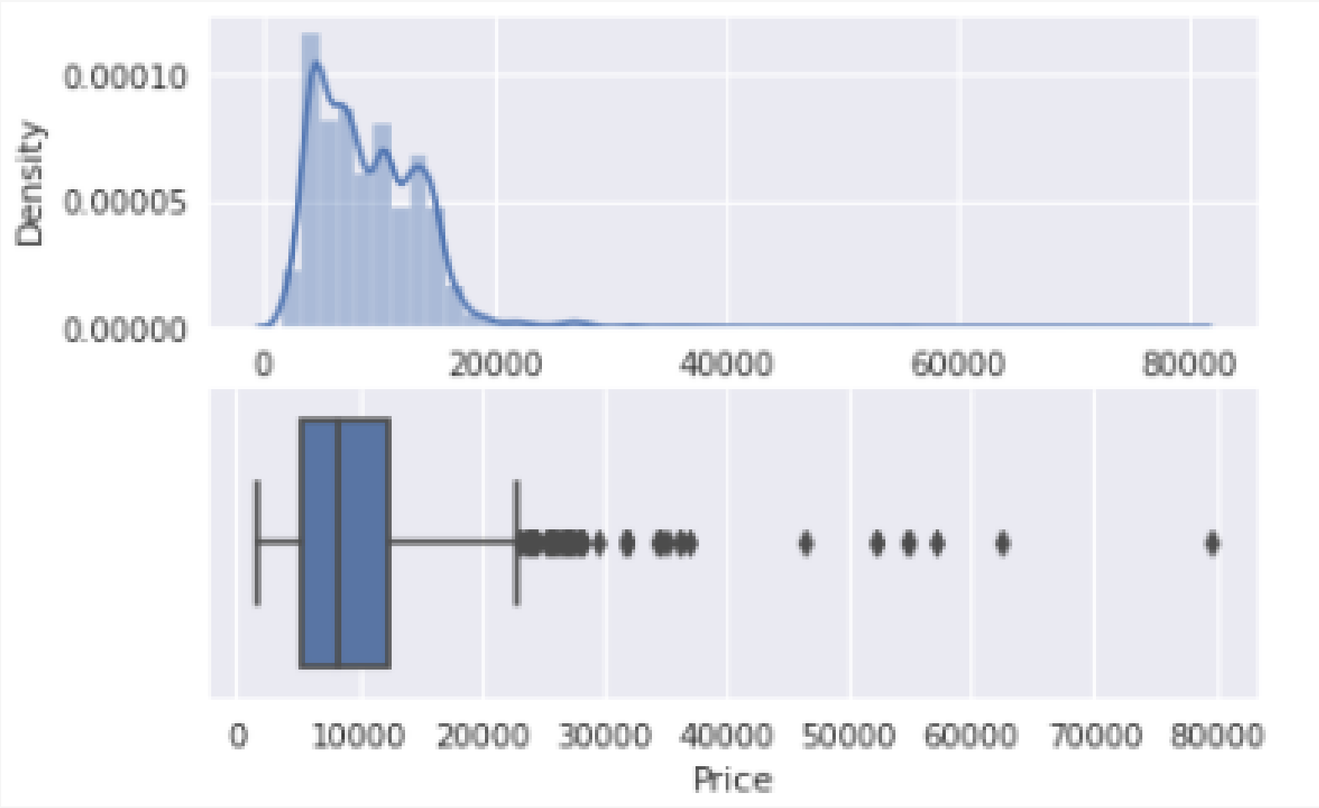
10 Categorical Data

1 Numerical Data

2 Missing Cell

220 Duplicate Row

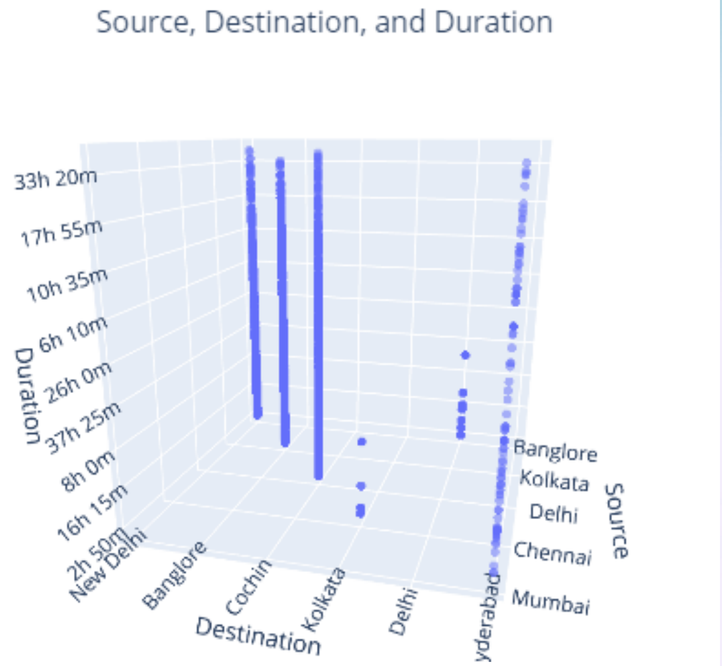
EXPLORATORY DATA

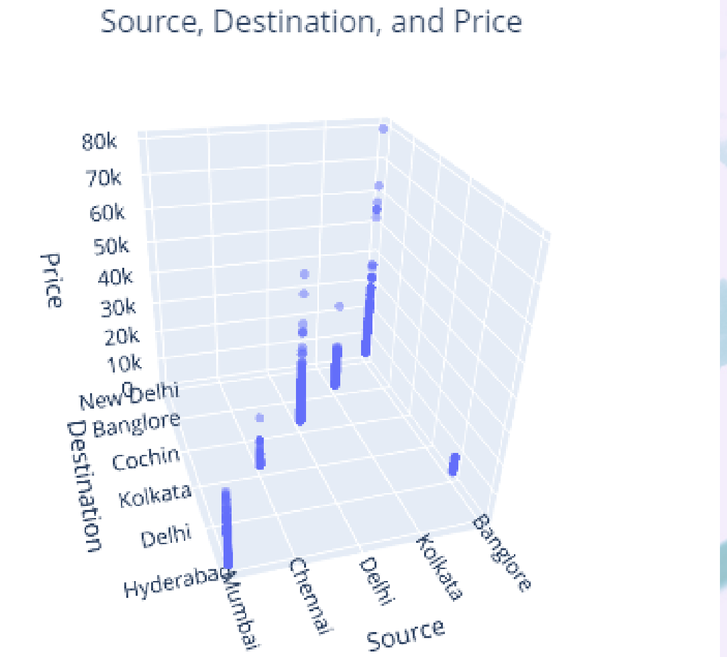


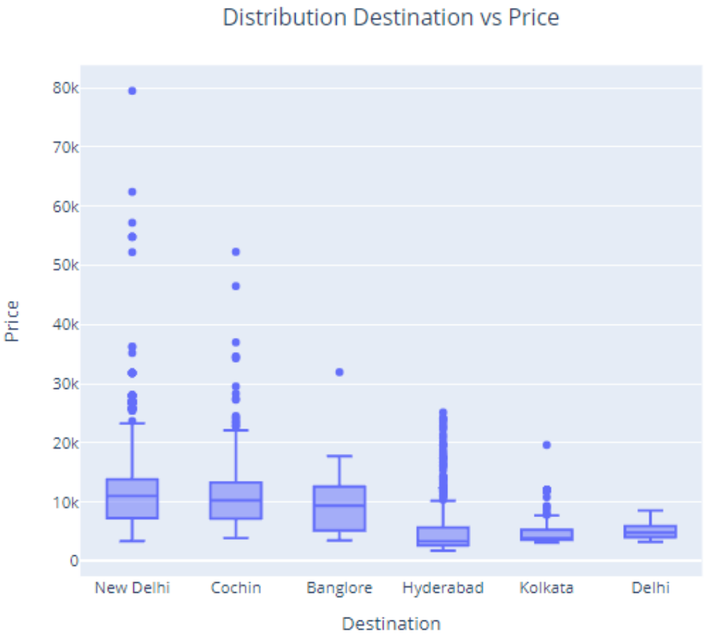
NUMERICAL DATA
PRICE AS TARGET
VARIABLE

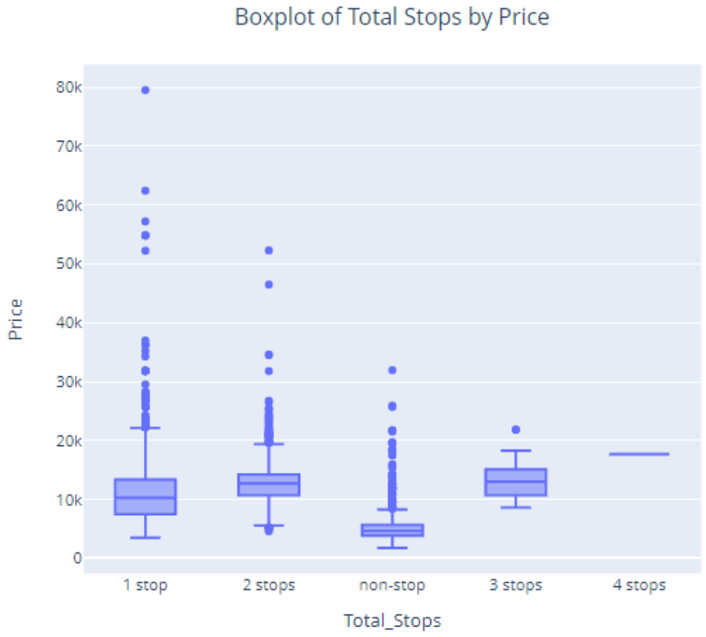
EXPLORATORY DATA

CATEGORICAL DATA
SOURCE,
DESTINATION



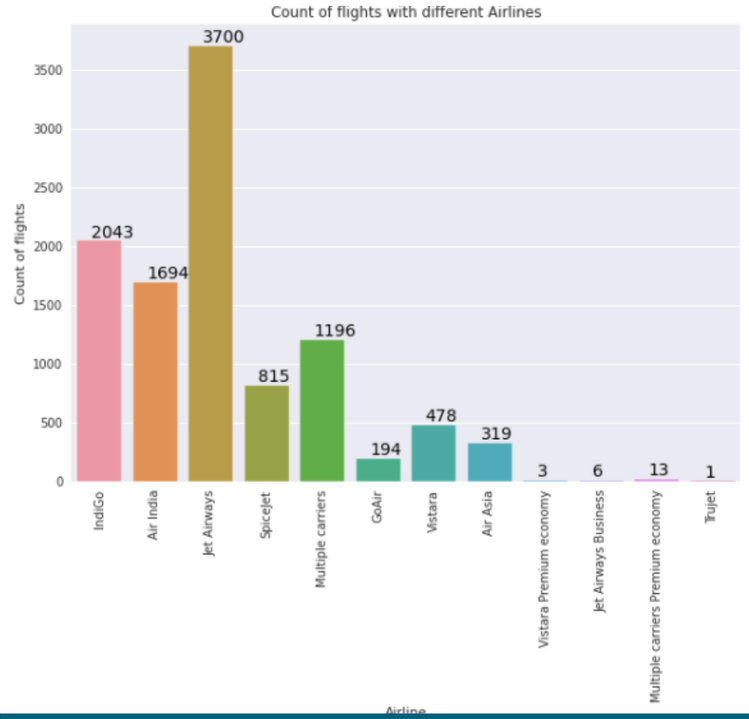






EXPLORATORY DATA

CATEGORICAL DATA
TOTAL STOP,
AIRLINES



CORELLATION

Between Column Price, Total_Stops, and Route

the target variable has a strong enough correlation on the total stops compared to the route variable. So for the next Route variable will not be used because it has been represented by the Total Stops variable

Data Pre-Processing

Handling:

- Missing Value
- Duplicate Data
- Outlier

Feature Engineering

Data Numerical

- Date of Journey
- Dep_time
- Arrival_time
- Duration

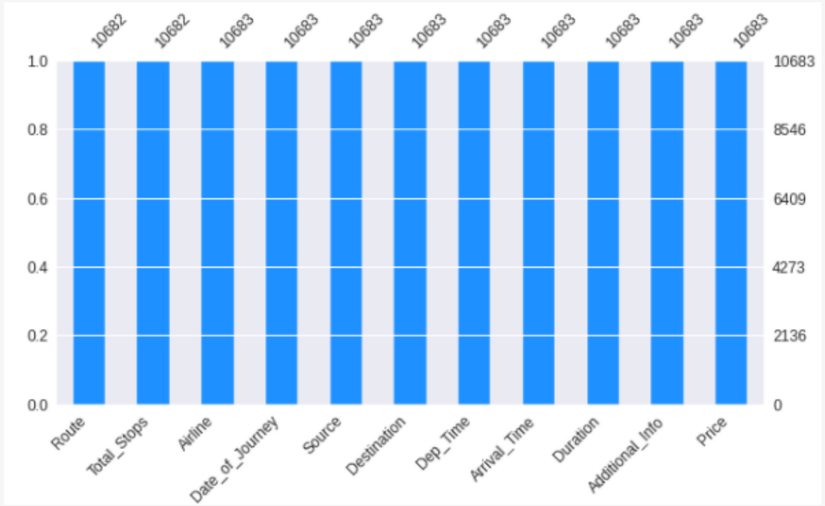
Feature Engineering

Data Categorical

- Airline
- Source
- Route
- Destination
- Total_Stops
- Additional_Info

Transformation Data

- Continous/ Numerical Data
- Categorical Data

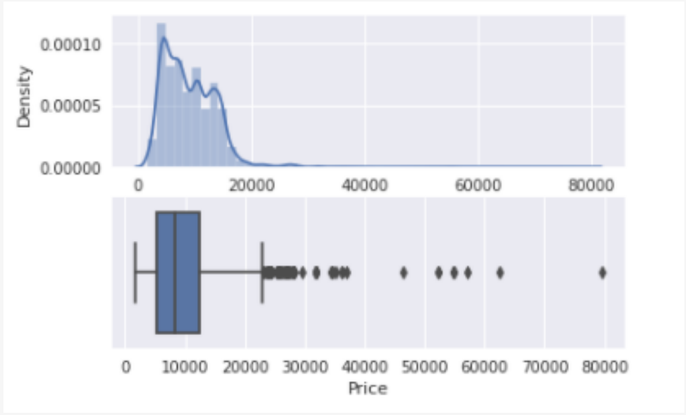


MISSING VALUE

```
jumlah row sebelum dibersihkan: 10682
```

```
jumlah row setelah dibersihkan: 10462
```

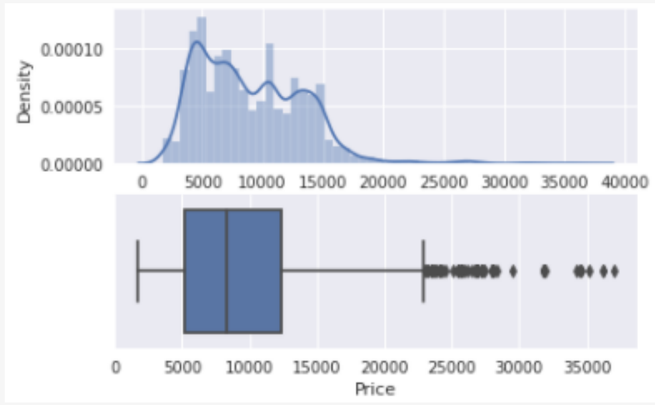
DUPLICATE DATA



Before

OUTLIER

Price



After

	Day_Journey	Month_Journey	Dep_Time_hour	Arrival_Time_hour	Dep_Time_minute	A
0	0.875000	0.181818	0.956522	0.043478	0.363636	
1	0.083333	0.000000	0.217391	0.565217	0.909091	
2	0.125000	0.727273	0.391304	0.173913	0.454545	
3	0.083333	1.000000	0.782609	1.000000	0.090909	
4	0.000000	0.000000	0.695652	0.913043	0.909091	
5	0.875000	0.454545	0.391304	0.478261	0.000000	
6	0.000000	1.000000	0.782609	0.434783	1.000000	
7	0.000000	0.000000	0.347826	0.217391	0.000000	
8	0.000000	1.000000	0.347826	0.434783	1.000000	
9	1.000000	0.363636	0.478261	0.826087	0.454545	

Transformation Numerical Data using Scaling-MinMax Scaler

Normalize data on numerical data so that the value range becomes (0,1)

Total_Stops	Airline_Air India	Airline_GoAir	Airline_IndiGo	Airline_Jet Airways
0	0	0	1	0
2	0	0	0	1
2	0	0	0	0
2	0	0	1	0
2	1	0	0	0

Transformation Categorical Data using OneHotEncoder & LabelEncoder

- Airline → OneHotEncoder
- Source → OneHotEncoder
- Destination → OneHotEncoder
- Additional Info → OneHotEncoder
- Total Stops → LabelEncoder

SPLIT DATA DEPENDENT AND INDEPENDENT

Price

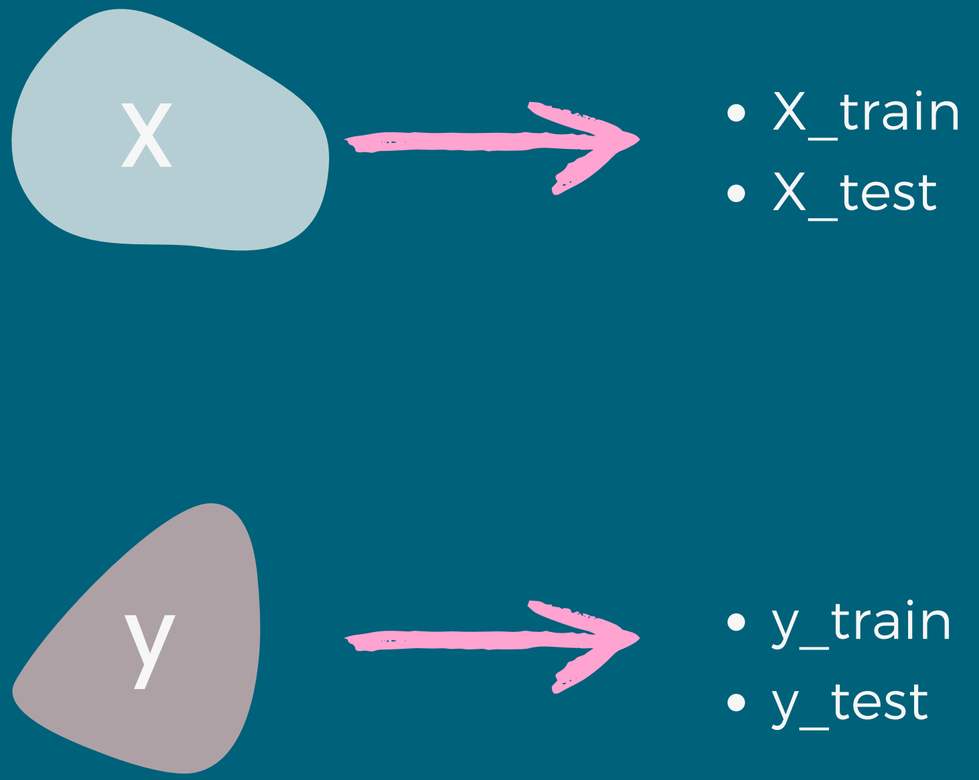
'Total_Stops', 'Airline_Air India', 'Airline_GoAir', 'Airline_IndiGo',
'Airline_Jet Airways', 'Airline_Jet Airways Business',
'Airline_Multiple carriers',
'Airline_Multiple carriers Premium economy', 'Airline_SpiceJet',
'Airline_Trujet', 'Airline_Vistara', 'Airline_Vistara Premium economy',
'Source_Chennai', 'Source_Delhi', 'Source_Kolkata', 'Source_Mumbai',
'Destination_Cochin', 'Destination_Delhi', 'Destination_Hyderabad',
'Destination_Kolkata', 'Destination_New Delhi',
'Additional_Info_1 Short layover', 'Additional_Info_2 Long layover',
'Additional_Info_Business class', 'Additional_Info_Change airports',
'Additional_Info_In-flight meal not included',
'Additional_Info_No Info',
'Additional_Info_No check-in baggage included',
'Additional_Info_Red-eye flight', 'Day_Journey', 'Month_Journey',
'Dep_Time_hour', 'Arrival_Time_hour', 'Dep_Time_minute',
'Arrival_Time_minute', 'Duration_hours', 'Duration_minute'

DEPENDENT

INDEPENDENT

TRAIN TEST SPLIT

- Training set 80%
- Testing set 20%



MODELING TIME!

MACHINE LEARNING

Because the target/dependent variable that the author wants is price (continuous numeric value) then this is a regression problem

LINEAR REGRESSION

Training score: 0.6444587235121174

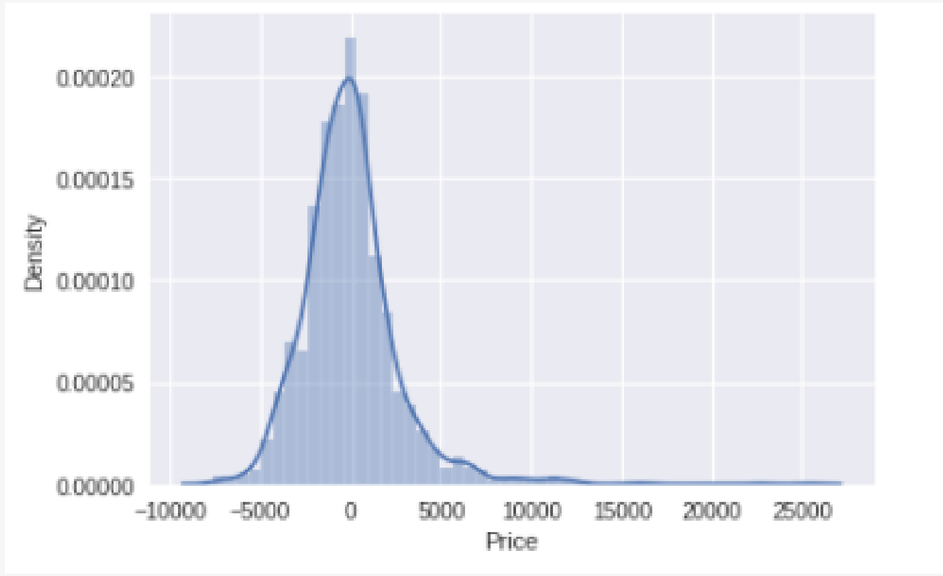
Test score: 0.64114843364973

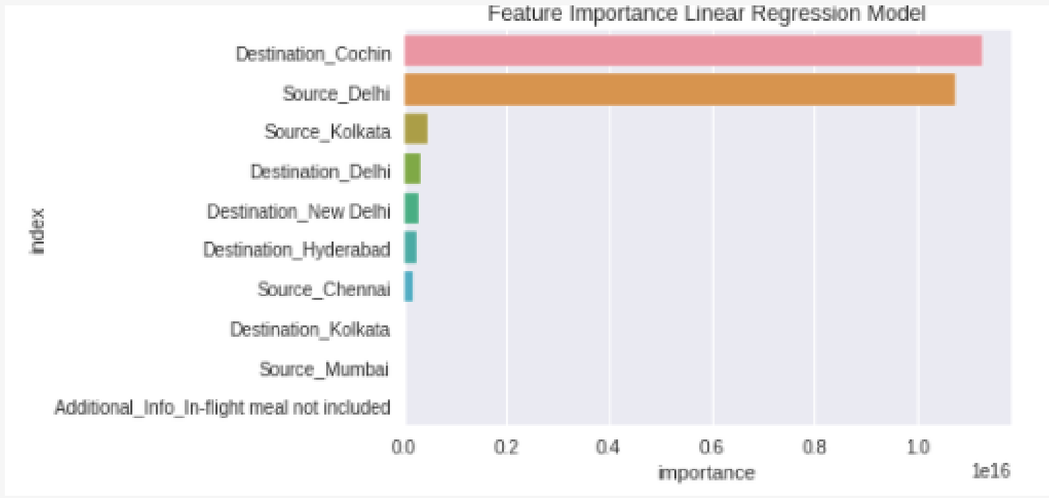
r2 score is 0.6411

MAE: 1849.7969421882465

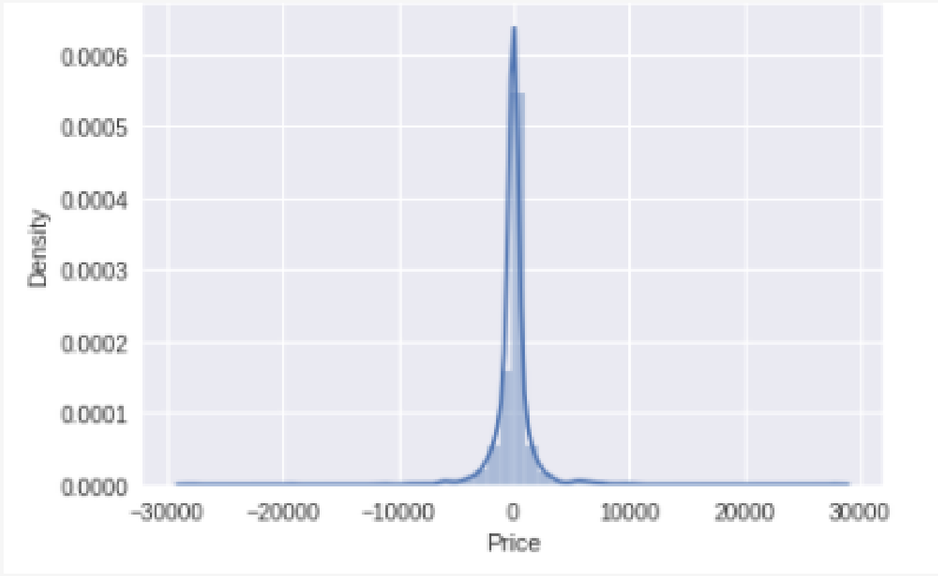
MSE: 7120851.471392738

RMSE: 2668.4923592532054





DECISION TREE REGRESSION



Training score: 0.995636081971856

Test score: 0.8387160631509332

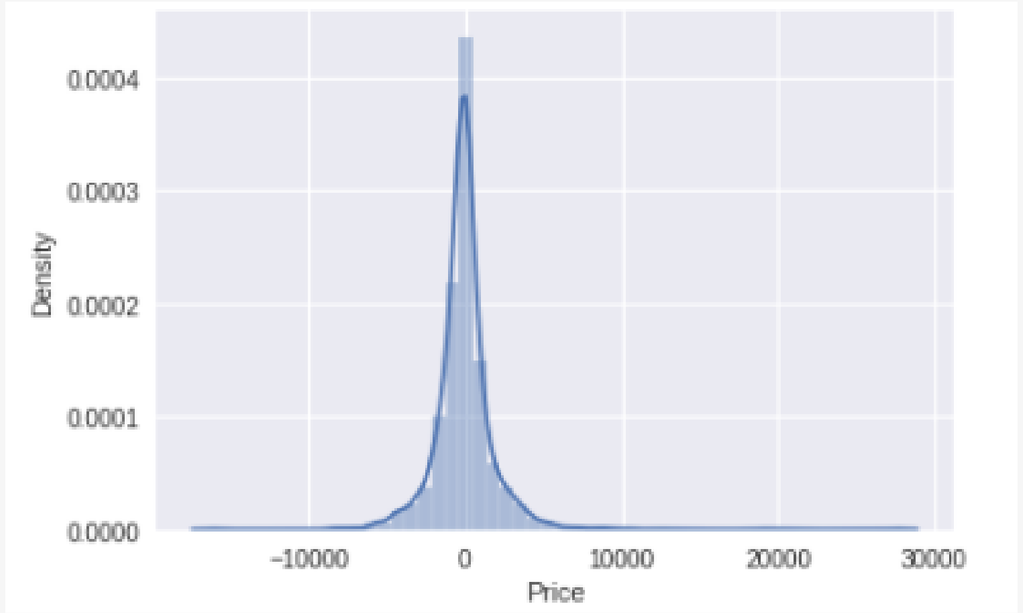
r2 score is 0.8387

MAE: 752.7490046185698

MSE: 3200428.9982746723

RMSE: 1788.9742866443532

KNEIGHBORS REGRESSOR



Training score: 0.87230987154623

Test score: 0.798421918948415

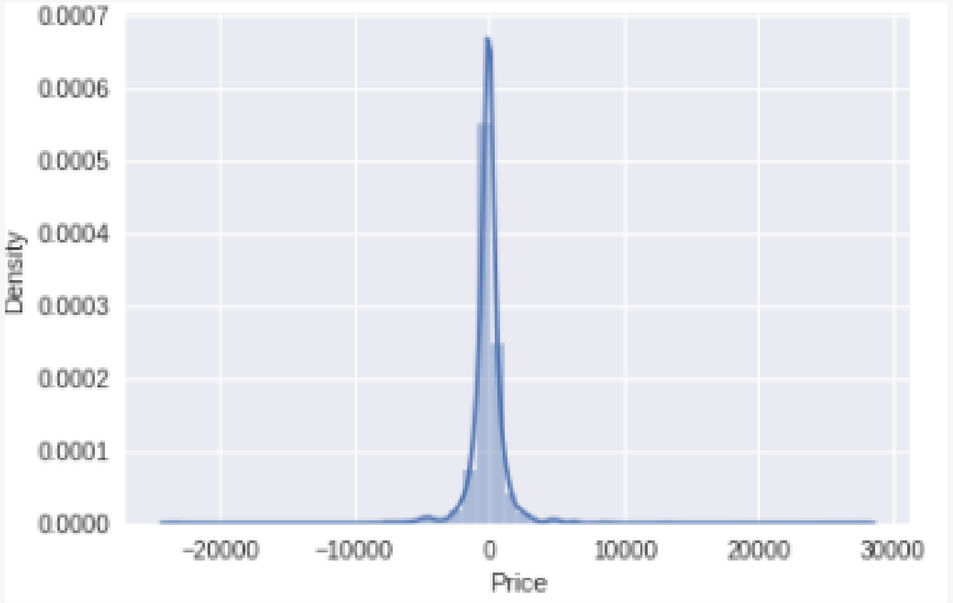
r2 score is 0.7984

MAE: 1173.6290492116577

MSE: 4000003.6495747734

RMSE: 2000.0009123934851

RANDOM FOREST



Training score: 0.9840821860775307

Test score: 0.8951472105139958

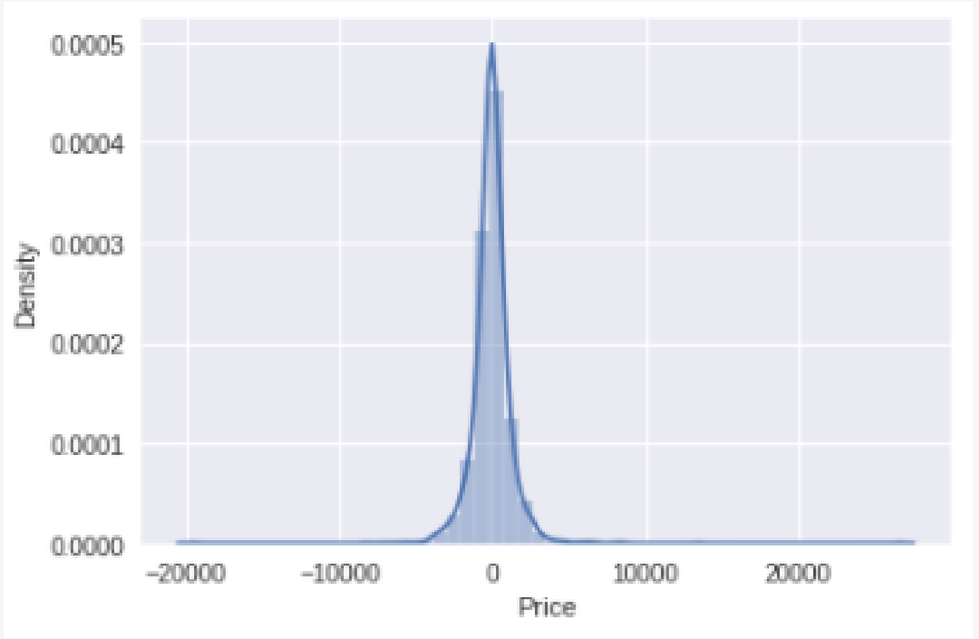
r2 score is 0.8951

MAE: 659.5246564170819

MSE: 2080640.6054871716

RMSE: 1442.4425830816183

CATBOOST REGRESSOR



Training score: 0.9458963171111978

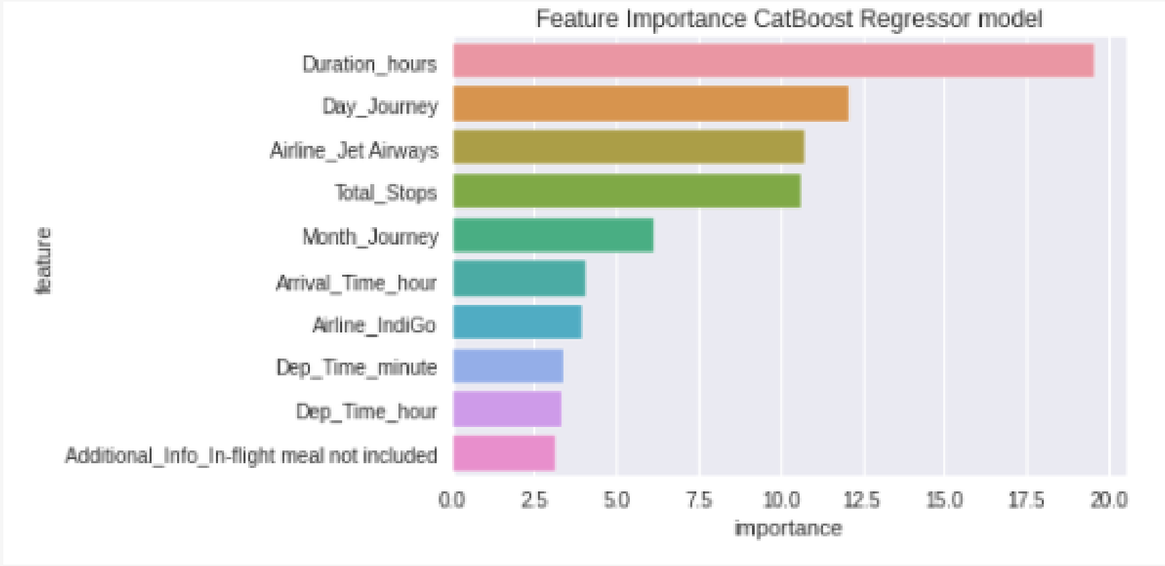
Test score: 0.8970579447769425

r2 score is 0.8971

MAE: 818.613771314273

MSE: 2042725.054424862

RMSE: 1429.2393272034121



CONCLUSION

MODEL MACHINE LEARNING

1

Based on several Regression Model, CatBoost Regressor has the highest R squared 0.89 and the smallest RMSE : 1429.23. Where this model can be used to make predictions from the Flight Fare even though it is not perfect. There are still many wasy to improve the model and evaluate the model further research

CONCLUSION

MODEL MACHINE LEARNING

2

The Decision Tree Regression, Random Forest Regression, and CatBoost Regressor models all have the highest importance feature, its the Duration hour feature, which means that the length of the trip/far flight greatly affects the price.

CONCLUSION

INSIGHT FROM DATA

3

Flight prices can also be affected by the type/type of Airlines. In the data we collect, there are economic and business/exclusive types. Business Airlines types such as Jet Airways Business have the highest price range compared to other airlines.

CONCLUSION

INSIGHT FROM DATA

4

Destinations Bangalore → New Delhi, Kolkata → Bangalore,
Delhi → Cochin is the destination with the most flights.
This can be caused by the 5 cities above are included in the big cities in the State of India, and New Delhi is the administrative center of the Indian government. Bangalore is the most densely populated and Delhi is the capital city of India.

CONCLUSION

INSIGHT FROM DATA

5

1-stop or 1-time transit has the highest price range compared to others, it can be said that total_stops does not have much effect on airline ticket prices