Predict Airline Fares

PERSONAL PROJECT

BY: THARINDU YAKKALA

Objective & Methodology

Analyze airline fare data to create a predictor for fares in Indian airlines.

Python used to load, clean, and create models.

Models used: Linear Regression, Decision Trees, KNN

RMSE used as model metric

Summary

Route and additional_info features not used

Route would result in 128 features added in, and additional_info has a lot of missing data.

When comparing Linear Regression, Decision Trees, and KNN, Decision Trees had the smallest root mean squared error.

With outliers removed, decision trees remained the best model, however KNN regression had the best improvement of in terms of RMSE.

In decision trees total_duration was the most importance feature, followed by Journey_day

Raw Data

- Raw data features
  - Airline: string
  - Date_of_Journey: datetime
  - Source
  - Destination
  - Route (removed)
  - Dep_Time/Arrival_Time
  - Duration
  - Total Stops
  - Additional Info (removed)

	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info
0	IndiGo	24/03/2019	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2h 50m	non-stop	No info
1	Air India	1/05/2019	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7h 25m	2 stops	No info
2	Jet Airways	9/06/2019	Delhi	Cochin	DEL → LKO → BOM → COK	09:25	04:25 10 Jun	19h	2 stops	No info

Data cleaning & Feature Extraction

Data Cleanup & Feature extraction: Datetime

Step1:

Set Date_of_Journey, Dep_Time,Arrival_Time to datetime to extract features

Step2:

From date_of_journey

- We get Day, Month, and Year

From arrival time and departure time

- We get hour and minute

Journey_Day	Journey_Month	Journey_Year	Dep_Hour	Dep_Min	Arrival_Hour	Arrival_Min
24	3	2019	22	20	1	10
1	5	2019	5	50	13	15

Data Cleanup & Feature extraction: Duration of flight

METHOD

- Current format is 2h 30m or 30m as a string.
- Using regex we can replace h with *60, and then do 2*60+30 and extract that into duration in minutes.

```
df['Duration']

0        2h 50m
1        7h 25m
2           19h
3        5h 25m
4        4h 45m
          ...
10678    2h 30m
10679    2h 35m
10680        3h
10681    2h 40m
10682    8h 20m
```

AFTER EXTRACTION

```
df['Total_Duration_Min']

]: 0         170
   1         445
   2        1140
   3         325
   4         285
            ...
   10678     150
   10679     155
   10680     180
   10681     160
   10682     500
```

Data Cleanup & Feature extraction: total stops

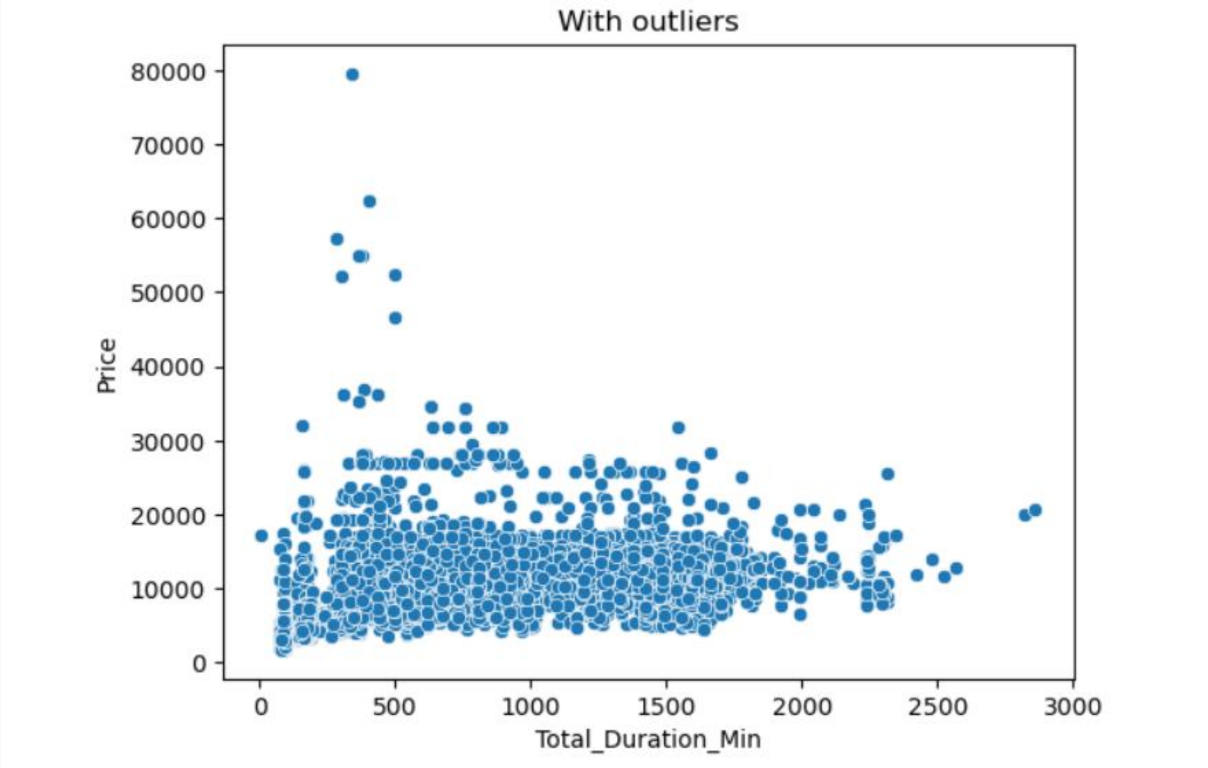
```
df['Total_Stops']
```

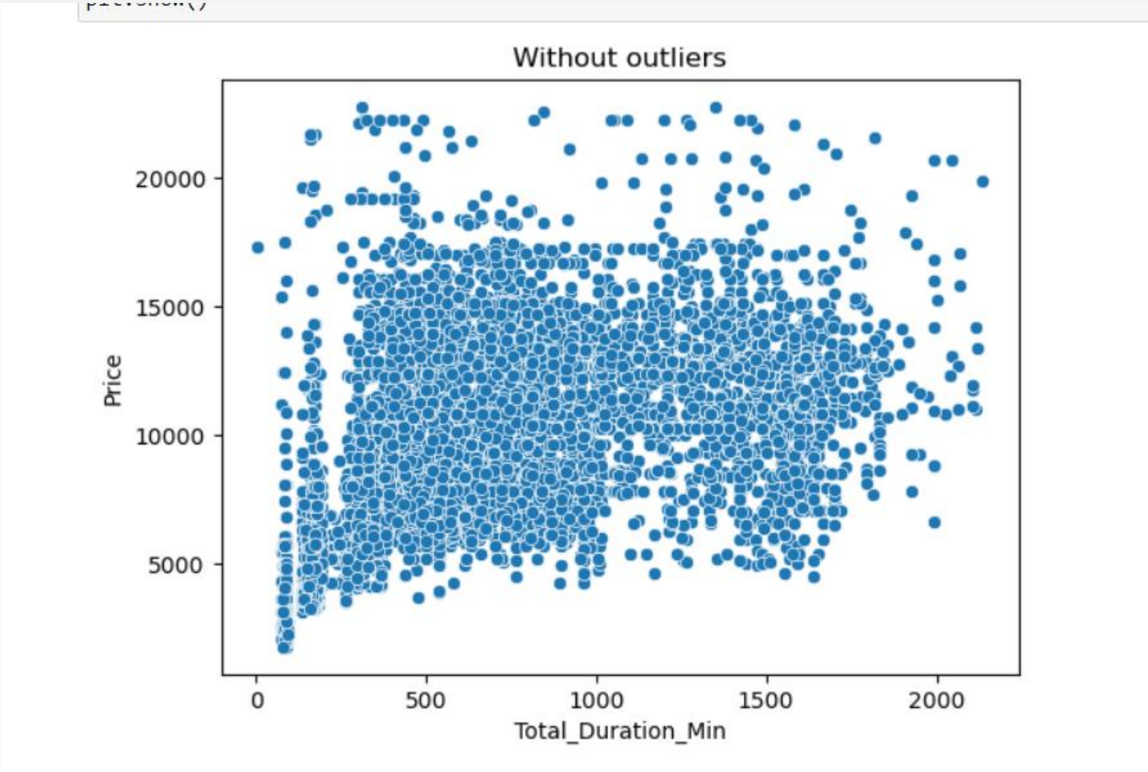
```
]: 0          non-stop
   1           2 stops
   2           2 stops
   3            1 stop
   4            1 stop
                  ...
   10678      non-stop
   10679      non-stop
   10680      non-stop
   10681      non-stop
   10682       2 stops
```

```
df['Total_Stops']
```

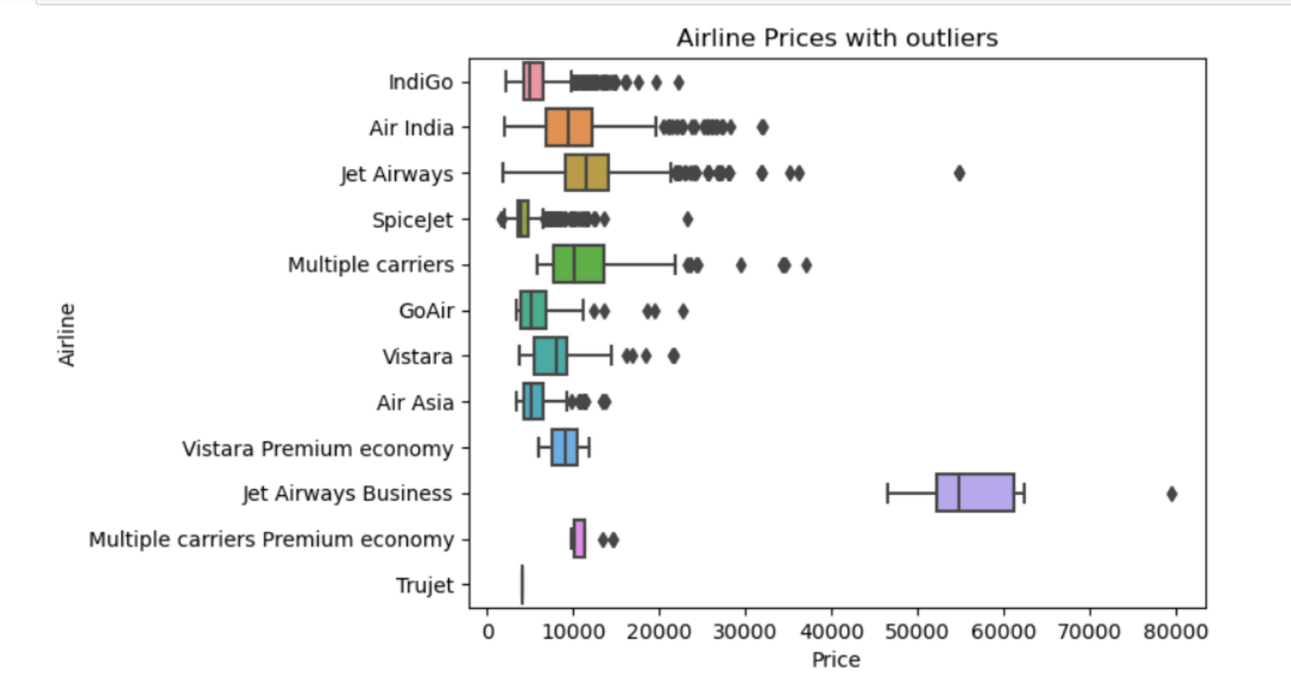
```
: 0        0
  1        2
  2        2
  3        1
  4        1
          ..
  10678    0
  10679    0
  10680    0
  10681    0
  10682    2
```

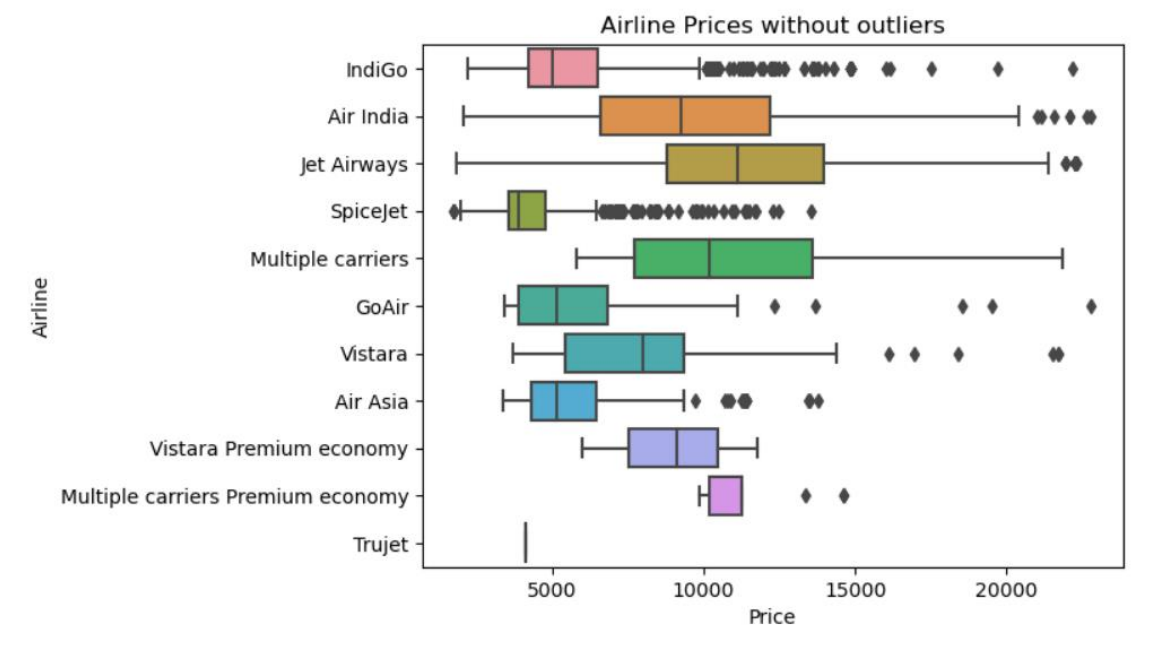
Price vs. Duration





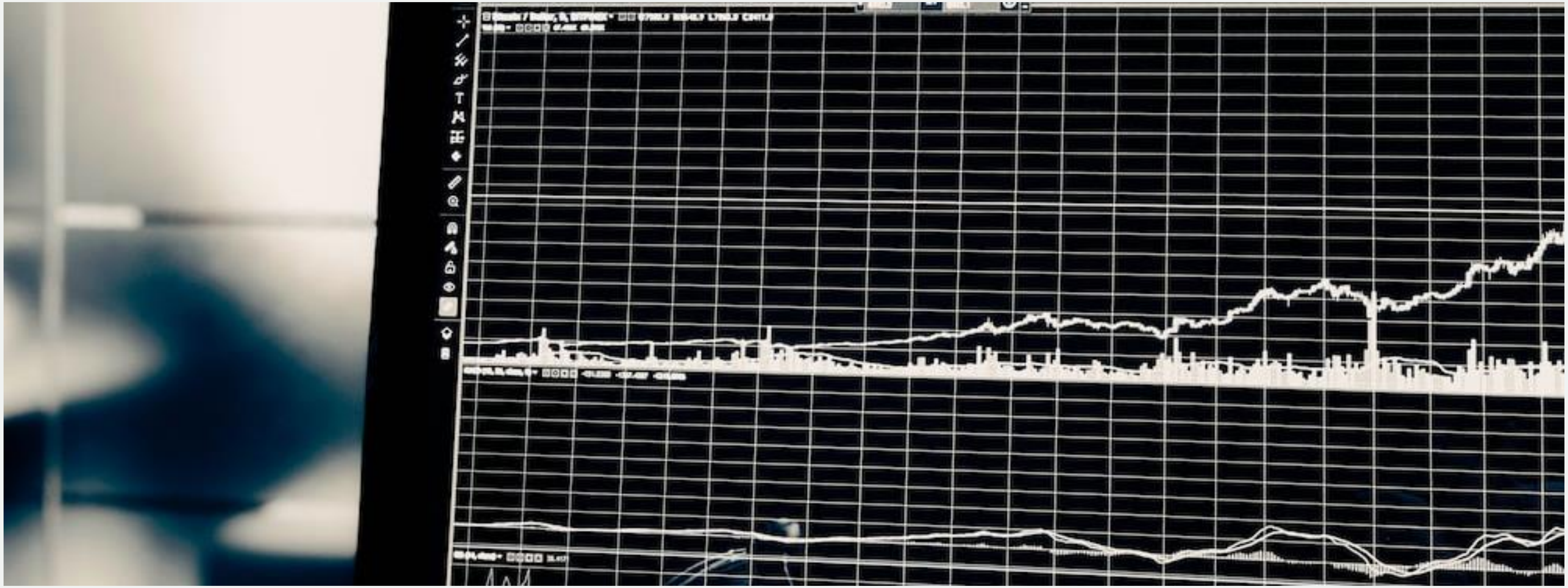
There is a stronger linear relationship when outliers are removed





Most airline prices are similar but Jet Airways Business is a big outlier

Models



Model RMSE comparison with outliers

Linear Regression: 2804.64

Decision Tree: 2539.68

KNN: 3019.20

Based on results, decision tree model is best used for airline fare preditions when comparing the models used.

Future considerations:

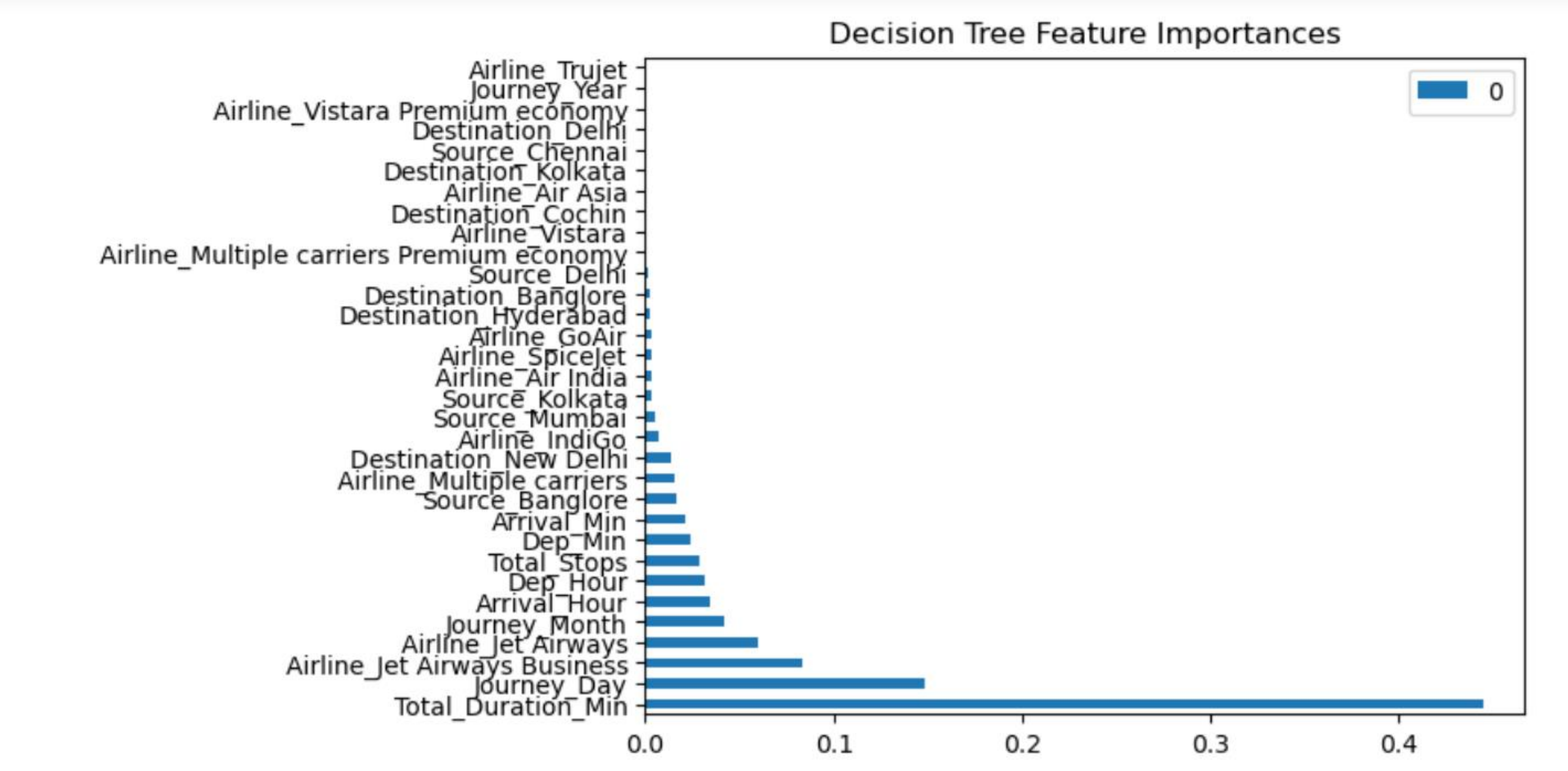
- Trying deep learning models

RMSE with outliers removed

Linear Regression: 2804.64 -> 2413.24 **13.96% improvement**

Regression Tree: 2539.68 -> 2210.22 **12.97% improvement**

KNN Regressor: 3019.20 -> 2404.67 **20.35% improvement**


Decision Tree Feature Importances
0
Airline_Trujet
Journey_Year
Airline_Vistara Premium economy
Destination_Delhi
Source_Chennai
Destination_Kolkata
Airline_Air Asia
Destination_Cochin
Airline_Vistara
Airline_Multiple carriers Premium economy
Source_Delhi
Destination_Banglore
Destination_Hyderabad
Airline_GoAir
Airline_SpiceJet
Airline_Air India
Source_Kolkata
Source_Mumbai
Airline_IndiGo
Destination_New Delhi
Airline_Multiple carriers
Source_Banglore
Arrival_Min
Dep_Min
Total_Stops
Dep_Hour
Arrival_Hour
Journey_Month
Airline_Jet Airways
Airline_Jet Airways Business
Journey_Day
Total_Duration_Min
0.0
0.1
0.2
0.3
0.4