Flight Data Analysis:

Flight Price Forecasting: Using Linear Regression and Regression Tree Models

Introduction

Throughout 2022, air travel encompassed around 7 billion passengers, with a significant volume of ticket sales occurring online. The complexity of selecting the optimal time to buy tickets— among a wide selection of airlines, departure times, and flight lengths—can be overwhelming for consumers.

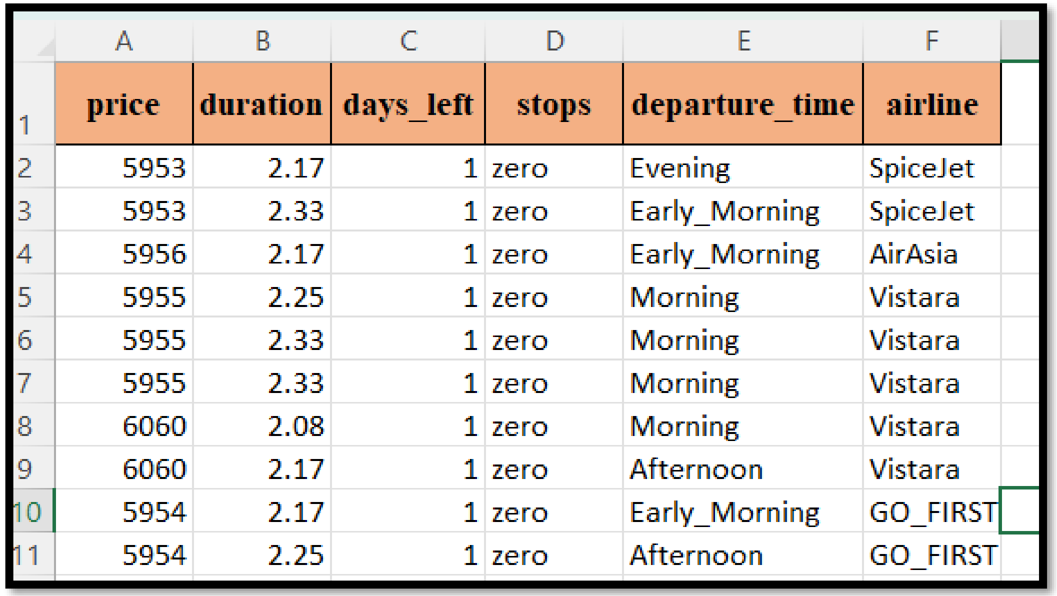
•Objective 1: **Identify the most significant predictors of flight pricing.**

•Objective 2: **Develop a predictive models that can be used to estimate flight prices based on some key predictor variables.**

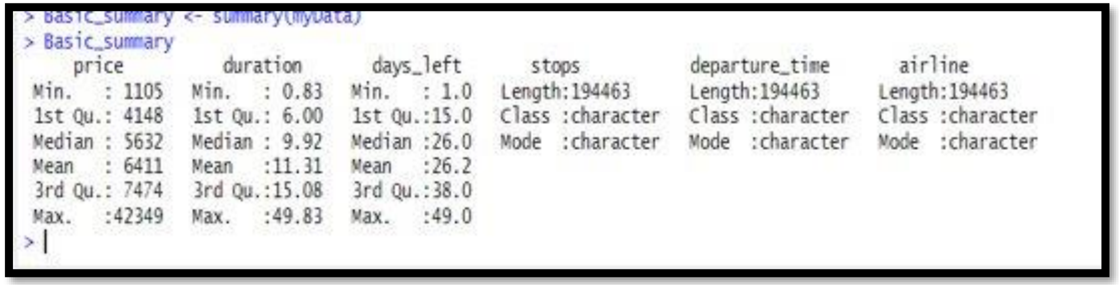
•Objective 3: **Provide actionable insights to consumers on flight pricing strategies to improve their experiences through better planning, cost savings, and enhanced travel decision-making.**

Data Overview

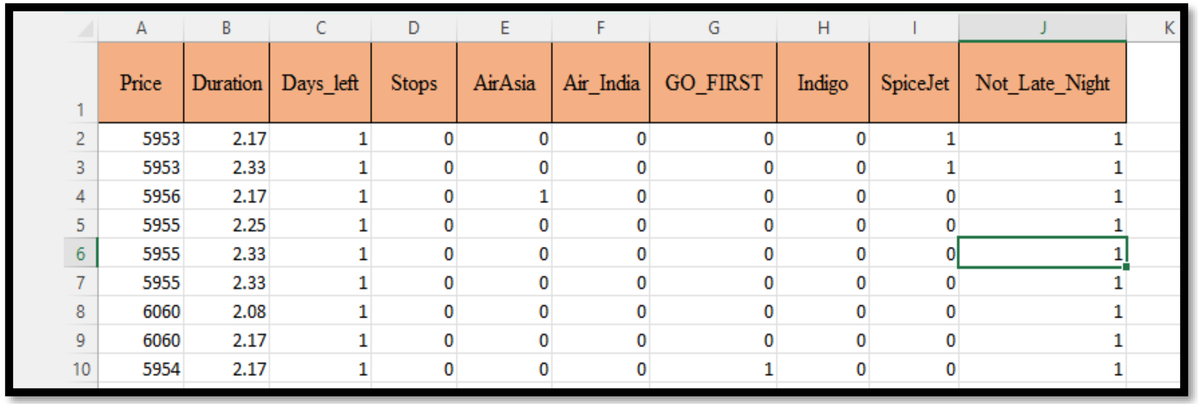
Data preparation and summary measures



	A	B	C	D	E	F
1	price	duration	days_left	stops	departure_time	airline
2	5953	2.17	1	zero	Evening	SpiceJet
3	5953	2.33	1	zero	Early_Morning	SpiceJet
4	5956	2.17	1	zero	Early_Morning	AirAsia
5	5955	2.25	1	zero	Morning	Vistara
6	5955	2.33	1	zero	Morning	Vistara
7	5955	2.33	1	zero	Morning	Vistara
8	6060	2.08	1	zero	Morning	Vistara
9	6060	2.17	1	zero	Afternoon	Vistara
10	5954	2.17	1	zero	Early_Morning	GO_FIRST
11	5954	2.25	1	zero	Afternoon	GO_FIRST



```
> Basic_summary <- summary(myData)
> Basic_summary
     price          duration        days_left         stops           departure_time        airline         
 Min.   : 1105   Min.   : 0.83   Min.   : 1.0   Length:194463      Length:194463      Length:194463     
 1st Qu.: 4148   1st Qu.: 6.00   1st Qu.:15.0   Class :character   Class :character   Class :character  
 Median : 5632   Median : 9.92   Median :26.0   Mode  :character   Mode  :character   Mode  :character  
 Mean   : 6411   Mean   :11.31   Mean   :26.2                                                           
 3rd Qu.: 7474   3rd Qu.:15.08   3rd Qu.:38.0                                                           
 Max.   :42349   Max.   :49.83   Max.   :49.0                                                           
> |
```



	A	B	C	D	E	F	G	H	I	J
1	Price	Duration	Days_left	Stops	AirAsia	Air_India	GO_FIRST	Indigo	SpiceJet	Not_Late_Night
2	5953	2.17	1	0	0	0	0	0	1	1
3	5953	2.33	1	0	0	0	0	0	1	1
4	5956	2.17	1	0	1	0	0	0	0	1
5	5955	2.25	1	0	0	0	0	0	0	1
6	5955	2.33	1	0	0	0	0	0	0	1
7	5955	2.33	1	0	0	0	0	0	0	1
8	6060	2.08	1	0	0	0	0	0	0	1
9	6060	2.17	1	0	0	0	0	0	0	1
10	5954	2.17	1	0	0	0	1	0	0	1

Prepared dataset, with correct variable type and dummy variables created.

Summary:

Average Price: INR 6411

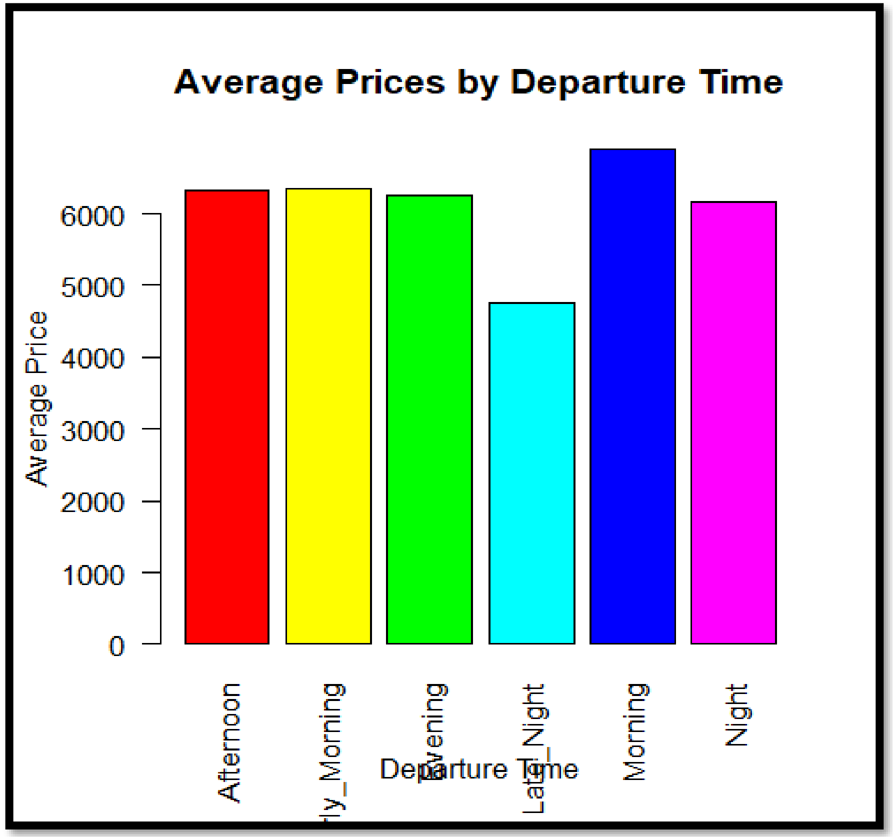
Average Duration: 11.3 Hours

Average Days left: 26

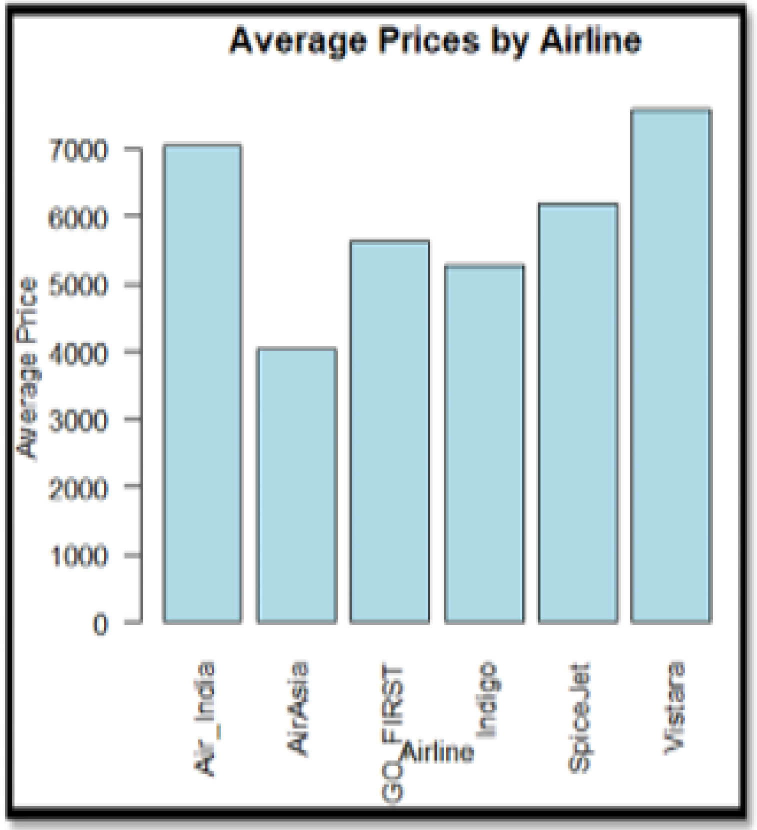
Stops : 85.65% of flights have 1 or more

99.41% are not late night flights

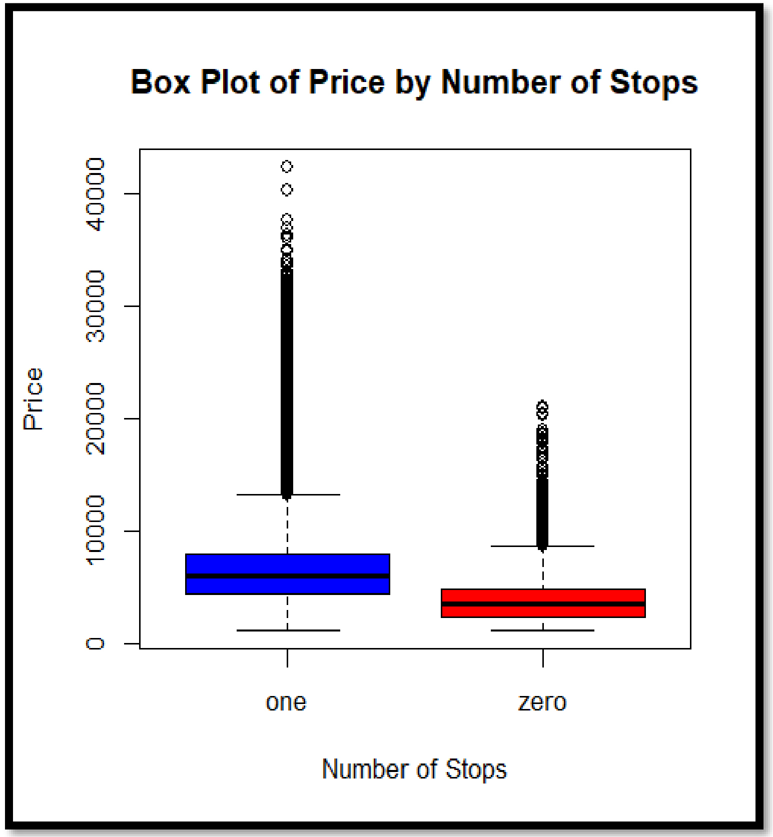
Data Exploration



By Departure Time, Late_Night have the lowest average price.

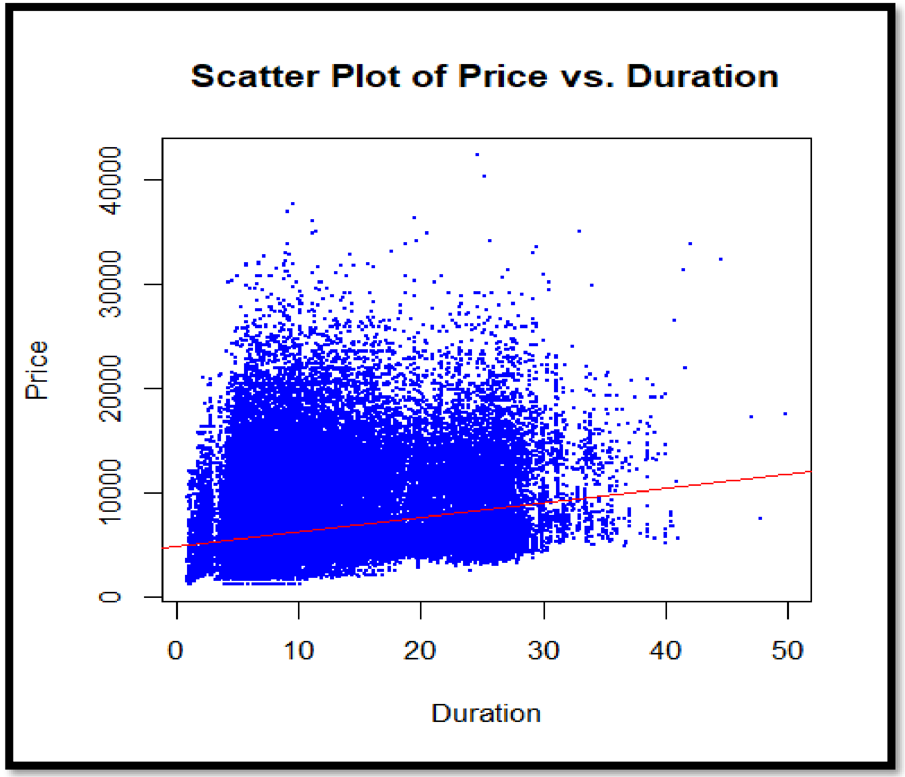


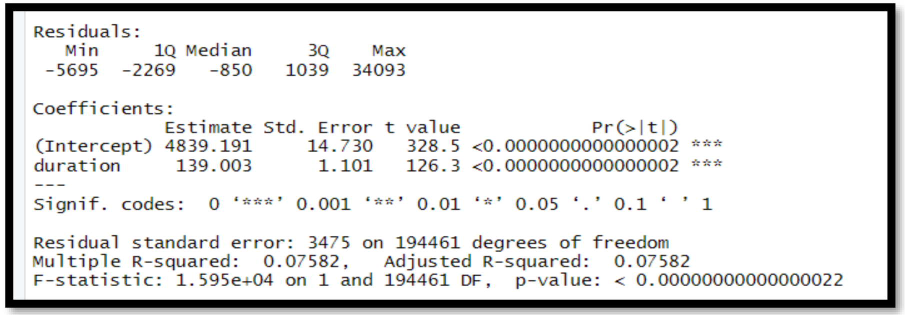
By Airlines, AirAsia have the lowest average price.



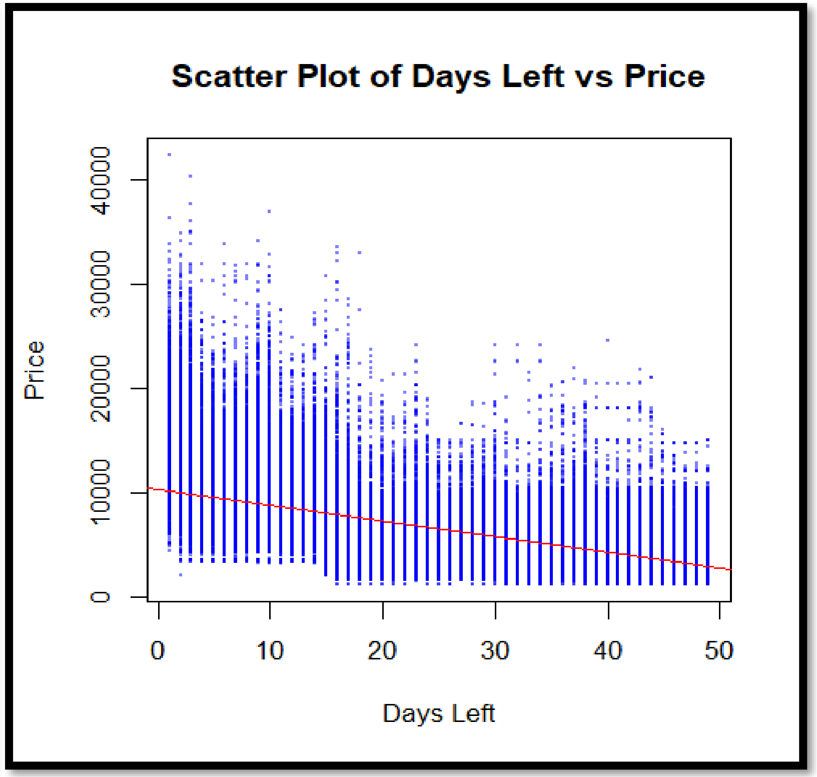
Flights with no stops are expected to be cheaper.

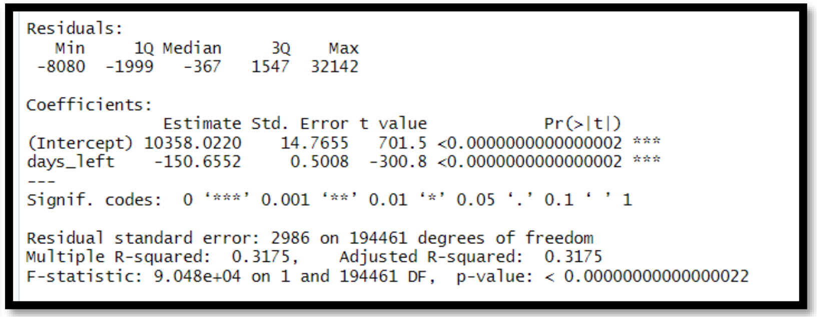
Data Exploration





```
Residuals:
   Min     1Q Median     3Q    Max 
 -5695  -2269   -850   1039  34093 

Coefficients:
            Estimate Std. Error t value            Pr(>|t|)    
(Intercept) 4839.191     14.730   328.5 <0.0000000000000002 ***
duration     139.003      1.101   126.3 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3475 on 194461 degrees of freedom
Multiple R-squared:  0.07582,	Adjusted R-squared:  0.07582 
F-statistic: 1.595e+04 on 1 and 194461 DF,  p-value: < 0.00000000000000022
```





```
Residuals:
   Min     1Q Median     3Q    Max 
 -8080  -1999   -367   1547  32142 

Coefficients:
              Estimate Std. Error t value            Pr(>|t|)    
(Intercept) 10358.0220    14.7655   701.5 <0.0000000000000002 ***
days_left    -150.6552     0.5008  -300.8 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2986 on 194461 degrees of freedom
Multiple R-squared:  0.3175,	Adjusted R-squared:  0.3175 
F-statistic: 9.048e+04 on 1 and 194461 DF,  p-value: < 0.00000000000000022
```

Days_Left Vs Price

R^2 = 0.3175

Adjusted R^2 = 0.3175

Standard Error = 2,986

Duration Vs Price

R^2 = 0.07582

Adjusted R^2 = 0.07582

Standard Error = 3,475

Model 1

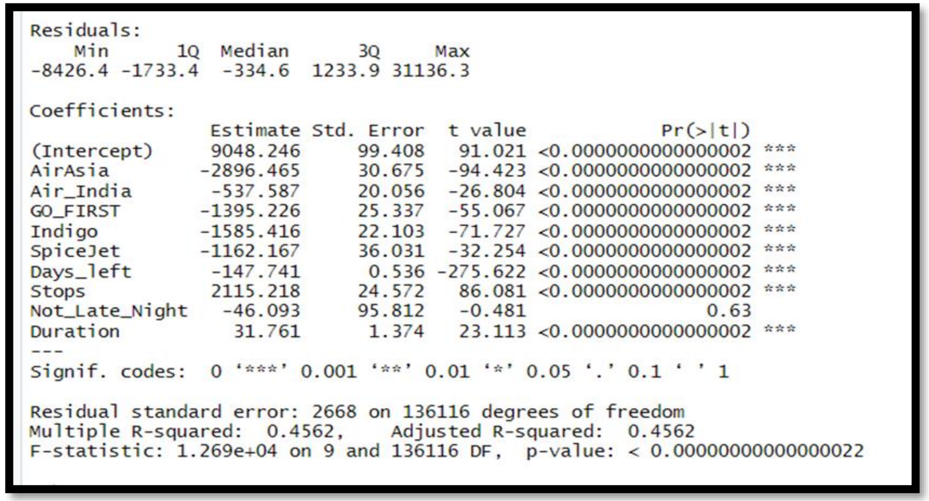
Linear Model (Random Sampling)

```
model_a <- lm(Price ~ AirAsia + Air_India + GO_FIRST + Indigo + SpiceJet +
Days_left, data  = trainSet)

model_b <- lm(Price ~ AirAsia + Air_India + GO_FIRST + Indigo + SpiceJet +
Days_left + Stops, data = trainSet)

model_c <- lm(Price ~ AirAsia + Air_India + GO_FIRST + Indigo + SpiceJet +
Days_left + Stops + Not_Late_Night, data = trainSet)

model_d <- lm(Price ~ AirAsia + Air_India + GO_FIRST + Indigo + SpiceJet +
Days_left + Stops + Not_Late_Night + Duration, data = trainSet)
```

Price *= 9,103.505 - 2892.594 ** ***AirAsia*** *- 537.368 ** ***Air_India*** *- 1386.746* **** GO_FIRST*** *- 1564.186 ** ***Indigo*** *- 1168.487 ** ***SpiceJet*** *- 154.186* ****Days_left*** *+ 2119.802 ** ***Stops*** *+ 31.203 ** ***Duration***

Partition Method	Model_a	Model_b	Model_c	Model_d
Standard Error	2,794.0	2,671.0	2,671.0	2,666.0
Coef R2	0.4031	0.4544	0.4544	0.4564
Adjusted R2	0.4031	0.4543	0.4543	0.4564

Best Results	ME	RMSE	MAE	MPE	MAPE
Model_d	0.096	2663.149	1925.71	-12.993	35.337

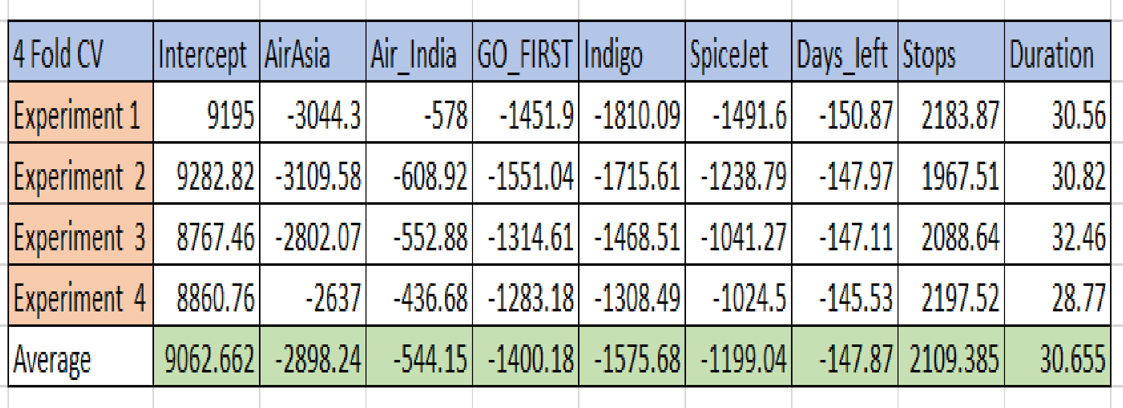


```
Residuals:
    Min      1Q  Median      3Q     Max 
-8426.4 -1733.4  -334.6  1233.9 31136.3 

Coefficients:
                Estimate Std. Error  t value             Pr(>|t|)    
(Intercept)     9048.246     99.408   91.021 <0.0000000000000002 ***
AirAsia        -2896.465     30.675  -94.423 <0.0000000000000002 ***
Air_India       -537.587     20.056  -26.804 <0.0000000000000002 ***
GO_FIRST       -1395.226     25.337  -55.067 <0.0000000000000002 ***
Indigo         -1585.416     22.103  -71.727 <0.0000000000000002 ***
SpiceJet       -1162.167     36.031  -32.254 <0.0000000000000002 ***
Days_left       -147.741      0.536 -275.622 <0.0000000000000002 ***
Stops           2115.218     24.572   86.081 <0.0000000000000002 ***
Not_Late_Night   -46.093     95.812   -0.481                 0.63    
Duration          31.761      1.374   23.113 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2668 on 136116 degrees of freedom
Multiple R-squared:  0.4562,	Adjusted R-squared:  0.4562 
F-statistic: 1.269e+04 on 9 and 136116 DF,  p-value: < 0.00000000000000022
```

Model 2

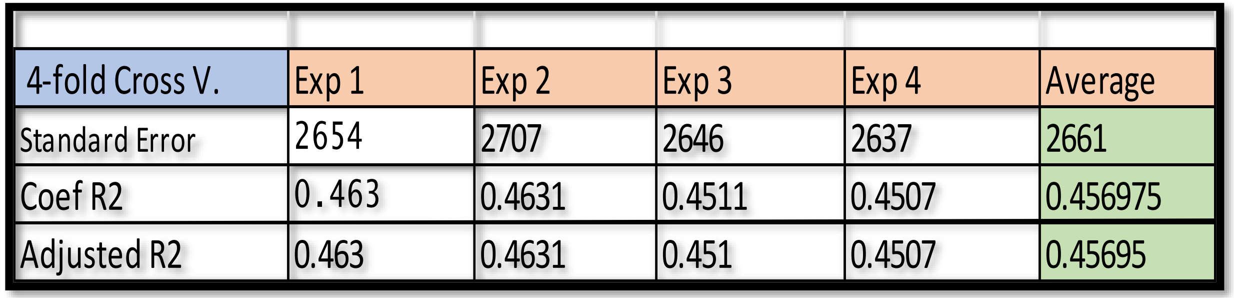
Using 4-fold Cross Validation

R Output

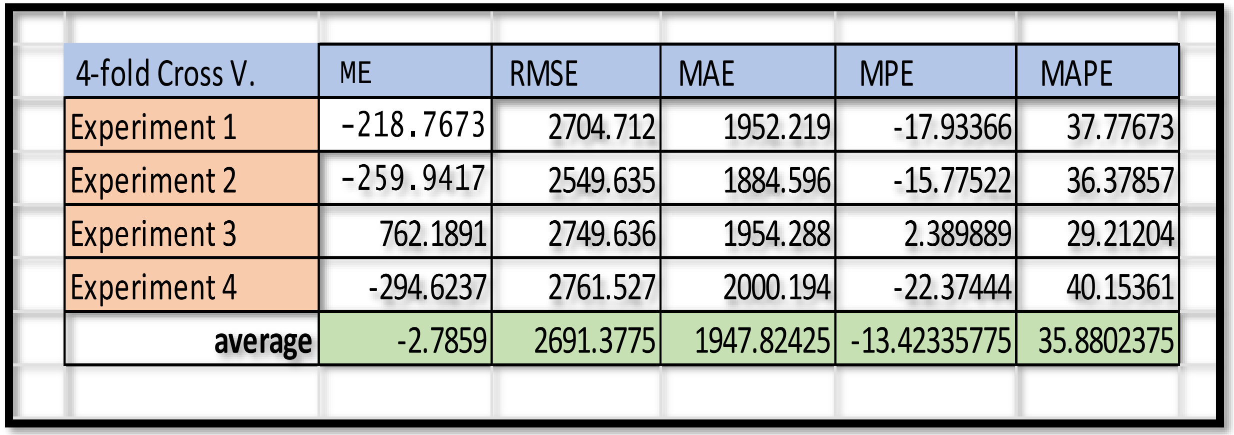


4 Fold CV	Intercept	AirAsia	Air_India	GO_FIRST	Indigo	SpiceJet	Days_left	Stops	Duration
Experiment 1	9195	-3044.3	-578	-1451.9	-1810.09	-1491.6	-150.87	2183.87	30.56
Experiment 2	9282.82	-3109.58	-608.92	-1551.04	-1715.61	-1238.79	-147.97	1967.51	30.82
Experiment 3	8767.46	-2802.07	-552.88	-1314.61	-1468.51	-1041.27	-147.11	2088.64	32.46
Experiment 4	8860.76	-2637	-436.68	-1283.18	-1308.49	-1024.5	-145.53	2197.52	28.77
Average	9062.662	-2898.24	-544.15	-1400.18	-1575.68	-1199.04	-147.87	2109.385	30.655

Price = *9026.662 - 2,898.235 * **AirAsia** - 544.145 * **Air_India** - 1400.182 * **GO_FIRST** - 1575.675 * **Indigo** - 1199.042 * **SpiceJet** - 147.87 * **Days_left** + 2109.385 * **Stops** + 30.655 * **Duration***



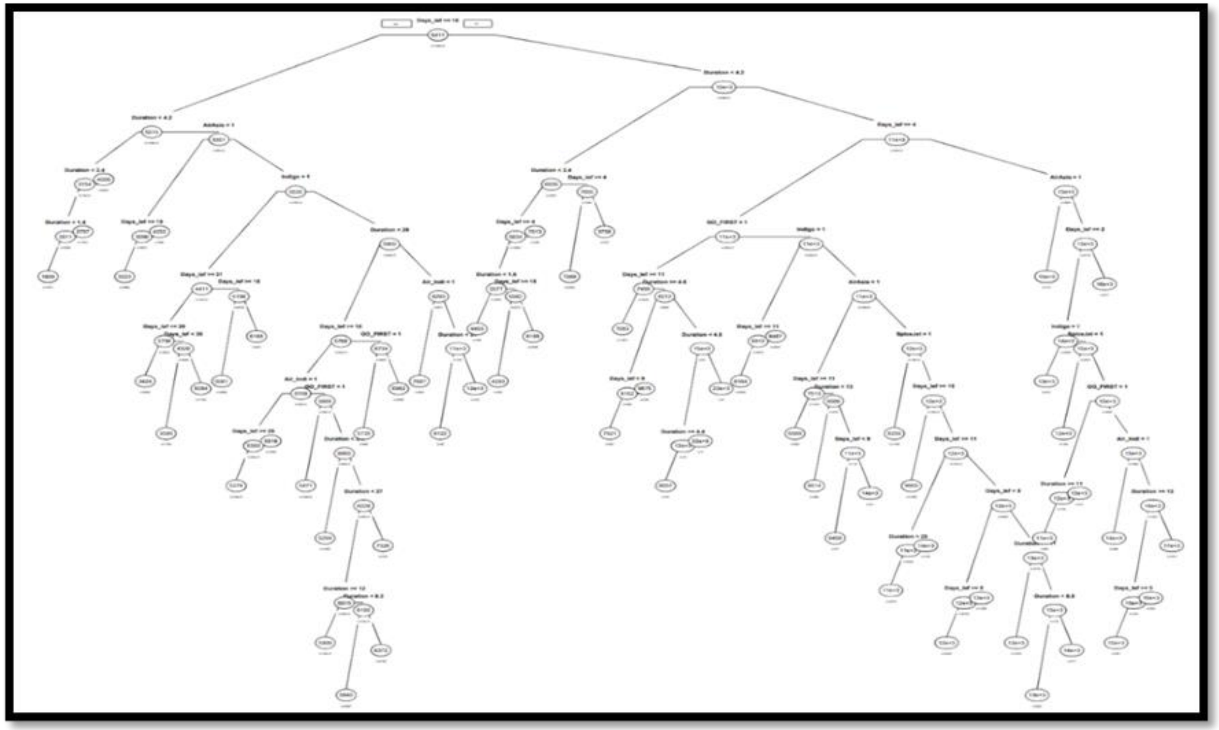
4-fold Cross V.	Exp 1	Exp 2	Exp 3	Exp 4	Average
Standard Error	2654	2707	2646	2637	2661
Coef R2	0.463	0.4631	0.4511	0.4507	0.456975
Adjusted R2	0.463	0.4631	0.451	0.4507	0.45695



4-fold Cross V.	ME	RMSE	MAE	MPE	MAPE
Experiment 1	-218.7673	2704.712	1952.219	-17.93366	37.77673
Experiment 2	-259.9417	2549.635	1884.596	-15.77522	36.37857
Experiment 3	762.1891	2749.636	1954.288	2.389889	29.21204
Experiment 4	-294.6237	2761.527	2000.194	-22.37444	40.15361
average	-2.7859	2691.3775	1947.82425	-13.42335775	35.8802375

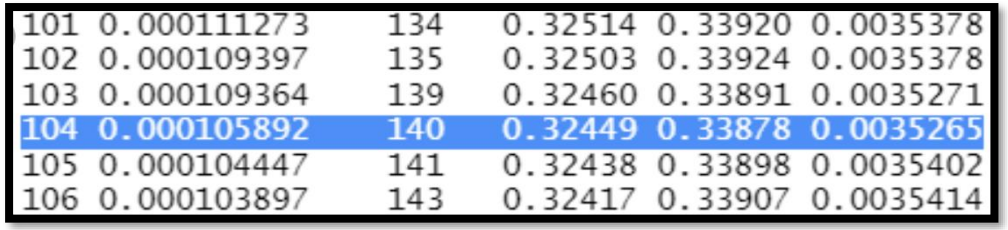
Model 3

Regression Tree



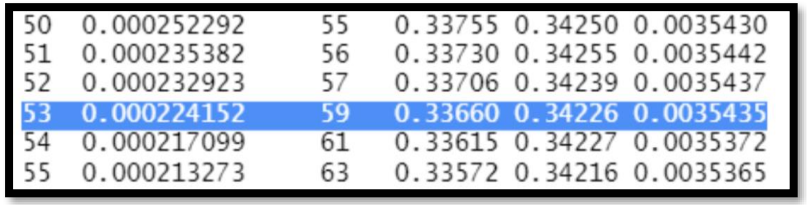
First split is based on Days_left
Second split is based on Duration

Lowest Error



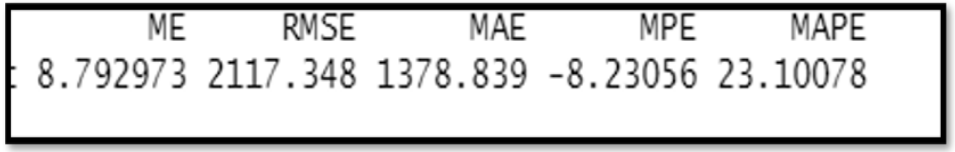
```
101 0.000111273      134    0.32514 0.33920 0.0035378
102 0.000109397      135    0.32503 0.33924 0.0035378
103 0.000109364      139    0.32460 0.33891 0.0035271
104 0.000105892      140    0.32449 0.33878 0.0035265
105 0.000104447      141    0.32438 0.33898 0.0035402
106 0.000103897      143    0.32417 0.33907 0.0035414
```

CP for best pruned tree



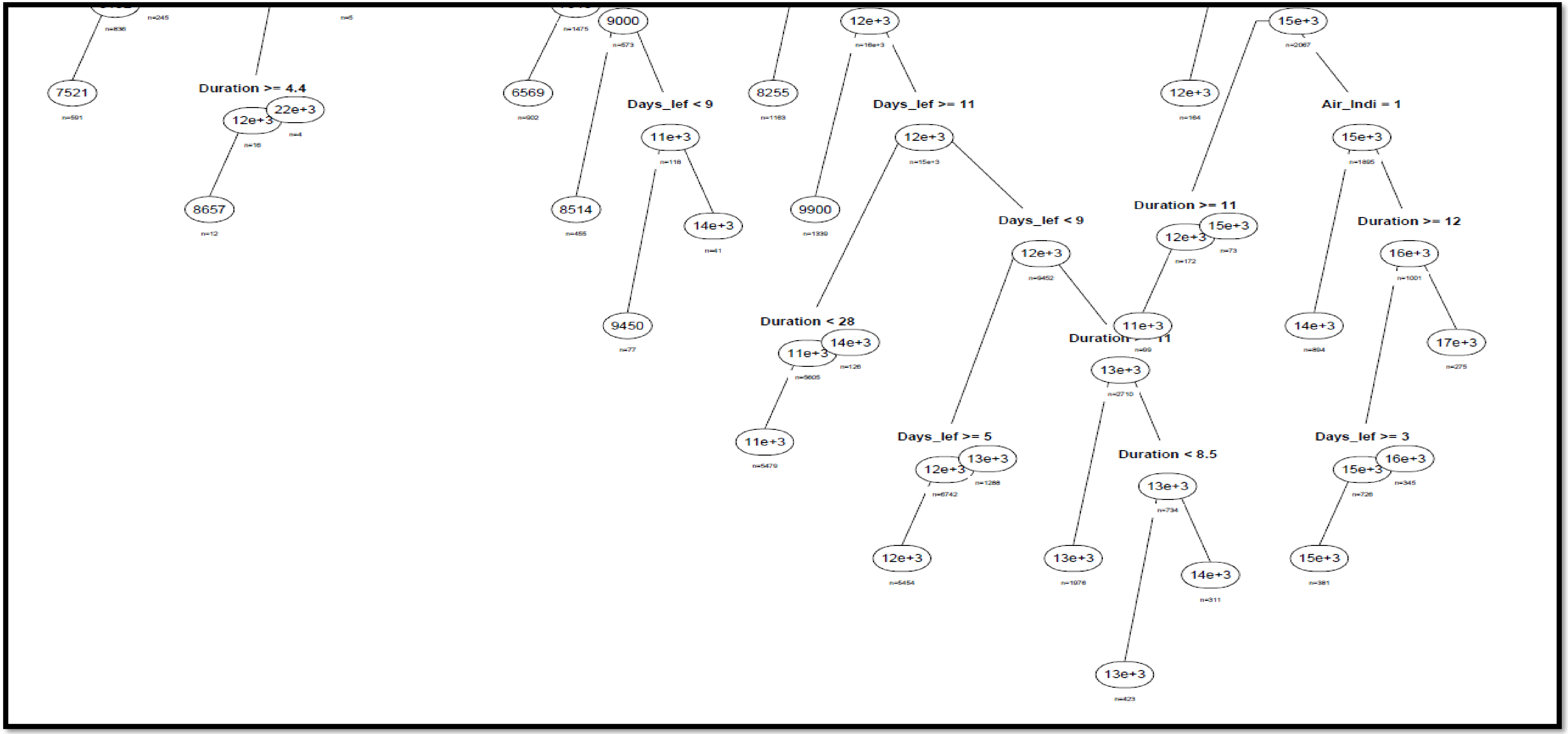
```
50  0.000252292      55    0.33755 0.34250 0.0035430
51  0.000235382      56    0.33730 0.34255 0.0035442
52  0.000232923      57    0.33706 0.34239 0.0035437
53  0.000224152      59    0.33660 0.34226 0.0035435
54  0.000217099      61    0.33615 0.34227 0.0035372
55  0.000213273      63    0.33572 0.34216 0.0035365
```

Results



```
      ME     RMSE      MAE      MPE     MAPE
 8.792973 2117.348 1378.839 -8.23056 23.10078
```

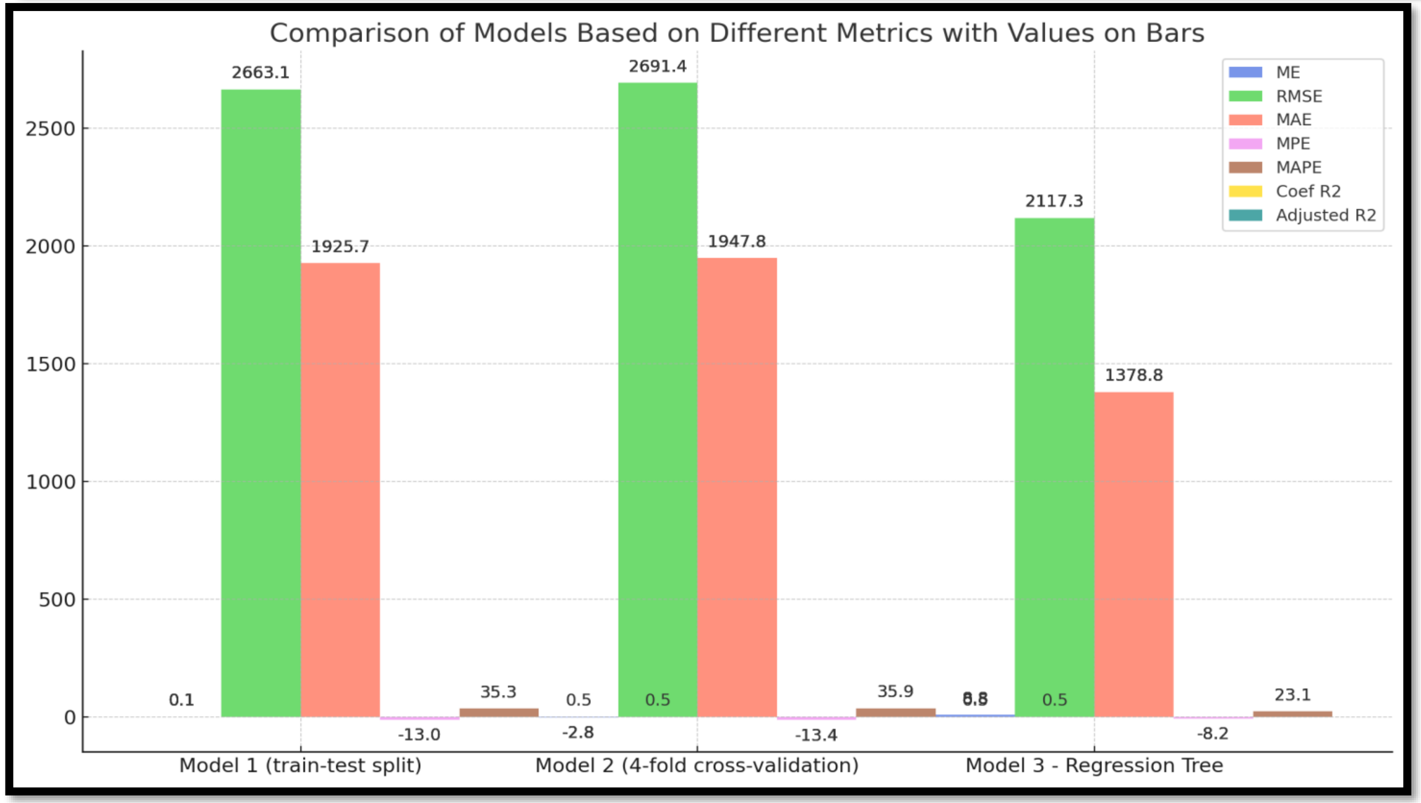
Model 3



n=245
n=836
n=5
n=1475
9000
n=573
12e+3
n=16e+3
15e+3
n=2067
7521
n=591
Duration >= 4.4
22e+3
12e+3
n=4
n=16
6569
n=902
Days_lef < 9
8255
n=1163
Days_lef >= 11
12e+3
n=164
Air_Indi = 1
11e+3
n=118
12e+3
n=15e+3
15e+3
n=1895
8657
n=12
8514
n=455
14e+3
n=41
9900
n=1339
Duration >= 11
15e+3
12e+3
n=73
n=172
Days_lef < 9
Duration >= 12
12e+3
n=9452
16e+3
n=1001
9450
n=77
Duration < 28
14e+3
11e+3
n=126
n=5605
11e+3
Duration
11
n=99
14e+3
n=894
17e+3
n=275
13e+3
n=2710
11e+3
n=5479
Days_lef >= 5
13e+3
12e+3
n=1288
n=6742
Duration < 8.5
13e+3
n=734
Days_lef >= 3
16e+3
15e+3
n=345
n=726
12e+3
n=5454
13e+3
n=1976
14e+3
n=311
15e+3
n=381
13e+3
n=423

Model Evaluation

Models Comparaison	ME	RMSE	MAE	MPE	MAPE	Standard Error	Coef R2	Adjusted R2
Model 1 (train-test split)	0.096	2663.149	1925.710	-12.993	35.337	2666.000	0.456	0.4564
Model 2 (4-fold cross Validation	-2.786	2691.378	1947.824	-13.423	35.880	2661.000	0.457	0.4570
Model 3 - Regression Tree	8.792973	2117.348	1378.839	-8.231	23.101	N//A	N/A	N/A



Based on the performance metrics:

- ***Model 1** would be preferable to Model 2.*
- ***Model 3** has a higher ME but it did better than other models on all measuring metrics*

Evaluations and Recommendation

- **Model 1:** Exhibits high accuracy with a Mean Error (ME) of 0.096, indicating precise predictions. It outperforms Model 2 in overall accuracy despite similar performance metrics.
- **Model 2**: Tends to overpredict with a negative ME of -2.786, suggesting systematic bias in its estimates. Despite a marginally better Adjusted R^2 of 0.457, its larger error metrics imply weaker predictive performance and potential outlier sensitivity.
- **Model 3:** Shows a consistent underprediction bias with a positive ME of 8.793. Nevertheless, it surpasses the other models with lower error metrics across the board, indicating it generally provides the most reliable forecasts

Recommendation 1: Scheduled with days left higher than 16

Recommendation 2: Departure time does not have a significant effect on the prices.

Recommendation 3: AirAsia is the cheapest Airline with Vistara being the more expensive.

Recommendation 4: Flights with no stops tend to be significantly cheaper than flights with stops.