
Tools
Wego
Steps
Questions
Predicting Airline Ticket Prices
Using Linear Regression Models

wego

Wego is the largest online travel marketplace in the Middle East, North Africa & Asia Pacific.

Wego travel app currently ranks within the top five of the most popular apps in the iOS and Play store travel category for Middle Eastern region.

Wego continuously collaborates with tourism boards of various countries, such as Singapore, South Korea, Jordan, Macao, and many others

Wego is used by millions of people every month.

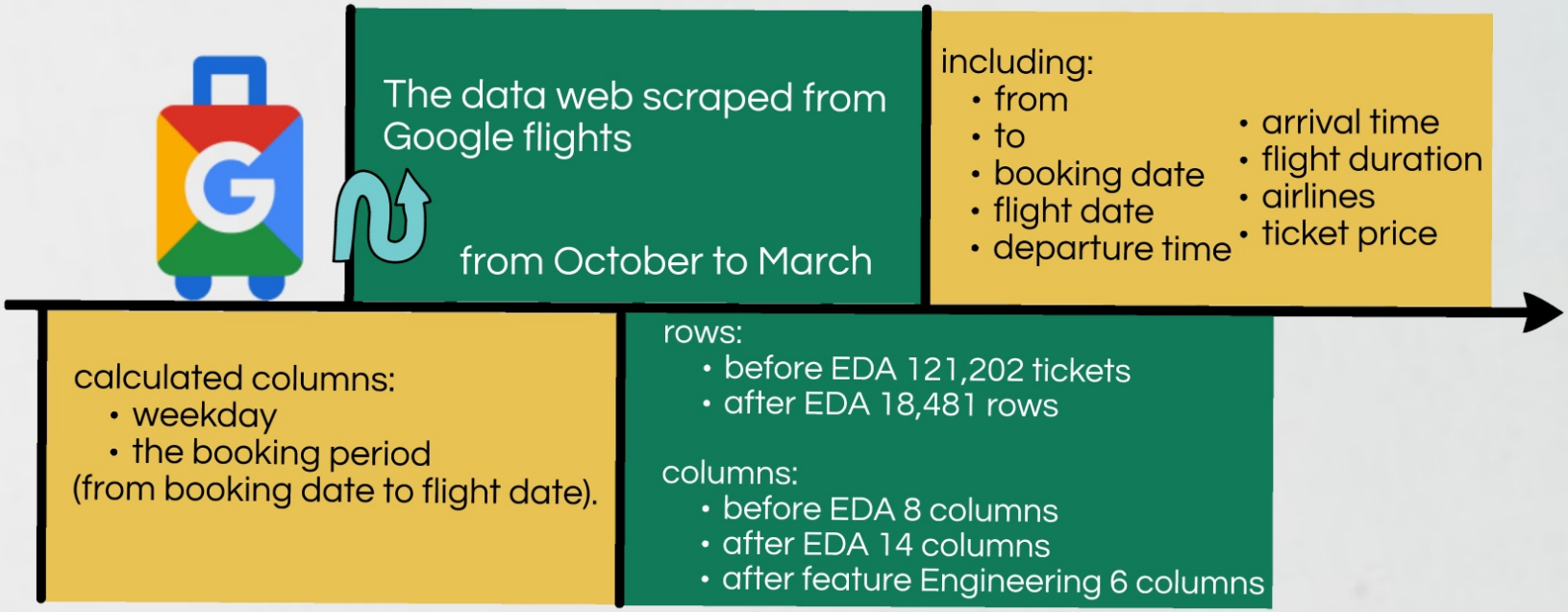
The Need & DATA

Need

The purpose of this project is to help our client provide the best tickets for their customers during Riyadh Season 2021

which is achieved by using linear regression models to predict Airline tickets to Riyadh during the Riyadh entertainment season 2021.

DATA


The data web scraped from Google flights
from October to March
including:
• from
• to
• booking date
• flight date
• departure time
• arrival time
• flight duration
• airlines
• ticket price
calculated columns:
• weekday
• the booking period
(from booking date to flight date).
rows:
• before EDA 121,202 tickets
• after EDA 18,481 rows
columns:
• before EDA 8 columns
• after EDA 14 columns
• after feature Engineering 6 columns

wego

Wego is the largest online travel marketplace in the Middle East, North Africa & Asia Pacific.

Wego travel app currently ranks within the top five of the most popular apps in the iOS and Play store travel category for Middle Eastern region.

Wego continuously collaborates with tourism boards of various countries, such as Singapore, South Korea, Jordan, Macao, and many others

Wego is used by millions of people every month.

The Need & DATA


Tools
Wego
Steps
Questions
Predicting Airline Ticket Prices
Using Linear Regression Models

Steps

1

Web scrapped the data from google flights of Saudi local cities:

- Jeddah
- Dammam
- Jazan
- Medina

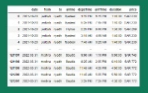
to Riyadh during Riyadh season period which is from October 2021 to March 2022

2

Load web scrapped data into a CSV file called tickets

3

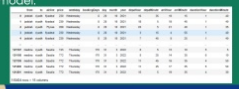
Downloaded the tickets data csv file into Jupiter notebook then Loaded the file into a pandas Data Frame.



pandas

4

Explored and got to know the data while understanding how to use it and which attributes should be considered as features and prepare to be used in the prediction model.

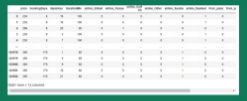


5

Cleaned the data and filtered it removing outliers and duplicates

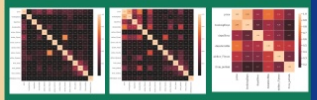
6

Converting categorical features to binary dummy variables during EDA



7

Checking for correlation and normalizing the data taking the highly correlated features with the target and removing some features that are highly correlated with each other



8

Splitting the data into sets: train 60%, validation20%, and test 20% in chronologically sorted order

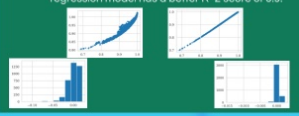
9

Fitting the Linear regression 6 models on train data using different features and comparing the score of each. Ending up with the best score of R^2: 0.9114945339772216 and Adjusted R Square: 0.9114705812123981, the intercept: 1.0312317965552298 and the following features: 'bookingDays', 'depaHour', 'durationMin', 'airline_Flynas', 'from_jeddah' with their coefficients as:

- bookingDays : -0.25
- depaHour : -0.11
- durationMin : -0.16
- airline_Flynas : -2.11
- from_jeddah : -0.66

10

The results ended up in a kind of a curve shape; so, a polynomial regression model was compared with the linear regression model on the validation set, where the polynomial regression model has a better R^2 score of 0.9.



11

Test on the test set to get a final evaluation of expected generalization performance.

```
Polynomial Score:  0.9997476598895741
Mean Squared Error:  7.238691009300055e-07
```

1

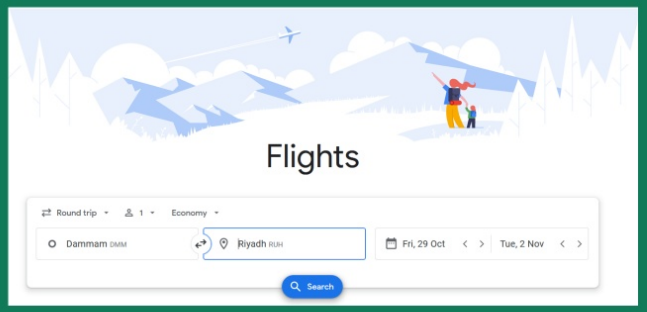
Web scrapped the data from google flights of Saudi local cities:

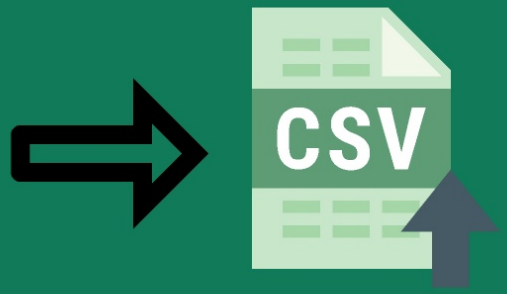
- Jeddah
- Dammam
- Jazan
- Medina

to Riyadh during Riyadh season period which is from October 2021 to March 2022

2

Load web scrapped data into a CSV file called tickets





3

Downloaded the tickets data csv file into Jupiter notebook then Loaded the file into a pandas Data Frame.

	date	from	to	airline	depaTime	arriTime	duration	price
0	2021-10-20	jeddah	riyadh	flyadeal	4:35 PM	6:15 PM	1:40:00	SAR 230
1	2021-10-20	jeddah	riyadh	flyadeal	6:05 PM	7:45 PM	1:40:00	SAR 230
2	2021-10-20	jeddah	riyadh	Flynas	8:05 PM	9:40 PM	1:35:00	SAR 269
3	2021-10-20	jeddah	riyadh	flyadeal	3:15 AM	4:55 AM	1:40:00	SAR 230
4	2021-10-20	jeddah	riyadh	flyadeal	7:45 AM	9:25 AM	1:40:00	SAR 230
...	...	...	...	...	...	...	...	...
121197	2022-03-31	medina	riyadh	Saudia	8:05 AM	1:10 PM	5:05:00	SAR 743
121198	2022-03-31	medina	riyadh	Saudia	8:25 AM	2:35 PM	6:10:00	SAR 772
121199	2022-03-31	medina	riyadh	Saudia	11:45 AM	4:35 PM	4:50:00	SAR 772
121200	2022-03-31	medina	riyadh	Saudia	11:45 AM	5:35 PM	5:50:00	SAR 772
121201	2022-03-31	medina	riyadh	Saudia	3:05 PM	7:35 PM	4:30:00	SAR 772

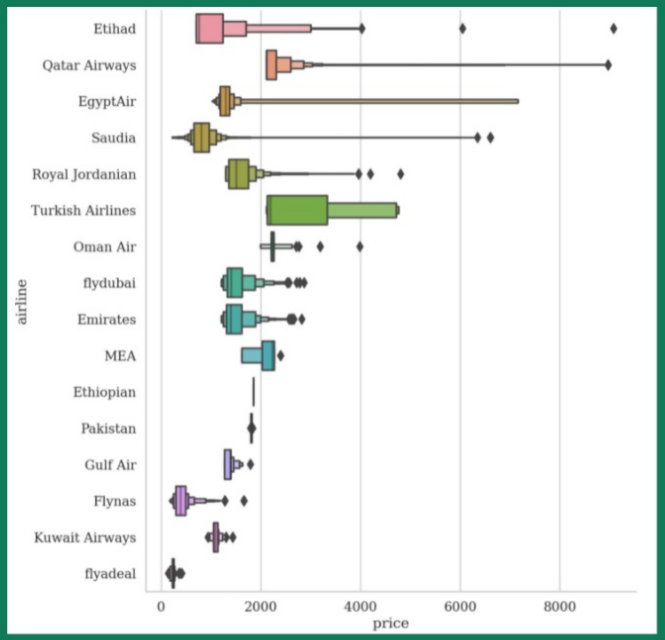
4

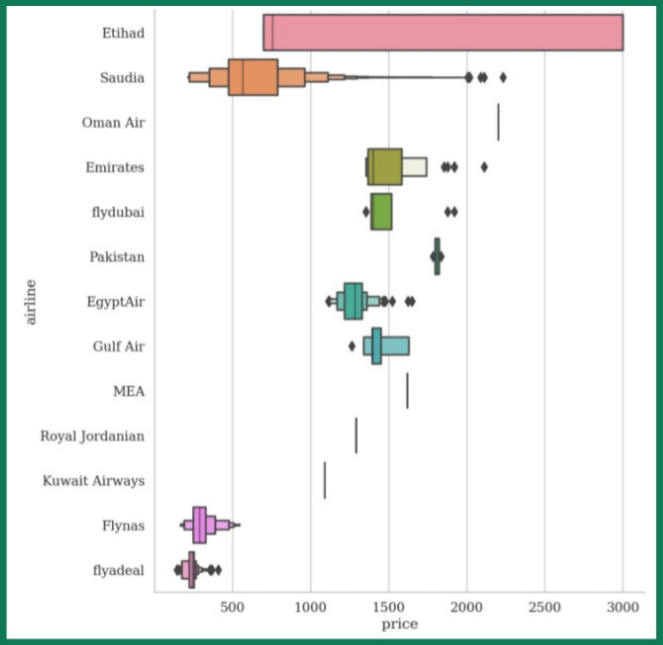
Explored and got to know the data while understanding how to use it and which attributes should be considered as features and prepare to be used in the prediction model.

	from	to	airline	price	weekday	bookingDays	day	month	year	depaHour	depaMinute	arriHour	arriMinute	durationHour	durationMinute
0	jeddah	riyadh	flyadeal	230	Wednesday	8	20	10	2021	16	35	18	15	1	40
1	jeddah	riyadh	flyadeal	230	Wednesday	8	20	10	2021	18	5	19	45	1	40
2	jeddah	riyadh	Flynas	269	Wednesday	8	20	10	2021	20	5	21	40	1	35
3	jeddah	riyadh	flyadeal	230	Wednesday	8	20	10	2021	3	15	4	55	1	40
4	jeddah	riyadh	flyadeal	230	Wednesday	8	20	10	2021	7	45	9	25	1	40
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
121197	medina	riyadh	Saudia	743	Thursday	170	31	3	2022	8	5	13	10	5	5
121198	medina	riyadh	Saudia	772	Thursday	170	31	3	2022	8	25	14	35	6	10
121199	medina	riyadh	Saudia	772	Thursday	170	31	3	2022	11	45	16	35	4	50
121200	medina	riyadh	Saudia	772	Thursday	170	31	3	2022	11	45	17	35	5	50
121201	medina	riyadh	Saudia	772	Thursday	170	31	3	2022	15	5	19	35	4	30

118404 rows × 15 columns

5 Cleaned the data and filtered it removing outliers and duplicates





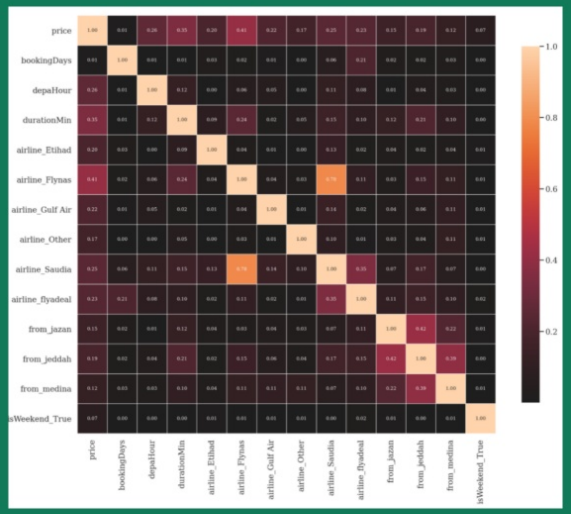
6 Converting categorical features to binary dummy variables during EDA

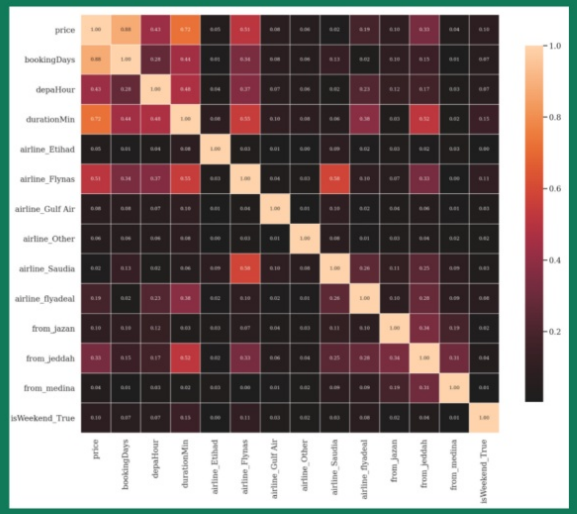
	price	bookingDays	depaHour	durationMin	airline_Etihad	airline_Flynas	airline_Gulf Air	airline_Other	airline_Saudia	airline_flyadeal	from_jazan	from_je
0	230	8	16	100	0	0	0	0	0	1	0	
1	230	8	18	100	0	0	0	0	0	1	0	
2	269	8	20	95	0	1	0	0	0	0	0	
3	230	8	3	100	0	0	0	0	0	1	0	
4	230	8	7	100	0	0	0	0	0	1	0	
...	...	...	...	...	...	...	...	...	...	...	...	...
121178	395	170	1	85	0	0	0	0	1	0	0	
121179	395	170	5	80	0	0	0	0	1	0	0	
121180	395	170	9	80	0	0	0	0	1	0	0	
121181	395	170	18	80	0	0	0	0	1	0	0	
121182	395	170	21	80	0	0	0	0	1	0	0	

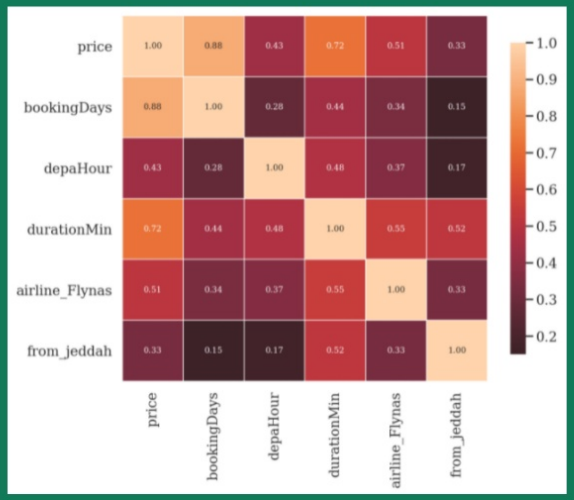
18481 rows × 14 columns

7

Checking for correlation and normalizing the data taking the highly correlated features with the target and removing some features that are highly correlated with each other







8

Splitting the data into sets: train 60%, validation20%, and test 20% in chronologically sorted order

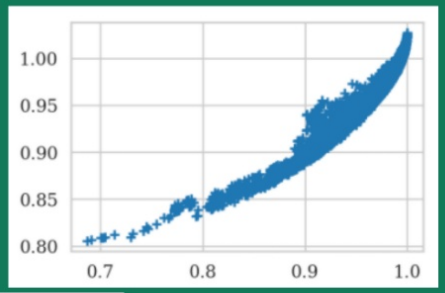
9

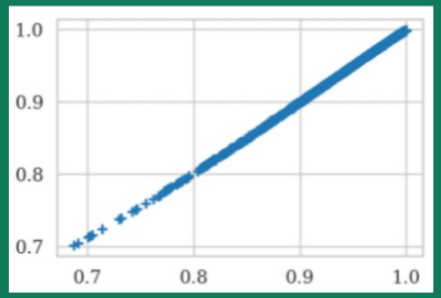
Fitting the Linear regression 6 models on train data using different features and comparing the score of each. Ending up with the best score of R^2: 0.9114945339772216 and Adjusted R Square: 0.9114705812123981, the Intercept: 1.031231756552298 and the following features: 'bookingDays', 'depaHour', 'durationMin', 'airline_Flynas', 'from_jeddah' with their coefficients as:

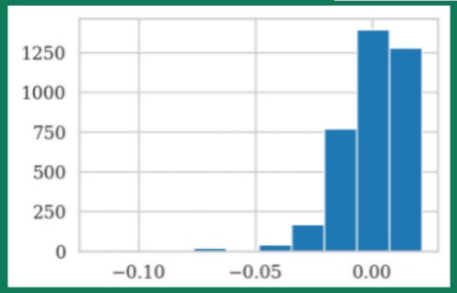
- bookingDays : -0.25
- depaHour : -0.11
- durationMin : -0.16
- airline_Flynas : -2.11
- from_jeddah : -0.66

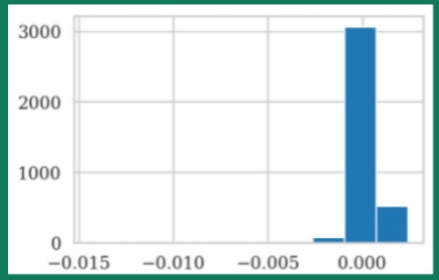
10

The results ended up in a kind of a curve shape; so, a polynomial regression model was compared with the linear regression model on the validation set, where the polynomial regression model has a better R^2 score of 0.9.









11

Test on the test set to get a final evaluation of expected generalization performance.

```
Polynomial Score:  0.9997476598895741
Mean Squared Error:  7.238691009300055e-07
```


Tools
Wego
Steps
Questions
Predicting Airline Ticket Prices
Using Linear Regression Models


Tools
• Python & Juypter Notebook
• Selenium web scrapping
• Numpy & Pandas for data manipulation
• Matplotlib and Seaborn for plotting
• Sklearn for machine learning
• Statsmodel for statistics description
• Prezi for presenting


Tools
Wego
Steps
Questions
Predicting Airline Ticket Prices
Using Linear Regression Models


Thank you 💐
Juri AlSayigh


Tools
Wego
Steps
Questions
Predicting Airline Ticket Prices
Using Linear Regression Models