AIRPLANE Price Prediction

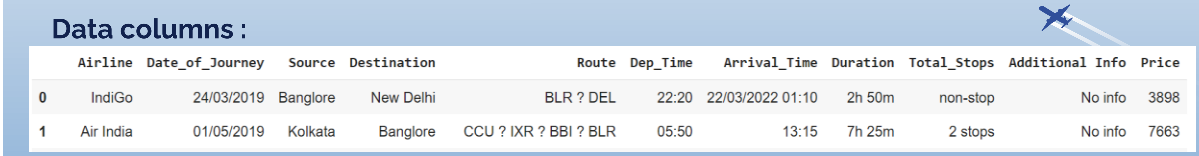
Rina Yasiun & Ohad Gutman

About The Data

The data set is about flights in India during 2019.

It's contains information over 10,000 domestic flights, direct and indirect, across different dates in the year.

Data columns :

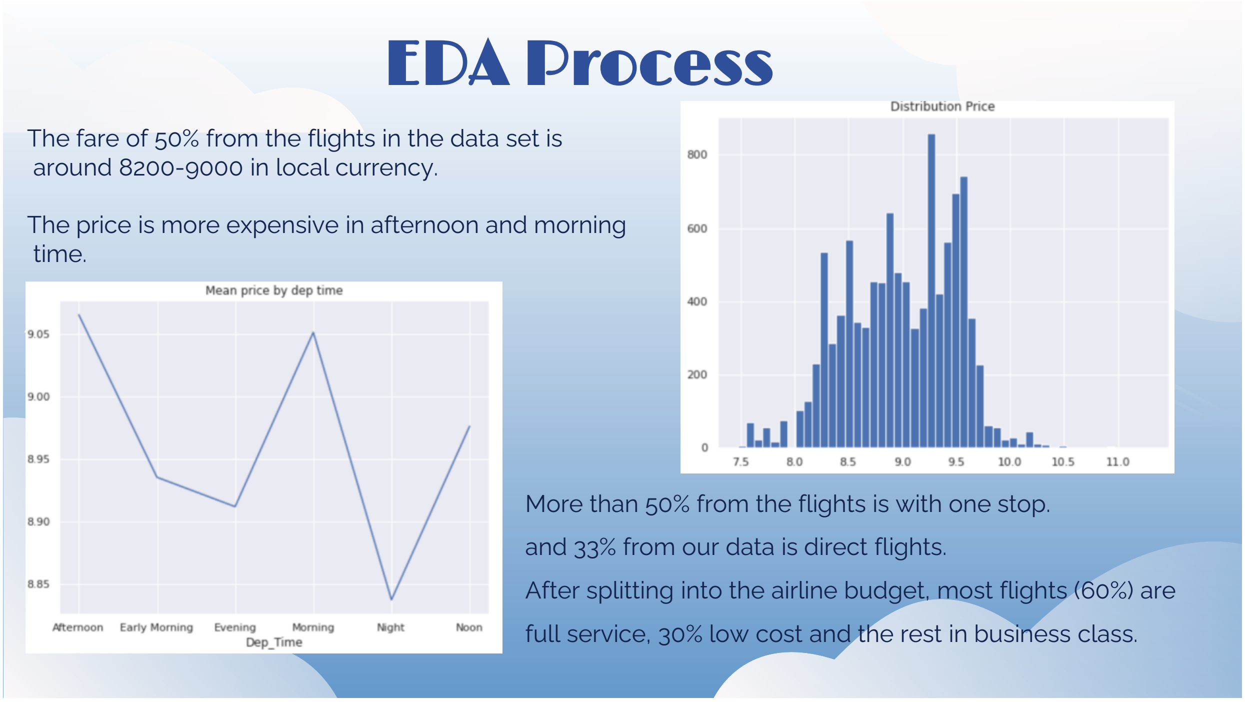


	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional Info	Price
0	IndiGo	24/03/2019	Banglore	New Delhi	BLR ? DEL	22:20	22/03/2022 01:10	2h 50m	non-stop	No info	3898
1	Air India	01/05/2019	Kolkata	Banglore	CCU ? IXR ? BBI ? BLR	05:50	13:15	7h 25m	2 stops	No info	7663

Business Problem : since flight prices today are unpredictable in this project we will try to find which factors effects directly on the flight price.

And predict the price flight according to data we have.

Data source: The data set took from Kaggle.



EDA Process

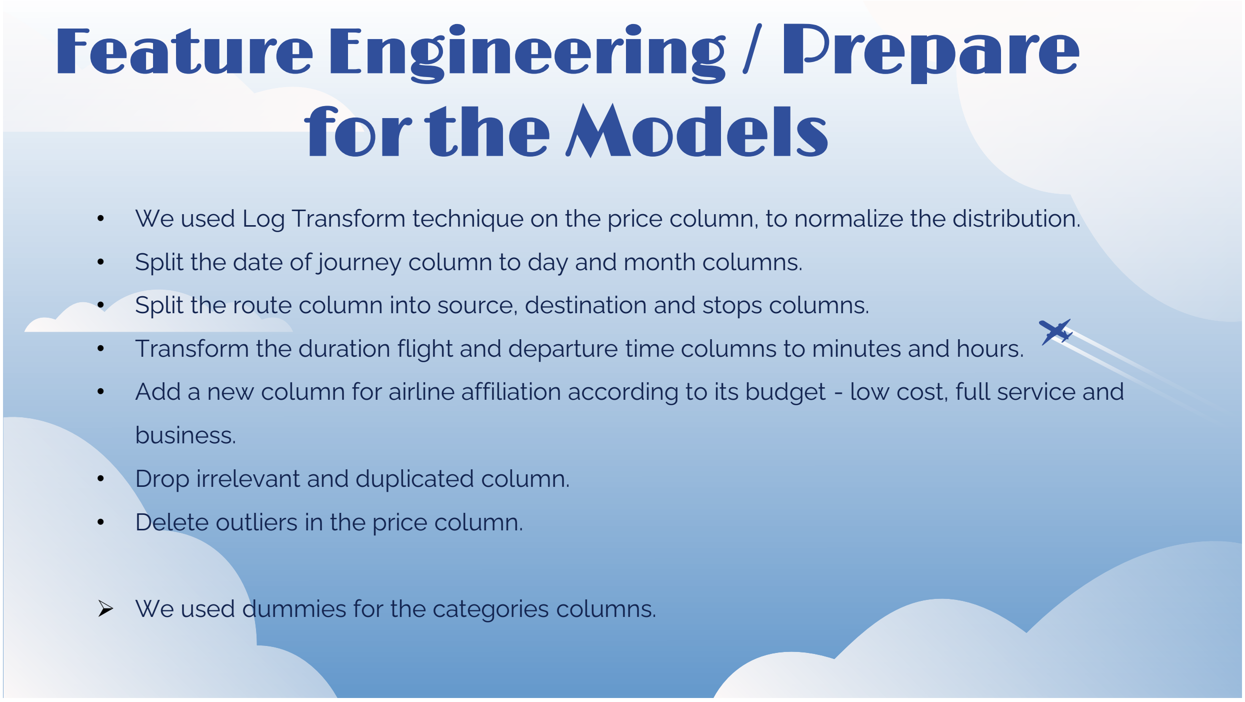
The fare of 50% from the flights in the data set is around 8200-9000 in local currency.

The price is more expensive in afternoon and morning time.

More than 50% from the flights is with one stop.

and 33% from our data is direct flights.

After splitting into the airline budget, most flights (60%) are full service, 30% low cost and the rest in business class.

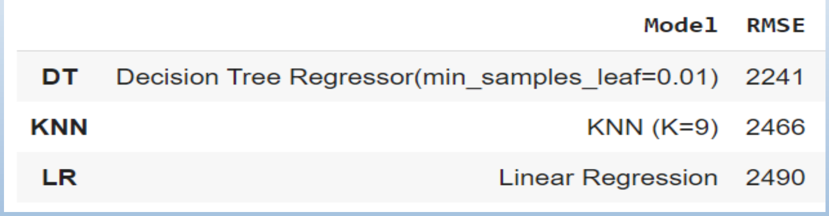


Feature Engineering / Prepare for the Models

- We used Log Transform technique on the price column, to normalize the distribution.
- Split the date of journey column to day and month columns.
- Split the route column into source, destination and stops columns.
- Transform the duration flight and departure time columns to minutes and hours.
- Add a new column for airline affiliation according to its budget - low cost, full service and business.
- Drop irrelevant and duplicated column.
- Delete outliers in the price column.

➢ We used dummies for the categories columns.

Results Of The Models



	Model	RMSE
DT	Decision Tree Regressor(min_samples_leaf=0.01)	2241
KNN	KNN (K=9)	2466
LR	Linear Regression	2490

After using a number of models and a lot of running, the best score is in the decision tree model.

The reasons that helped to improved the results are that we used of splitting additional columns and the use of standard and Minmax scalars .s

Thank you