MACHINE LEARNING

PROJECT: AIRLINE PRICE PREDICTION

REPORT

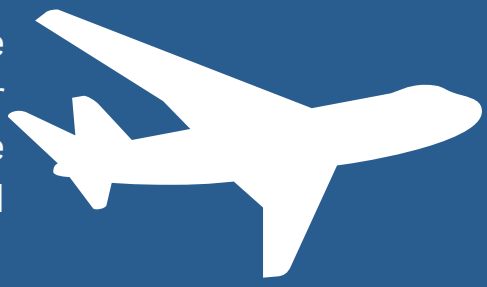
23/11/2023

Prepared by :

Pratham Singhal 2021082
Akshat Kumar 20210230
Hitesh Kumar 2021257
Harsh 2020061

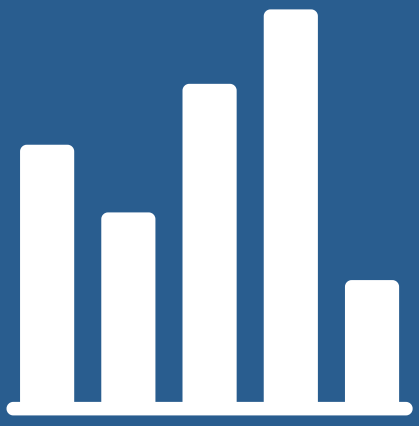
Introduction

Airlines and aviation industry is on a high boom since the invention of the airplanes by Wright brothers. For both the commercial as well as the general purpose driven commutation of human beings, it have showed immense growth and competition for fare prices.



Objective

Our objective for this project has been to study the given dataset of different airlines plying from different cities, with varying number of stops and duration, with 2 options of comfort level for the customer with fluctuating fare amounts. After rigorous study of the data, train a Machine Learning Model, which can be plied successfully to applications like Make My Trip, Goibibo, etc., to predict the fare prices given the preferences by the customer, and tell which airplane best suits them and what is its best predicted fare.



Research Questions

- How does price vary with Airlines?
- What is the behavior of the price of the airline with the number of days remaining in the flight?
- How does price fluctuates with arrival and departure time of the flight?
- How does the departure and arrival cities affect the price?
- What remain to be the strong deciding factor of the price?
- How does the number of stops enroute affect the price of the flight?

Dataset

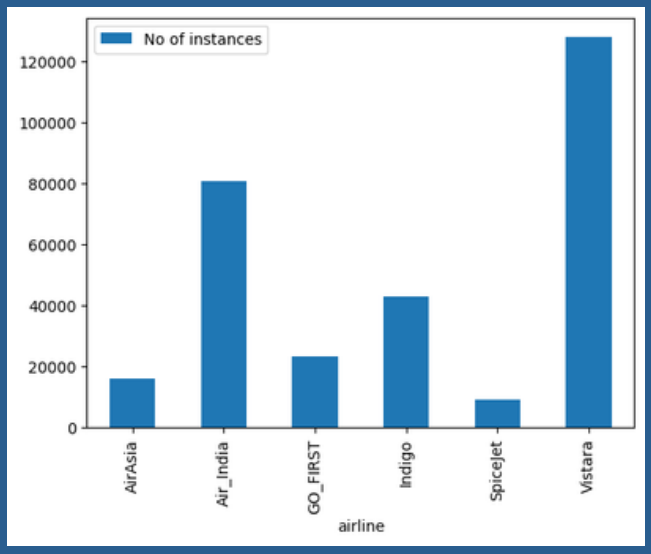
Dataset contains information about flight booking options from the website Easemytrip for flight travel between India's top 6 metro cities. There are 300261 datapoints and 11 features in the cleaned dataset.

Features

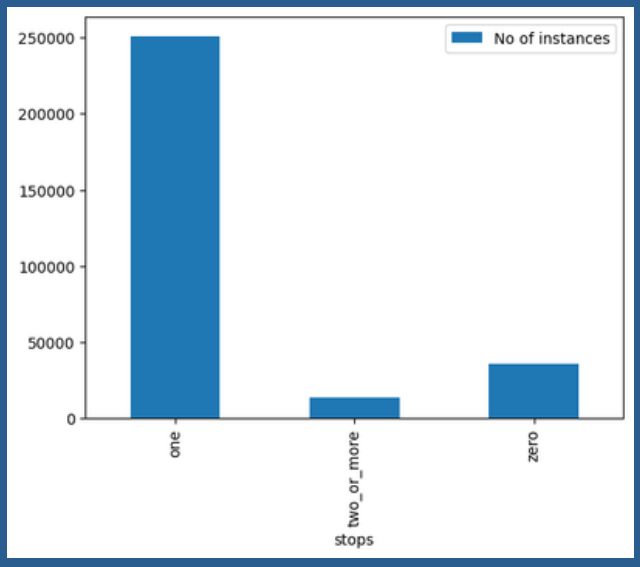
There are 11 features in the dataset, as follows:

Sno.	Feature	Description
1	AIRLINE	The name of the Airline Company
2	FLIGHT	Stores the code name of the airplane, which is a Categorical data.
3	SOURCE CITY	Name of the city from where the airplane takes off from.
4	DEPARTURE TIME	Time of departure from the source city, divided into segments like, morning, evening, afternoon, night. This is also a Categorical Data.
5	STOPS	The number of stops the plane takes enroute while in flight. Categorical Data with encoding as zero, one and two_or_more.
6	ARRIVAL TIME	The time of arrival at the arrival city, same Categorical division like departure time.
7	DESTINATION CITY	Name of the city where the plane has to land.
8	CLASS	Different comfort level class; mainly only economic and business class.
9	DURATION	Total flight time measured in hours.minutes; it is a numerical value
10	DAYS LEFT	Total number of days left from takeoff, that customers can take account of while booking the flight.
11	PRICE	Fare amount of the flight, which is in rupees, which is our dependent variable in the data.

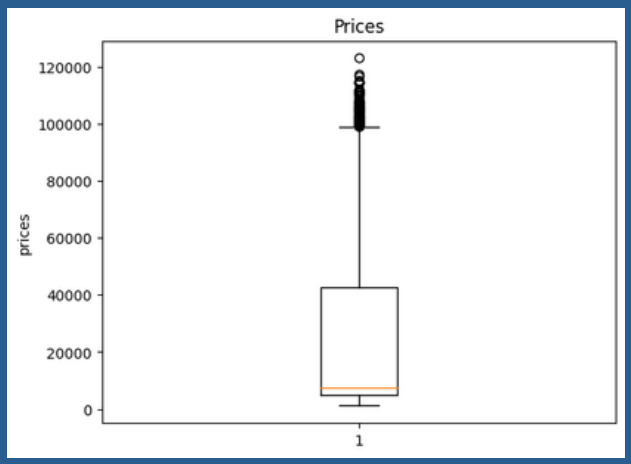
Data Visualization



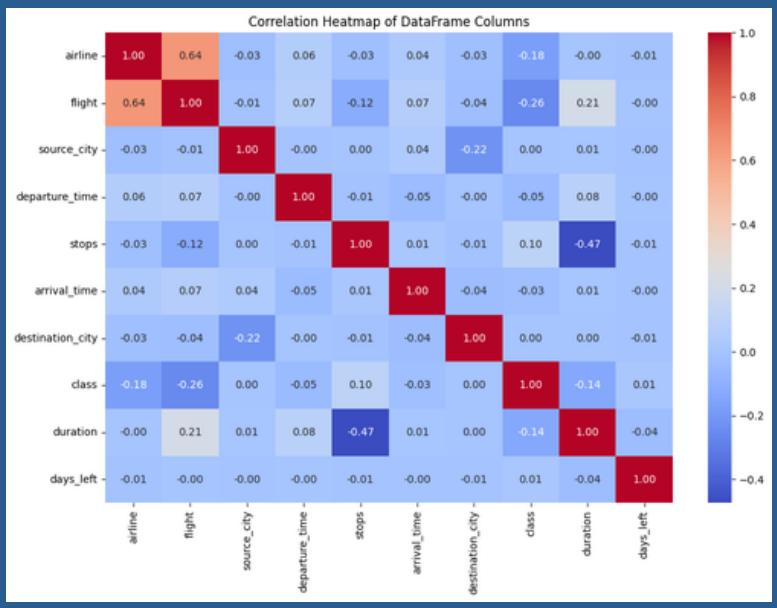
The graph here represent the number of instances of each available airline plying from different cities. Studying this we found that Vistara is one of the most sought after airline for the customers here with more than one lakh twenty thousand instances.



This graph represent the instance of flights with different number of stops. This indicates that the flights plying are of more in numbers which has one stop in between.



The box plot here represent the variation in Prices individually and from here we can notice the mean, maximum and minimum amount of fare. This gives us an image that how price distribution over the dataset looks like.



This is a correlation Heatmap, where we can notice the relation between different variables and its effect on determining the price

Data Preprocessing & Feature Engineering

- Removing the NA values from the dataset.
- Differentiating the Categorical and Numerical features.
- Label Encoding of the categorical data. We have assigned different numbers to different classes in Categorical Variables which has made it easy to train our models on the preferred features.
- Normalization of the numerical data. In order to discard the anomalies in the variation of the numerical data, we normalized these features so that our model does not count the bias created from them.
- Bootstrapping. Observing our dataset, we found that certain features involve high bias for certain values in the feature. hence we added certain more datapoints in those classes of features where the number was less using bootstrap that is repeating the number of the same instances. This is done because we do not want out model to loose out any class while training.

ML Models

Sno.	Model	R2	MAE
1	LINEAR REGRESSION	0.904	4622.18
2	LASSO	0.900	4833.36
3	RIDGE	0.904	4617.12
4	SGD REGRESSOR	0.904	4640.27
5	NAIVE BAYES	0.934	3021.48
6	SVM	0.310	15927.34
7	DECISION TREES	0.982	892.61
8	RANDOM FOREST	0.989	865.05
9	POLYNOMIAL(DEG=2,3,4,5)	0.94	3666.12

Result

Polynomial regression, Naive Bayes, Random Forest and Decision Trees are the best performing models in our study. We found that Random Forest gave us the most accuracy out of the other models which we performed training.

Observing the trends in our Model, we found it necessary to do K-fold Cross Validation for each model to check which one is better off.
Random Forest gave us the highest accuracy.

Conclusion

aIn conclusion, the flight price prediction ML project successfully developed a robust model for accurate ticket price forecasts. Thorough data preprocessing, including cleaning and feature engineering, laid the foundation for model training. Exploration of multiple machine learning algorithms, coupled with hyperparameter tuning, resulted in optimized predictive performance. Addressing imbalanced data, particularly for airlines like Vistara, involved the strategic application of oversampling and class weighting. Evaluation metrics, including precision, recall, and F1-score, validated the model's effectiveness. The deployment of this model promises valuable real-time predictions, with ongoing monitoring and retraining considerations for adapting to dynamic market conditions. The project not only met its objectives but also provided insights applicable to the broader domain of travel forecasting and data-driven decision-making.