ANALYZING AND PREDICTING FLIGHT PRICES USING MACHINE LEARNING

A Data Science Project using Python and Machine Learning

Edward Jonathan

OUTLINE

INTRODUCTION

Edward Jonathan

+62878-7869-2180

edjonathannnnn
@gmail.com

LinkedIn

01

Experience

GTM (Go-To-Market)
OPPO Indonesia

Business Development
JD.ID

Project Manager
JD.ID

02

Education

Data Science Bootcamp
dibimbing.id

Bachelor of Accounting
University of Tarumanagara

BUSINESS UNDERSTANDING

01 PROBLEM

Flight prices are highly volatile due to factors like demand fluctuations, seasonality, airline pricing strategies, and external events (e.g., fuel price changes, holidays). Consumers and businesses lack reliable tools to predict optimal booking times, leading to overpayment or missed opportunities. A data-driven flight price prediction model can address this by leveraging machine learning to forecast prices accurately.

02 OBJECTIVE AND GOALS

- Primary Goal: Develop a model that predicts flight prices with high accuracy to help users make cost-effective booking decisions.
- Secondary Goals:
  - Identify key pricing influencers (e.g., flight duration, route popularity).
  - Provide real-time price estimates for dynamic decision-making.

Success Criteria (Measurable Goals)

- Model Performance:
  - MAPE (Mean Absolute Percentage Error) < 10% or MAE < 1500 on test data.
- Business Impact:
  - Reduce average overpayment by 15% for users following model recommendations.

BUSINESS UNDERSTANDING

03 DATA COLLECTION AND METHODOLOGY

Octoparse scraping tool was used to extract data from the website. Data was collected in two parts: one for economy class tickets and another for business class tickets. A total of 300261 distinct flight booking options was extracted from the site. Data was collected for 50 days, from February 11th to March 31st, 2022. Data source was secondary data and was collected from Ease my trip website.

04 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 combined dataset.

DATA UNDERSTANDING

FEATURES

DATASET HAS **300261 ROW & 11 COLUMNS**

Kaggle

01

AIRLINE

The name of the airline company is stored in the airline column. It is a categorical feature having **6 different airlines**.

02

FLIGHT

Flight stores information regarding the plane's **flight code**. It is a categorical feature.

03

DEPARTURE CITY

City from which the flight takes off. It is a categorical feature having **6 unique cities**.

04

DEPARTURE TIME

This is a derived categorical feature obtained created by grouping time periods into bins. It stores information about the departure time and have **6 unique time** labels.

05

STOPS

A categorical feature with **3 distinct** values that stores the number of stops between the source and destination cities.

06

ARRIVAL TIME

This is a derived categorical feature created by grouping time intervals into bins. It has **6 distinct time** labels and keeps information about the arrival time.

DATA UNDERSTANDING

FEATURES

DATASET HAS **300261 ROW & 11 COLUMNS**

Kaggle

07 **DESTINATION CITY**

City where the flight will land. It is a categorical feature having **6 unique cities**.

08 **CLASS**

Categorical feature that contains information on seat class; it has two distinct values: **Business and Economy**.

09 **PRICE**

Target variable stores information of the t**icket price.**

10 **DURATION**

A continuous feature that displays the overall **amount of time** it takes to travel between cities in hours.

11 **SEASON**

A categorical feature that explains the **season** of the year.

TOOLS USED

PYTHON

Programming Language

GOOGLE COLLAB

Data Cleaning
Data Manipulation
Data Analysis & Visualization
Machine Learning Modeling

DATA CLEANING & MANIPULATION

Dataset was collected from Kaggle, the data itself is still divided into 2 file based on the flight class (Business and Economy), all of it are textual hence requires transforming steps for further analysis

01
COMBINE BOTH DATAFRAMES
Combining both dataframes, both data has the same column, added an extra columns [‘class’] consist of Business and Economy.

02
EXTRACT TEXT USING REGEX
Extracting required text from [‘stop’] column due to some error at parsing hence needed to be extracted.

03
REMOVE DUPLICATES, UNKNOWN VALUES
Remove some duplicates data and there’s no unknown values detected.

04
MERGING CH CODE AND NUM CODE
Merge the column flight number and code to create a new Flight Code column. e.g. [‘AI’][‘868’] → [‘AI 868’].

DATA CLEANING & MANIPULATION

05
CREATE COLUMN SEASON
Create a Season column based on the month they're flying.

07
CONVERT DATATYPES
Convert date → datetime, time_taken → float, & price → int.

06
EXTRACT TEXT FROM TIME TAKEN
Extract text from time taken for each flight and then change it into hr.mins e.g. 2h 45m → 2.75 (45/60=75).

08
OUTLIER & NULL VALUE HANDLING
Removed Starair and Trujet due to small amount of data that might hinder Machine Learning performance, some of the Null value are removed.

After data cleaning and manipulation, the total number of records was reduced from
300261 → 300153

EXPLORATORY DATA ANALYSIS

```
<class 'pandas.core.frame.DataFrame'>
Index: 300153 entries, 0 to 300260
Data columns (total 12 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   date            300153 non-null  datetime64[ns]
 1   airline         300153 non-null  object
 2   from            300153 non-null  object
 3   time_taken      300153 non-null  float64
 4   stop            300153 non-null  object
 5   to              300153 non-null  object
 6   price           300153 non-null  int64
 7   class           300153 non-null  object
 8   season          300153 non-null  object
 9   flight_code     300153 non-null  object
 10  departure_time  300153 non-null  object
 11  arrival_time    300153 non-null  object
```

MAX PRICE

123.071 ₹

MIN PRICE

1.105 ₹

AVG PRICE

7.425 ₹

AVG FLIGHT TIME

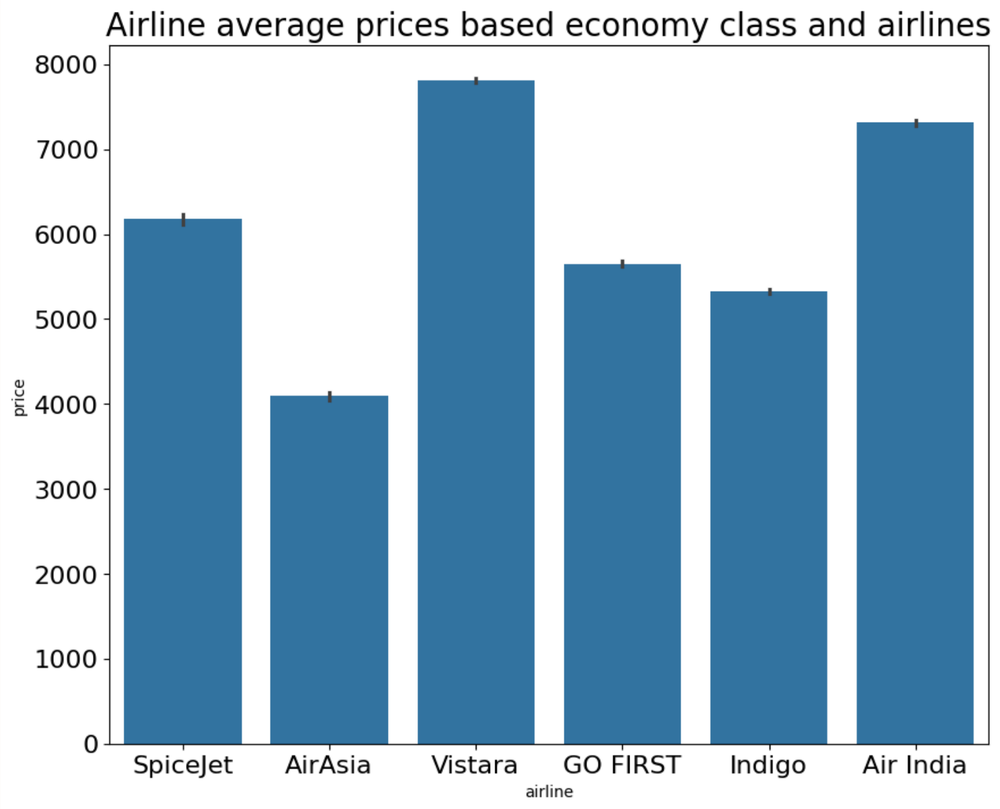
12 HRS

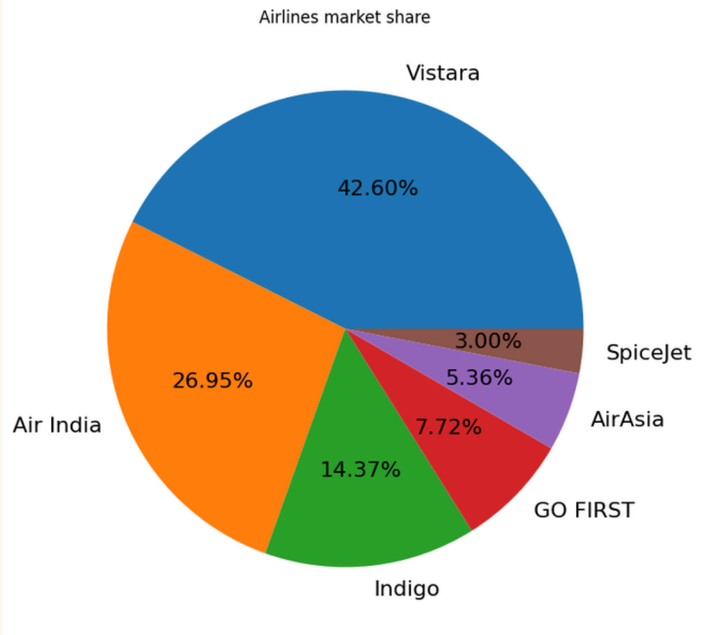
- DATASET CONTAINED 300.153 FLIGHT RECORDS
- MAXIMUM PRICE RECORDED IS 123.071 ₹ AND LOWEST IS 1.105 ₹
- AVERAGE FLIGHT TIME IS 12 HRS

WE CAN ASSUME THE CURRENCY ARE IN RUPEE

EXPLORATORY DATA ANALYSIS (EDA)

01



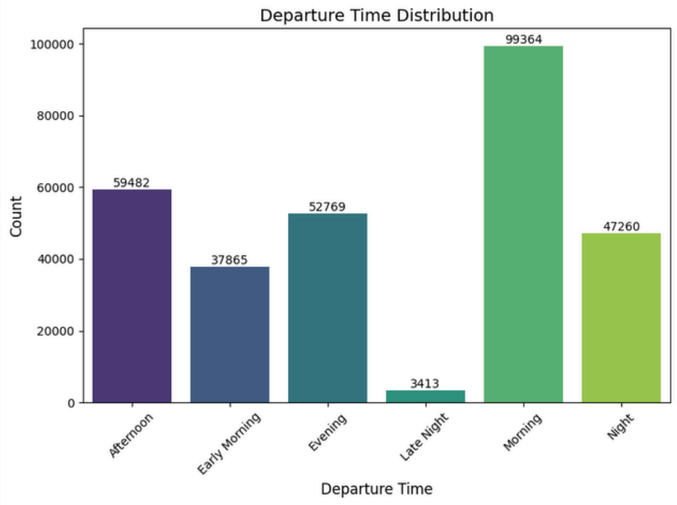


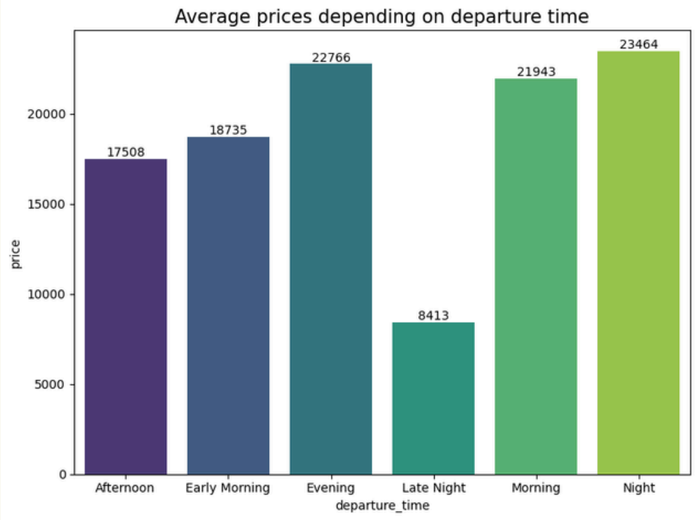
WHAT IS THE FAVORITE AIRLINES? & HOW DOES PRICE VARY WITH AIRLINES?

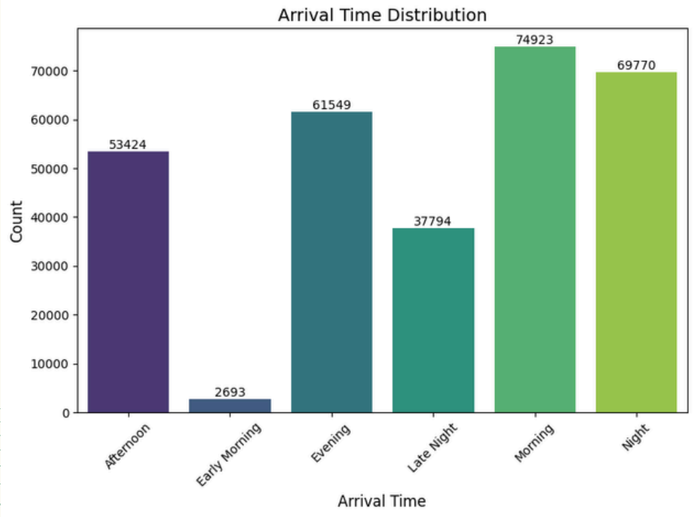
The majority of people would prefer to use Vistara and Air India as their airlines to fly with, with Vistara has 42.6% market share and Air India has 26.95%. Both of them combined total are 70% of total records. People might choose the Vistara and Air India regardless of the higher price.

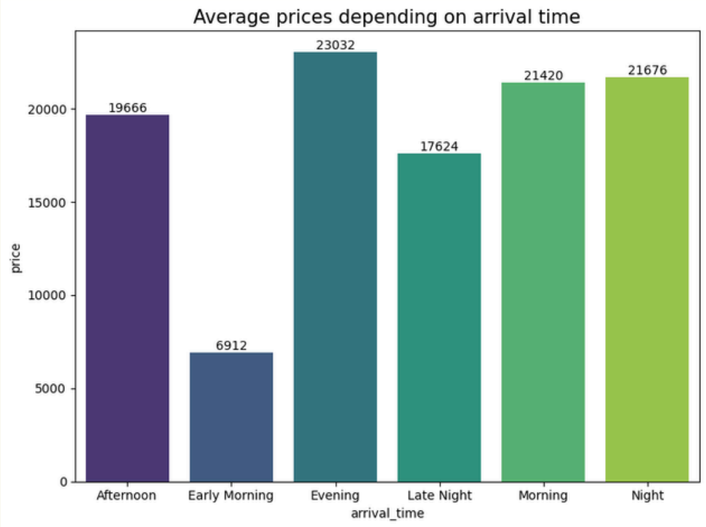
EXPLORATORY DATA ANALYSIS (EDA)

02







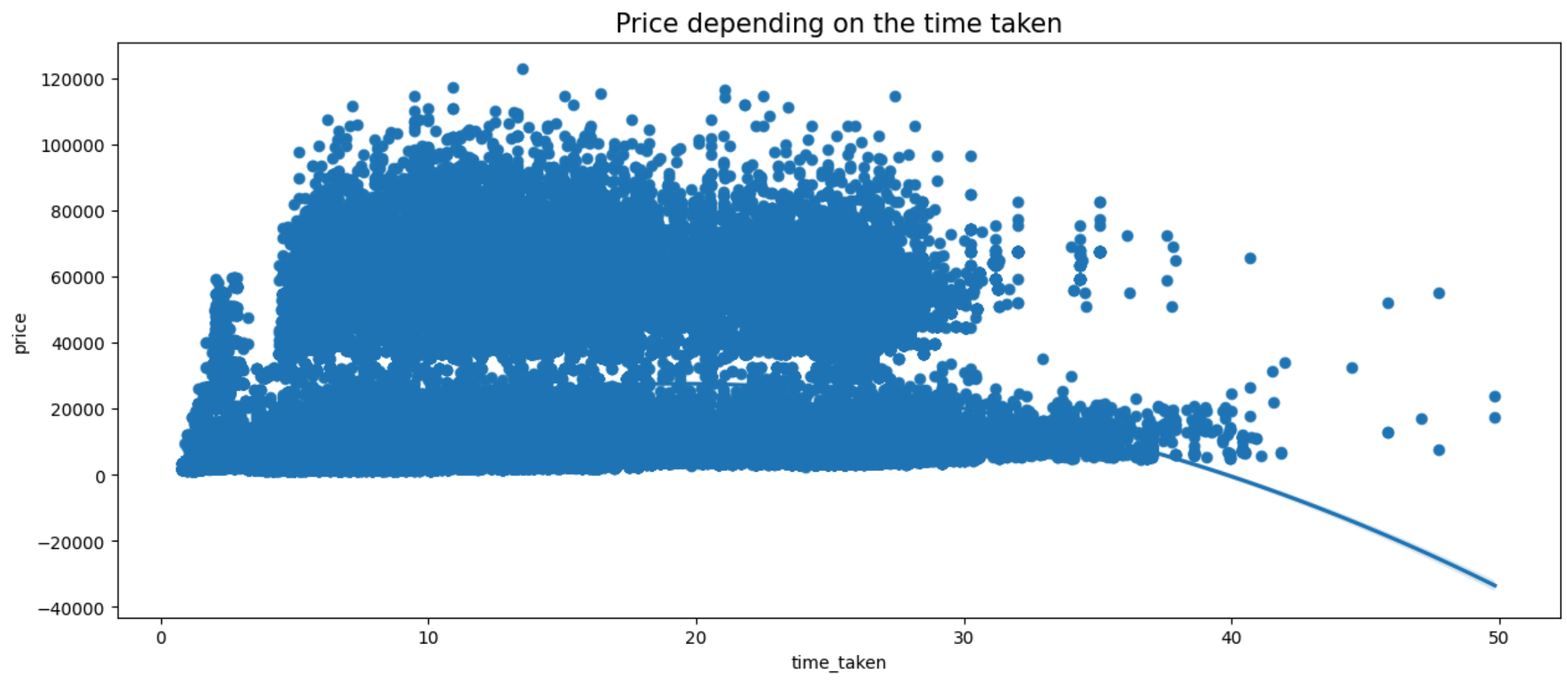


DOES TICKET PRICE CHANGE BASED ON THE DEPARTURE TIME AND ARRIVAL TIME?

Most people prefer flying in the morning (7 AM - 11 AM), with demand nearly double other times. Yet, morning prices rank third-highest, while night flights are the most expensive (23.464 INR), likely due to lower demand. This suggests travelers on long trips depart in the morning and arrive at night.

EXPLORATORY DATA ANALYSIS (EDA)

03

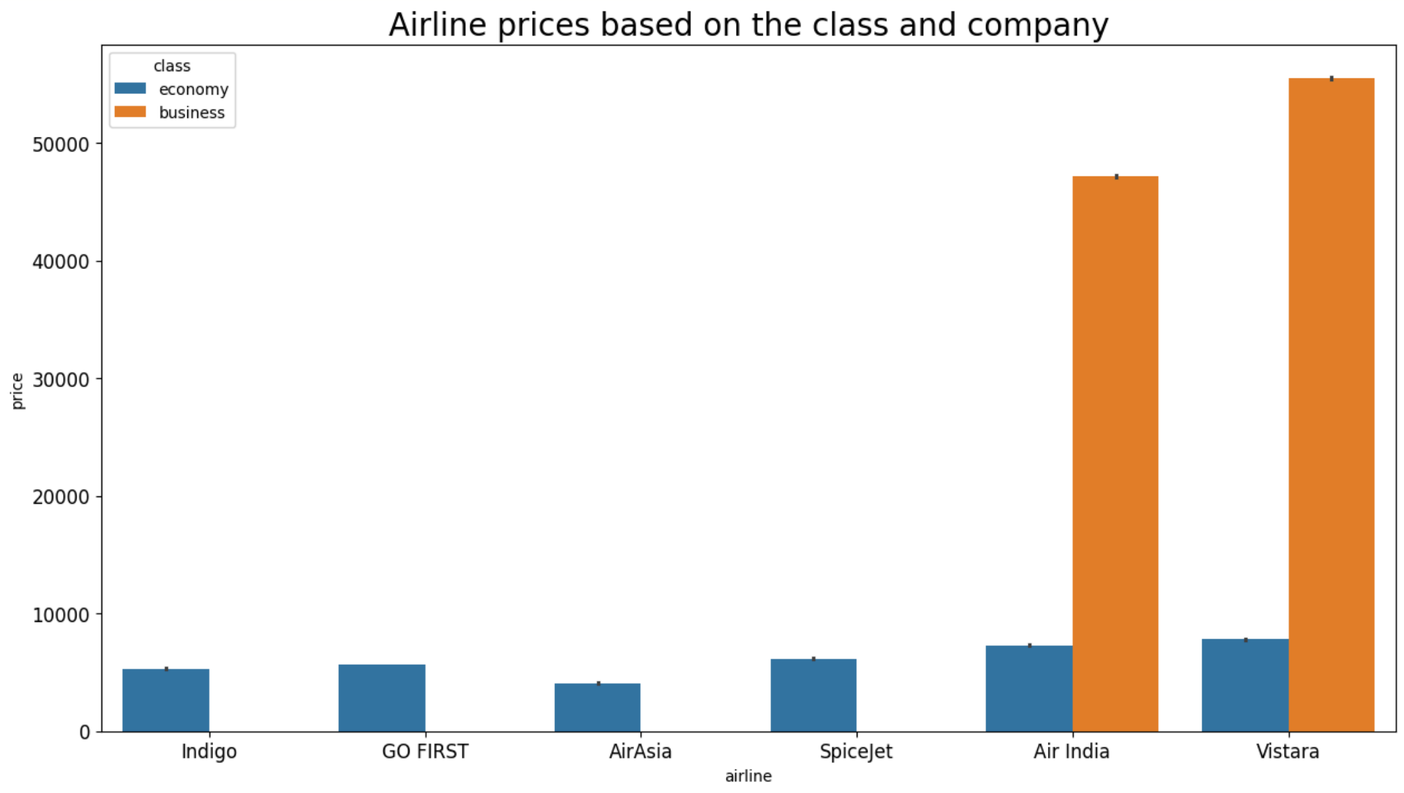


DOES THE PRICE CHANGE WITH THE DURATION OF THE FLIGHT?

- Prices generally increase with flight duration (longer flights cost more), but the relationship seems inconsistent.
- Below 5hrs flight, seems to have more stable pricing.
- Flight above 10hrs and below 30hrs usually have more higher range of price.

EXPLORATORY DATA ANALYSIS (EDA)

04



HOW DOES THE TICKET PRICE VARY BETWEEN ECONOMY AND BUSINESS CLASS?

Business class flights are exclusively offered by two airlines: **Air India** and **Vistara**.
The price difference between classes is significant, with business tickets costing nearly five times more than economy fares.

01 PREDICTIVE ANALYSIS

DELETING USELESS COLUMNS

Drop Flight_Code & Date for better effective machine learning

```
X_train = X_train.drop(columns=['flight_code','date'])

X_test = X_test.drop(columns=['flight_code','date'])
```

MACHINE LEARNING

Machine learning models were applied to estimate used car prices based on key features such as departure time, airlines, arrival time, stops, flight time, origin & destination. Next step is splitting the data.

DATA SPLITTING

80% TRAINING

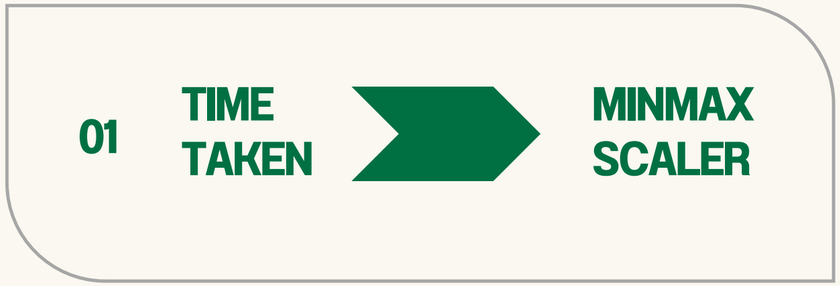
20% TEST

02 FEATURE ENCODING

Feature Encoding is the process of converting variables into numerical format so machine learning models **can interpret** them.

ORDINAL ENCODER:

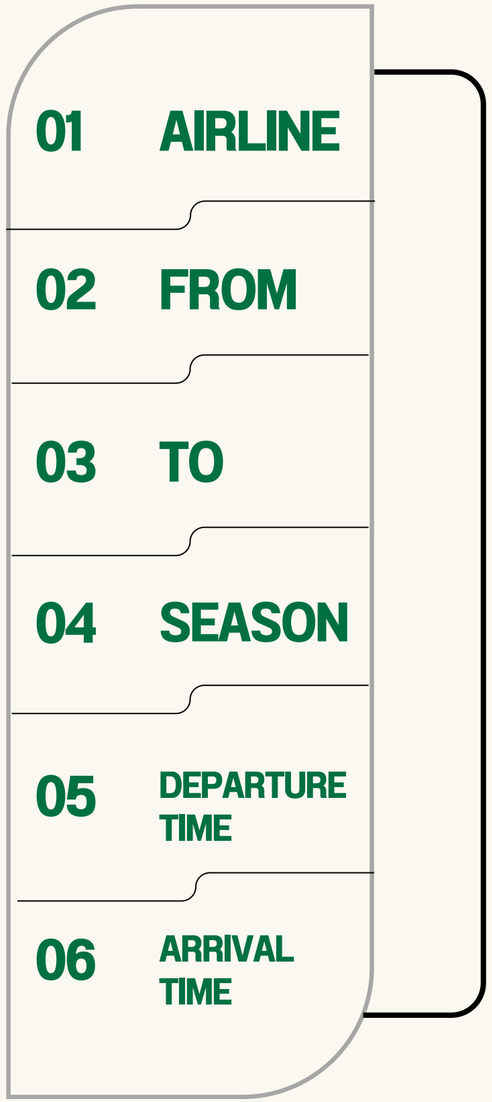
Converts categorical data into integers while preserving order (e.g., "low", "medium", "high" → 0, 1, 2).



MINMAX SCALER:

Normalizes numerical features to a fixed range (typically 0 to 1) by scaling values based on their minimum and maximum.

02 FEATURE ENCODING

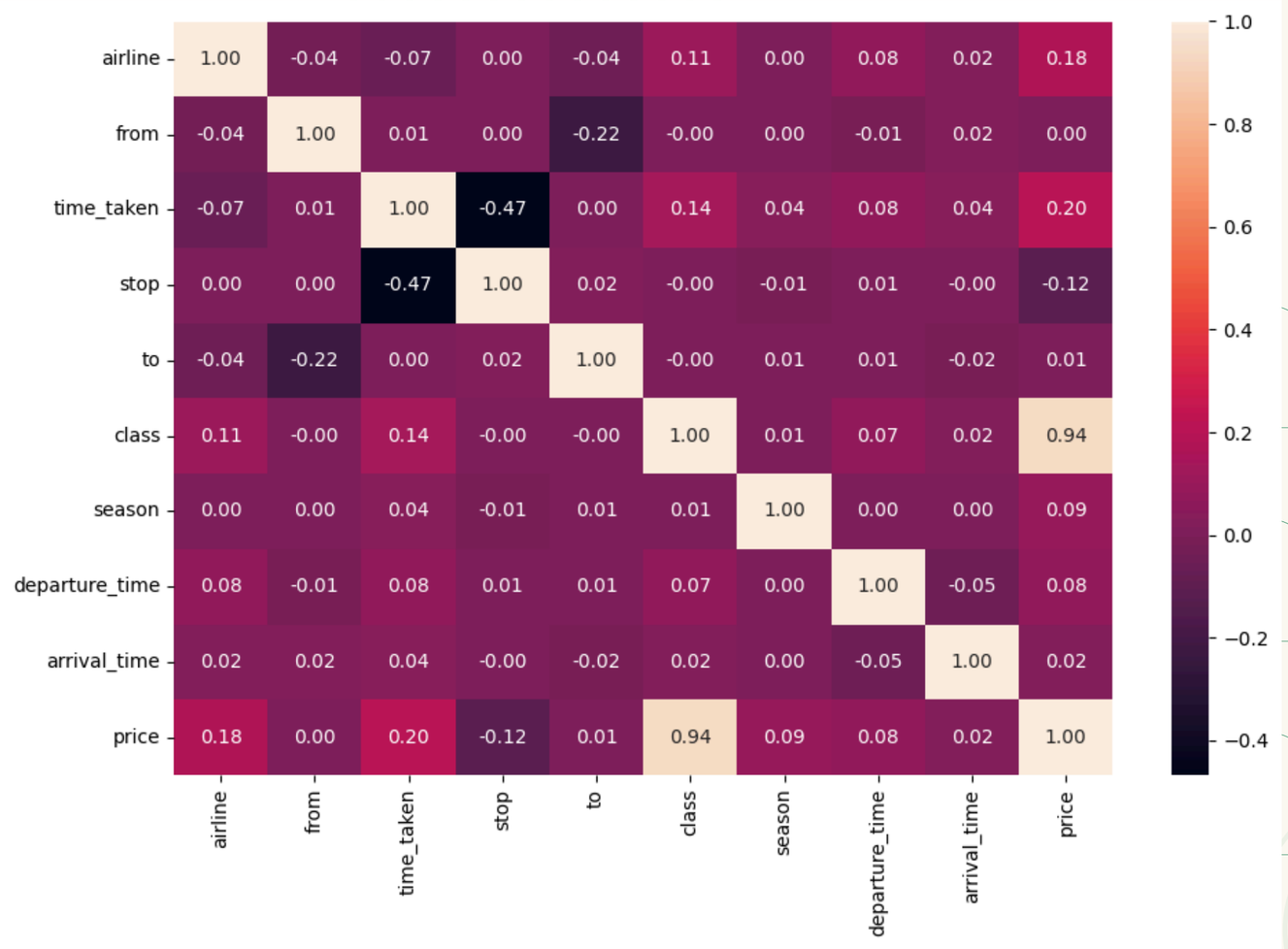


LABEL ENCODER

LABEL ENCODER:

Converts categorical labels into numerical values (e.g., "cat", "dog" → 0, 1), but without implying order.

03 CORRELATION HEATMAP



- The correlation map highlights key relationships between flight variables. The **strongest correlation** is between **price and class** (0.94), meaning higher travel classes (e.g., business vs. economy) are almost always linked to higher ticket prices.
- Price also has **weaker but notable** ties to airline (0.18), time_taken (0.20), and season (0.09), suggesting that certain airlines, longer flights, and seasonal demand may slightly increase costs.
- A **significant negative correlation** exists between time_taken and stop (-0.47), implying flights with more stops tend to be shorter overall—possibly due to optimized layovers.
- Most other variables, like from, to, and departure_time, show **near-zero correlations** with price, indicating they have little direct linear influence on pricing. The diagonal values (1.00) simply confirm each variable's perfect correlation with itself, which is expected.

Key Takeaway:

In summary, **class is the dominant price predictor**, while other factors like stops and flight duration play secondary roles.

MODELING

To evaluate predictive performance, four machine learning models were applied

LINEAR REGRESSION	For baseline performance and model interpretability.
CATBOOST REGRESSOR	A machine learning model that predicts numbers by finding patterns in data, working well with minimal setup
RANDOM FOREST	An ensemble method that handles non-linear relationship better.
XGBOOST REGRESSOR	An optimized gradient-boosting algorithm that can handle non-linear relationship better, that improves each iterative.

MODEL COMPARASION

Model	MAE	RMSE	MAPE(%)	R-Squared
Linear Regression	4564.53	49,249,468	42%	90.40%
CatBoost Regression	2362.29	17,602,003	18%	96.60%
Random Forest	1409.91	9,141,751	12%	98.20%
XGBoost Regressor	2157.25	13,834,869	18%	97.30%
CatBoost Regression (Tuned)	2182.18	15,545,698	16%	97.00%
Random Forest (Tuned)	1417.83	9,117,210	12%	98.20%
XGBoost Regressor (Tuned)	1795.04	10,578,190	15%	97.90%

01

RANDOM FOREST

Has the **lowest** prediction error with MAPE (12%) & MAE (1.409 INR) achieving the highest accuracy (R-square = 98.2%) between the other 3 models.

02

LINEAR

These modelperform the **weakest**, has the highest MAPE (42%), but able to achieve accuracy (R-square = 90.4%)

03

XGBOOST REGRESSOR

Perform **2nd best** while having prediction error slightly higher than Random Forest with MAE (1.795 INR) and MAPE (15%), and having accuracy (R-square = 97,9%)

04

CATBOOST REGRESSOR

Perform **3rd best** while having prediction error much higher than XGBoost and Random Forest with MAE (2.182 INR) and MAPE (18%), and having accuracy (R-square = 96.6%)

CONCLUSION

This project mission is to built a machine learning capable of predicting price based on the variable: routes, flight time, origin, destination, class.

- The dataset of **300261** records of flight was cleaned and engineered to extract key features (routes, flight time, origin, destination, class, etc.). Vistara and Air India dominates the market by having around 78% share of the market.
- Price is strongly affected by **class** & **flight duration**.
- Four machine learning model were tested: Linear Regression, CatBoost Regressor, Random Forest, & XGBoost Regressor. **Random Forest** performed the best with MAE = 1.409₹, MAPE = 12% and R-squared = 98.2%, indicates that the model has high accuracy on predicting price.

RECOMMENDATION & FUTURE WORK

Feature enhancement: Integrate more varibles such as Weekend or Weekday flight, how many days left before the flight date and oil crude prices.

Model Optimization: Explore more advanced model (e.g. LightGBM, Neural Networks).

APPENDIX

DATASET

Kaggle

SOURCE CODE

Google Collab

MODEL

THANK YOU!