COURSE CERTIFICATE

Apr 9, 2022

YOHANES SETIAWAN

has successfully completed

Supervised Machine Learning: Regression

an online non-credit course authorized by IBM and offered through Coursera

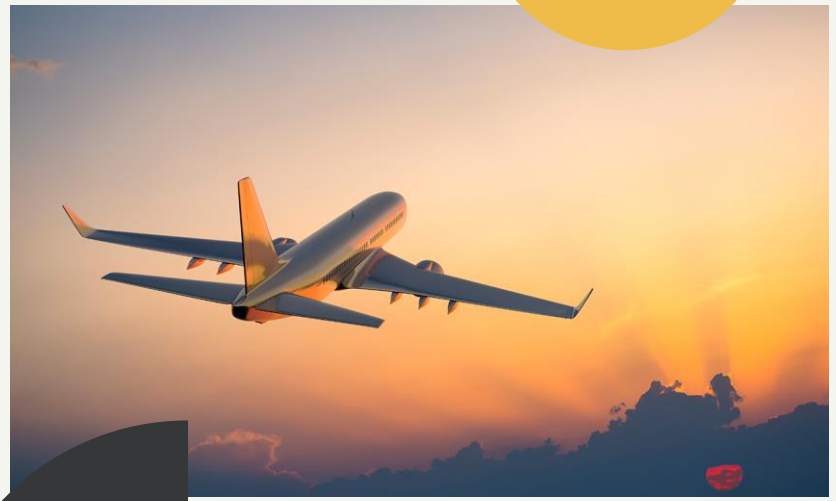
Mark J Grover
Digital Content Delivery Lead
IBM Data & AI Learning

Miguel Maldonado
Machine Learning Curriculum Developer
Data and AI Learning

Verify at:
https://coursera.org/verify/2DP8SQ9UGZYN

Coursera has confirmed the identity of this individual and their participation in the course.

Flight Ticket Price Prediction using Linear Regression Models



Yohanes Setiawan

Yohanes Setiawan

Hi, I am Yohanes Setiawan.

I am interested in Data Science and Analytics.

I graduated from Institut Teknologi Sepuluh Nopember in 2020 with a master's degree in informatics and Universitas Airlangga in 2017 with a bachelor's degree in mathematics.

I have curiosity to find insights in a dataset. Furthermore, I love writing. Therefore, I proudly present this report of regression findings to those who want to go deeply in a hands-on project.

Happy reading!

Business Understanding

Introduction

- Easemytrip is an internet platform for booking flight tickets, and hence a platform that potential passengers use to buy tickets
- A thorough study of the data will aid in the discovery of valuable insights that will be of enormous value to passengers

Problem Statement

Passengers are difficult to calculate a range of ticket price to make better plan for their trip

Goal

To give feedbacks to passengers in India for their best trip planning and predict the ticket price based on given features in Easemytrip application

Research Questions

- How is the price affected when tickets are bought in just 1 or 2 days before departure?
- Does ticket price change based on the departure time and arrival time?
- Does price vary with Airlines?
- How does the ticket price vary between Economy and Business class?
- How to predict flight ticket price?

Objective Statements

- Get insight about the effect of ticket price which are bought in just 1 or 2 days before departure
- Get insight about ticket price change based on the departure time and arrival time
- Get insight about variations of ticket price with airlines
- Get insight about variation of ticket price between economy and business class
- Conduct research to find the best model of flight price prediction using Linear Regression Models for passengers in India

Analytical Approach

- Descriptive analysis
- Graph analysis
- Table analysis
- Predictive analysis (Regression Problem)

Data Understanding & Exploratory Data Analysis

Data Understanding

- Data source was secondary data and was collected from Ease my trip website
- Source: https://www.kaggle.com/shubhambathwal/flight-price-prediction
- 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 Understanding: Dataset Information

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 300153 entries, 0 to 300152
Data columns (total 12 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   Unnamed: 0        300153 non-null  int64
 1   airline           300153 non-null  object
 2   flight            300153 non-null  object
 3   source_city       300153 non-null  object
 4   departure_time    300153 non-null  object
 5   stops             300153 non-null  object
 6   arrival_time      300153 non-null  object
 7   destination_city  300153 non-null  object
 8   class             300153 non-null  object
 9   duration          300153 non-null  float64
 10  days_left         300153 non-null  int64
 11  price             300153 non-null  int64
dtypes: float64(1), int64(3), object(8)
memory usage: 27.5+ MB
```

This means there is no missing value in the dataset.

Data Understanding: Dataset Information

```
[ ]  pd_flight.duplicated().sum()

     0
```

This means there is no duplicated data in the dataset.

Exploratory Data Analysis: Descriptive Statistics

Numerical columns:

	Unnamed: 0	duration	days_left	price
count	300153.000000	300153.000000	300153.000000	300153.000000
mean	150076.000000	12.221021	26.004751	20889.660523
std	86646.852011	7.191997	13.561004	22697.767366
min	0.000000	0.830000	1.000000	1105.000000
25%	75038.000000	6.830000	15.000000	4783.000000
50%	150076.000000	11.250000	26.000000	7425.000000
75%	225114.000000	16.170000	38.000000	42521.000000
max	300152.000000	49.830000	49.000000	123071.000000

Exploratory Data Analysis: Descriptive Statistics

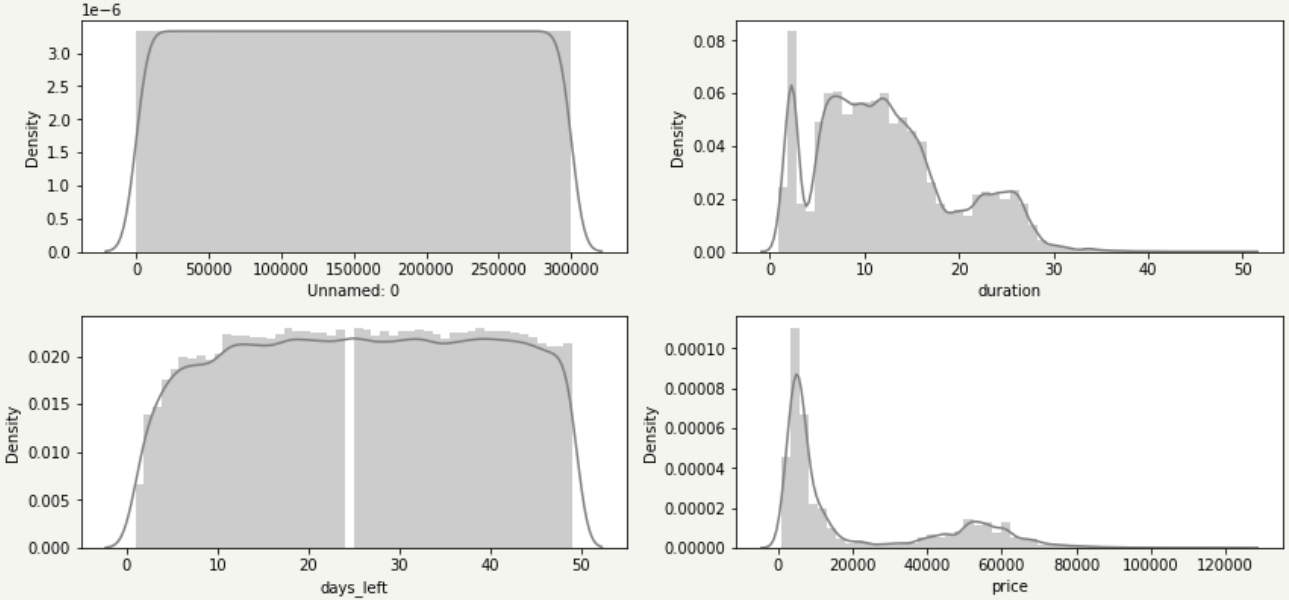
Categorical columns:

	airline	flight	source_city	departure_time	stops	arrival_time	destination_city
count	300153	300153	300153	300153	300153	300153	300153
unique	6	1561	6	6	3	6	6
top	Vistara	UK-706	Delhi	Morning	one	Night	Mumbai
freq	127859	3235	61343	71146	250863	91538	59097

Exploratory Data Analysis: Descriptive Statistics

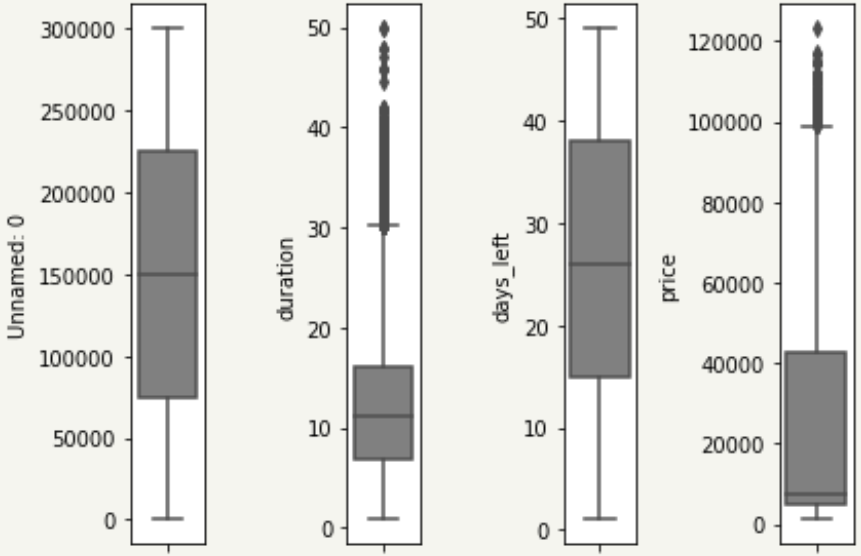
- The most favorite airline is Vistara
- The passengers mostly depart from Delhi
- The most likable destination city is Mumbai
- The passengers are likely to choose one stop for time efficiency
- Morning departure time has been the best time for passengers
- The passengers like to arrive at night
- There are 1561 unique values in column "flight". Therefore, it should be removed to avoid redundancy feature

Exploratory Data Analysis: Univariate Distribution



- Uniform distribution: "Unnamed: 0" and "days_left"
- Skewed: "duration", "price"

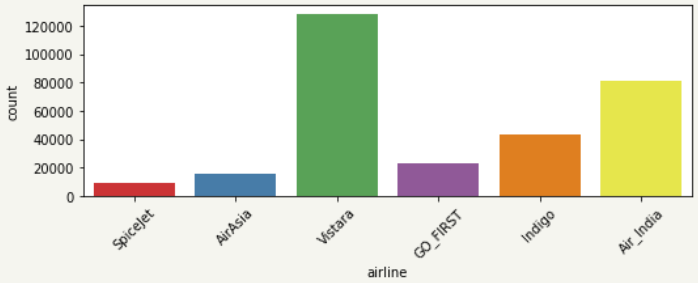
Exploratory Data Analysis: Box Plot

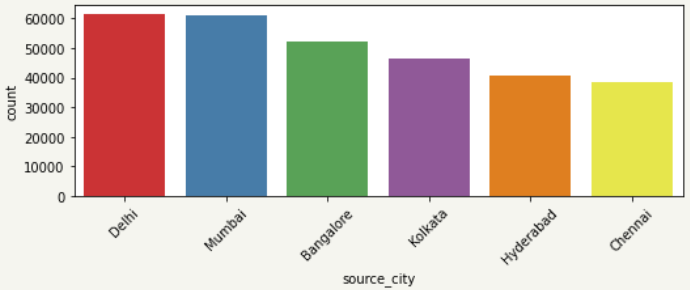


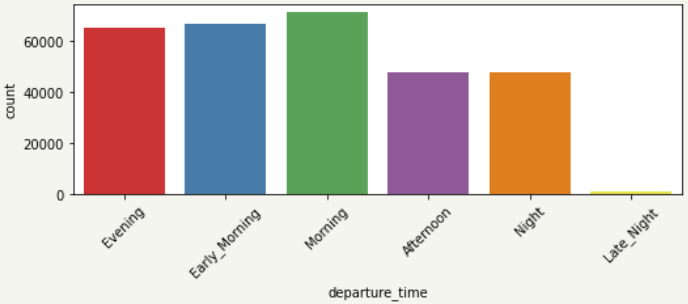
Exploratory Data Analysis: Box Plot

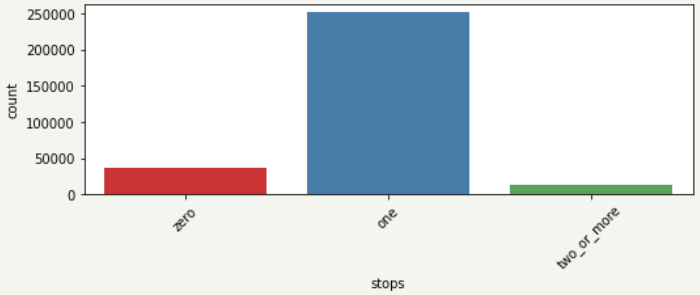
- The only independent feature/column with outlier is "duration".
- "price" is the dependent feature with outlier. However, this can be ignored because "price" is our target variable.
- Passengers with duration between cities for more than 30 hours are considered as outliers

Exploratory Data Analysis: Categorical Plot

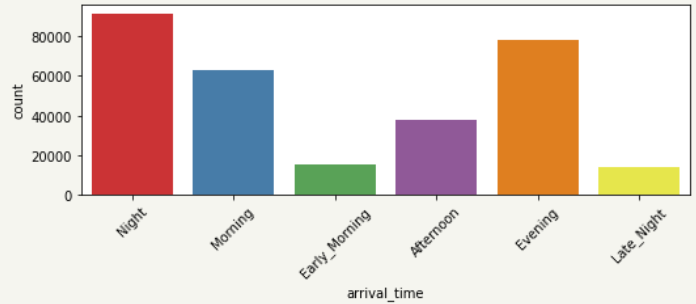


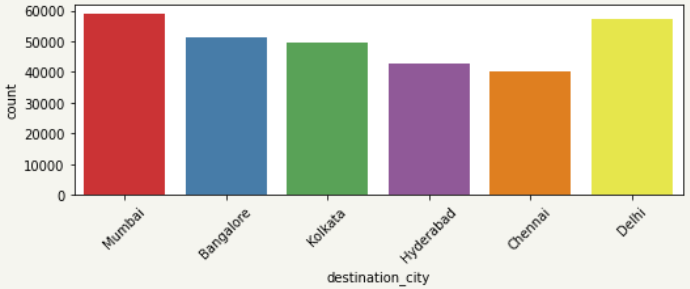


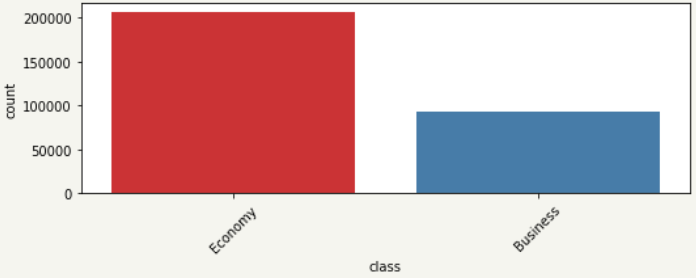




Exploratory Data Analysis: Categorical Plot



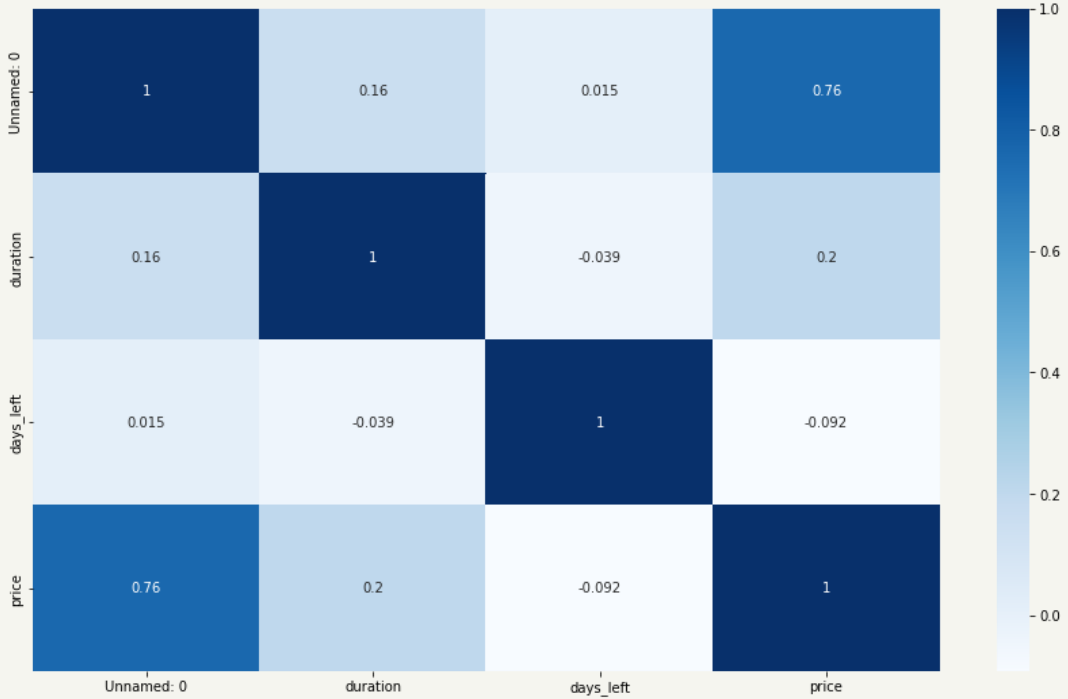




Exploratory Data Analysis: Categorical Plot

- Top 3 best seller airline in Easemytrip : 1. Vistara, 2. Air India, 3. Indigo
- Passengers do not like two or more stops between the source and destination cities
- Economy class has been the best seller in Easemytrip

Exploratory Data Analysis: Correlation Analysis



Exploratory Data Analysis: Correlation Analysis

- Column "Price" is highly correlated with "Unnamed: 0". This means the higher values in "Unnamed: 0", also the higher values in "Price" as the target variable.
- However, the "Unnamed: 0" is the ID of the customer, which is sorted from the lowest bought ticket price until the highest bought ticket price.
- Therefore, column "Unnamed: 0" will affect the predicted model seriously and need to be removed.

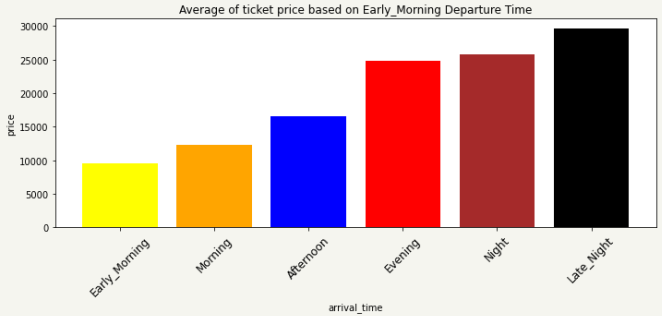
Get Insights

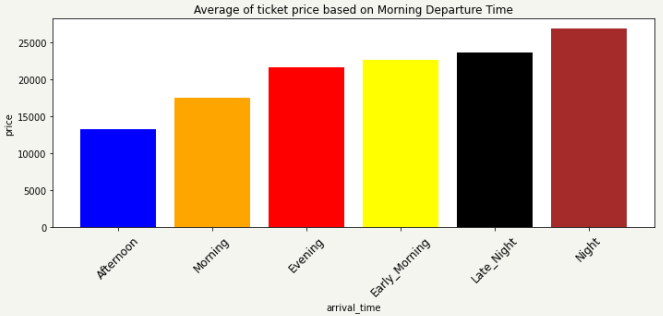
How is the price affected when tickets are bought in just 1 or 2 days before departure?

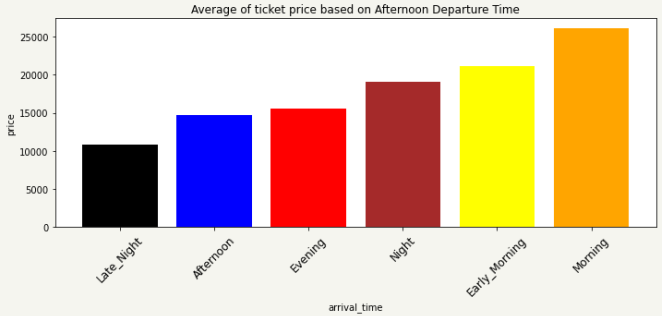
How is the price affected when tickets are bought in just 1 or 2 days before departure?

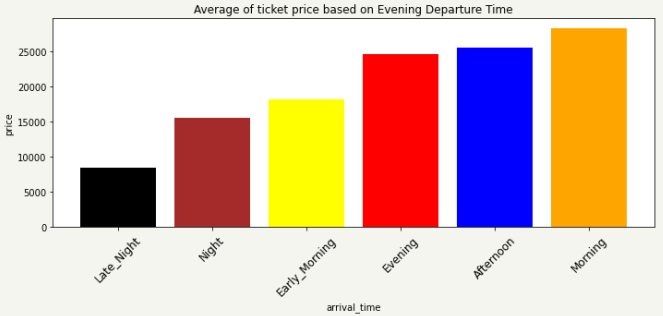
- The ticket price becomes higher when the ticket is booked in 1-2 days before departure and comes to the peak when it is booked 3 days before departure
- The ticket price becomes lower when it is booked around 25-30 days before departure

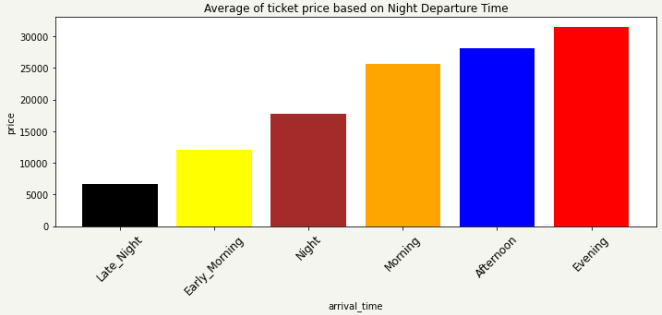
Does ticket price change based on the departure time and arrival time?

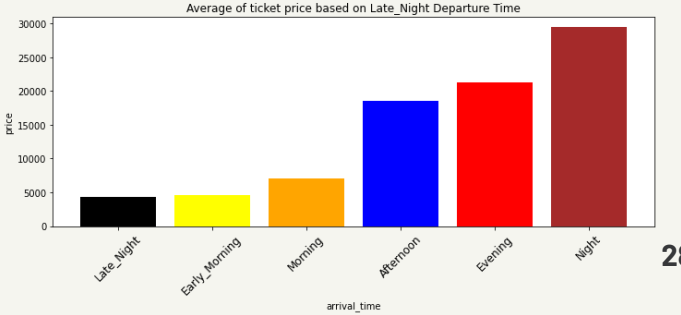




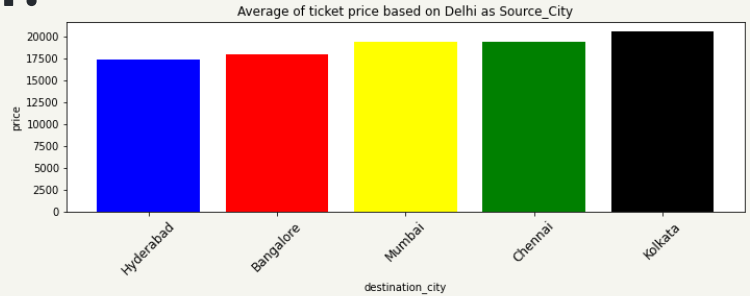


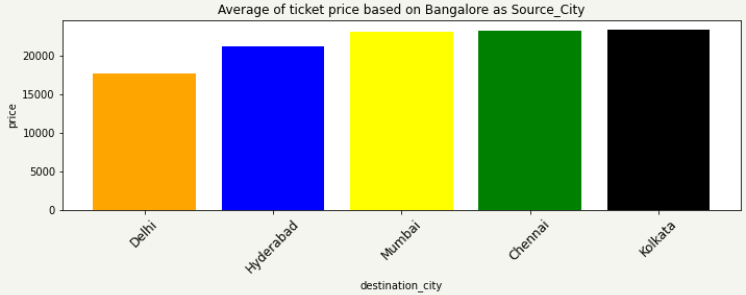


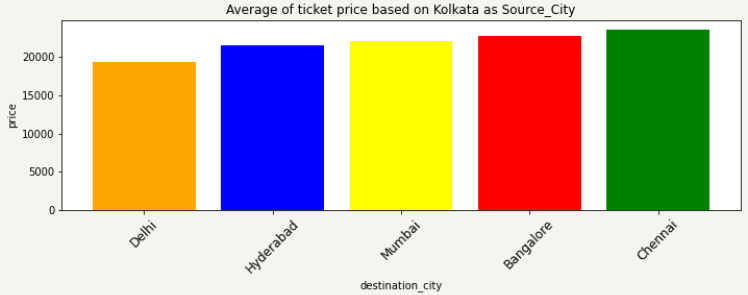


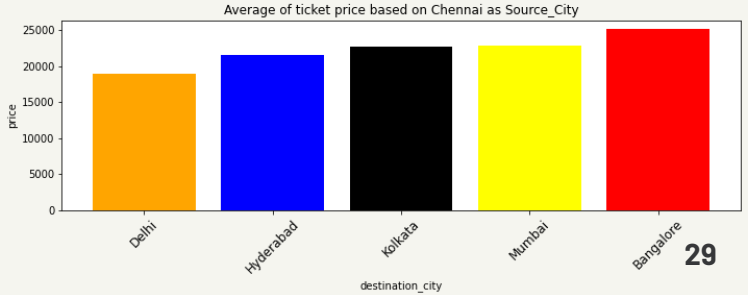


How the price changes with change in Source and Destination?



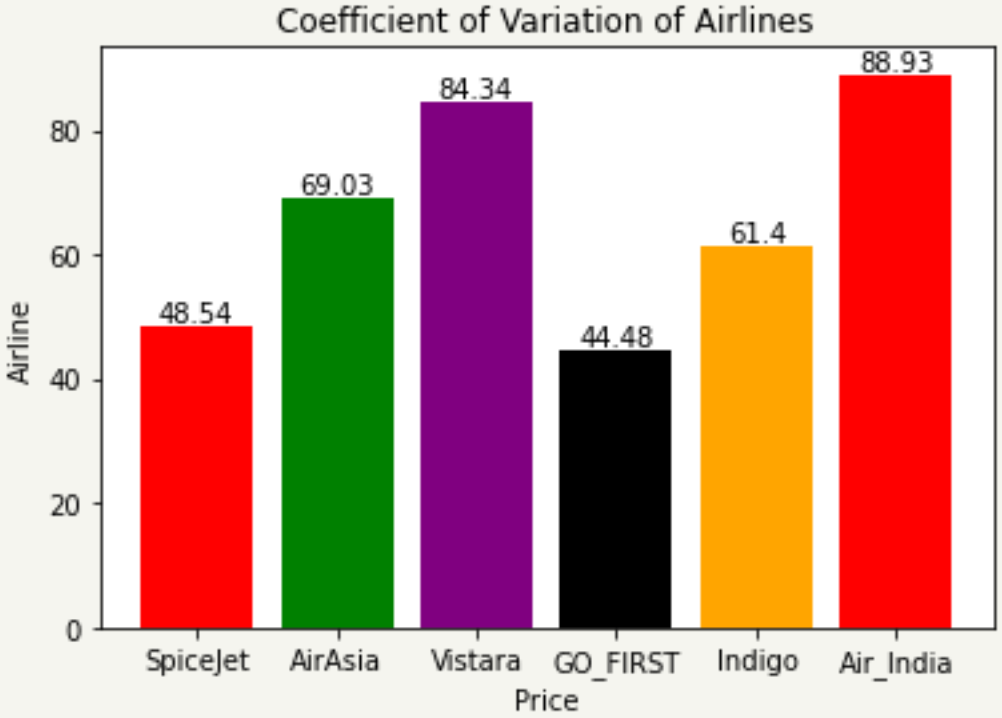






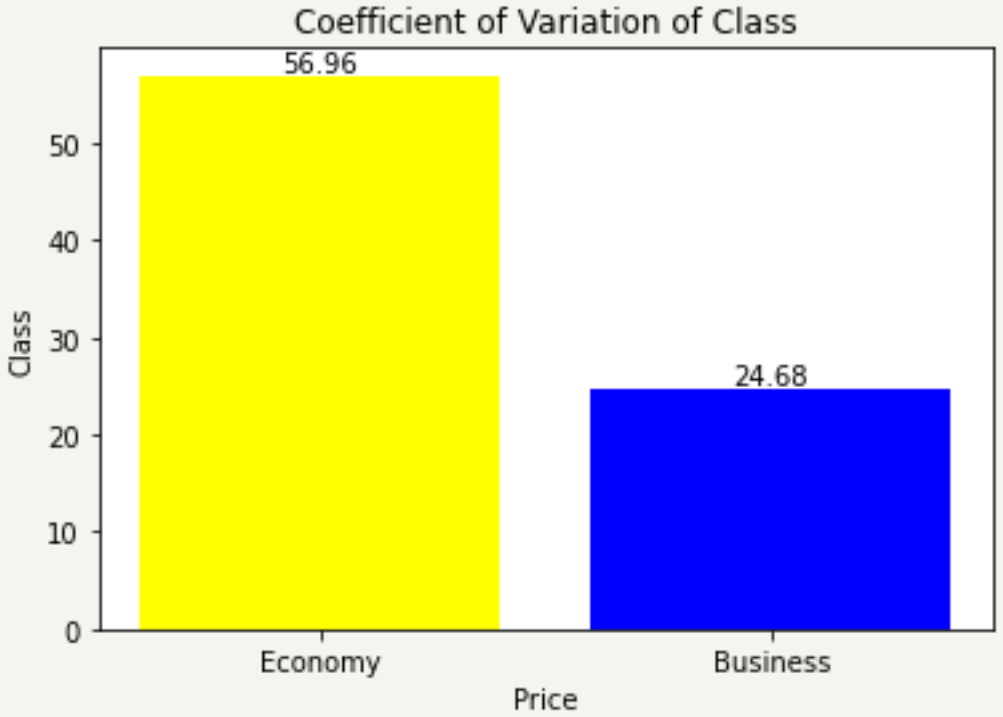
Does price vary with Airlines?

- The ticket prices vary with airlines.
- Air India is the most varying airline because of its highest CV.



How does the ticket price vary between Economy and Business class?

		price	
		min_price	max_price
airline	class		
AirAsia	Economy	1105	31917
Air_India	Business	12000	90970
	Economy	1526	42349
GO_FIRST	Economy	1105	32803
Indigo	Economy	1105	31952
SpiceJet	Economy	1106	34158
Vistara	Business	17604	123071
	Economy	1714	37646



How does the ticket price vary between Economy and Business class?

- Economy class varies more than business class because all of airlines have economy class.
- Air India and Vistara are airlines which have between economy and business class, while the others only have economy class. This answers previous analysis that finds higher CV of Air India and Vistara because of its high variation of price.
- AirAsia, Air India, Go First and Indigo have similar minimum ticket price in economy class
- The most expensive airline is Vistara

Does ticket price change based on the number of stops between the source and destination cities?

	price		
	mean_price	min_price	max_price
duration_bins			
Low	19630.553992	1105	123071
Medium	25507.349486	2477	116562
High	15313.040314	4802	82729

- From lowest to highest price based on the mean: High, Low, Medium
- Low duration has been the highest maximum of the ticket price
- High duration affects the lowest minimum of the ticket price

Feature Engineering & Data Preparation

Feature Engineering

- Handling Outlier using Interquartile Range Analysis: columns “duration”
- Removing Irrelevant Feature(s): column “Unnamed: 0” and “flight”
- Categorical Encoding using One Hot Encoding for categorical columns

Data Preparation

- Training-Testing Split: 70% training data and 30% testing data
- Feature Scaling with StandardScaler()
- Box Cox Transformation for Target Variable

Modelling

Modelling

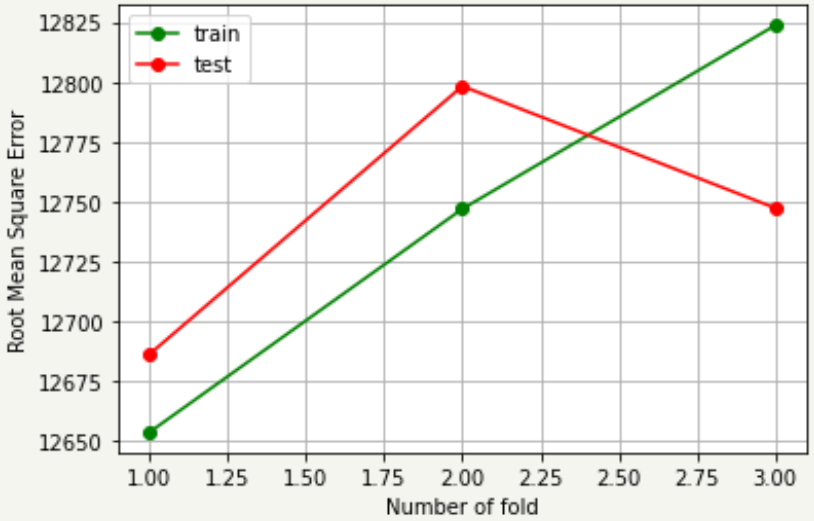
- I used 3 linear regression models: Multiple Linear Regression, Polynomial Regression, and Lasso Regression
- The model will be evaluated by 3-Fold Validation to check whether the model is overfitting or underfitting.

Modelling:
Multiple Linear Regression

- Generalization of simple linear regression for more than one predictor variable.
- Two multiple linear regression are compared: With and Without Box Cox Transformation

Modelling:
Multiple Linear Regression

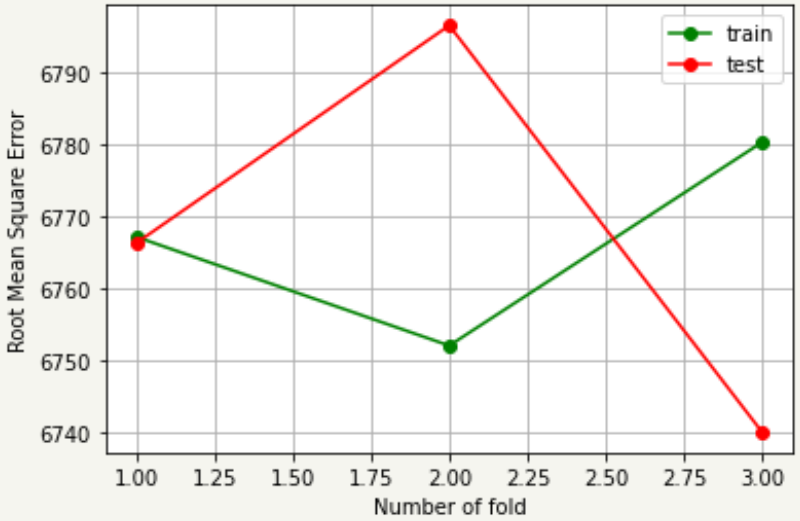
With Box Cox Transformation



The model from Multiple Linear Regression with Box Cox Transformation is not overfitting or underfitting.

Modelling:
Multiple Linear Regression

Without Box Cox Transformation

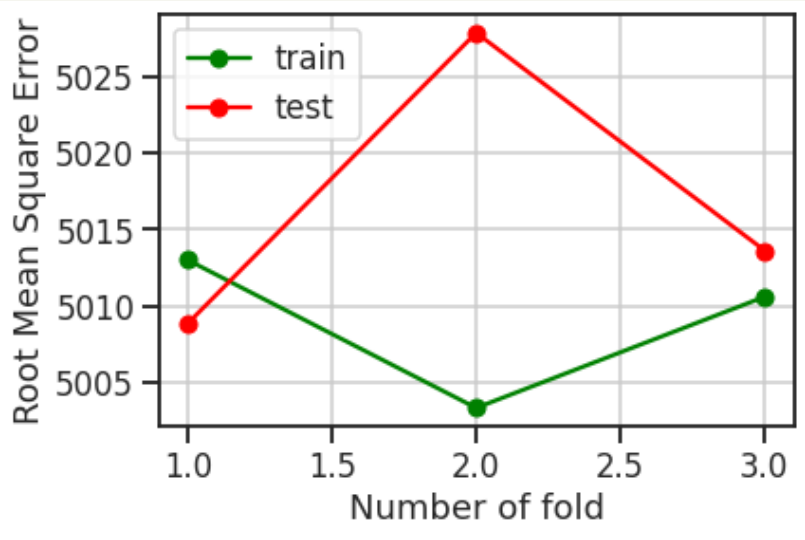


- The model from Multiple Linear Regression without Box Cox Transformation is not overfitting or underfitting.
- Because the error from Multiple Linear Regression without Box Cox Transformation is lower, it will be chosen to compare with other linear regression models.

Modelling:
Polynomial Regression

- Linear Regression with Polynomial Features
- I choose the maximum degree of polynomial = 2 because the dataset have too many features such that the polynomial transformation of the features will not affect too many addition to the features of the dataset.

Modelling:
Polynomial Regression

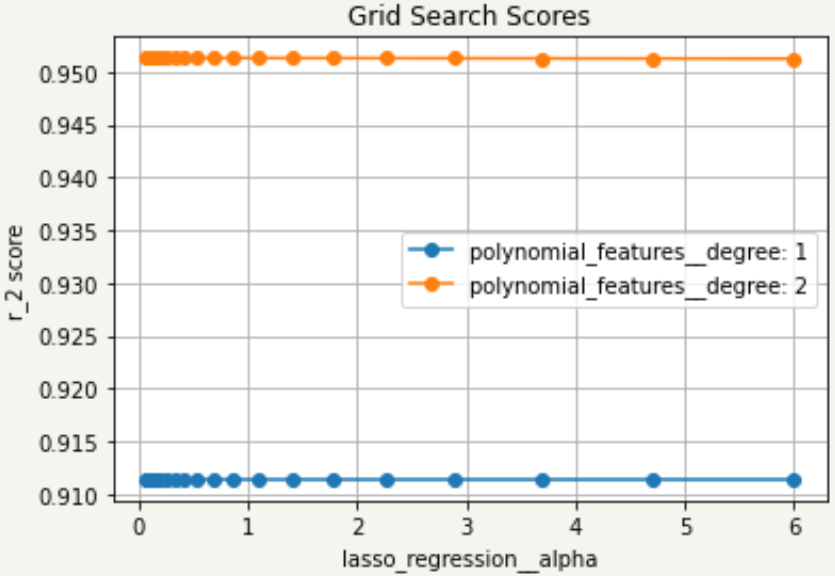


- The model from Polynomial Regression is not overfitting or underfitting
- Because the model with polynomial features are better than without polynomial features, then polynomial features will be added in Lasso Regression.

Modelling:
Lasso Regression

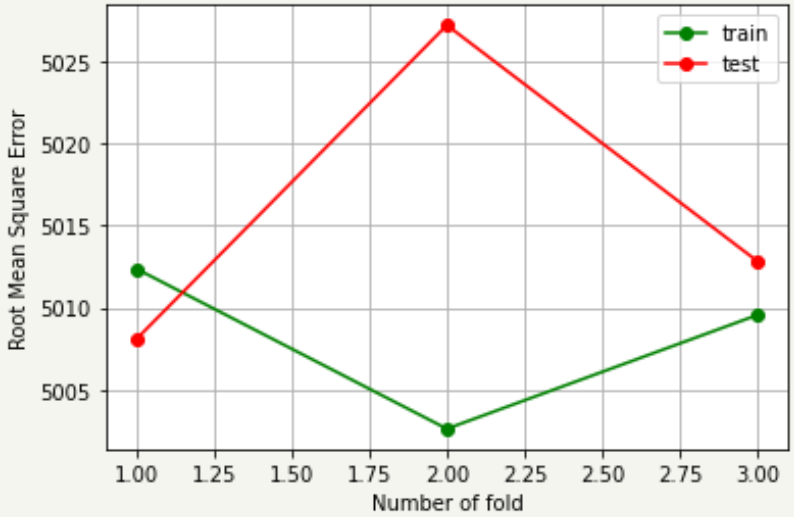
- Linear regression which performs shrinkage regularization with automatically selecting features
- Lasso has one important hyperparameter, that is alpha. To find the best alpha, I used hyperparameter tuning (GridSearchCV)

Modelling:
Lasso Regression

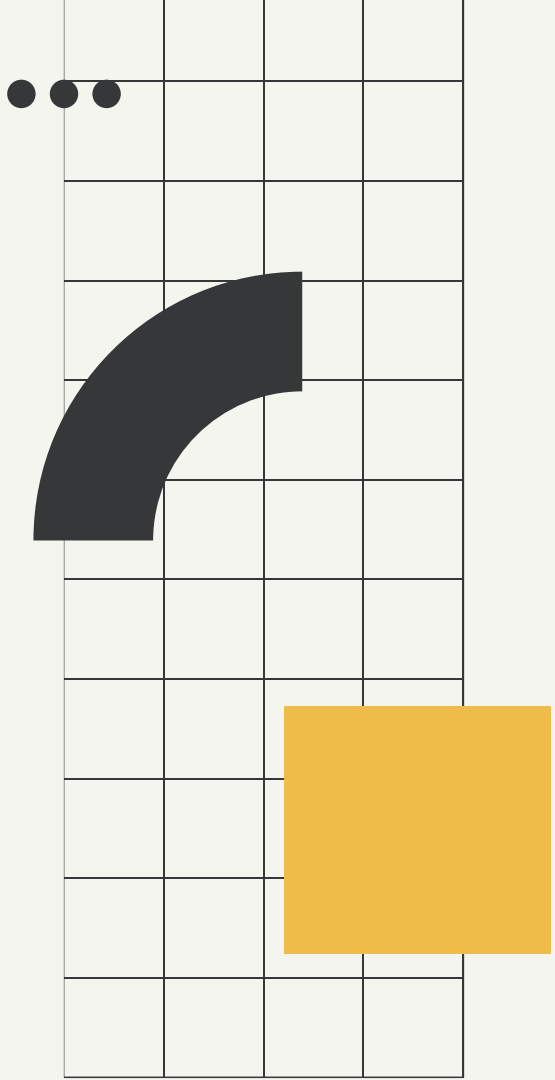


- It can be seen that degree of polynomial features = 2 has the higher R Squared Score
- The best hyperparameter: alpha = 0.20158909717702692 and polynomial_degree=2

Modelling:
Lasso Regression



The model from Lasso Regression is not overfitting or underfitting



Model Evaluation

Model Evaluation

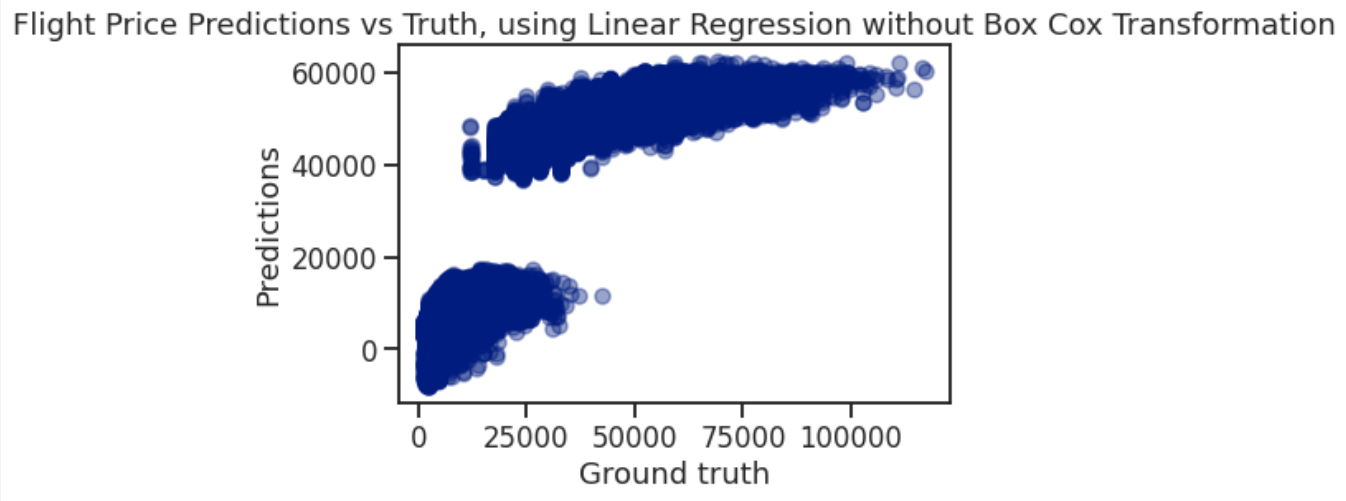
- I plot the scatter plot for every model and evaluate every model using R Squared (R2) Score.
- In addition, Root Mean Square Error (RMSE) is used to measure error between predicted and actual values.

$$R^2 = 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}},$$
$$= 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}.$$

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

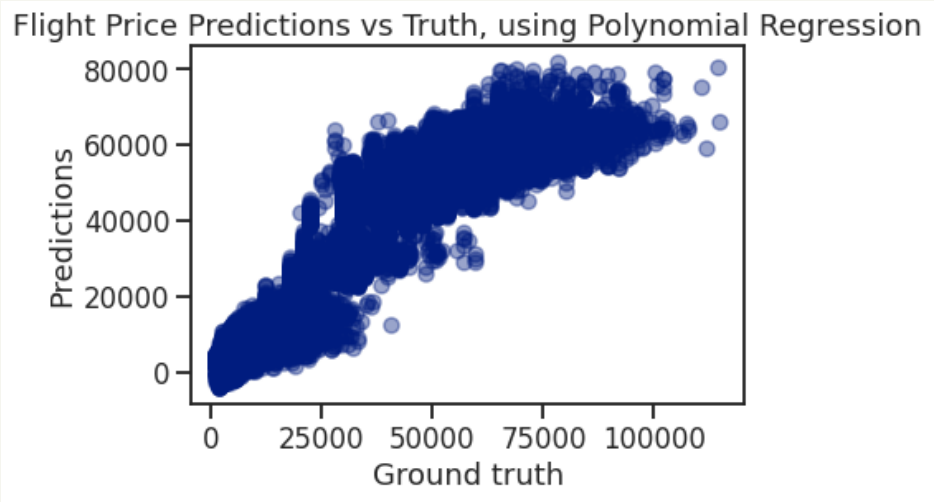
$\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ are predicted values
$y_1, y_2, \ldots, y_n$ are observed values
n is the number of observations

Model Evaluation: Multiple Linear Regression



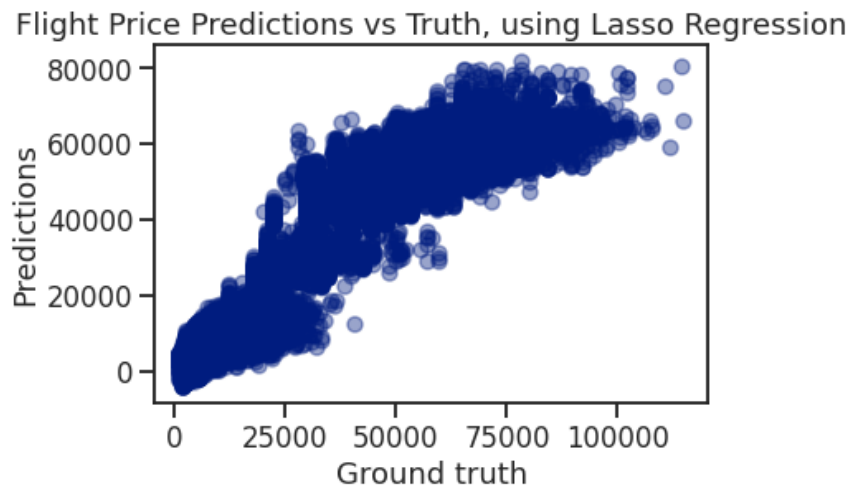
As we have seen that Multiple Linear Regression has failed to capture flight ticket price in range [0,50000] which is the lower price

Model Evaluation: Polynomial Regression



The model from Polynomial Regression is good enough to predict ticket prices.

Model Evaluation: Lasso Regression



Lasso regression model tends to similar with Polynomial Regression. It can captures model very well. However, in order to compare between Polynomial and Lasso, I need to check the evaluation between models.

Model Selection

Type	RMSE	R2 Score
Multiple Linear Regression	6774.180186	0.911250
Polynomial Regression	5011.836200	0.951421
Lasso Regression	5010.349536	0.951450

- In summary, Lasso Regression has been chosen to be the best linear regression model to predict flight ticket price because it has the lowest RMSE and highest R2 Score
- For future predicitons, Lasso model will be saved in a pickle form which is ready to be deployed

”

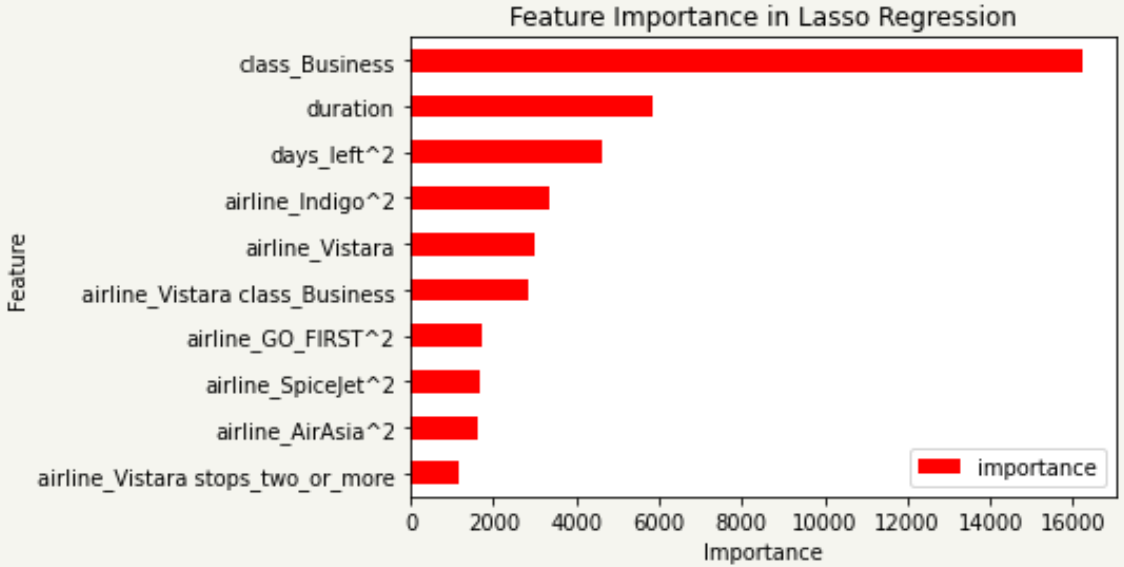
Get Insights From Selected ML Model

Insights from Lasso

- Length of features (with Polynomial Features, not Lasso) is 741 features.
- Length of features **after Feature Selection in Lasso** is 197 features.
- I have many redundant features through polynomial features which is automatically removed by Lasso for better prediction results.

Insights from Lasso

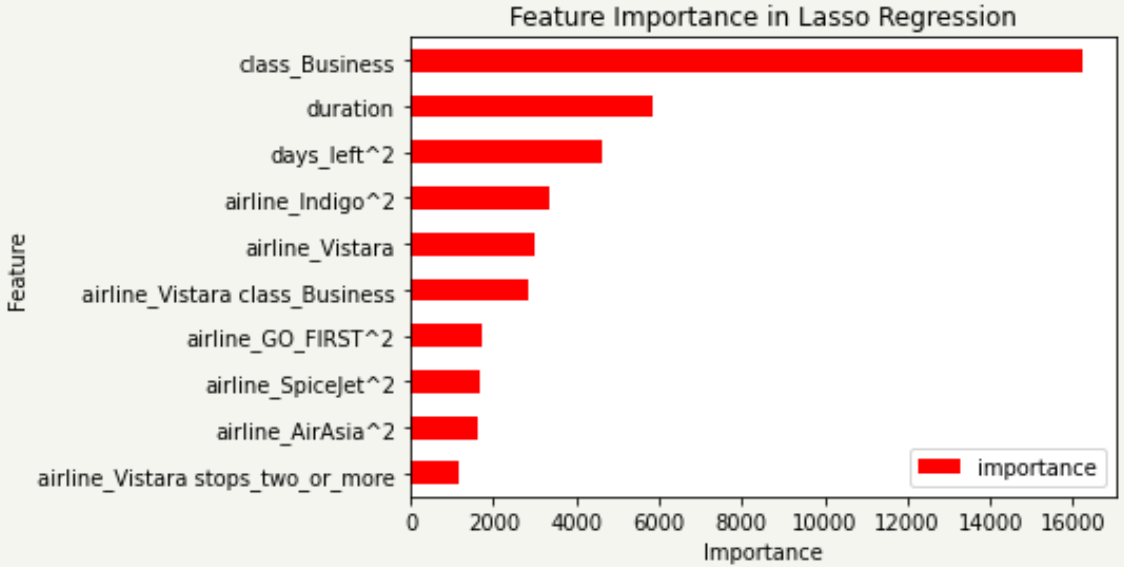
Top 10 Feature Importance for Flight Ticket Price Prediction



Business Class has been the most important feature in Lasso's prediction because it is a special class which exists only in Air India and Vistara. The chosen class determines the ticket prices in prediction.

Insights from Lasso

Top 10 Feature Importance for Flight Ticket Price Prediction



Duration is important to determine the ticket price. Therefore, customer should consider the overall amount of time it takes to travel between cities before ordering flight tickets.

Summary of Findings and Suggestions

- If passengers want cheaper tickets, they should buy around 25-30 days before departure. Buying tickets 1-2 days before departure is only for emergency
- The best departure for cheaper tickets happens when passengers choose Late Night or Early Morning as departure and arrival time
- Bangalore, Chennai, and Kolkata are the top 3 highest ticket price

Summary of Findings and Suggestions

- Delhi and Hyderabad are considered as top 2 lowest ticket price
- If passengers want to try business class in a cheaper mode, then they should choose Air India. But, if passengers choose the most likable business class with best facilities (and higher ticket price absolutely), they may choose Vistara.
- If passengers want cheaper ticket, they should choose flight higher duration between cities. The more less duration between cities, the higher ticket price should be.

Summary of Findings and Suggestions

- The standard linear regression model is severely under performing on low and high valued tickets, while the polynomial and ridge models are smoothly fit across the entire range of ticket prices.
- Lasso regression with alpha between 0 and 1 has been the best alpha for modelling by searching from hyperparameter tuning (GridSearch)
- Lasso regression tends to remove more features in using the polynomial features
- Showing the list of features the model believes are the most important in predicting the ticket price to give insights

Thank you!

https://github.com/yohset95/TicketPrice_Prediction

Credit: