Predict The Flight Ticket Price

Jyoti Dhayal
B21EE029

Abstract

This paper discusses the issue of airfare. A set of characteristics defining a typical flight is chosen for this purpose, with the assumption that these characteristics influence the price of an airline ticket. Flight ticket prices fluctuate depending on different parameters such as flight schedule, destination, and duration. As a result, having a basic understanding of flight rates before booking a vacation will undoubtedly save many individuals money and time. Analysing the dataset to get insights about the airline fare and the features of the dataset are applied to the seven different machine learning (ML) models which are used to predict airline ticket prices, and their performance is compared. The goal is to investigate the factors that determine the cost of a flight. The data can then be used to create a system that predicts flight prices.

I. Introduction

In today's world, airlines attempt to control flight ticket costs in order to maximize profits. Most people who fly regularly know the best times to buy cheap tickets. However, many customers who are not good at booking tickets fall into the discount trap set by the company, causing them to spend their money. The main goal of airline companies is to make a profit, while the customer is looking for the best purchase. Customers frequently aim to purchase tickets far in advance of the departure date in order to prevent price increases as the departure date approaches. Due to the great complexity of the fare models used by airlines, it is very difficult for a customer to buy an airline ticket at a very low price because the price is constantly fluctuating. Airlines can lower their ticket prices when they need to create a market and when tickets are harder to obtain. These tactics consider a number of financial, marketing, commercial, and social factors that are all linked to ultimate flight pricing. They might be able to get the most profit possible. As a result, costs may be influenced by various factors. The price model used by airlines is so complex that prices fluctuate constantly, making it very difficult for customers to buy tickets at very low prices. The proposed framework draws some predictions about the price of the flight based on some features such as what type of airline it is, what is the arrival time, what is the departure time, what is the duration of the flight, source, destination and more. The framework achieves a high prediction accuracy with 0.869 adjusted R squared score on the testing dataset.

A. Dataset

The file dataset.csv is used as the dataset. The train dataset contains 10683 rows where each row represents details of a flight with 11 columns containing :

- ID: Contiguous sample number
- Airline: The name of the airline
- Date of Journey: The date of the journey
- Source: The source from which the service begins
- Destination: The destination where the service ends
- Departure Time: The time when the journey starts from the source.
- Arrival Time: Time of arrival at the destination.
- Duration: Total duration of the flight.
- Total Stops: Total stops between the source and destination.
- Additional Info: Additional information about the flight
- Price: The price of the ticket

The dataset has been split into train and test with test size of 0.3

II. METHODOLOGY

A. Overview

There are various classification algorithms present out of which we shall implement the following

- Random Forest Classification
- KNN
- Logistic Regression
- SVM
- MLP
- Gaussian Naive Bayes

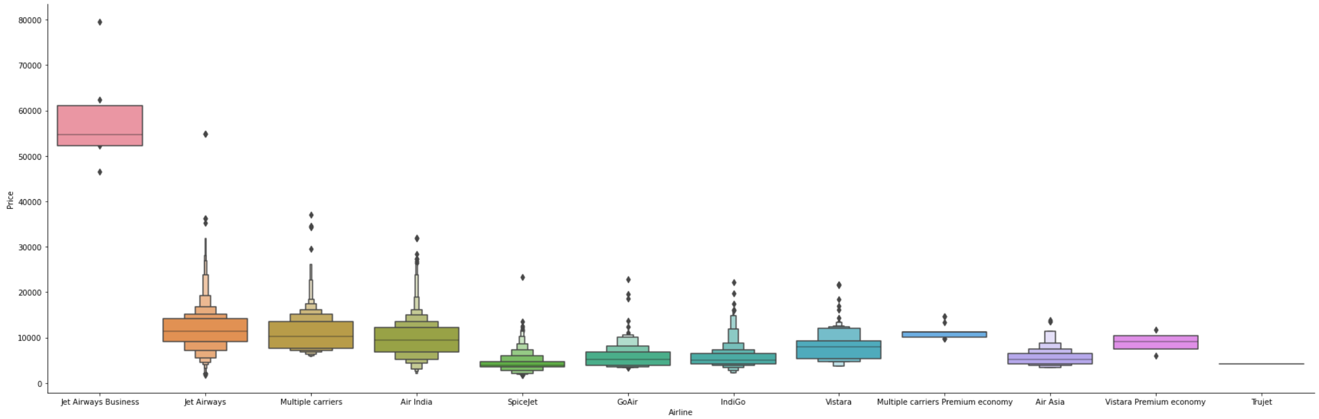
We also make use of PCA and LDA for dimensionality reduction and feature selection.

B. *Exploring the dataset and pre-processing*

The data was checked for null values and it was found that some values of routes and total stops were missing and they were filled in based upon similar data available in the dataset.

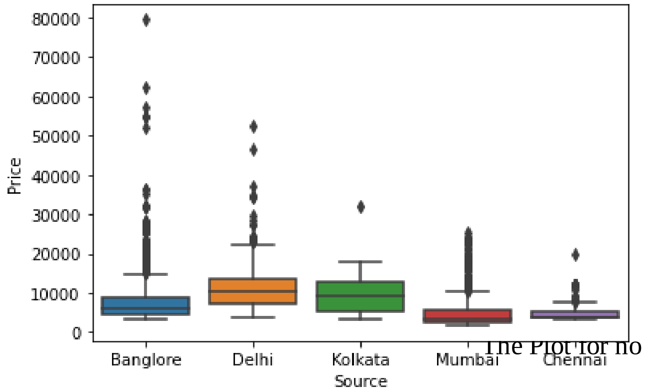
The total duration of the flights were converted from hour and minutes format to total duration in minutes of the flight.

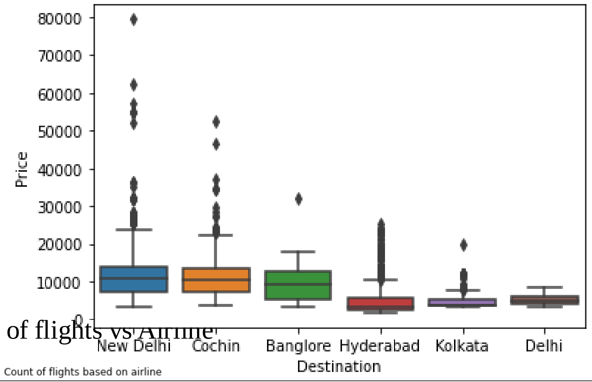
The Plot for airline vs price



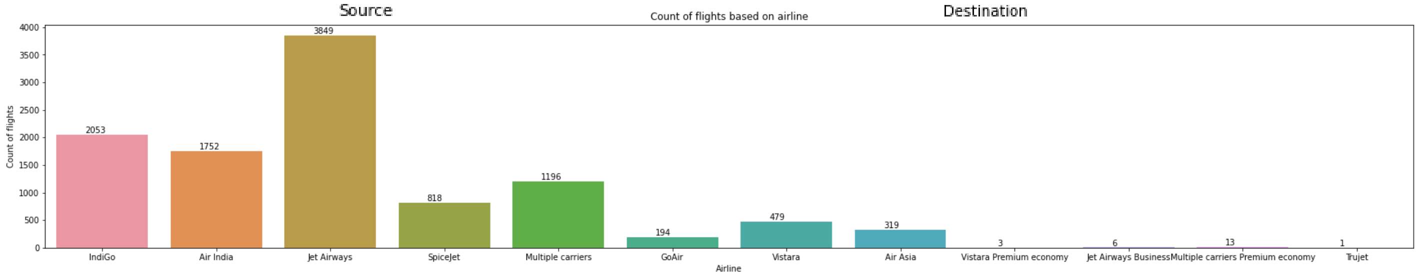
The Plot for Source of flight vs price

The Plot for Destination of flight vs price



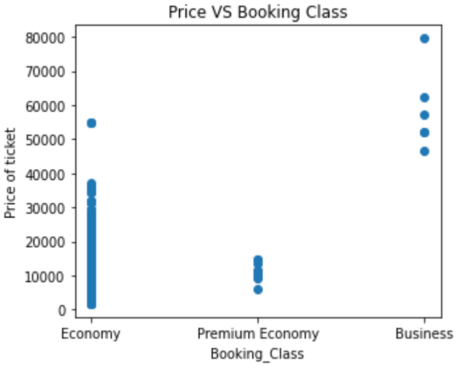


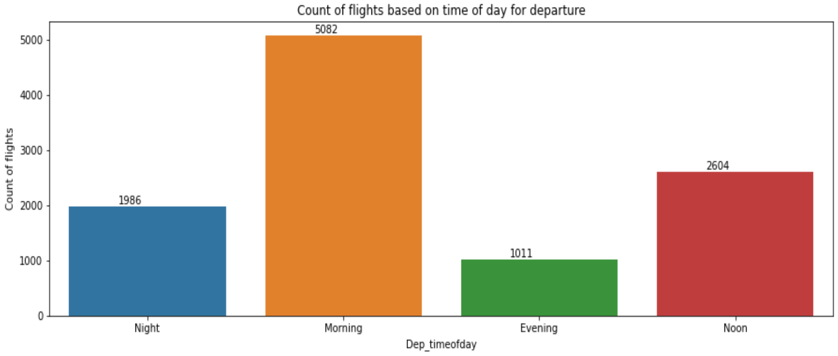
The Plot for no of flights vs Airline

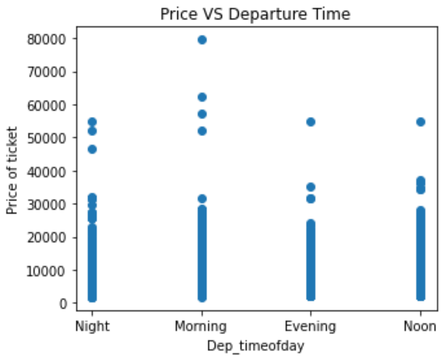


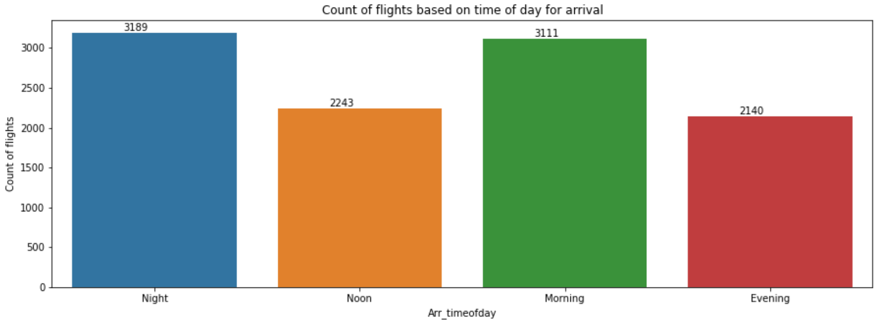
The flights were divided in different booking classes based upon the airlines companies.

The flight timings were divided into 4 sections of the day, morning, evening, noon and night.

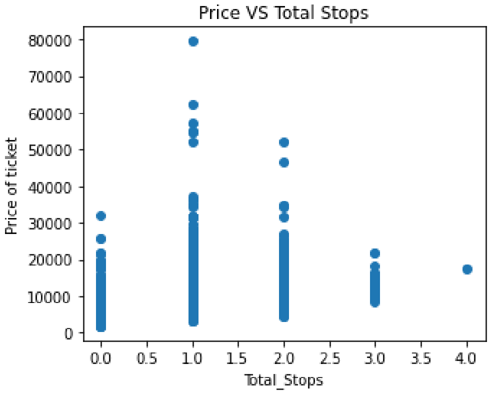








The total number of stops were converted from string format to integer format.



The categorical data was then converted to indicator variables using get_dummies function from pandas library.

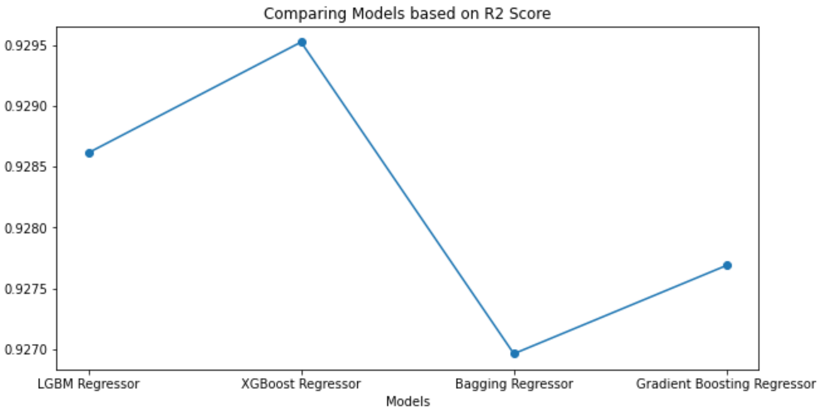
C. *Implementation of classification algorithms*

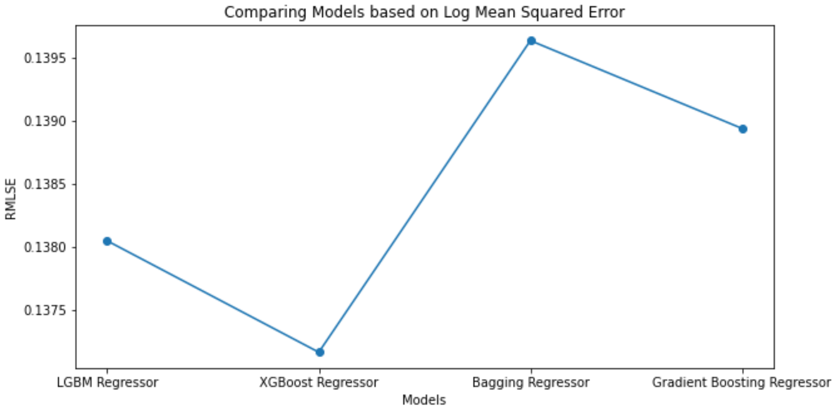
1. *LGBM Regressor:* It is a gradient boosting framework that uses tree based learning algorithms. It is designed to be distributed and efficient.
2. *XGBoost Regressor:* It is a powerful approach for building supervised regression models.The objective function contains loss function and a regularization term. It tells about the difference between actual values and predicted values, i.e how far the model results are from the real values. Here the most common loss functions in XGBoost for regression problems, reg:linear, is used.
3. *Bagging Regressor:*It fits base regressors each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction.
4. *Gradient Boosting Regressor:* In gradient boosting, each predictor corrects its predecessor's error. There is a technique called the Gradient Boosted Trees whose base learner is CART (Classification and Regression Trees).

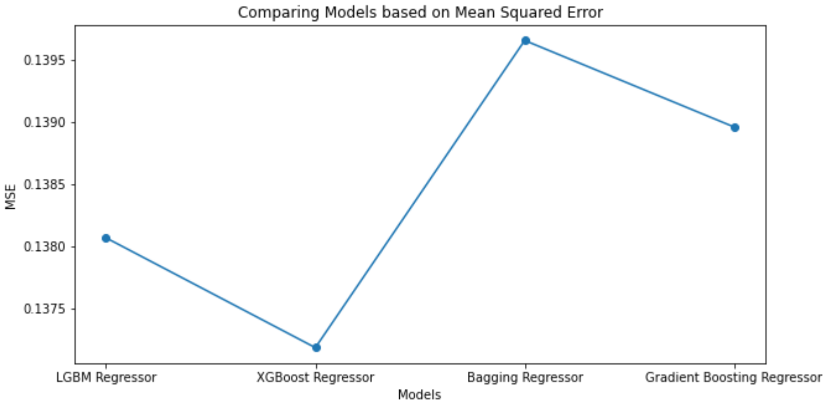
III. PREDICTIONS

Model	LGBM Regressor	XGBoost Regressor	Bagging Regressor	Gradient Boosting Regressor
RMSE	0.138071068697	0.13718817481	0.139657578546	0.13896206992
RMLSE	0.138051690985	0.13716860971	0.139638486325	0.13894206331
R2 Score	0.928613478622	0.92952352026	0.926963515270	0.92768916176

IV. EVALUATION OF MODELS







V. RESULTS AND ANALYSIS

The table shows that all classifiers had nearly equally efficient performance. Root mean squared error is 0.138 for the predictions. The root-mean-square error is used to measure the differences between values predicted by a model or an estimator and the values observed. the R2 error for the predicted values was found to be 0.92. R-Squared is also termed the standardized version of MSE. R-squared represents the fraction of variance of the actual value of the response variable captured by the regression model rather than the MSE which captures the residual error.

Therefore XGBoost model is preferred because it gives the max R2 score and the least error.