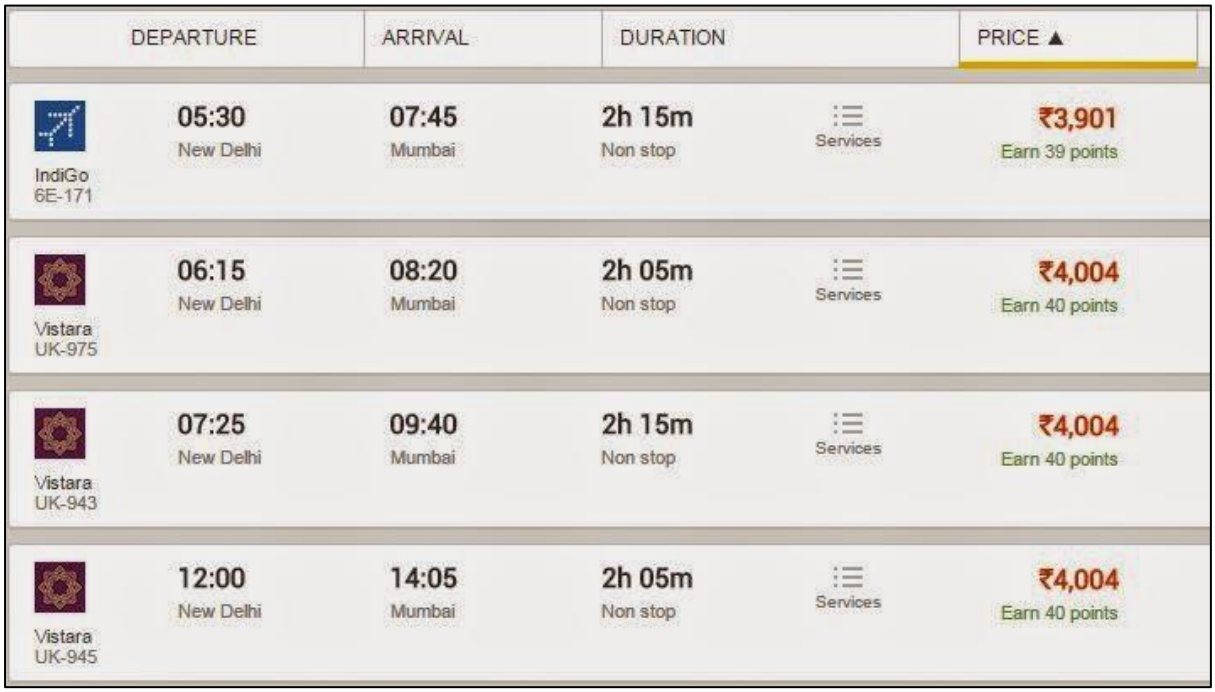
FLIGHT PRICE PREDICTION



	DEPARTURE	ARRIVAL	DURATION		PRICE ▲
IndiGo 6E-171	05:30 New Delhi	07:45 Mumbai	2h 15m Non stop	Services	₹3,901 Earn 39 points
Vistara UK-975	06:15 New Delhi	08:20 Mumbai	2h 05m Non stop	Services	₹4,004 Earn 40 points
Vistara UK-943	07:25 New Delhi	09:40 Mumbai	2h 15m Non stop	Services	₹4,004 Earn 40 points
Vistara UK-945	12:00 New Delhi	14:05 Mumbai	2h 05m Non stop	Services	₹4,004 Earn 40 points

Submitted by:

UJJWAL PRATIK

Contents

ACKNOWLEDGMENT

I sincerely thanks to the Data Trained Faculty for the guidance. They have covered the topics like Machine Language, Python & SQL. Under their guidance I learned a lot about this project. I had also taken help from YouTube & online videos.

INTRODUCTION

ABOUT FLIGHT/AIRLINES

Airline industry is one of the most sophisticated in its use of dynamic pricing strategies to maximize revenue, based on proprietary algorithms and hidden variables. Therefore, it is challenging for consumers to predict the price change in the future I. With the information of the airfare available online, buyers are trying to track the prices of the flight over a certain period and anticipate the price change in the future. However, it turns out to be rather difficult to predict the price of the flight precisely only by observation.

PROBLEM STATEMENT

Flight ticket prices can be something hard to guess, today we might see a price, check out the price of the same flight tomorrow, it will be a different story. We might have often heard travellers saying that flight ticket prices are so unpredictable. Here you will be provided with prices of flight tickets for various airlines between the months of March and June of 2019 and between various cities.

ABOUT DATASET

We have two dataset training data with 10683 records and testing data with 2671 records.

DATA ANALYSIS

COLUMNS IN THE DATAFRAME

- Airline
- Date of Journey
- Source
- Destination
- Route
- Dep Time
- Arrival time
- Duration
- Total stops
- Additional info
- Price (Target)

OBJECTIVE

- Our main objective is to predict Flight fare.
- All the parameters will be analysed through Machine Learning algorithms like Linear Regression, Lasso Regression, Ridge Regression, Elastic Net Regression etc which will help to predict flight fare.

DATA DESCRIPTION

- There are two dataset training and test data.
- The source of data is taken from GitHub.
- (https://github.com/dsrscientist/Data-Science-ML-Capstone-Projects)

METHEDOLOGY

- It gives insights of the dependency of target variables on independent variables using machine learnings techniques to determine the temperature because it gives the best outcome.

METRIC USAGE

a. Linear Regression.
b. Lasso Regression.
c. Ridge Regression.
d. Elastic Net Regression.
e. K Nearest Regressor.
f. Random Forest Regressor.

SYSTEM REQUIREMENTS

Hardware and Software Requirements and Tools Used

a) Hardware Requirement:
   i. Intel core i5
   ii. 8 GB Ram

b) Software Requirement:
   i. Python 3.x with packages:
      1. Pandas: Data analysis and manipulation tool
      2. NumPy: Provide support for mathematical functions, random number etc.
      3. Matplotlib: is a low-level graph plotting library in python that serves as a visualization.
      4. Seaborn: is a library mostly used for statistical plotting in python.
      5. Scikit-Learn: is an open-source Python library that has powerful tools for data analysis and data mining.

APPROACH

IDENTIFICATION OF POSSIBLE PROBLEM-SOLVING APPROACHES

- R2 score: is used to evaluate the performance of a linear regression mode.
- Linear Regression: Logistic regression is fast and relatively uncomplicated, and it is convenient for you to interpret the results.
- Lasso: The Lasso is a linear model that estimates sparse coefficients with l1 regularization.
- Ridge: Ridge regression is an extension of linear regression where the loss function is modified to minimize the complexity of the model.
- Elastic Net: is a linear regression model trained with both l1 and l2 -norm regularization of the coefficients.
- Cross-Validation-Score: a model that would just repeat the labels of the samples that it has just seen would have a perfect score but would fail to predict anything useful on yet-unseen data.
- Grid Search CV: This function helps to loop through predefined hyperparameters and fit your estimator (model) on your training set.
- Mean Squared Error: this metric gives an indication of how good a model fits a given dataset.
- Root Mean Squared error : is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed.
- Z-score: Z-score is also known as standard score gives us an idea of how far a data point is from the mean.
- Label Encoder: Label Encoding refers to converting the labels into numeric form.
- K Nearest Regressor: It observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.
- Standard Scaler: Standard Scaler. Standard Scaler helps to get standardized distribution, with a zero mean and standard deviation of one (unit variance).
- Random Forest Regressor: A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

TESTING OF IDENTIFIED APPROACH(Algorithms)

a. Train Test Split
b. Linear Regression
c. Lasso Regression
d. Ridge Regression
e. Elastic Net Regression
f. Grid Search CV
g. Cross Validation
h. K Nearest Regressor
i. Random Forest Regressor

KEY FOR SUCCESS IN SOLVING PROBLEM UNDER CONSIDERATION

- Analysed data for any unique values.
- Extraction information from some columns and made another variable from them.
- Analysed data for distribution.
- Caparison between two variables.
- Checked outliers through z-score method.
- checked skewness present in the dataset.
- Done Standard Scaling.
- Cross validate the r2 score from overfitting.
- Hyper Parameter tuning using Grid Search CV

CONCLUSION

This study shows that it is feasible to predict the airline ticket price based on historical data. One possible way to increase the accuracy can be combining different models after carefully studying their own performance on each individual bin. Additionally, as the learning curve indicates, adding more features will increase the accuracy of our models. However, limited by the current data source that we have, we are unable to extract more information of a particular flight. In the future, more features, such as the available seat, the departure time of a day, and whether the departure day is a holiday or not, can be added to the model to improve the performance of the predicting model.

AFTER PREDICTING RESULT ON TEST DATA

	Unnamed: 0	Fare
0	0	13312.709083
1	1	13221.519094
2	2	8685.651056
3	3	9925.388285
4	4	4867.743362
...	...	...
2666	2666	11551.064639
2667	2667	10274.497170
2668	2668	14463.182208
2669	2669	9150.427473
2670	2670	10925.677035

LEARNING OUTCOMES OF THE STUDY IN RESPECT OF DATA SCIENCE

This study gives me opportunity for lots of learning starting from various types of plotting like histograms, boxplot, scatterplot, line chart and many more graphs. These graphs helped me to analyse different aspects of data like outlier, skewness, correlation etc. It also helped me to learn how to apply various model techniques on data and enable predications.

REFRENCES

- Data trained course videos.
- Google Search
- YouTube
- GitHub
- UCI Machine learning repository