FLIGHT price prediction

Priyadarsi Das

**ACKNOWLEDGMENT**

References:

* Data Trained Course Materials / Videos / Projects
* Udemy course: Data Science Bootcamp by Jose Portilla
* Youtube Video Links:

<https://www.youtube.com/watch?v=yoLpcelanpI&t=1908s&ab_channel=KrishNaik>

<https://www.youtube.com/watch?v=6WDFfaYtN6s&list=PLZoTAELRMXVPwYGE2PXD3x0bfKnR0cJjN&ab_channel=KrishNaik>

INTRODUCTION

Business Problem Framing

Anyone who has booked a flight ticket knows how unexpectedly the prices vary. The cheapest available ticket on a given flight gets more and less expensive over time. This usually happens as an attempt to maximize revenue based on - 1. Time of purchase patterns (making sure last-minute purchases are expensive) 2. Keeping the flight as full as they want it (raising prices on a flight which is filling up in order to reduce sales and hold back inventory for those expensive last-minute expensive purchases) So, you have to work on a project where you collect data of flight fares with other features and work to make a model to predict fares of flights.

STEPS 1. Data Collection

You have to scrape at least 1500 rows of data. You can scrape more data as well, it’s up to you, More the data better the model

In this section you have to scrape the data of flights from different websites (yatra.com, skyscanner.com, official websites of airlines, etc). The number of columns for data doesn’t have limit, it’s up to you and your creativity. Generally, these columns are airline name, date of journey, source, destination, route, departure time, arrival time, duration, total stops and the target variable price. You can make changes to it, you can add or you can remove some columns, it completely depends on the website from which you are fetching the data.

2. Data Analysis

After cleaning the data, you have to do some analysis on the data. Do airfares change frequently? Do they move in small increments or in large jumps? Do they tend to go up or down over time? What is the best time to buy so that the consumer can save the most by taking the least risk? Does price increase as we get near to departure date? Is Indigo cheaper than Jet Airways? Are morning flights expensive?

3. Model Building

After collecting the data, you need to build a machine learning model. Before model building do all data pre-processing steps. Try different models with different hyper parameters and select the best model.

Follow the complete life cycle of data science. Include all the steps like

1. Data Cleaning

2. Exploratory Data Analysis

3. Data Pre-processing

4. Model Building

5. Model Evaluation

6. Selecting the best model

Analytical Problem Framing

Mathematical/ Analytical Modeling of the Problem

The following models were used on the dataset:

* GradientBoosting
* RandomForest
* ExtraTree
* XGB
* Bagging
* Huber
* ElasticNet
* BayesianRidge
* KNeighbors
* Ridge
* Lasso
* LinearRegression
* SGD
* AdaBoost
* DecisionTree
* RANSAC

Data Sources and their formats

Variables

'Departure Time',

'Travel Time',

'Stops',

'Arrival Time',

'Price',

'Airline',

'Departure Place',

'Arrival Place'

Steps

Web scrapping Data from www.makemytrip.com

Perform Exploratory Data Analysis

Handle Null Values

Check for outliers

Check co-relation of variables

Check for skewness

Find best random state

Train Test Split

Model Building, Feature Selection, Feature Extraction

Hyper-parameter Tuning

Saving the model

EDA Observations

Most of the flights are with 1 stops or 0 stops. Very few flights are with no stops.

Most of the flight tickets cost between 5000 to 10,000

Indigo has the highest number of flights, lowest number of flights are for Air India

Departure: More number of flights are from New Delhi, followed by Mumbai. Pune flights have the lowest number of flights.

Arrival: Most number of flights arrive at Mumbai followed by New Delhi. Pune has lowest number of flights.

Morning flights are costiest, evening flights are the cheapest.

Price increases with travel time and number of stops.

Price for flights arriving in the evening is high.

Air India has the highest price followed by Indigo.

ExpressJet has the lowest pice followed by Air India.

Departure: Pune flights are costliest followed by New Delhi.

Departure: Mumbai flights are cheapest followed by Chennai.

Arrival: Madurai flights are costliest followed by Coimbatore.

Arrival: Mumbai flights are cheapest followed by Chennai.

Hardware, Software and Libraries Used for the project

Hardware used – Intel i5 Laptop, 12 GB RAM, 2 GB Graphics, 240 GB SSD

Software used – Anaconda Jupyter Notebook

Packages used – numpy, pandas, matplotlib, seaborn, sklearn, scipy

Model/s Development and Evaluation

Model Building without Pipelines - Results

Linear Regression

R2 Score: 0.40

It is seen that applying lasso does not improve the score

Catboost Regressor

R2 Score 0.54

Decision Tree Regressor

R2 Score: 0.67

----------------------------------------------------

Hyperparameter tuning on Catboost Regressor gives us

R2 Score: 0.79

Model Building with Pipelines – Results

Standard Scaler + Different Models gives the following results:

('RandomForest', 0.7737133155510799),

('XGB', 0.7721565966304781),

('Bagging', 0.768521813976206),

('ExtraTree', 0.7652555880472789),

('DecisionTree', 0.7366634098728845),

('KNeighbors', 0.7285571834918082),

('GradientBoosting', 0.7146707134287512),

('AdaBoost', 0.5342314187180923),

('Lasso', 0.35535718041359826),

('BayesianRidge', 0.35533442615845495),

('Ridge', 0.35509277295127145),

('SGD', 0.35505771094734817),

('LinearRegression', 0.35494103981942765),

('Huber', 0.3476090221928619),

('ElasticNet', 0.33525059588248646),

('RANSAC', -1.5679017564873194)

Applying Feature Selection

SelectKBest and f\_regression is used to select 5 best parameters

* ('RandomForest', 0.7226028304238262),
* ('XGB', 0.7196365403688934),
* ('Bagging', 0.7190536716896716),
* ('ExtraTree', 0.7171207711315791),
* ('DecisionTree', 0.6660167591680335),
* ('KNeighbors', 0.6577487250984307),
* ('GradientBoosting', 0.5951089692825902),
* ('AdaBoost', 0.47507606618230164),
* ('Lasso', 0.32226999333253586),
* ('BayesianRidge', 0.3222148887349419),
* ('Ridge', 0.32190115146847426),
* ('LinearRegression', 0.3217400746156712),
* ('SGD', 0.3214719656793145),
* ('Huber', 0.31585986886787304),
* ('ElasticNet', 0.3082572973295982),
* ('RANSAC', -0.9887182671427327)

Applying Feature Extraction

Standard Scaler and PCA are used to perform Feature Extraction

('RandomForest', 0.7694795730293345),

('XGB', 0.7669908088760066),

('ExtraTree', 0.7641640308254056),

('Bagging', 0.75990267570043),

('DecisionTree', 0.7327265753845841),

('KNeighbors', 0.7246983217598224),

('GradientBoosting', 0.7093315823998628),

('AdaBoost', 0.5168338066329536),

('BayesianRidge', 0.3553760069061956),

('Lasso', 0.35537570670271235),

('Ridge', 0.35537457677369233),

('LinearRegression', 0.3553740893800761),

('SGD', 0.35488732875976003),

('Huber', 0.3480517004642746),

('ElasticNet', 0.31621858983520335),

('RANSAC', -0.6812128283845481)

Hyper parameter Tuning of Final Models

Models and R2 Score with GridSearchCV

'RandomForest', 0.77

**'GradientBoosting', 0.795**

‘Random Forest’ : 0.77

‘XGB’ : 0.79

Conclusion

Base models with pipelines will all features gives better results than applying Feature Selection & Feature Extraction.

Gradient Boosting is selected as the final model at R2 Score of 79.5 with Hyper-parameter tuning.