Web Scraping - Airline Price Analysis

Sri R Sankaranarayanan
DSC680 | Applied Data Science | Project 2 | Milestone 1
https://github.com/rengsankar1986/DSC680

Topic :-

This project domain is Airline Transportation. The Objective of this project is to perform Web Scraping on either of Kayak.com and expedia.com or both to predict flight prices. My target variable will be the price (continuous numeric value).

I am planning to execute the Python Script every few hours, compare prices since the inception of executing the script and email me the results in frequency.

Business Problem :-

Airline companies use complex algorithms to calculate flight prices given various conditions present at t2hat particular time. These methods take financial, marketing, and various social factors into account to predict flight prices. It is difficult for airlines to maintain prices since prices change dynamically due to different conditions. That's why we will try to use machine learning to solve this problem. This can help airlines by predicting what prices they can maintain. It can also help customers to predict future flight prices and plan their journey accordingly.

Datasets :-

I will be scraping through Expedia.com or kayak.com websites for current pricing data in real-time (web scraping technique explained below), below will be the datasets that I am going to be procuring in real-time.

Airline: So this column will have all the types of airlines like Indigo, Jet Airways, Air India, and many more.

Date_of_Journey: This column will let us know about the date on which the passenger's journey will start.

Source: This column holds the name of the place from where the passenger's journey will start.

Destination: This column holds the name of the place to where passengers wanted to travel.

Route: Here we can know about that what is the route through which passengers have opted to travel from his/her source to their destination.

Arrival_Time: Arrival time is when the passenger will reach his/her destination.

Duration: Duration is the whole period that a flight will take to complete its journey from source to destination.

Total_Stops: This will let us know in how many places flights will stop there for the flight in the whole journey.

Additional_Info: In this column, we will get information about food, kind of food, and other amenities.

Price: Price of the flight for a complete journey including all the expenses before onboarding.

Data Science Methods :-

a. I would be scraping data using Python Selenium. Below are the steps for procuring real-time JSON dataset from Expedia.com

   1. Construct the URL of the search results from Expedia- Here is one for the available flights listed from Dallas to Chennai, India – https://www.expedia.com/Flights-Search?flight-type=on&mode=search&trip=oneway&leg1=from%3ADallas+%28DFW+-+Dallas-Fort+Worth+Intl.%29%2Cto%3AChennai+%28MAA+-+Chennai+Intl.%29%2Cdeparture%3A8%2F16%2F2022TANYT&options=cabinclass%3Aeconomy&passengers=children%3A2%5B6%3B2%5D%2Cadults%3A2%2Cseniors%3A0%2Cinfantinlap%3AY&fromDate=8%2F16%2F2022&d1=2022-08-16
   2. Download HTML of the search result page using Python Requests.
   3. Parse the page using LXML – LXML lets you navigate the HTML Tree Structure using Xpaths. We have predefined the XPaths for the details we need in the code.
   4. Save the data to a JSON file.

b. Importing the training data
c. Data pre-processing
d. Importing test data and data pre-processing
e. Creating Independent and Dependent variables
f. Features Selections
g. Model Building
h. Hyperparameter tuning

Ethical Considerations:-

I am presenting my user-agent string in the Expedia or Kayak website when procuring datasets so the website administrator understands the intention of the Python requests. We will not import or use any personal data and have already confirmed with the websites that it is allowed for Web Scraping. The price data is publicly available data, hence we will use that as our target variable. The rate of Web scraping allowed will be taken into consideration when running the requests.

Challenges/Issues:-

Answering bot captcha challenge questions may become challenge in using sites like Expedia or Kayak. I have other alternative sites like Skyscanner.com or momondo.com if the primary sites do not allow Web Scraping.

References :-

https://towardsdatascience.com/how-to-scrape-flight-prices-with-python-using-selenium-a8382a70d5d6 --> Web Scraping with Python using Selenium.

https://www.analyticsvidhya.com/blog/2021/06/flight-price-prediction-a-regression-analysis-using-lazy-prediction/ → Lazy prediction Regression Analysis

https://www.analyticsvidhya.com/blog/2022/01/flight-fare-prediction-using-machine-learning/ → I will be performing EDA for getting data insights, feature engineering, Data Visualization and then followed by ML algorithms.

https://medium.com/mlearning-ai/flight-fare-predictions-b1150fdc45a3 --> Good comparative case study similar to my project.

https://www.kaggle.com/datasets/nikhilmittal/flight-fare-prediction-mh --> Sample Dataset will be used from Kaggle.com for performing initial Data Visualization.

https://medium.com/@fneves/if-you-like-to-travel-let-python-help-you-scrape-the-best-fares-5a1f26213086 --> Good Case study on how Web Scraping is done.

https://www.altexsoft.com/blog/flight-price-predictor/ → Training models to find the best priced flight details per itinerary.

https://www.eff.org/document/preliminary-injunction-american-airlines-v-farechase-inc --> Ethical considerations

https://www.zyte.com/learn/web-scraping-best-practices/ → Ethical usage using Web Scraping best practices.

https://www.todaysoftmag.com/article/3288/ethical-and-legal-considerations-in-web-scraping-at-scale --> Ethical challenges in Web Scraping understood.