Dashboard / ALB Datathletes Home / DatAthletes 4.0 (June 2020) 🔒 📎

✏ Edit   ☆ Save for later   👁 Watch   ⤴ Share   ⋯

📘 Confidential & restricted

Challenge

Created by Ramesh Kumar MANICKAM, last modified by Divyam SINHA on Jul 01, 2020


Overview

This version of Datathletes comes with a fresh challenge and great opportunity to solve a problem whose solution can be productionized.

Gebeco based in Kiel, is a tour operator for study and adventure travel. See an introduction of the Gebeco business by the account manager @ Unknown User (jjahnke)

Gebeco sells packaged trips , sometimes they sell product up to 2 years in advance without actually booking the ticket, as GDS only allows 1 year in advance purchase. So in that case they want to optimize the booking date to buy the ticket at the lowest price.
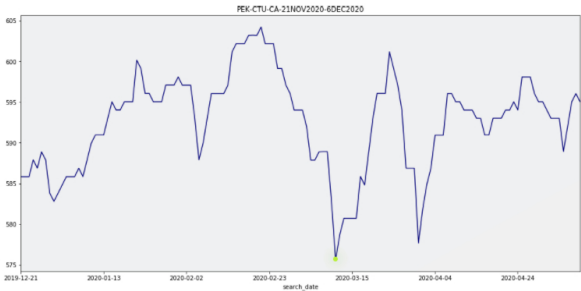
In above example it can be seen that for a cell (origin:PEK , destination:CTU , airline:CA , departure date:21-NOV-2020 , return date:06-DEC-2020) the best time to book the ticket is 11-MAR-2020 (highlighted by green)

We have recorded prices every day over the past months for 53324 cells. The problem we want to solve is finding, for each cell, the best date to book. Given past prices of each cell between the 20-DEC2019 and 10-APR2020 what will be the best date to book between 11-APR-2020 and 04-JUN-2020?

Data: Past prices

For each cell (I.e. every Origin-Destination-Airline-DepartureDate-ReturnDate) prices are provided for search dates between 20-DEC-2019 and 10-APR-2020 (there might be some missing as in any real data set).

- orig - origin airport
- dest - destination airport
- airline - carrier code
- dep_date - departure date of trip (YYYY-MM-DD)
- ret_date - return date of trip (YYYY-MM-DD)
- search_date - date on which search was made (YYYY-MM-DD)
- price - price of ticket in euros

That makes around 110 rows per cell (one per search date)

orig	dest	airline	dep_date	ret_date	search_date	price
ADD	NBO	ET	2020-06-24	2020-07-07	2019-12-21	230.34
ADD	NBO	ET	2020-06-24	2020-07-07	2019-12-22	230.34
ADD	NBO	ET	2020-06-24	2020-07-07	2019-12-23	230.34
ADD	NBO	ET	2020-06-24	2020-07-07	2019-12-24	232.05
ADD	NBO	ET	2020-06-24	2020-07-07	2019-12-25	231.90
ADD	NBO	ET	2020-07-01	2020-07-14	2019-12-21	230.34
ADD	NBO	ET	2020-07-01	2020-07-14	2019-12-22	230.34
ADD	NBO	ET	2020-07-01	2020-07-14	2019-12-23	230.34
ADD	NBO	ET	2020-07-01	2020-07-14	2019-12-24	232.05
ADD	NBO	ET	2020-07-01	2020-07-14	2019-12-25	231.90
ADD	NBO	ET	2020-07-01	2020-07-14	2019-12-27	230.85
ADD	NBO	ET	2020-07-01	2020-07-14	2019-12-29	230.77
ADD	NBO	ET	2020-07-01	2020-07-14	2019-12-31	229.72
ADD	NBO	ET	2020-07-01	2020-07-14	2020-01-01	203.36
ADD	NBO	ET	2020-07-08	2020-07-21	2019-12-21	230.34
ADD	NBO	ET	2020-07-08	2020-07-21	2019-12-22	230.34
ADD	NBO	ET	2020-07-08	2020-07-21	2019-12-23	230.34
ADD	NBO	ET	2020-07-08	2020-07-21	2019-12-24	232.05
ADD	NBO	ET	2020-07-08	2020-07-21	2019-12-25	231.90
ADD	NBO	ET	2020-07-08	2020-07-21	2019-12-27	230.85
ADD	NBO	ET	2020-07-08	2020-07-21	2019-12-29	230.77
ADD	NBO	ET	2020-07-08	2020-07-21	2019-12-31	203.72
ADD	NBO	ET	2020-07-08	2020-07-21	2020-01-01	203.36
ADD	NBO	ET	2020-07-08	2020-07-21	2020-01-02	203.16
ADD	NBO	ET	2020-07-08	2020-07-21	2020-01-03	203.24

Here are the data:

https://amadeusworkplace-my.sharepoint.com/:u:/g/personal/divyam_sinha_amadeus_com/EQWOpnTbCU1MhPFZlryHjpcBSubwggF7m5yqix9AfM3TLQ?e=jcw8LR
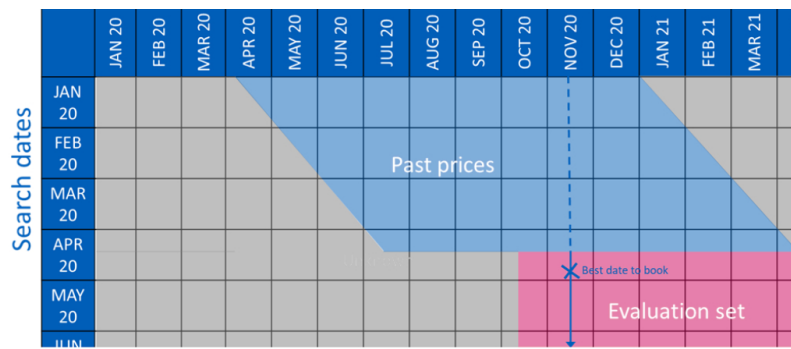
Note : Size of the data was a bit large to be handled on Confluence page (52mb) , please download the file from the above link

Goal

Let's say we are on 11-APR-2020 and that we have to book a ticket for every cell (I.e. every Origin-Destination-Airline-DepartureDate-ReturnDate) before the 04-JUN-2020. What is the optimal booking dates for each cell?

FRA-IST-TK   Departure dates

(this schema represents the set of prices for Origin:FRA , Destination:IST , Airline:TK. It doesn't show the return date dimension)

Result submission

For each cell (I.e. every Origin-Destination-Airline-DepartureDate-ReturnDate) the **best booking date** should be provided, i.e. the search date with the lowest expected price. The best booking date must be **between the 11-APR-2020 and the 04-JUN-2020**.

This is format to be used to submit results:

- orig - origin airport
- dest - destination airport
- airline - carrier code
- dep_date - departure date of trip   (YYYY-MM-DD)
- ret_date - return date of trip   (YYYY-MM-DD)
- best_booking_date - date predicted to be best time to book (YYYY-MM-DD)
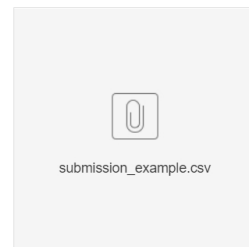
Example:

orig	dest	airline	dep_date	ret_date	best_booking_date
PEK	CTU	CA	2020-12-07	2020-12-22	**2020-04-15**
PEK	CTU	CA	2020-12-13	2020-12-29	**2020-05-05**

Cells to predict:

We do not expect prediction on all the cells that are in the past_prices file, only some of them. Hereafter is a sample submission file. It contains the list of cells for which we expect a prediction of the best booking date. It has the best booking date as empty (last column). You can replace the last column with your own prediction.

Note: when selecting the cells to be predicted we especially focused on departure dates far in the future: prices' evolutions are less erratic in the far future and the customer would not want to wait the last minute to book the tickets he sold a long time ago. Cells to be predicted only have departure date after 09OCT20.

submission_example.csv

Reminders

- Date format should always be YYYY-MM-DD
- The best date to book should be within the range 2020-04-11 to 2020-06-04
- There should be one row per cell: the number of rows should be same as in the submission example. We expect predictions for all the cells in the above file but not for other cells (even if they are available in the past_prices file).
- Please avoid putting some dummy dates in file (only generic model output is considered as a valid submission)

Evaluation

Thanks to the prices from 11-APR-2020 to 04-JUN-2020 we will evaluate the accuracy of the best booking date predictions.

Evaluation metrics will be **Mean Percentage Revenue Loss** incurred by Gebeco which will evaluate how close one has predicted best time to book.

Mean Percentage Revenue Loss = Mean(((|Price at predicted best time to book - Price at actual best time to book|)/Price at actual best time to book)*100)

So it doesn't necessarily matters to be far from the optimal date if the price at that date is closed to the lowest price.

Note: We may have data missing for days. In that case the considered predicted best time to book will be the closest day for which data is recorded. It is transparent to the participants. For example if the predicted best date to book for a given cell provided by a participant is 08-MAY-2020 but we don't have recorded the price for that date (because of some failure in data collection job) we will convert the date to the 07-MAY-2020 and estimate the results accordingly. For sure we collected nothing (for any cell) from 08 to 10-MAY-2020 and there might some other missing data for some cells here and there.

Max. ~~25~~ 30 submissions are allowed now.

Final evaluation

Top 10 teams on leader board by end of competition will be presenting the solution overview to the esteemed jury .

With further objective evaluation winners will be announced

Guidelines for Jury round :

1. Top 10 on final Leaderboard need to submit the code (that generated the best score) through email to DG-BLR-DATATHLETES-Coreteam@amadeus.com by the same submission   timeline : 6th July – 5 PM (GMT)

2. Jury round constitutes of 10-15 minutes for each team (Code Walkthrough/Solution Overview and Q&A) and subsequent validations/demo/Q&A runs as needed. Please prepare your   presentation of approaches .

3. An email communication will be sent out to the Top teams for their respective slots (8th-9th July) . Requesting availability & flexibility from Top Teams.

4. Final winners will be decided by Jury after Validation !!

No labels 🏷️

1 Comment

Marek VAN LEEUWEN
For your interest, in Travel Channels, beside Amadeus' own "FareOptimizer" product, the Online segment is working on a partnership with Whizar in Israel, offering a similar production ready solution for optimizing the booking date.

Reply   Like   4 people like this   Jul 17, 2020

Write a comment...

▲ ATLASSIAN