OTB Data Science Interview Tasks: Autumn, 2018

Description:

You are given a data set from a simple simulation of an online travel company's web-traffic and sales performance.

The company is offering different holiday product types, such as beach hotels, city break hotels, cottages, flight-only bookings, hotel-plus-flight packages, luxury holiday packages.

There is a dedicated web page for each product and the associated traffic and performance for each product is recorded.

The company also has historic records of its online and TV advertising activities, along with records of external economy- related and weather-related data.

In this simulation, the progression (evolution) of traffic and sale volumes of each product type is modelled in a type-specific way. This progression (evolution) is influenced by company's advertising activities and by external macro-economic and weather factors.

Note: the structure of each model used to generate traffic and sales volumes for different product types is identical across product types.

Tasks:

Using either the Python or R (or both) computer languages (Matlab is accepted, though less desirable) you should complete the following three tasks:

1. Your most important task is to use the records of historic traffic (sessions) and sales (bookings) volumes for each product (in the file otb_interview_task__train__product_time_series.csv) along with the records of macro-economic , weather, and advertising indices (in the file otb_interview_task__train__meta_time_series.csv) to build your own tools and models to explain what impact the latter (macro-economic , weather, and advertising indices) have on the traffic and sales volumes of each product type.
   In the next stage of the interview, you will be asked to present verbally your findings and there will be a discussion about the tools, algorithms, and models you have used and/or developed. Therefore, you should bring along your code, as well.

2. Implicit in the previous task is the assumption that you should be able to identify the product types (groups) that were simulated and be able to match each product to its corresponding type.
   This product-to-type mapping is not provided to you. Also, there is no information available to you about what the product types (groups) or how many there are.
   It is one of your tasks to use the series provided in the training data sets along with the models and tools you have built to identify the number of distinct product types and of what

type each product is. That is, you are expected to fill in the `product_type` column in the `otb_interview_task__test__product_groupings.csv` file.
An example of a submission file is given, as well (the file `otb_interview_task__SAMPLE_SUBMISSION__product_groupings.csv`)

3. Your final task is to use the tools you have built and the insights you have gained to forecast future traffic and sales volumes.
   As part of this you will have to fill in the `sessions` and the `bookings` columns in the file `otb_interview_task__test__product_time_series.csv`.
   Again, an example of a submission file is given in `otb_interview_task__SAMPLE_SUBMISSION__product_time_series.csv`.

Data:

The following is a list of the CSV files included in the package, along with a short description of the columns (fields) in each file.

`otb_interview_task__train__meta_time_series.csv:` This file holds historic records of various macro-economic, weather, and advertising activity indices (all data is simulated, not real). It has the following columns:

`Date:` The date of records.

`Consumer Confidence Index:` Captures how consumers in the country feel about the state of the economy.
The values are relative to a base period. For example *1.0* means it is the same as the value in the base period, while *0.95* means it is *5%* lower than the value at the base period. Thus, all values should be bigger than or equal to zero. Negative values will be rare, but if you find any, you should make them zeros.

`Exchange Rate:` A currency basket index.
The values are relative to a base period. For example *1.0* means it is the same as the value in the base period, while *0.95* means it is *5%* lower than the value at the base period. Thus, all values should be bigger than or equal to zero. Negative values will be rare, but if you find any, you should make them zeros.

`Online Visibility:` An indicator of the daily average position of company's ads in the search results of major online search engines. A higher value indicates a higher position and thus a better visibility.
The values are relative to a base period. For example *1.0* means it is the same as the value in the base period, while *0.95* means it is *5%* lower than the value at the base period. Thus, all values should be greater than or equal to zero. Negative values will be rare, but if you find any, you should make them zeros.

`TV Ad Reach:` An indicator of whether there were any TV ads running on a given day and of how many people saw the ads. These are scaled values of the actual numbers. All values should be bigger than or equal to zero. Negative values will be rare, but if you find any, you should make them zeros.

Weather Index: An indicator of how the weather compares to the historic seasonal average, usually in terms of temperature and cloud density. For example, "*average*" means that the temperature and the cloud density where the typical values for the season while "*better than average*" indicates both that the temperature was significantly closer to the ideal 20-23 degrees Celsius than the typical seasonal value and that there were fewer clouds than usually observed for this season (and the opposite for "*worse than average*").

otb_interview_task__train__product_time_series.csv: This file holds historic records of traffic and sales volumes for each product. It has the following columns:

product_id: A unique identifier for each product.

date: The date of records.

sessions: The number of total sessions (visits to the web-site) on the given date for the given product.

bookings: The number of total booking (purchases) on the given date for the given product.

otb_interview_task__test__product_groupings.csv: This is one of the files that you need to fill in, as described in the *Tasks* part above. It has the following columns:

product_id: A unique identifier for each product.

product_type: An empty column that should be filled in with the estimated product types.

otb_interview_task__SAMPLE_SUBMISSION__product_groupings.csv: This file shows how the file above should look like. It has the following columns:

product_id: A unique identifier for each product.

product_type: Estimated product types (randomly generated).

otb_interview_task__test__product_time_series.csv: This is one of the files that you need to fill in, as described in the *Tasks* part above. It has the following columns:

product_id: A unique identifier for each product.

date: The date of records.

`sessions:` An empty column that should be filled in with the forecasted number of total sessions (visits to the web-site) on the given date for the given product.

`bookings:` An empty column that should be filled in with the forecasted number of total bookings (purchases) on the given date for the given product.

otb_interview_task__SAMPLE_SUBMISSION__product_time_series.csv:
This file shows how the file above should look like. It has the following columns:

`product_id:` A unique identifier for each product.

`date:` The date of records.

`sessions:` Forecasted number of total sessions (visits to the web-side) on the given date for the given product (randomly generated).

`bookings:` Forecasted number of total bookings (purchases) on the given date for the given product (randomly generated).

Deadline:

One week (7 calendar days) after receiving the task description and data.

Evaluation:

The accuracy of traffic and sales volume forecasts and product groupings is far less important than the ability to tell how the macro-economic, weather, and advertisement indices impact these volumes.

The major thing we would like to measure is your ability to analyse the presented data and gain quantifiable insights from it. The forecasts are used as a way of objectively measuring the validity of the insights presented.

We also expect that you have sufficient experience in building data-driven mathematical and computational models of real-life systems and in manipulating large sets of data. Therefore, we would like to see the written source code and test them live in the interview meeting.