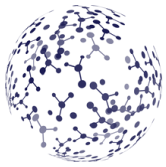

DATASCIENCEGO VIRTUAL

Hackathon Case

The Challenge

You have been hired by the U.S.Environmental Protection Agency (EPA) to analyze data from different air quality monitoring sites in California. The EPA wants to help scientists and the government understand the air quality in protected ecosystems that are sensitive to pollution. To this end, they have set up remote sensing stations and have collected a number of pollution measurements. These measurements provide information on the level of risk that environmental damage may occur.

Your job is to understand the data, gather insights about what is happening over time. You may want to focus on an in-depth analysis of the relationships between the measures, the effects over time or between sites. You are also free to explore the data and generate any meaningful insight that will aid scientists and the government into understanding what is happening with these ecosystems' air quality. The important part is that you tell a story that is actionable and meaningful.

As part of storytelling, you can create a visualization that communicates a message about what is happening with the air quality. The visualization can be a figure, set of figures, animation or interactive graphic. While these should be visually appealing, it is more important that it communicates a message to its audience.

The EPA has identified that the total nitrate (NO3) is a metric that they are interested in studying. However, some sites lose this information from time to time. They are interested in a predictive model that well allow them to estimate the nitrate values when they are missing.

While you are free to complete all three tasks, but that is not what is expected. You can select a single task or multiple tasks. Ideally, your team will produce a story that provides insights into the data or a model that will address the issue of the missing nitrate values.

The EPA values collaboration as part of its standard work process. It recognizes that individuals can contribute great value to other teams that are working on different projects. This cross-team work generally results in significant improvements in the work product that is delivered and the value that the project has to the organization. Thus, there is an award for the person that is most helpful to other teams.

About the data

This hackathon provides a dataset which should be used as the basis of your analysis. However, this is an open data hackathon meaning that any publicly available data can be used to augment the main dataset.

Description of the data:

Air quality sites have been established in nine locations in California, US. These sites are located in or near rural areas that have sensitive ecosystems. Data is collected on ambient levels of pollutants where urban influences are minimal. The data collection was initiated in 1986 but not in all locations. The monitoring provides the data that is needed to assess and report on geographic patterns and long-term temporal trends in ambient air pollution and dry atmospheric deposition.

The data was generally collected every hour and it has been aggregated and it is reported as the average over the week. This is real world data and therefore, there are missing observations for extended periods of time for some sites. The file air_status.csv contains the observations. Each row is uniquely identified as by the SITE_ID and DATEON and contains all the pollutant measurements and other features that were recorded for the week starting on the date give in DATEON at the site that is identified by SITE_ID.

A data dictionary (see the file data_dictionary.csv) is provided. It defines the features and a description of what they measure and their units, if applicable. The file site.csv contains the SITE_ID and a label that describes the actual location.

Several columns in the air_status.csv file contain encoded information. These would be such things as standardize comments, temperature source information, QA codes, etc.. The file codes.csv provides a dictionary of how to interpret these values.

Predictions:

There is no requirement that your team has to submit a predictive model. However, if you choose to do so, you will be predicting the value of TOTAL_NO3 for a given SITE_ID and DATEON. The file test.csv is to be used to make predictions. Your model can supplement this data as you see fit. The predictions are to be submitted in a single CSV file that has the following format:

SITE_ID,DATEON,TOTAL_NO3
CON186,2003-08-26,X.XXXX
CON186,2004-04-27,Y.YYYY
CON186,2004-07-27,Z.ZZZZ

The first row is to be a header containing three columns with those exact names. Each prediction is to be on a separate line. The date is to be in YYYY-MM-DD format and the predicted value is to have four values (no more) after the decimal place.

Models will be evaluated on the Sum of Squared Error (SSE). If not, all predictions are included, your team will be disqualified from the Best Model competition. If your submitted file does not adhere to the previously stated formatting, it will receive a 5% penalty in its SSE metric.

GitHub Repository

dsgovirtual/DSGO-Virtual-Oracle-Hackathon

This is the official folder of the DSGO Virtual Hackathon, October 2020 -...

GitHub

Expected Submission

- Submission is expected to be made by 3 pm PST in the #Hackathon-Submissions channel in Slack
- Submission can be a Webapp or a Github repository, including the notebook that implements the full lifecycle of data preparation, model creation and evaluation. Source code and information used for the challenge in CSV format.
- PDF presentation (5 slides max) with your observations, predictions, and conclusions. At the end of the activity, your team will have 5 minutes to present the results.

The notebook should contain:

- Any steps you take to prepare the data
- You assumptions
- Training of your model