SPRINGBOARD DATA SCIENCE CAREER TRACK

IDEAS FOR CAPSTONE PROJECT #1

SILVIA MAIONE

1. Predicting the conditions of street trees in New York City

This project is based on a Tree survey conducted in New York City in 2015 by the Parks and Recreation department. The idea is to provide a model that can help to identify what type of trees thrive or not in specific areas and what factors affect their status. With this type of information, the department can decide what actions need to be taken to guarantee good tree health and where to plant new individuals and what type. Specifically, the goal of the model is to predict the status of a tree (dead or alive) from some of the features as the ones collected in the survey. Most of them are in the form of categorical data, but they can be converted. The availability of geographical coordinates offers an opportunity for visualization on a map.

Data:

https://data.cityofnewyork.us/Environment/2015-Street-Tree-Census-Tree-Data/uvpi-gqnh

2. Predicting usage patterns of the Citi Bike network of New York City

The city of New York has an extensive bike sharing network. A decision about an increase of the number of available bicycles and even a possible network extension relies on being able to predict the usage of such network. By knowing where and when most of the trips happen, it's possible to evaluate if more bicycles are needed or the existing ones might need maintenance more often.

This project aims at providing means to identify the most popular trips and stations based on the season and time of the day, customers' type, age, gender and trip duration. It relies on historical data downloadable from the Citi Bike website:

https://s3.amazonaws.com/tripdata/index.html

A description of the data features is given at the following link:

https://www.citibikenyc.com/system-data

A list of the stations in Json format is also available:

https://feeds.citibikenyc.com/stations/stations.json

3. Predicting solar power generation

In this project the goal is to build a model of electricity generation from solar energy based on radiation and time data relative to one Country (Italy in this case). The load is also included in the dataset (both actual and predicted), together with the production from solar energy; while time and weather-related information can be extracted from the timeseries (season, day, and time of the day), also included in the same file. Radiation data (diffuse and direct) is available in a separate data set. This type of model would be useful in energy management systems that, based on demand and conditions, distribute the load between different generation systems.

Data:

https://www.kaggle.com/arielcedola/solar-generation-and-demand-italy-20152016

This data has only the time series and power generated but the same resource has weather data too.

https://data.open-power-system-data.org/time_series/2017-07-09

https://data.open-power-system-data.org/weather_data/2019-04-09