SPRINGBOARD DATA SCIENCE CAREER TRACK

CAPSTONE PROJECT #1 PROPOSAL

SILVIA MAIONE

Problem: Predicting the health of street trees in New York City

According to a census conducted in 2015 by volunteers, staff and partners of the New York City Parks and Recreation department, there are almost 700k trees in the whole five boroughs' metropolitan area. Being able to identify what type of trees thrive or not in specific areas and what factors affect their status allows proactive decision making. Beneficiary of this type of information are both the city's Park and Recreation department and the administrators. In fact, they can decide what actions need to be taken to guarantee trees' good health and to prevent those that aren't in good conditions from deteriorating. Also, they can make an informed decision on where to plant new individuals and of what type.

The goal of this project is to provide a model to predict the health of a tree (good, fair or poor), knowing certain characteristics such as some of the features collected in the survey.

Data

The data used in this project can be found at the following page:

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

It consists of over 680k rows and 45 columns. The rows correspond to trees entries, the columns contain various forms of identification of a tree, including species' name; status of the tree (dead or alive, general health); information on size and other factors describing the surroundings (presence of guards, root or trunk problems, etc.). Most of the data is of object or text type. Geographical information is also available (latitude, longitude). The target variable is chosen to be the general health of a tree, and it can have three values, i.e. good, fair or poor. It is available as one of the columns of the dataset.

A detailed description of the various fields is available in pdf format in the same page.

Methodology

The problem at the core of this project is of the supervised multi-class classification type, since the target variable is the tree's health, which can have three values, as described in the data section.

The first step will be identifying and dealing with any missing data, doing any needed conversion of categorical variables and normalization or scaling.

Then exploratory data analysis will be conducted to identify possible correlation between the features and the outcome. Plotting techniques will be used to visualize important aspects or features. This part of the projects will also give the opportunity to answer other questions, for example how the trees types are distributed across the boroughs, if there's a specific type more common in one area, etc. Those fields that are identified as not being relevant will be removed from the actual data set used to build the model.

Once the predictors will be chosen, multi-classification modeling will be performed. A couple of modeling techniques will possibly be considered, i.e. logistic regression (one vs all) and neural network.

Deliverables

The final deliverables will consist of the Python code, a paper and a slide deck.