Data Science Career Track
Capstone Project 1 - Proposal

Thibault Dody
08/03/2017

Topic #1 - Potholes. Potholes everywhere...

1. Problem

Using existing data from the city of Boston, I would like to evaluate the following factors leading to an uneven distribution of the potholes over the city:

- What are the main factors leading to a fast/slow repair?
- How is the frequency of appearance of pothole related to the weather?
- Does the features of a neighborhood impacts the frequency of its requests and the duration prior to the repair? (i.e. industrial, residential,)
- Can we predict future trends using machine learning?

2. Clients

The client of this project is the Massachusetts Department of Transportation (MassDOT). The purpose of this project is to optimize an existing tool available to report potholes in the entire Commonwealth. Currently, this tool is used to treat requests on a case by case basis. The goal of this project is to develop a tool to predict repair time based on the location and time of a request. In addition to the predictive aspect, trends discovered during this study can be used to optimize resources and planning.

3. Data

In order to evaluate other impacting factors, I would like to combine the following datasets:

Data	Provider	Location	Dates	Link
Potholes report - Closed cases	City of Boston	Boston	07/01/2011 to 06/27/2017	(a)
Monthly weather data (Temperatures, Precipitations)	National Centers for Environmental Information (NOAA)	Logan Airport weather station	01/01/2017 to 06/01/2017	(b)
Boston neighborhoods data	Boston Maps	Boston	2014	(c)

(a): https://data.cityofboston.gov/City-Services/Closed-Pothole-Cases/wivc-syw7/data
(b): Upon request on NOAA site
(c):https://www.arcgis.com/home/item.html?id=71bec6dadfbb462dbdbb0293a6b10be2#data

4. Project Outline

The first step of the project will consist of the construction and preparation of my dataset. After exploring both datasets, additional features will be added to the pothole dataset. These new features can either be extracted from the pothole dataset itself or created from the weather dataset. During this process, the choice may be made to look for additional feature not presented in the datasets listed in Section 3.

The second phase of the project will focus on discovering and verifying trends. Since the pothole issue is a well-known problem, several of its causes are already well known. However, the main purpose of the project is to expand the knowledge of the issue in order to improve the repair time. The tools used during this phase will be mainly visual and statistics.

Finally, various machine learning algorithms will be used to try to predict the time to patch a pothole based on the key features obtained previously.

During each step, an effort will be made to quantify the limitations of the findings and the options to improve the work.

5. Deliverables

The deliverables for this project are defined as follow:

1. Comparison report of the occurrence and delays of repairs per neighborhood.
2. Comparison report of the occurrence and delays of repairs as a function of the weather
3. A predictive model used to estimate the time of repair of a pothole

The results of this project will be presented in a report supplemented by the code used to perform the analysis.