Data Science and Machine Learning Capstone Project

Map of NYC plotted using locations of all yellow taxi pickups, 2010-2015.

OBJECTIVES

- Showcase data science and machine learning skills to solve NYC311 problems.
- Use data science methodologies to define and formulate a problem.
- Use data analysis tools to ingest a dataset, clean it, and wrangle it.
- Use data visualization skills to visualize the data and extract meaningful patterns to guide the modelling process.
- Use machine learning skills to build a predictive model in order to improve functional efficiency.

LANGUAGES: Python

PLATFORMS: IBM cloud, Watson Studio

FIELDS: Data Science, Machine Learning

DATASETS

We will use two datasets from the Department of Housing Preservation and Development of New York City to address their problems (data can be downloaded by using SODA API).

- NYC 311 Complaint Dataset

  https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9.

- Pluto Dataset for Housing

  https://data.cityofnewyork.us/City-Government/Primary-Land-Use-Tax-Lot-Output-PLUTO-/xuk2-nczf.

Problem Statement

The people of New Yorker use the 311 system to report complaints about the non-emergency problems to local authorities. Various agencies in New York are assigned these problems. The Department of Housing Preservation and Development of New York City is the agency that processes 311 complaints that are related to housing and buildings.

In the last few years, the number of 311 complaints coming to the Department of Housing Preservation and Development has increased significantly. Although these complaints are not necessarily urgent, the large volume of complaints and the sudden increase is impacting the overall efficiency of operations of the agency.

Therefore, we will help to manage the large volume of 311 complaints they are receiving every year.

The answers to the following questions will be supported by data and analytics:

1. Which type of complaint should the Department of Housing Preservation and Development of New York City focus on first?
2. Should the Department of Housing Preservation and Development of New York City focus on any particular set of boroughs, ZIP codes, or streets (where the complaints are severe) for the most important types of complaints?
3. Do the most important types of complaints have an obvious relationship with any particular characteristic or characteristics of the houses or buildings?
4. Can a predictive model be built for a future prediction of the possibility of complaints of the type that you identified as the most important types of complaints?

As the lead data scientist to provide the answers to these questions, I will follow the standard approach of data science and machine learning in order to find the answers.