Machine Learning for Predicting Ward Success in Chicago

The City of Chicago is segmented into 50 wards, where each ward has its own alderman that acts as a liaison between their community and the Chicago area as a whole. The fifty aldermen make up the city council which is responsible for making important decisions that affect its communities regarding things like healthcare, welfare, and taxation. Ward health can be categorized by many things: happiness of its residents, crime rates, number of new residents moving into the neighborhood, number of businesses, etc. For the scope of this project we will focus most on crime rates and business license data. We are specifically interested in whether there is a correlation between ward health and the monetary contributions made to the aldermen.

Getting Started

To start, you will need to do one of two things. First, you can gather each of the data sets attached in the Data GitHub file and join them together using the Data By Ward and Dataset Combination python scripts. To make this process simpler, we have uploaded the final combined dataset, Data By Ward, into the Data GitHub folder as well. Most of our data is from the the *City of Chicago Data Portal,* but we did access ward demographics from *The Chicago Data Guy*. I will detail a bit of the specifications of our data below:

- Lobbyist Data - Gifts
    - 106 KB
    - 848 observations
    - 13 attributes
- Lobbyist Data - Employers
    - 709 KB
    - 5,677 observations
    - 13 attributes
- Lobbyist Data - Compensation
    - 3 MB
    - 28,958 observations
    - 11 attributes
- Lobbyist Data - Lobbyists by Client
    - 2 MB
    - 5,857 observations
    - 7 attributes
- Ward Offices
    - 12 KB
    - 50 observations

    - 16 attributes
- Business Licenses - Current
    - 24 MB
    - 70,671 observations
    - 34 attributes
- Crimes 2001 - Present
    - 2 GB
    - 7,333,802 observations
    - 22 attributes
- Microlending
    - 32 KB
    - 252 observations
    - 14 attributes
- Estimated Ward Populations
    - 23 KB
    - 848 observations
    - 13 attributes

Prerequisites

Jupyter Notebook

- Install: https://jupyter.org/install.html

Google Drive

- Access: https://www.google.com/intl/en/drive/

Google Colab

- Access: https://colab.research.google.com/notebooks/intro.ipynb

Tableau

- Install: https://www.tableau.com/products/desktop/download

Knowledge of Gephi

- Access: https://github.com/KiranGershenfeld/VisualizingTwitchCommunities

Data File Descriptions

1. Contributions-EDA.ipynb
    a. Contains exploratory data analysis on the contributions dataset
    b. Contains exploratory data analysis on contributions plus ward dataset
2. Dataset Combinations.ipynb
    a. Contains initial attempts at combining multiple datasets into one
3. data by ward.ipynb
    a. Contains the code for creating the main dataset we used for our model building

    b. Contains a bit of exploratory data analysis on this final dataset
4. EDA Business Licenses.ipynb
    a. Contains initial exploratory data analysis on the business license dataset
5. EDA.ipynb
    a. Contains various exploratory data analysis on crimes data and business licence data
6. InitialModelBuilding.ipynb
    a. Contains the base models for our project
    b. Linear regression, lasso regression, ridge regression, random forest, gradient boost, ADA boost, and sequential neural network
7. InitialModelBuilding - With PCA.ipynb
    a. Contains an implementation of the base models from the InitialModelBuilding.ipynb notebook, but adds PCA dimension reduction to those models, where applicable

Authors

Olivia Fenwick
Miguel Gutierrez
Meghan Rokas
Aliya Zhdanov