Machine Learning Engineer Nanodegree

Capstone Proposal

Alison O. Gaby
May 24th, 2018

Proposal

For my project, I aim to predict housing prices in Queens, NY. using the New York City Property Sales dataset generously hosted and curated by Enigma Public. This dataset is partially based on Rolling Sales Data provided by the NYC OpenData initiative (found here: http://www1.nyc.gov/site/finance/taxes/property-rolling-sales-data.page) in addition to historical Annualized Sales (found here: http://www1.nyc.gov/site/finance/taxes/property-annualized-sales-update.page).

Domain Background

This project was inspired by the Boston Housing project that was a part of Udacity's Machine Learning Nanodegree, in which students estimated the prices of housing in a suburb of Boston, MA in 1978 using supervised learning techniques. The project raised several questions regarding the applicability of using supervised learning techniques to other geographical areas, or to areas where there is more variability in housing affordability or population density.

Problem Statement

Estimating housing prices can be a difficult problem. There are multiple factors one must take into account beyond taking into account many different variables. In our project using the Boston Housing set, we explored using different regression methods to get accurate housing price predictions. With that particular data set, we were able to get fairly accurate results, but does this model "travel"? Can we use this on data from a different local? Let us explore.

Datasets and Inputs

Given the data available thanks to the NYC Open Data Initiative and Enigma Public, I can test out whether the techniques used will provide accurate predictions. As the data provided in from NYC Open Data does not include the same features for rooms per home, student-teacher ratio, percentage of low income residents, or median sale price, I will attempt to select features that correlate highly with `sale_price`, in addition to using `earned_income` (percentage receiving earned income credit by Zip Code), and `air_qual` (air quality as measured by concentraitions of Sulfur Oxide present).

While this method is fairly simplistic, and there are services like Zillow and Trulia available for regular users, I would like to see if my estimates using these techniques are comparable.

- General Information about this dataset can be found here: https://public.enigma.com/datasets/new-york-city-property-sales/fd6efa37-2dcd-4294-8795-a0e6044f15b4
- Rolling Sales Data for all Boroughs of New York, NY: http://www1.nyc.gov/site/finance/taxes/property-rolling-sales-data.page
  - For Queens, NY (clicking link with prompt download of xls file): http://www1.nyc.gov/assets/finance/downloads/pdf/rolling_sales/rollingsales_queens.xls
  - Glossary of Terms used: http://www1.nyc.gov/assets/finance/downloads/pdf/07pdf/glossary_rsf071607.pdf
- Annualized Sales Update: http://www1.nyc.gov/site/finance/taxes/property-annualized-sales-update.page
  - For Queens, NY (clicking link with prompt download of xls file): http://www1.nyc.gov/assets/finance/downloads/pdf/rolling_sales/neighborhood_sales/queens_sales_prices.xls
- Earned Income Credit by Zip Code obtained from IRS
  - Individual Income Tax ZIP Code Data: https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-2015-zip-code-data-soi
- Air Quality by Zip Code
  - Neighborhood Air Quality Concentrations: Sulfur Dioxide (SO2) : https://catalog.data.gov/dataset/air-quality-ef520
- Actual data used can be found here: https://public.enigma.com/datasets/new-york-city-property-sales/fd6efa37-2dcd-4294-8795-a0e6044f15b4

Dataset Info:

- The New York City Property Sales Dataset - Based on Rolling Sales Data provided by the City of New York's Department of Finance's collection of property listings that sold in the last twelve-month period. Total Records: 60,295, Total fields: 63. Of the 63 fields in the dataset, the below columns have been included in this project:
  - zipcode
  - land*square*feet

    - gross*square*feet
    - sale_price
    - community_district
    - school_district
    - floor*area*total_buildings
    - floor*area*residential
    - maximum*allowable*residential_far
    - x_coordinate
    - y_coordinate

- Annualized Sales Update- This is a collection of yearly sales information of properties sold in New York City between the years 2005 to 2016. (Only data from the years 2013-2016 was used).

    - This dataset includes the following attributes, grouped by neighborhood, and type of home:
        - NEIGHBORHOOD
        - TYPE OF HOME
        - NUMBER OF SALES
        - LOWEST SALE PRICE
        - AVERAGE SALE PRICE
        - MEDIAN SALE PRICE
        - HIGHEST SALE PRICE

- Tax Information - Earned Income Credit by Zip was extracted from this dataset, and mapped to the existing housing data

- Neighborhood Air Quality Concentrations: Sulfur Dioxide (SO2) - this data was extracted from the linked dataset, and mapped from each UHF42 section to each zip code, then mapped to the original housing data.

Solution Statement

In order to predict sales prices in Queens, NY, I will map `earned_income` or Percentage of Earned Income and air quality measurements (`air_qual`) with our existing New York Property Sales. Then, I will select which features correlate most heavily with `sale_price` and apply them to a simple linear regression model as a benchmark for comparison.

Benchmark Model

As a benchmark, we will use a simple linear regression model for comparison. In addition, we will compare our estimations to historical data provided in the Annualized Housing datasets, which include highest, lowest, median, and mean sale prices for the years 2013-2016 (only results for Single Family Dwellings will be used for any comparison).

Evaluation Metrics

As an evaluation metric, I've chosen the to use the R^2 method or the Coefficient of Determination. As an evaluation method for regression, we will use this method for testing how accurately our model is by how close it reaches 1 (out of values 0 to 1, with 1 being the most accurate possible). As our model is a regression, I've chosen this method for accuracy as our method is based on how accurately we can predict roughly normally distributed`sale_price` data (normally distributed data is assumed when using the R^2 method for scoring).

Project Design

In order to approach this problem, I will first take a look at the data set that I have for housing in Queens. What features are similar? How can we decide which features to use? The Boston dataset has features which are highly predictive pre-selected. I will compare features in this dataset in order to find the most appropriate substitutes.

After selecting the most predictive features, I will then compare the dataset to the historical annualized data. How does this distribution compare? Does our data need to be scaled?

Next, we will fit our data to a simple decision tree regression to test its accuracy as a baseline. How does it compare to the accuracy of the Boston Housing project's model?

In order to improve accuracy, let us fit our data to an ensemble method called AdaBoost. How does this accuracy improve? We will then adjust to account for overfitting.

Let us compare our results in this project with that of that Boston dataset. Are our results the same? Have they improved? If not, I will propose different methods to test in future experiments.