Module 2 Summary- Snizhana Kurylyuk

Abstract

A client reached who is thinking of buying a single family home and doesnt know which area is best. They wanted me to build them a regression to allow them to not overpay for a home and provide analysis on the types of homes they would find in each neighborhood. My objective was to Web Scrape Redfin for various home features and analyze the home prices based on those features. Then build a predictive regression model to predict the price of a home. After refining my model, I built dashboards to visualize and communicate my results. Which I will present to my client with my predictions.

Design

Lasso and Ridge lambda tuning models were applied to standardized data, as well as Step Ordinary Least Squares modelling, removing influential points and outliers. A Lasso model was chosen for final testing in order to avoid overfitting and data loss while achieving a maximal fit.

Data

1000+ data points were scraped from redfin targeting 22 cities in Illinois. The variables I scrapped included Price/Sq.Ft, Lot Size, Year Built, Tax Annual Amount, Tax Year, Tax Exemptions, bedrooms, bathrooms, sq_footage, price, address, City, Zip Code, and State.

I extracted data from 2018 Tax Returns provided by the IRS database. I used that data to pull adjusted gross income by zip code and incorporated that into my analysis.

Algorithms

Featured Techniques

- Feature Engineering & Selection
- Supervised Machine Learning
- Regression

Feature Engineering

- Converting categorical features to binary dummy variables
- Using Imputation to drop data points that were missing significant information
- Limiting Outliers by reducing my range of certain features
    - I examined graphical discontinuity in my prices, and square footage
- Categorical column grouping to identify trends by city

Models

- Linear Regression
- Lasso Regression- To remove unnecessary features

Module 2 Summary- Snizhana Kurylyuk

- Ridge Regression- To reduce effects of collinear features
- Polynomial Regression
- Data split 80/20 train.test
- Statsmodels OLS Summaries levergaged for preliminary feature selection

Tools

- NumPy and Pandas for data manipulation
- Matplotlib and Seaborn for plotting
- SKLearn, Statsmodels for modeling
- Beautiful soup for web scraping and data ingestion

Communication

My project was completed in line with time requirements and can be viewed here.