Predicting NYC House Prices at Neighborhood Level using HPM (Hedonic Pricing Model)

Outline

Part I. Motivation

Imagine our client is an REIT in New York City and they want us to help them predict the fair transaction price of a property before it is sold. Our task is to construct a real-estate pricing model in neighborhood level.

Part II. Methodology

I. Overview

We are going to use hedonic pricing model on the data collected from the empirical worl d, and apply advanced machine learning tools to optimize its performance, also will leve rage map visualization, which is a plus to make outcomes more intuitive.

II. Introduction to Hedonic Pricing Model

The main idea of hedonic pricing model is that people regard the contribution of each at tribute to the total price of the property as separate values. In other words, they are addible.The general model is R = f (P, N, L, C). We can use linear, log or semi-log regr essions to get results.

III. Map Visualization

Concat neighborhood level data with json formatted coordinate data.
i. Advantage: Visualize outcomes more explicitly.
ii. Tools leveraged: 1)defined functions in 2D array converting. 2) matplotlib, seaborn libraries in Python 3) Folium library in Python dynamically to help zoom in and out.

IV. Machine Learning

We here use the term to describe ① a diverse collection of high-dimensional models for s tatistical prediction, combined with ② so-called “regularization” methods for model se lection and mitigation of overfit, and ③ efficient algorithms for searching among a vast number of potential model specifications.

Another reason that we think machine learning is applicable to this project is that it helps us to deal with multicollinearity and overfitting that may exist in the model.

Part III. Collecting Data and Variable Descriptions

I. Data Sources

i. Open Government Databases:
1) Advanrage: More reliable, less costly.
2) Sources used in this research: NYC Open Data, NYC Department of Finance (2018-2019), NYC Planning Labs, NYC coordinates geojson file, US Census Bureau, NYC City Planning, NYC census tract data in Neighborhood Tabulation Areas (NTAs) Level.

ii. Web Scraping (APIs):
Scrape down venue characters of NYC neighborhood via Foursquare API:

- for lat, lng, neighborhood in zip(nyc_neighborhood['Latitude'], nyc_neighborhood['Longitude'], nyc_neighborhood['Neighborhood']):
  url = "https://api.foursquare.com/v2/venues/explore?client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}".format(CLIENT_ID, CLIENT_SECRET, VERSION, lat, lng, r, limit)
  results = requests.get(url).json()['response']['groups'][0]['items']

II. Variables Discriptions

i. Property Characteristics

Categories	Name	Descriptions
Property Structure		
	Sale Price	Transaction Price
	Land Square feet	House Square Feet
	Building class category	e.g.one/two family dwelling
Location	Borough	Administrative districts
Neighborhood	Neighborhood	more detailed districts each come with their own atmosphere
House Age	Age	The house age since year built

ii. Demographic (38 columns)

Total Population, Men, Women, Hispanic, Asian, Income, Income Per Capita, Drive, Transit, Employed, Professional…

iii. Venue categories (386 columns)
Airport Terminal, Chinese Restaurant, Grocery Stores, Whisky Bar, Yoga Studio, Gym...

Part IV. Basic Data Analysis

I. Data Cleaning and Visualization

i. Use matplotlib and seaborn libraries in Python to plot histograms for numerical variables to see distributions, and bar plots for categorical variable distributions.
ii. Duplicates, errors and typos.
iii. Insights: heterogeneous buyers → non-linear relationship
iv. Dealing with missing values and outliers.

II.Geographic visualization

i. Sales Prices:
Define functions on NTAs, combined with matplotlib to show relationship between variables. e.g. we use the function to look into distributions of indexes like Asian population, median household income, income per capita, percentage of people taking public transportation to work, percentage of residents taking professional jobs.

ii. Insights from map visualiztion:
From the median household income map, we could clearly see that Manhattan East 96th Street is generally the wealthiest. Also, comparing the two plots we can also tell that Manhattan have fewer people per household than other boroughs.
The professional job distribution map precisely implies that a huge amount of residents in Manhattan are occupied with professional jobs which generally require higher education and people in return get higher salaries, which greatly echoes what we find in Sales Prices.

iii. Venue categories
1) Venue data of NYC neighborhood scrapped via Foursquare API:
- results = requests.get(url).json()

2) Print the top 10 most common venues in each neighborhood after regrouping.
3) Leverage Folium library in Python to dynamically look into each venues.

Part V. Machine Learning

I. Feature Engineering

i. Preparation

In order to run a hedonic pricing model, property characteristics need to be combined with external neighborhood venue data.

Here we conduct one hot encoding to turn categorical variables into numeric.

- venues_type = pd.get_dummies(venues_df[['Category']], prefix='', prefix_sep='')

ii. Correlation Heatmap

iii. train test splits 80 to 20

iv. Data Standardizing

We need to scale all features to between 0 and 1. Here we apply data standardizing to get zero mean centring and unit scaling.

- train_mean = x_train.mean()
  train_std = x_train.std()
  x_train = (x_train - train_mean)/train_std

II. Model Evaluations

Root mean squared error

R-squared

The smaller rmse, the higher r-squared, the better the model.

III.Optimizing Models

i. OLS

ii. PCR (Principal Component Regression) to reduce dimensions of features

iii. Ridge

iv. Random Forest Regression

v. XGBoost

GridSearchCV used in parameter tuning.

IV.Choosing Winning Model

Choose XGBoost model based on performance.

Part VI. Conclusion and Future Possibilities

I. Contribution

i. In empirical world, data-driven real estate companies are surging on the market. One way we could apply to practice is that when one client tells us his prefers for the house surroundings, for example, restaurant or beauty&spas, or more schools, we could recommend an interval the house price may possibly fall in, which helps save time and improve efficiency. For sellers, vise versa.

ii. Clusters could be used in further studies for a similar spot recommendation. Even more creative techniques, for example, ensemble of these models, stacking, could be applied.

II. Work in the Future

i. Combine with time series analysis, cross-validation.
ii. Add more detailed factors, like crime rate, number of middle schools, even in a block level.

Part VII. References

1. Dubin, R. A., & Goodman, A. C. (1982). Valuation of education and crime neighborhood characteristics through hedonic housing prices. Population and Environment, 5(3), 166-181. doi:10.1007/BF01257055
2. Gibbons, S., & Machin, S. (2008). Valuing school quality, better transport, and lower crime: Evidence from house prices. Oxford Review of Economic Policy, 24(1), 99-119. doi:10.1093/oxrep/grn008
3. Gu, S., Kelly, B., & Xiu, D. (2018). *Empirical asset pricing via machine learning* (No. w25398). National Bureau of Economic Research.

4. Herath, S., & Maier, G. (2010). The hedonic price method in real estate and housing market research: a review of the literature.