Project Proposal: Predicting Manhattan Home Sales

Basic List of Goals and Objectives

By the completion of this project, I will have selected and trained a model that predicts housing prices in Manhattan, NYC. While the original dataset of property sales I'm using contains columns that will be used as independent variables (i.e. land square feet), I am also taking into consideration environmental factors that put a premium on house prices. In particular, I will look at 5 of these variables: high schools, healthy stores (including bodegas), farmers' markets, parks, and subway stations. Using libraries like *Boruta* and *randomForest*, I can get figure out which variables are most important in determining house prices (a process called feature selection or engineering) and start model training to find my final model. Throughout the process, I will be creating maps and graphs relevant to each step, but the most important graphs will be after I have determined my model and am able to map predicted housing prices to compare with actual housing prices. This, of course, will all be for Manhattan, but I want to also see how well the model works for another borough. In general, I'll be attempting to answer the following questions:

1) What factors influence housing prices in Manhattan the most?
2) What model best predicts housing prices?
3) How do the predictions compare with the actual prices?
4) How do the predictions fare in another borough?
5) What are the predictions for new data?

This will also be my first project wherein I'll be implementing machine learning algorithms, and so I will rely heavily on the textbook and online resources to learn about the appropriate libraries and functions. In doing so, I hope that by the end of the project I have a better understanding of the intersection between geocomputation, geodata science, and machine learning.

Background

The main reason I chose this project topic was to better understand how real estate companies like Zillow estimate homes that are not even on the market. These companies and real estate agents likely use a wide array of models to come up with their respective predictions, especially since every type of property and location has specifics. Manhattan (and New York City, in general) is infamous for its insanely high living costs, a majority part of which is due to housing and rent. Though the final model will be most useful for people who are thinking of putting their houses on the markets, it will also be useful for those who are considering moving to Manhattan and want to find the best valued home.

Furthermore, housing prices are an important indicator of the economy. The data I am using is on rolling sales from the last 12 months (April 2019 – March 2020). The coronavirus outbreak will change or is already changing the real estate market and so building a model of data from this timeline will hopefully give insight into housing prices in the midst of this outbreak.

Data Sources

I will be using data from these sources:

1) The NYC Department of Finance provides data on property sales from April 2019 to March 2020 for all NYC boroughs (Manhattan, Bronx, Brooklyn, Queens, and Staten Island). These datasets have 21 columns including borough, neighborhood, building class category, tax class at present, block, lot, easement, building class at present, address, apartment number, zip code, residential units, commercial units, total units, land square feet, gross square feet, year built, tax class at time of sale, building class at time of sale, sale price, and sale date. https://www1.nyc.gov/site/finance/taxes/property-rolling-sales-data.page
2) The NYC Open Data site provides data that I will be merging with the original Manhattan property sales dataset. I will be testing what variables beyond the ones in the original dataset (i.e. number of units) can predict property sale price.

These variables include distances from the property to schools (public and private), to health stores, to farmers' markets, to parks, and to MTA subway stations. These datasets can be found at the following links:

a. Schools: https://data.cityofnewyork.us/Education/2019-2020-School-Locations/wg9x-4ke6
b. Healthy Stores: https://data.cityofnewyork.us/Health/Recognized-Shop-Healthy-Stores/ud4g-9x9z
c. Farmers' Markets: https://data.cityofnewyork.us/dataset/DOHMH-Farmers-Markets/8vwk-6iz2
d. Parks: https://data.cityofnewyork.us/Recreation/Open-Space-Parks-/g84h-jbjm
e. Subway Stations: https://data.cityofnewyork.us/Transportation/Subway-Stations/arq3-7z49

Geocomputational Process

I expect to use geocomputational operations from both inside and outside the textbooks. For example, one of the first and most important steps I do is geocoding home addresses. After cleaning them (i.e. removing dashes, addresses less than 5 in length, adding "Manhattan" to the end of each), I used the *geocode_OSM()* function from the *tmapstools* library to geocode each address. With the *dplyr* library, I used *merge()* to inner join resulting geocoded addresses dataframe with my original Manhattan home sales dataframe, as we learned in Chapter 4.

After, I will have again clean (and possibly geocode) the other datasets (schools, healthy stores, farmers' markets, parks, subway stations) before merging them. Before merging them, I will have to calculate the distances between each home and independent variable. I will use functions like *st_buffer(), st_is_within_distance()*, and *st_within()* instead of calculating the actual Euclidean distances between each and every home and variable (that would be too many calculations and also not that useful). I can, for instance, calculate how many schools are within a 1km buffer of a home, generate a map to visualize that, and create a continuous variable. An alternative could be to implement a K-nearest-neighbors algorithm and create a variable based on the mean distance between homes and, say, the 4 closest parks.

Before diving into feature selection, I also want to explore the relationships between each independent variable with home sales (the dependent variable) through graphs. These graphs will include those plotting home sales against distance to each independent variable and calculating correlation, creating density maps, buffered maps, and more.

During feature selection/engineering, I will use try few modelling techniques including *randomForest* and *Boruta* libraries. After I have narrowed down which variables contribute the most to predicting house sales, I can start model training by fitting various models from Poisson to Lasso and XGBoost and picking a model that has the smallest mean absolute error (MAE). Chapter 11 ("Statistical Learning") of the textbook contains information on how to (spatially) cross-validate models and in Chapter 14 ("Ecology"), Lovelace provides an example in which he uses random forests. I'll be using these chapters and other (to be cited) resources from the web for the feature selection, model selection, and cross-validation steps, as they're all very complicated and new to me.

Lastly, I will have to analyze the selected model results with more plots and maps. For instance, I can map the predicted house prices in, say, SoHo and compare it with a map of the actual house prices. Since this model will be one built off of data from Manhattan houses, I will test its predictions of houses in other boroughs like Brooklyn or even other types of properties. I will also feed the model new data to see predictions for housing prices in places that aren't provided in the original dataset (i.e. new property listings I find or data on sales from 2018).

I recently finished cleaning the property data and have uploaded the results in my final project repository: https://htmlpreview.github.io/?https://raw.githubusercontent.com/lcao21/GIS3_Final_Project/master/Data%20Cleaning%20and%20Wrangling.html. I am currently working on creating a variable out of the schools dataset (i.e. counting how many schools are within a buffer of each home) and will likely use the same method for the other 4 datasets of healthy stores, farmers' markets, parks, and subway stations; my current progress, in general, can be tracked here: https://github.com/lcao21/GIS3_Final_Project

Timeline of Work

Week	Work	Details	Deliverables
Week of May 3	Clean data; exploratory analysis; feature selection pt. 1	Clean data (i.e. remove rows with NA values; geocoding); create maps and graphs to explore the data (i.e. which neighborhoods have the highest home sales); merge the 5 additional datasets with original sales dataset	Project proposal
Week of May 10	Feature selection pt.2	Create graphs and analyze which variables are most important in predicting prices with different methods (i.e. Boruta; Random Forest)	
Week of May 17	Model training	Train different models and select one model with the smallest error (i.e. mean absolute error)	
Week of May 24	Project data report; map results	Mapping results of prediction vs actual	Project data report
Week of May 31	Create more maps; use model to create and map predictions in the other boroughs; predict new sales	How well does the model built off of Manhattan's homes work for the other boroughs (Bronx, Brooklyn, Queens, and Staten Island)? What does the model tell us about prices for X? (X being new data)	
Week of June 7	Write up final project and organize GitHub repository		Final project in GitHub repository