Summary of Two Bedroom Rent Price Prediction in Manhattan, NY

Metis - Project 2 - Logistic Regression with Web Scraping
Gregory Lull Oct, 2019

Project Design

Having moved from NY recently and going through the process of finding an apartment inspired me to see if I could predict rent prices. I thought that this would be helpful to know for people who are moving to a new city and estimating what their budget needs to be, instead of scrolling hundreds of results. Ultimately if this proved viable it would be interesting to see how different features affect prices in different cities if the data was expanded.

I chose NY because there's a website called Streeteasy.com that I know is used very often by New Yorkers seeking to rent. This website is pretty comprehensive in terms of unit features (bed/bath, sqft, washer/dryer, etc.), and building features (age of the building, doorman, elevator, etc.). Initially I had issues using selenium and beautiful soup to scrape the website because the Streeteasy blocked these tools, but I decided to try using a chrome extension of my own creation because I believed it would be unblockable since it is my own original code.

For the analysis I only used the most basic linear regression and didn't have time to get to more advanced techniques such as regularization, or creating more automated pipelines.

Looking Back

What I would've done differently would be to spend way less time perfecting the scraping and instead spend the majority of my time on the analysis. I joined the bootcamp looking to understand data science and its theory and tools, but instead fell back on my previous professional experience.

The MVP milestones that the instructors set for us was very helpful, and it is something I should've stuck to. If I had built my pipeline to just take in data instead of leaving the pipeline to the last minute I would've been able to try out more modeling tools, and have a model I'm much more confident and proud of.

In terms of the data I collected I would've liked to had more continuous independent variables since most of my features were dummy variables. I think I should've also used a little bit of the Google Place API to find out things about the neighborhood, such as restaurant density, grocery stores, bars, etc.

Appendix

Tech Stack:

IDEs: Jupyter notebook, Visual Studio Code

Python

- Analysis: Pandas, SciKit Learn, numpy
- Visuals: matplotlib

Javascript

- Chrome extension code for scraping

Model Features

These are the features that I ended up with:

Feature	Type
Baths	number
Laundry In Unit	boolean
Gym	boolean
Doorman	boolean
Elevator	boolean
Total Subway stations nearby	number
Closest station distance	number
Laundry in building	boolean
Building pre war	boolean
Live-in super	boolean
Building postwar	boolean
Dishwasher in unit	boolean