Valentina Rizzati

Axelrod launches in NYC

May 05, 2021

Question

I am a Data Scientist at Axelrod Development Group (Axelrod), one of the major Residential Real Estate developers in Los Angeles. Axelrod is characterized by the adoption of a very data-driven approach to creating, selecting and developing the most profitable real estate projects.

In 2022, Axelrod is planning to allocate a significant portion of its new investment fund to launching the NYC market. For this purpose, the Chief Investment Officer asked me to analyze the NY Real Estate market on the sellers side and:

1. **Interpretative goal:** understand what are the factors driving the largest impact on property prices so that the Investment team will know what features (e.g. zip code, number of rooms) to prioritize when developing new projects
2. **Predictive goal:** build a model that will allow the Investment Team to get an estimate of property price by simply inputting the features (e.g. 11249 for zipcode, and 5 for number of rooms)

Data Description

The property data will be acquired by web scraping Streeteasy or Zillow. After having examined both sites, I've noticed that the first site is easier to scrape but the second site has somewhat better and more structured information. My approach will be to try scraping Zillow first and, if that proves too challenging in the allotted time, I will scrape Streeteasy.

The geographical scope is New York City (5 boroughs).

Once the data has been scraped, I will proceed with running the linear regression model.

In equation format, the regression model I am planning to build will be presented as follows:

$$\textbf{Price}_p = \beta_0 + \beta_1\,\textbf{num\_bedrooms} + \beta_2\,\textbf{num\_bathrooms} + \ldots\ldots + \beta_k\,\textbf{zipcode}$$

If I were to scrape Zillow, the model's target and features will be the ones presented below:

Element	Variable	Data Type	Comment
y	Actual Property Price	Numerical Continuous	
y_p	Estimated Property Price	Numerical Continuous	
x_1	Number of bedrooms	Numerical Discrete	Potential multicollinearity with bathrooms
x_2	Number of bathrooms	Numerical Discrete	Potential multicollinearity with bedrooms
x_3	Square footage	Numerical Continuous	
x_4	Zip code	Categorical Nominal	Potentially need to cluster in *areas*
x_5	Parking	Categorical Nominal to be transformed in Categorical Binary	• 0 where value is "None" or "0 spaces" • 1 otherwise
x_6	New Construction	Categorical Binary	
x_7	Year Built	Numerical Discrete	Irrelevant for interpretation goal because cannot be optimized by Investment team
x_8	Type	Categorical Nominal	
x_9	Number of Schools	Numerical Discrete	
x_{10}	Walk Score	Numerical Discrete	
x_{11}	Transit Score	Numerical Discrete	
x_{12}	Pool	Categorical Binary	

3

If I were to scrape Streeteasy, the model's target and features will be the ones presented below:

Element	Variable	Data Type	Comment
y	Actual Property Price	Numerical Continuous	
y_p	Estimated Property Price	Numerical Continuous	
x_1	Number of bedrooms	Numerical Discrete	Potential multicollinearity with bathrooms
x_2	Number of bathrooms	Numerical Discrete	Potential multicollinearity with bedrooms
x_3	Square footage	Numerical Continuous	
x_4	Zip code	Categorical Nominal	Potentially need to cluster in *areas*
x_5	Pets allowed	Categorical Binary	
x_6	Doorman	Categorical Binary	
x_7	Elevator	Categorical Binary	
x_8	Washer/Dryer	Categorical Binary	
x_9	Dishwasher	Categorical Binary	
x_{10}	Parking	Categorical Binary	
x_{11}	Gym	Categorical Binary	
x_{12}	View Type	Categorical Nominal	
x_{13}	Terrace	Categorical Binary	
x_{14}	Year Built	Numerical Discrete	Irrelevant for interpretation because cannot be optimized by Investment team
x_{15}	Type	Categorical Nominal	
x_{16}	Number of Schools	Numerical Discrete	Potential multicollinearity

			with zip code and number of colleges, parks and museums
x_{17}	Number of Colleges	Numerical Discrete	Potential multicollinearity with zip code and number of schools, parks and museums
x_{18}	Number of Parks	Numerical Discrete	Potential multicollinearity with zip code and number of colleges, schools and museums
x_{19}	Number of Museums	Numerical Discrete	Potential multicollinearity with zip code and number of colleges, parks and schools
x_{20}	Outdoor Space	Categorical Binary	
x_{21}	Swimming Pool	Categorical Binary	

Due to the high amount of features, the model might be susceptible from overfitting, which I will test for in the modeling part.

Feature engineering might have to be executed on some variables like zip code.

One row of data (i.e. observation) is represented by one property and the full dataset will be made of ca. 20,000 observations, which correspond to the properties currently listed on the market in NYC.

Tools

For the purpose of scraping the data I will use the Python package Beautiful Soup and the web browser automation tool Selenium.

For the data clearing, EDA, and modeling part I will use Scikit-learn, Pandas, Numpy, Matplotlib, and Seaborn.

If needed, for more interactive visualizations, I will use plotly or Tableau.

MVP Goal

As an MVP I intend to present a preliminary version of the regression formula, with a first description of the model output from both an interpretative and predictive perspective.

A couple of relevant visualizations will accompany the descriptive component of the MVP.