Predicting Housing Prices in Santiago, Chile

Second Capstone Project
Robert Walker

The Objective

The goal of this project is to scrape data on housing prices from the web and use this to build models that can accurately predict the prices of homes based on their general features. The models will also identify the most important factors that determine the price of a home. The imagined client is a new real estate agency in Santiago, Chile that needs a systematic approach to selecting prices for homes. They have requested a model that can take features of a house as parameters, and return a recommended price, as well as a project report, and a slide presentation. As part of the report/presentation, the agency would also like to know which features most strongly affect price.

The Data

To build the dataset for the project, we will be scraping the website https://chilepropiedades.cl/. It contains thousands of listings of new and used homes for sale in Chile. These listings are updated constantly, so it will be important to save a copy of the raw data once it is scraped for the purpose of reproducibility.

Methodology

Data Wrangling: Web Scraping
The first part of the data wrangling stage will involve using web scraping to acquire data needed for the project from the website mentioned above. The libraries required for this step are requests and BeautifulSoup. The web scraping will involve making a request to the website to obtain the HTML source code, and then parsing it with BeautifulSoup. After that the relevant information can be extracted by searching for tags that contain it. This will be accomplished by writing functions that perform the web scraping and return the results as pandas data frames. Afterward, all of the data frames will be combined into one, and the data frame will be saved as a csv file.

Data Wrangling: Data Cleaning
Next, the data will have to be cleaned. All of the data scraped from HTML is in string form, so much of it will have to be converted to floats or integers to make analysis possible. This step will also involve dropping rows that have errors or too many null values. Finally, the listings are in different currency types, so it will be necessary to choose one currency as our price unit and convert the others to it. Once finished, the cleaned data will be saved as yet another csv file.

Exploratory Data Analysis

The first step in the EDA stage will be to calculate some statistics for each *comuna*, such as average price, variance of price, number of houses for sale, etc. The next step will be to begin looking at how each feature interacts with price, probably producing a series of scatter plots, as well as a heatmap to see how the features interact with each other. Since many of the features roughly correspond to "size", it may also be useful to do PCA and see how a general size feature relates to price across different *comunas*. If two houses of the same size, located in different *comunas*, can have very different prices, then *comuna* will have to be included when building machine learning models.

Machine Learning

In the machine learning phase of the project, the data will first be divided into training and test sets. Next we will experiment with different types of regression models to predict the price. Ideally, this will involve building pipelines, and using grid search along with cross-validation to fine tune the models. In the end, one model will be selected based on performance on the test data. Once a model is selected, it could be demonstrated by selecting a row from the dataset at random, training the model on the rest of the data, and having the model predict a price which could be compared to the ground truth. This model will be presented to the client along with a report and a slide deck that summarize the process and key findings.