Problem Set 2: Making Money with ML?
"It's all about location location location!!!"
Ciencia de Datos para la toma de decisiones en Economía

Due Date: Friday, September 16 at 4 pm

1 Introduction

Zillow's fiasco inspired this problem set[1]. Zillow developed algorithms to buy houses. However, their models considerably overestimated the price of homes. This overestimation meant losses of about 500 million for the company and an approximate reduction of 25% of their workforce.

In this problem set, we will try to avoid(?) a similar fiasco while buying houses in Barranquilla.

The data set for this problem set comes from https://www.properati.com.co. The data contains information on listing prices as well as features of the properties on sale.

1.1 General Instructions

The main objective is to construct a predictive model of asking prices. From Rosen's landmark paper "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition" (1974), we know that a vector of its characteristics, $C = (c_1, c_2, \ldots, c_n)$, describes a differentiated good.

In the case of a house, these characteristics may include structural attributes (e.g., number of bedrooms), neighborhood public services (e.g., local school quality), and local environmental amenities (e.g., air quality). Thus, we can write the market price of the house as:

$$P_i = f(c_{i1}, c_{i2}, \ldots, c_{in})$$

However, Rosen's theory doesn't tell us much about the functional form of f. In this problem set, you will explore different models to yield the best prediction possible.

[1]For more info, see the following article here.

Please turn a pdf document to i.sarmiento@uniandes.edu.co. The document should not be longer than 6 (six) pages and include at most 6 exhibits (tables and/or figures). These exhibits are not counted in the six pages. You are welcome to add an appendix, but the main document must be self-contained. Specifically, a reader should be able to follow the analysis in the paper and be convinced it is correct and coherent from the main text alone, without consulting the appendix.

The document must contain the following sections:

- Introduction. Take this section as an opportunity to "sell" your predictive model, showing the advantages/disadvantages of your chosen model and expected performance.
- Data. Besides the variables included in the data set, here you are required to expand it, at a minimum you have to add four extra variables:
  - At least 2 of these models should include predictors coming from external sources; both can be from open street maps.
  - At least 2 predictors coming from the title or description of the properties.

  As always, treat this section as an opportunity to present a compelling narrative to justify or defend your data choices. I expect a deep analysis that helps the reader understand your data, its variation, and the justification for your choices. Use your professional knowledge to add value to this section. Do not present it as a "dry" list of ingredients. Describe it accordingly with descriptive stats, graphs, etc. At a minimum, you should include:
  - A table with descriptive statistics
  - A map
- Model and Results. When presenting your predictive model include:
  - An explanation of the variables used to train this model, remember to use the variables you added in the previous section.
  - A detailed explanation on how it was trained, the selection of hyper-parameters, and any other relevant information about the model.
  - A discussion of your evaluation measure.
- Conclusions and recommendations. In this section, you briefly state the main takeaways of your work.