Problem Set 2: Making Money with ML?
"It's all about location location location!!!"

1 Introduction

A new start-up dedicated to buying and selling properties just hired you and your team to develop a predictive model. Their objective is to buy the most properties in the neighborhood of Chapinero in Bogotá, Colombia while spending as little as possible.

The company has a sample of individual property data on Bogotá from https://www.properati.com.co. However, information about properties in Chapinero is mostly missing.

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

There are two expected outputs:

1. A `.pdf` document.
2. Submissions with your team's predictions in Kaggle. To join the competition use the following link.

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 amenities (e.g., crime, air quality, etc). Thus, we can write the market price of the house as:

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

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

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.

The document must contain the following sections:

- Introduction. The introduction briefly states the problem and if there are any antecedents. It briefly describes the data and its suitability to address the problem set question. It contains a preview of the results and main takeaways.

- Data[2]. In this problem set, you are required to add expand the variables in your data (remember to expand the training and testing data), at a minimum you have to add six extra variables:

  - At least 4 predictors coming from external sources; these can be from open street maps.
  - At least 2 predictors coming from the title or description of the properties.

  When writing this section up, you must:

  1. Describe the data, it's sutability for the problem, and the sample construction process, including how the data was cleaned, combined, and how new variables were created.
  2. Include a descriptive analysis of the data. At a minimum, you should include a descriptive statistics table and two maps with its interpretation. However, I expect a deep analysis that helps the reader understand the data, its variation, and the justification for your data choices. Use your professional knowledge to add value to this section. Do not present it as a "dry" list of ingredients.

- Model and Results. This section presents the model with the best score submitted for evaluation. When writing this section up, 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.
  - A section comparing the performance of the best-scored submission to other submissions submitted to Kaggle. These submissions must include specifications trained using Linear Regression, Ridge, Lasso, Elastic Net, CART, Random Forest, and Boosting models. Please indicate in your submission file the name of the used model.

- Conclusions and recommendations. In this section, you briefly state the main takeaways of your work.

[2] This section is located here so the reader can understand your work, but it should probably be the last section you write. Why? Because you are going to make data choices in the estimated models. And all variables included in these models should be described here.

2 Additional Guidelines

- Predictions have to be submitted on Kaggle. Check the competition website for more information.

- Turn a `.pdf` document in Bloque Neón. The document should not be longer than 8 (eight) pages and include, at most, 8 (eight) exhibits (tables and/or figures). Bibliography and exhibits don't count towards the page limit. 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 include a link to your GitHub Repository.
  - The repository must follow the template.
  - The README should help the reader navigate your repository. A good README helps your project stand out from other projects and is the first file a person sees when they come across your repository. Therefore, this file should be detailed enough to focus on your project and how it does it, but not so long that it loses the reader's attention. For example, Project Awesome has a curated list of interesting READMEs.
  - Include brief instructions to fully replicate the work.
  - The main repository branch should show at least five (5) substantial contributions from each team member.
  - The code has to be:
    - Fully reproducible.
    - Readable and include comments. In coding, like in writing, a good coding style is critical. I encourage you to follow the tidyverse style guide.

- Tables, figures, and writing must be as neat as possible. Label all the variables included. If you have something in your figures or tables, I expect they are addressed in the text. Tables must follow the AER format.