Problem Set 3: Making Money with ML?
"It's all about location location location!!!"

1 Introduction

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

The company has access to property-level data from Bogotá sourced from Properati. However, information specific to Chapinero is largely missing.

Critically, the company is determined to avoid a repeat of Zillow's costly mistake.[1] Zillow developed algorithms to automate home buying, but their models significantly overestimated property values. As a result, the company lost nearly USD 500 million and laid off approximately 25% of its 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 of this problem set is to construct a predictive model for asking prices of residential properties. Drawing on Rosen's foundational paper, "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition" (1974), we understand that a differentiated good—such as a house—can be described by a vector of its characteristics, $C = (c_1, c_2, \ldots, c_n)$.

In the context of housing, these characteristics typically 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 a function of these features:

---

[1] For more on this case, see: The $500MM Debacle at Zillow Offers

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

However, Rosen's theory doesn't tell us much about the functional form of f. This opens the door to explore a variety of predictive models. Your task is to evaluate different approaches and determine which one yields the best prediction accuracy

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 suitability 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 best-performing model submitted to the Kaggle competition. Each team must submit at least **ten (10) predictions**, using at least one model from **each** of the following algorithm families:

  Linear Regression, Elastic Net, CARTs, Random Forest, Boosting, Neural Networks, SuperLearners.

  The analysis should include:

---

[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.

- A clear statement of the variables included in the function estimated for the **best-performing model**. For example:

$$P_i = f(\text{Area}_i, \text{Bedrooms}_i, \text{Distance to Parks}_i) + u_i$$

  where P_i is the asking price, and the function $f(\cdot)$ captures the relationship between the outcome and both standard variables (such as size and number of bedrooms) and engineered variables (such as proximity to green areas).

- A detailed explanation of the **algorithm** used to estimate $f(\cdot)$ in the **best-performing model**. This should include the choice and tuning of hyperparameters, the validation strategy (e.g., cross-validation), and any additional methodological decisions that contributed to the model's performance.

- An analysis comparing the results from regular cross-validation and spatial cross-validation. Discuss how each approach affects model performance in Kaggle. Describe how spatial CV was implemented (e.g., splitting the data by UPZs, neighborhoods, or grid cells), and interpret any differences in the results.

- A comparison between the **best-performing model** and 10 of the highest-scoring submissions from your team. This comparison should highlight differences in performance on the Kaggle leaderboard and offer explanations for why some models outperformed others — whether due to differences in specification, feature selection, training strategy, or algorithm choice.

- A discussion of feature importance in the **best-performing model**. Identify which variables contributed most to predictive performance and provide empirical evidence for their role

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

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 10 (ten) pages and include, at most, 10 (ten) 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.