Problem Set 3: Making Money with ML?

"It's all about location location location!!!"

MECA 4107

Due Date: July 26 at 23:59 on Bloque Neón

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 Chapinero, Bogotá; and in El Poblado, Medellín.

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. The data are available in the `data.zip` file and contain three files: training data, testing data, and a submission template. You must submit your predictions for the testing data following the submission template.

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.

There are two expected outputs:

1. A `.pdf` document.
2. And a `.csv` file with predictions.

The document should not be longer than 3 (three) pages and include at most 6 exhibits (tables and/or figures). 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 (remember to expand the training and testing data), 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. Walk the reader through your reasoning of how you "cleaned" the data and expanded it with additional data. Describe it accordingly with descriptive stats, graphs, etc. At a minimum, you should include:

  - A table with descriptive statistics
  - Two maps (one for each main city), you can choose what information to show

  Use your professional knowledge to add value to this section. Do not present it as a "dry" list of ingredients.
- 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.

2 Additional Guidelines

I expect the following things from the problem set, omission of any of these guidelines will be penalized.

- Turn in your document and a submission `.csv` file with predictions in bloque neón.
  - An example of how the submission file should look is in the data folder: `submission_template.csv`. This file includes 2 columns, one with the variable that identifies a property and one with your predicted price.
  - I will judge predictions based on those who spend less money and can buy the most properties. That is if you over-price a property that adds to your tab, if you under-price, you save. However, if the predicted price is under-priced by more that $40mill.$ COP (approx. 10 thousand dollars), it will be penalized. In this case, the sale would not take place, i.e., you won't be able to buy the property.
  - Please follow the following convention for your `.csv` file name. The file name must include the name `predictions`, followed by your teammates' last names, and the number of variables that you used for prediction, all separated by underscores, for example, `predictions_gomez_matinez_sarmiento.csv`
  - I will assign bonus points based on relative rankings.
- Tables, figures, and writing must be as neat as possible. Label and describe all variables, statistics, etc., included in your figures and tables.
- The document should point and include a link to your GitHub Repository.
- Follow the repository template.
- The repository should include a README file. 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.
- The repository should have at least five (5) substantial contributions from each team member.
- Your code should be readable and include comments. In coding, like in writing, a good coding style is critical for readable code. I encourage you to follow the tidyverse style guide.