IBM APPLIED DATA SCIENCE CAPSTONE

Predicting land prices in Buenos Aires

EZEQUIEL A. PÁSSARO
epassaro@carina.fcaglp.unlp.edu.ar

1 Introduction

The goal of this project is to build a machine learning model for predicting land prices in the City of Buenos Aires. This work will be of special interest for real state investors and will help them in the process of making better business decisions.

Our main objectives are:

- Identify the most expensive neighborhoods through data analysis and visualization tools.
- Build a simple regression model for predicting land prices in the City of Buenos Aires.

2 Data acquisition

All the necessary data can be found at the City of Buenos Aires Open Data website[1] and through the *Foursquare API*:

- **Barrios:** a GeoJSON file with the geographic data of Buenos Aires Neighborhoods.
- **Terrenos:** a CSV file with information about pieces of land and their market value in Argentinian Pesos (ARS) and US Dollars (USD) for the year 2018. It also contains information about *latitude*, *longitude*, *price per square meter in USD*, *neighborhood* and more.

  This dataset is complete and cleaning tasks are not needed.
- **Foursquare API:** we will ask for the number of venues near a piece of land in a radius of 500m.

3 Methodology

First we found which neighborhoods are the most expensive according to data from *Terrenos* dataset by computing the *mean price per square meter per neighborhood*.

1. Recoleta (5968 USD/m2)
2. Palermo (5465 USD/m2)
3. San Nicolás (4692 USD/m2)
4. Belgrano (4490 USD/m2)
5. Colegiales (3143 USD/m2)
6. Balvanera (3139 USD/m2)
7. Caballito (2972 USD/m2)
8. San Telmo (2759 USD/m2)
9. Villa Crespo (2738 USD/m2)
10. Almagro (2699 USD/m2)

[1] https://data.buenosaires.gob.ar

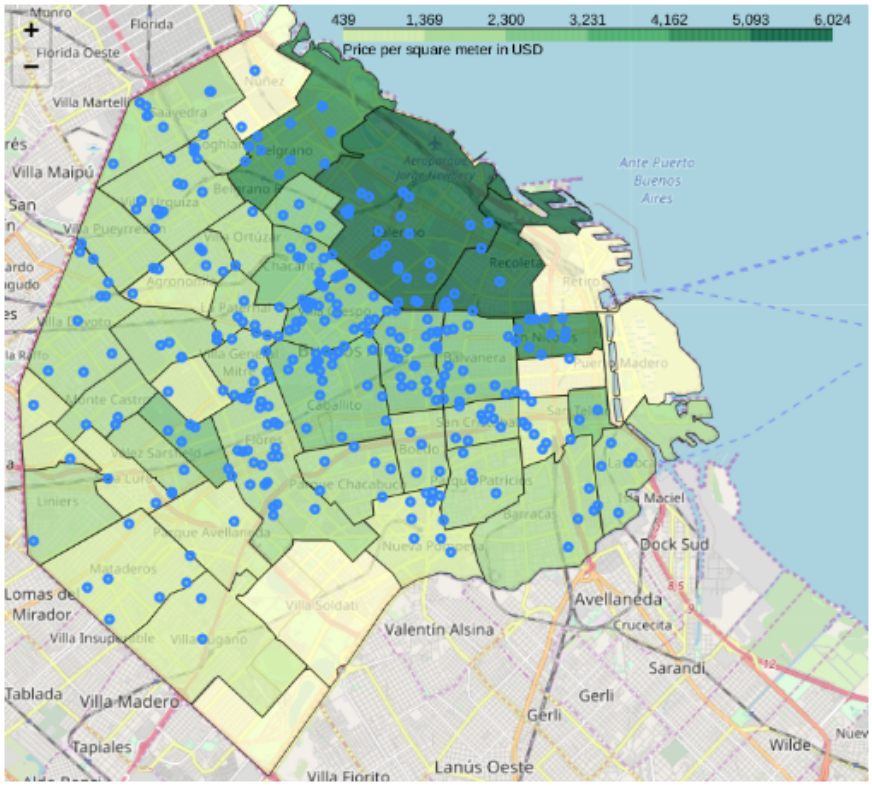

Figure 1: Price of lands per squared meter per neighborhood

We found that with a logarithmic transformation prices of lands are almost normal distributed. A scatter plot of *total square meters vs. price* in a log space suggests a linear regression could be the right model to choose.

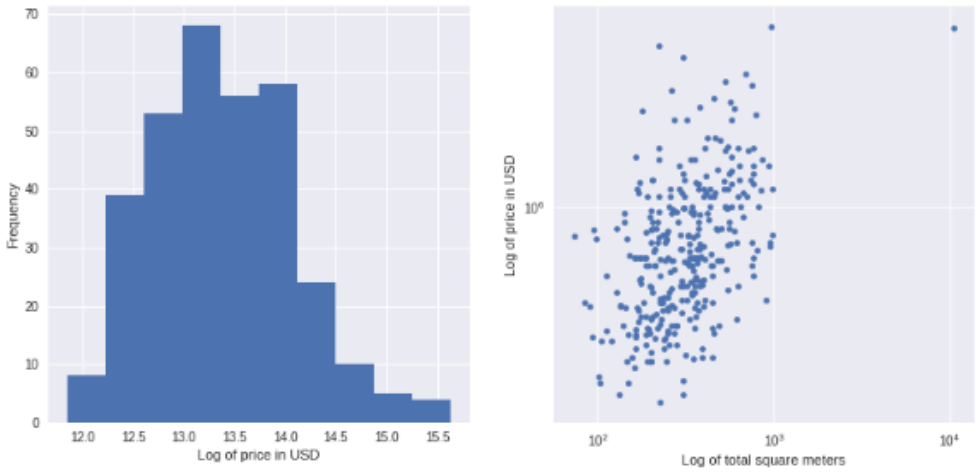

Figure 2: Distribution of prices

A good choice for the metric is the *Root Mean Squared Logarithmic Error* (RMSLE) which penalizes an under-predicted estimate example greater than an over-predicted estimate.

4 Results

After fitting four different models we obtained the following RMSLE metrics:

Model (Features)	No transform	Log-Log
Linear Regression (total square meters)	0.72	0.67
Multilinear regression (total square meters, venues near)	0.60	0.53

Values for *slope* and *intercept* were (0.59, 9.99) respectively in *linear regression (log-log)*.

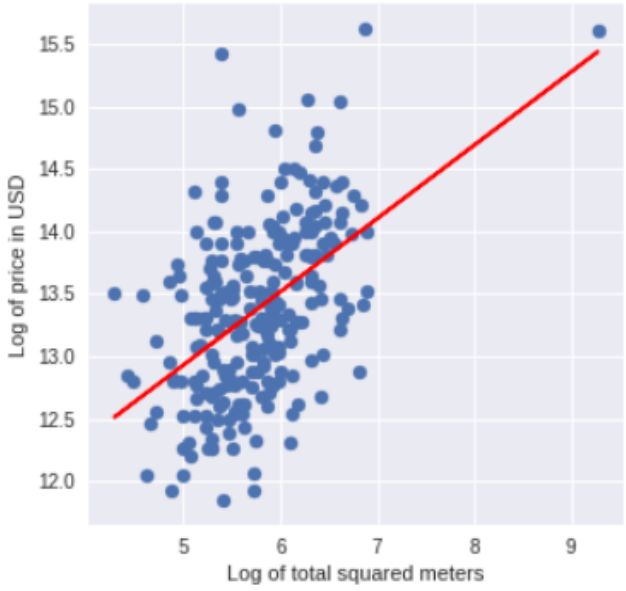

Figure 3: Linear regression

5 Discussion

It would be very convenient for future adjustments to gather more data about lands with a areas between 1.000 and 10.000 squared meters.

6 Conclusions

Despite being simple models, linear and multi-linear regression did a very good job on predicting prices. Also data provided by Foursquare improved our model predictive capabilities.