Housing Prices in Argentina

Group 9 - Part 3 Project Progress Report

JonPaul Ferzacca
CSPB 4502 - Peterson - Data Mining
University of Colorado Boulder
jofe1383@colorado.edu

Joey Musholt
CSPB 4502 - Peterson - Data Mining
University of Colorado Boulder
jomu7038@colorado.edu

ABSTRACT

In the Argentine real estate sector, our project utilizes a large dataset with over one million data points, encompassing diverse property listings. Our primary objectives include predicting property prices based on attributes like bedroom count and size, exploring the relationship between listing duration and price fluctuations, and deciphering the intricate interplay between property value and location. These insights empower local governance for urban planning and guide data-informed housing investment decisions.

Building upon prior work, including the Kaggle project by msorondo, we enhance our understanding of housing dynamics in Argentina. The dataset's key columns provide a comprehensive foundation for our analysis.

Our project encompasses Data Cleaning, Data Preprocessing, and Data Integration phases. Data Cleaning addresses missing data points, Ad Type inconsistencies, and potential data exclusion. Data Preprocessing involves currency conversion and feature engineering. Data Integration enriches our dataset with external data for enhanced analysis.

Utilizing tools such as GitHub, Python, and relevant libraries, we employ a comprehensive evaluation framework. This includes assessing predicted vs. labeled prices, overall model metrics, subset and temporal analyses, geospatial evaluations, and feature importance analysis.

Our project contributes to an enhanced understanding of the Argentine real estate landscape, offering valuable insights for stakeholders.

KEYWORDS

Real Estate Data Analysis, Predictive Modeling, Housing, Market Trends, Data Cleaning, Data Preprocessing, Data Integration, Geographic Information Systems, Property Price Prediction, Feature Engineering, Machine Learning, Urban Planning, Data Enrichment

1 Description

In the dynamic Argentine real estate sector, our extensive dataset comprises a comprehensive .csv file, encompassing a staggering one million individual data points, capturing a multitude of property listings. Our project seeks to unravel intriguing inquiries, such as predicting property prices utilizing various attributes like bedroom count and size, investigating the connection between listing duration and price fluctuations, and deciphering the intricate relationship between property value and location. The profound insights unearthed through this data-driven exploration will not only empower local authorities with valuable information for urban planning but also equip real estate investors with the knowledge to make astute housing investments.

1.1 Prior Work

One notable example of prior work in this field is the Kaggle project by msorondo, which presents a comprehensive housing price predictor for Buenos Aires. This project serves as a valuable reference, showcasing effective methodologies and techniques for addressing similar challenges in the Argentine real estate landscape. By leveraging insights from these prior endeavors, our project aims to contribute further to the understanding of housing dynamics in Buenos Aires, offering fresh perspectives and enhanced predictive models.

1.2 Data Sets[1]

id - Notification Identifier: This identifier is not unique, as each notification may be updated by the real estate agency, creating new records with the same id but different registration and cancellation dates.

operation_type - Operation Type: This column categorizes the type of operation, which, in this dataset, is primarily sales. (Note: If all entries are sales, this column may be redundant and can be considered for removal).

l2 - Administrative Level 2: Typically represents the administrative division at the provincial level.

l3 - Administrative Level 3: Usually corresponds to the city or locality within the administrative hierarchy.

lat - Latitude: Geographic coordinate indicating the north-south position of the property.

lon - Longitude: Geographic coordinate representing the east-west position of the property.

price - Published Price: The price listed in the advertisement for the property.

property_type - Property Type: Specifies the type of property, which can include options like House, Apartment, or PH (Preservation House).

rooms - Number of Rooms: The count of rooms available within the property, a useful metric in the context of Argentine real estate.

bathrooms - Number of Bathrooms: Indicates the quantity of bathrooms within the property.

start_date - Advertisement Start Date: The date when the advertisement was first created.

end_date - Advertisement End Date: The date when the advertisement concluded or was removed.

created_on - Notice Creation Date: The date when the initial version of the notice was generated.

surface_total - Total Area (m²): The overall area of the property in square meters, encompassing both covered and open spaces.

surface_covered - Covered Area (m²): The area of the property that is enclosed and covered, measured in square meters.

title - Advertisement Title: The title or headline of the advertisement, providing a concise description of the property.

description - Advertisement Description: A more detailed textual description of the property, providing additional information beyond the title.

ad_type - Advertisement Type: Categorizes the type of advertisement, differentiating between Property listings and Development/Project-related ads.

1.3 Proposed Work

In the course of our project, we will undertake a comprehensive data preparation process encompassing three key phases: Data Cleaning, Data Preprocessing, and Data Integration.

Data Cleaning: Our initial focus will be on ensuring the quality and consistency of the dataset. This entails the cleaning of missing data points, with specific attention to Longitude and Latitude, which are vital for spatial analysis. Additionally, we will address inconsistencies in the Ad Type variable and consider the potential exclusion of any data related to Uruguay. To enhance data reliability, we will also conduct outlier detection and removal procedures.

Data Preprocessing: In this phase, we will transform the dataset to make it suitable for analytical modeling. This involves converting price values to a uniform currency for accurate comparisons. We will engage in feature engineering to extract valuable insights from the dataset, including date extraction to better understand temporal trends. Furthermore, we will convert categorical variables into numerical formats, facilitating their integration into machine learning algorithms.

Data Integration: The final phase will involve integrating external datasets to augment our understanding of the Argentine real estate landscape. This will include map integration for

spatial analysis and the incorporation of currency conversion rates to account for fluctuations. Leveraging location data, we will calculate distances to important landmarks and city centers, providing valuable context for property listings and their respective locations. These integration efforts will enrich our dataset and pave the way for more insightful analyses and predictive modeling.

1.4 Evaluation

To comprehensively assess our project's outcomes, we will employ a diverse set of evaluation methods:

Predicted vs. Labeled Price: We will measure the accuracy of our predictive models by comparing their estimated property prices to the actual labeled prices in the dataset, providing a direct evaluation of pricing accuracy.

Overall Model Metrics: Utilizing standard machine learning metrics such as accuracy, precision, recall, and the F1-score, we will gauge the models' performance in classifying and predicting property prices, considering true and false positives.

Subsets and Temporal Analysis: We will evaluate model performance within specific data subsets (e.g., property types, locations) and conduct temporal analysis to uncover trends and assess predictive consistency over time.

Geospatial Analysis: A geospatial assessment will measure models' accuracy in predicting property prices based on proximity to landmarks, aiding in understanding spatial predictive capabilities.

Feature Importance: We will identify key influencing factors (e.g., room count, location) by assessing feature importance in our models.

By applying this comprehensive evaluation framework, we aim to gain valuable insights into project impact while optimizing predictive accuracy for the Argentine real estate market.

1.5 Tools

We will use GitHub to host all files related to the project but the data itself, which has a large file size for which GitHub is unsuited.

For the cleaning, preprocessing, integration, etc., we will be working with Python. Specifically, NumPy for computational performance and general use, Pandas for importing and manipulating the data through DataFrames, and Sklearn to import models.

For visualization and graphing of our data and model, we will use MatPlotLib and Seaborn. For live visualizations, we may also use Tableau.

1.6 Milestones

Completed So Far:

11/10 - Cleaning & Preprocessing ☑

11/24 - First pass of Model & data integration ☑

Remaining:

12/1 - Final model & draft report/presentation ▢

12/7 - Final report/presentation ▢

1.7 References

[1] msorondo. (2021, March 17). *2 million rows of data on homes for sale*. Kaggle. https://www.kaggle.com/datasets/msorondo/argentina-venta-de-propiedades/data

1.8 Results So Far

The first step to finding meaningful results is data cleaning. To that end, we have converted dates using the datetime library, removed missing values, and integrated the ARS

to USD conversion dataset to potentially improve predictive power by controlling for the volatility of the Argentine Peso.

Additionally, we've scoped out the distribution of some of the continuous features in order to get a sense of the dataset's landscape. Figure 1 shows the distribution of listing dates. There is a clear imbalance in the data towards earlier listings (possibly Covid-related based on the timing), which combined with currency volatility may cause issues down the line.

As price is our class label, getting a sense of its distribution is also crucial, as can be found in Figure 2. As expected, there are many more listings at lower price levels. Our model will have to account for this fact to avoid predictions that only work for one end of the price spectrum. Interestingly, some ~0.1% of listings are for 0 ARS, which will have to be investigated further or potentially removed as erroneous data.

The most important chart for investigating our continuous features is the correlation matrix. Log versions of each have also been included in order to pre-emptively scope out non-linear relationships. This can be found in Figure 3. On an individual feature level, price appears to be quite unrelated to the others. However, log_price has much better correlation values, suggesting a non-linear relationship between many of the features. Still, since the dataset covers all of Argentina, from rural areas to mid-size cities to Buenos Aires, it will be necessary to use some of the discrete features to help segment the prices between markets.

Fig 1: Distribution of Listing Creation Date

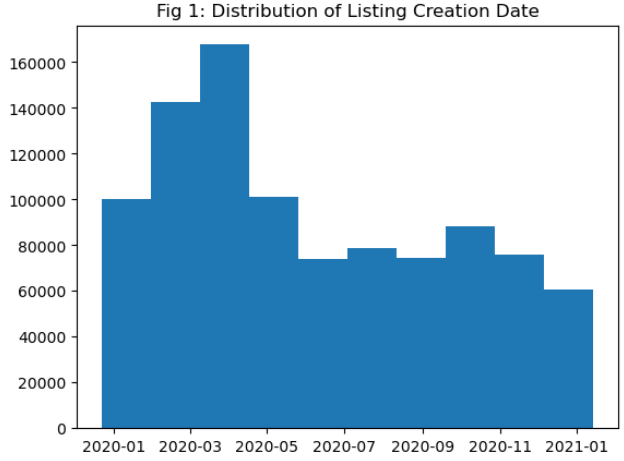

Fig 2: Distribution of Listing Price in ARS

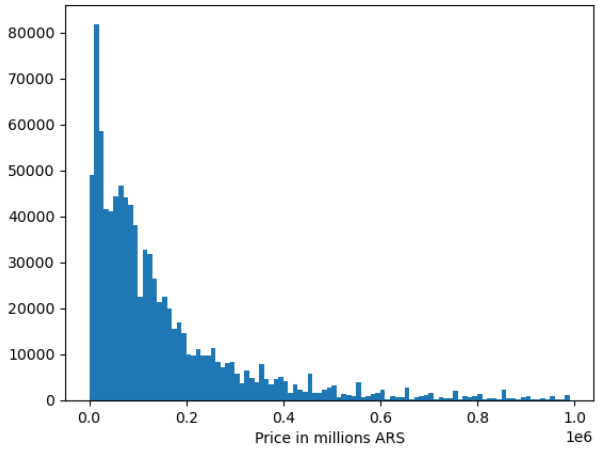

Fig 3: Correlation Matrix