Housing Prices in Argentina

Group 9 - Part 4 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 Introduction

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.

The overarching goal of this data mining project is to predict the price of real estate listings, but an important part of the explainability of the resulting model is the relative importance of different features in determining the price. This is expressed in both the predictive power of each feature as well as the actual magnitude of the relationship. An interesting question beyond “What should the price of this listing be based on certain real estate characteristics?” is “What is the impact of having an additional square meter of space on a listing’s price?” Other similar questions can also be addressed through mining the data.

The question of which features are most

relevant to determining listing price is important because it reveals what prospective buyers see as valuable in a property. This could be used to shape what developers include in new construction or enhancements to existing properties by maximizing price at a given cost or assist a buyer trying to get the best property within their budget: maybe that extra room isn't worth the associated extra cost, for instance.

The primary model function of predicting price is important because it can help property sellers set a price that is in line with the market or help buyers identify whether a given listing price is high or low relative to the market. It can also help to quantify unique circumstances of a property that are not captured by this dataset's features. For instance, proximity to restaurants or parks might increase the price of a listing, which could be measured as the gap between the listing price and the predicted listing price. These intangible benefits are real but harder to quantify, and a price prediction model can control for certain features like square meterage, location, or room count to better isolate those intangible benefits.

The final reason a population-wide pricing prediction model is important is that it represents the state of the market at a point in time (a year or so in our dataset). Comparing this model to a similar analysis in the future could give insight into what has changed about what the Argentine real estate market values. For instance, declining average household size could result in an increase in price for units having fewer bedrooms. Taking a snapshot of the market's current preferences through this data mining exercise will help contextualize future preferences.

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. Given Argentina stretches far from north to south, there is a lot of variation in this variable relative to longitude. Additionally, the length of a degree of longitude is determined by the latitudinal position, especially at points far from the equator such as Argentina.

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 Techniques

In the course of our project, we undertook a comprehensive data preparation process encompassing two phases: Data Cleaning and Data Preprocessing.

Data Cleaning: Our initial focus was on ensuring the quality and consistency of the dataset. This entailed cleaning missing data points (NaNs especially). The dataset has many missing values in the numerical variables, but few listings had none at all. As a result, we dropped those missing too many numerical variables and imputed the missing values for those that remained using the median of the dataset for the given missing feature. We also needed to convert many of the raw data types from strings to floats as well as dates from strings to datetimes. Lastly, we dropped columns ‘ad_type’, ‘l1’, ‘l4’, ‘l5’, ‘l6’, ‘title’, ‘description’, ‘currency’, and ‘price period’ as these were either unsuitable for data mining (within the scope of this course) due to being written descriptions or having sparse data that was too irregular to use for predictive purposes.

Data Preprocessing: In this phase, we transformed the dataset to make it more suitable for analytical modeling. This involved feature engineering and transformations of existing features. We created a log version of each feature to help capture non-linear relationships, using np.log1p() to handle zero values. We also created two time interval features: ‘listing_duration’ using the end_date and start_date of each listing and ‘listing_age’ to capture the age of the listing relative to the current day. These additional features help put context to the raw features already available in the data and add predictive power to the model. To make use of location data, we created the feature ‘distance_to_central’, calculating the distance in kilometers from the listing to the center of Buenos Aires. This location was chosen as it is the largest city in Argentina and a large portion of the listings were located there. This distance should stand in for location importance, as listings far from the city may be less desirable and fetch a lower price.

Model Building: Once the data had been cleaned and preprocessed, we selected a linear regression model using the numerical variables as well as one-hot encoded dummy variables for ‘property_type’ and ‘operation_type’. The target variable was log_price to account for non-linearities in the dataset. A linear regression was chosen for its explainability and resistance to overfit. The impact of each feature is quantified and, as opposed to classification methods, the price can be precisely predicted numerically rather than predicting a price’s range bucket.

1.4 Evaluation

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

Predicted vs. Labeled Price: We measured 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. For log_price, we have a mean squared error of 1.008.

1.5 Tools

We used 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 worked 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 used MatPlotLib and Seaborn.

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 Key Results

Before digging into the model's predictive results, we will review the results of initial data exploration. 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 accounts for this fact through predicting log_price to avoid predictions that only work for one end of the price spectrum. Interestingly, some ~0.1% of listings are for 0 ARS.

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 our assumption of a non-linear relationship between many of the features was correct. Still, since the dataset covers all of Argentina, from rural areas to mid-size cities to Buenos Aires, we used the engineered feature distance_to_central to try to account for this diversity of location.

The Error Distribution Chart demonstrates that the residuals are normally distributed, giving us confidence in the model. The residuals are heavily clustered around 0.0, with minimal tails, indicating that the error is inherent to the variance in the data rather than a defect of the model.

The R-Square of our model is 50.6%, which is fairly high for a dataset of this kind. This means that the model eliminates roughly half the variance of the dataset, the remainder likely being due to features outside the scope of the listing dataset. These could include specifics of a property's appearance, neighborhood, proximity to other nearby landmarks, stores, or parks, etc.

1.9 Applications

We can use the knowledge gained from mining this dataset to both predict prices and analyze the relative impact of different features on property price. The features were all statistically significant in terms of P-values and were largely positive in magnitude. This allows for comparing the relative impact of different features on listing price and provides useful context for a listing's price relative to predicted price. In the future, similar analyses could be conducted to measure changes in what features have changed in relative importance in the Argentine real estate market.

Fig 3: Correlation Matrix

Fig 2: Distribution of Listing Price in ARS

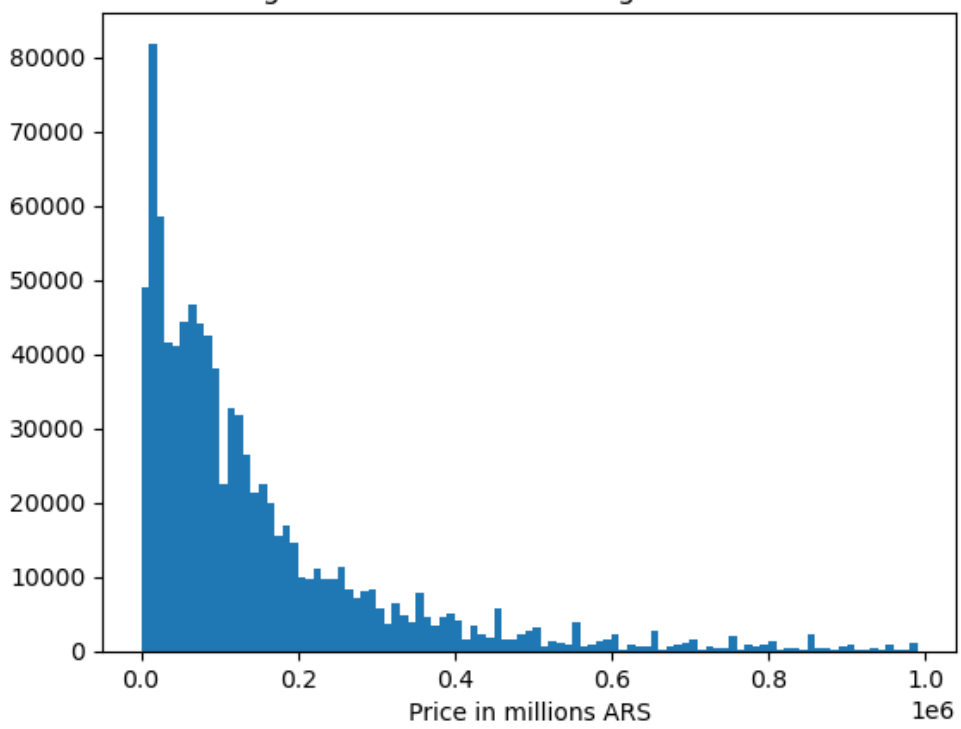
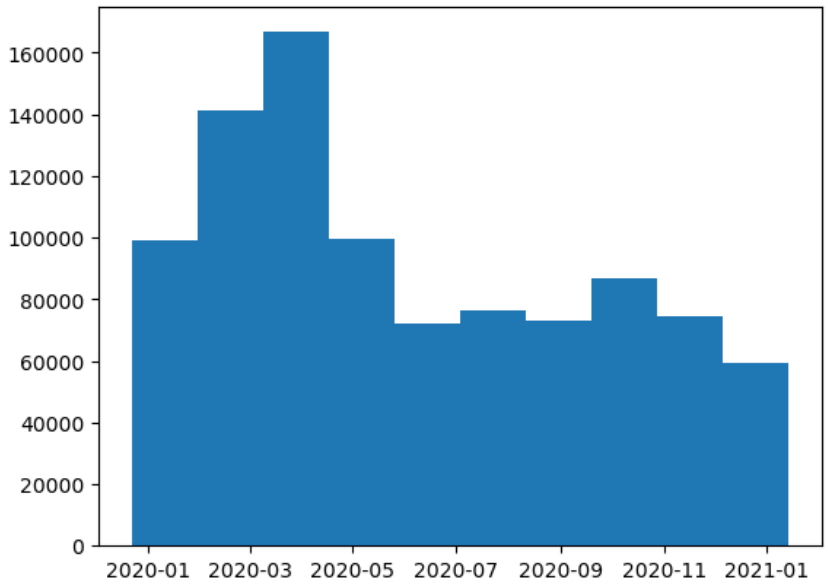


Fig 1: Distribution of Listing Creation Date



Coefficients:

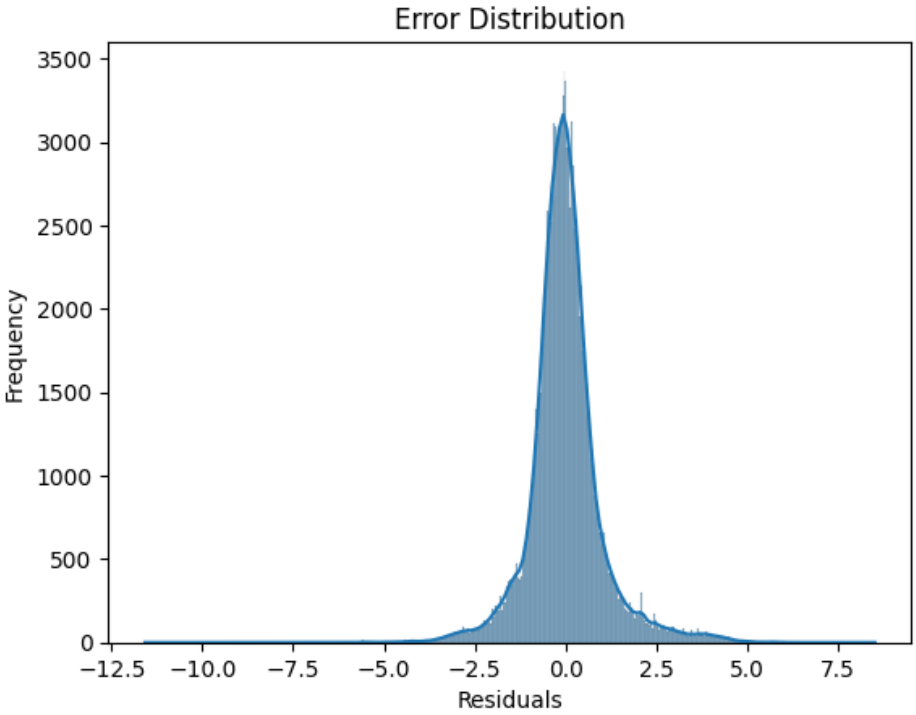
const	1.165277e-04
lat	3.448589e-02
lon	-2.088955e-02
rooms	3.419274e-02
bedrooms	1.091814e-03
bathrooms	2.934902e-01
surface_total	1.035122e-05
surface_covered	3.938042e-10
log_rooms	-3.985399e-02
log_bedrooms	5.673228e-02
log_bathrooms	-6.487932e-02
log_surface_total	1.917993e-02
log_surface_covered	-9.654320e-03
listing_duration	-5.675962e-03
listing_age	2.889762e-02

dtype: float64

P-values:

const	0.000000e+00
lat	1.408543e-92
lon	2.027306e-24
rooms	8.678894e-45
bedrooms	2.448840e-01
bathrooms	0.000000e+00
surface_total	4.903484e-102
surface_covered	3.967332e-01
log_rooms	2.107988e-08
log_bedrooms	3.280916e-43
log_bathrooms	1.336386e-13
log_surface_total	4.586359e-111
log_surface_covered	4.910636e-23
listing_duration	0.000000e+00
listing_age	0.000000e+00

dtype: float64


Error Distribution
Frequency
Residuals
3500
3000
2500
2000
1500
1000
500
0
−12.5
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5