Problem Set 3*†

Sergio Sandoval‡ Maria Fernanda Blanco
Juan Gutierrez Juan Fernando Barberi

"Its all about location location location!!!"

1 Introduction

Colombia's real estate sector has faced significant challenges in recent years, with the housing market contracting by 3.1 percentage points between 2022 and 2023. However, recent data indicates a partial recovery, with a 1.1 percentage point increase in the second quarter of 2024 (DANE, 2024). These fluctuations, driven by shifts in interest rates and public policy, underscore the need for innovative tools like machine learning models to analyze market dynamics and estimate property prices with precision.

This study focuses on developing a predictive model to estimate housing prices in the Chapinero locality of Bogotá, Colombia, for a start-up specializing in property transactions. The objective is to optimize property acquisitions while minimizing costs, despite challenges such as incomplete data and market volatility. Robust methodologies are essential to bridge these gaps and ensure reliable predictions. The case of Zillow's "Zillow Offers" program, which suffered over $500 million in losses due to overestimations and an inability to adapt to market changes, serves as a critical reminder of the importance of dynamic recalibration and rigorous validation in predictive modeling.

Machine learning has revolutionized housing price prediction, with models like XGBoost consistently demonstrating high accuracy and computational efficiency in diverse settings. Studies by Jha et al. (2020) and Sharma et al. (2024) underscore XGBoost's effectiveness in predicting housing prices in Florida and Iowa, respectively, while Vargas-Calderón and Camargo (2019) highlighted its adaptability in Bogotá by integrating textual and numerical data. These findings position XGBoost as a powerful tool for real estate price prediction, offering actionable insights for navigating the complexities of real estate markets.

Colombia's real estate market poses significant challenges, particularly in localized areas like Chapinero, Bogotá, where incomplete data and complex market dynamics complicate accurate property price estimation. This study tackled the problem by developing a predictive model for a start-up aiming to optimize property acquisitions while minimizing costs. To address these challenges, the analysis combined a comprehensive dataset encompassing structural features (e.g., bedrooms, bathrooms, area), spatial characteristics (e.g., distances to amenities and socioeconomic strata), and textual data extracted from property descriptions. These diverse data sources provided a rich foundation for addressing the multifaceted nature of property pricing in an urban context.

The strategy relied on rigorous data construction and the application of multiple machine learning models, including Linear Regression, Elastic Net, LightGBM, Neural Networks, and a Super Learner ensemble. LightGBM emerged as the most effective standalone model, while the Super Learner demonstrated the value of integrating diverse approaches for superior accuracy. The findings underscore the importance of robust data preprocessing and model diversity in addressing complex real estate problems. While the study successfully provided actionable insights, it also highlighted areas for improvement, such as incorporating richer temporal data and overcoming computational limitations for certain models. These results reinforce the potential of machine learning as a powerful tool for navigating the complexities of evolving real estate markets.

*Big Data and Machine Learning, Universidad de los Andes.
†Excluding tables and figures, the document is less than 8 pages.
‡Link of the repository: https://github.com/setosandoval/BDML_Team1_2024.git

2 Data

2.1 Source and Description

The dataset used in this analysis originates from Properati, covering property listings in Bogotá, Colombia, between 2019 and 2021. The data is divided into a training set, which includes information on properties across Bogotá, and a test set, which focuses exclusively on properties located in Chapinero, neighborhood for which price data is missing. The dataset includes key variables such as price, bedrooms, bathrooms, surface area, and covered surface area. However, aside from the variable indicating the number of bedrooms, most of these features exhibit significant levels of missing data, making imputation a necessary step. This presents a central challenge in the analysis, as the goal is to construct a predictive model capable of estimating property prices in Chapinero while relying on incomplete data from the training set.

The analysis leverages textual information from the variable "description", which contains detailed narratives about each property listing. This textual data not only facilitates the imputation of missing values but also serves as a rich source for creating additional predictive variables through natural language processing. Furthermore, given the geographic coordinates available in the dataset, spatial variables, such as the distance to key urban amenities like parks, schools, and public transportation, can also be incorporated. This aligns with the hedonic pricing framework proposed by Rosen (1974), which posits that a property's price is a function of its structural characteristics, neighborhood features, and accessibility to local amenities. By integrating these diverse data sources, this study aims to construct a robust predictive model to address the unique challenges of this problem.

2.2 Location Variables

The spatial characteristics of properties play a critical role in determining their market prices, as suggested by the hedonic pricing framework. In this study, we utilized the geographic location of each property to incorporate a wide range of external spatial variables. First, we employed the official boundaries of Bogotá's urban planning zones (UPZ), sourced from Bogotá's open data portal, Datos Abiertos Bogotá. The city is divided into 117 UPZs, and each property was assigned to its corresponding UPZ based on its coordinates. This allowed us to include a fixed-effect variable for UPZ, capturing neighborhood-specific location effects. Similarly, using the same geographic information, we assigned each property to one of Bogotá's 20 localities, enabling the creation of another fixed-effect variable to reflect broader regional influences on property prices.

Further granularity was achieved through the use of stratified block data. From Datos Abiertos Bogotá, we accessed detailed information on more than 40,000 urban blocks, each categorized by socioeconomic stratum. While including block-level fixed effects was computationally infeasible, we utilized this data to calculate several predictive features for each property, such as the average socioeconomic stratum of its nearest block, the average household size, the number of residential properties, and population density per block. Additionally, data from the Departamento Administrativo Nacional de Estadística (DANE) provided complementary information at the block level, including cadastral values and commercial property valuations. These variables serve as proxies for local economic conditions and housing market dynamics, offering additional insights into the determinants of property prices.

Finally, to further enrich the dataset, we integrated spatial data from OpenStreetMap to compute the distance of each property to 12 key urban amenities: banks, public transportation stations, parks, cycling paths, major roads, hospitals, shopping malls, supermarkets, police stations, restaurants, schools, and universities. The inclusion of these amenities captures the accessibility and desirability of each location, which are critical factors influencing property demand and pricing. By combining these spatial variables with neighborhood-level and block-level data, we created a robust framework to account for the multi-dimensional impact of location on property prices, ensuring that this fundamental determinant is thoroughly addressed in the predictive modeling process.

2.3 Text Variables

The descriptive text provided in the “description" variable offered a rich source of information to enhance the predictive power of the dataset. To extract meaningful insights, the text was first subjected to a thorough cleaning and normalization process. This included removing special characters, punctuation, and stop words, as well as standardizing the case and lemmatizing words to their base forms. These preprocessing steps, performed using libraries such as those provided by the Departamento Nacional de Planeación (DNP), ensured consistency across the dataset while reducing noise and redundancy. Lemmatization was particularly important as it grouped inflected forms of words under a single root, making the data more compact and interpretable. These steps were crucial for both the numerical extraction and the text-to-feature transformation processes.

The first approach focused on extracting key numerical variables explicitly mentioned in the property descriptions, such as the number of bedrooms, bathrooms, parking spaces, floors, and surface area. This process aimed to recover critical structural features of the properties that were missing or incomplete in the structured data. The second approach involved transforming the cleaned descriptions into a Bag of Words (BoW) representation, designed to capture the presence of descriptive terms that provide valuable information about property attributes. Specific words such as “amplio” (spacious), “vista” (view), “cocina” (kitchen), and “terraza” (terrace) were of particular interest, as they are often linked to features that enhance property value. For this purpose, we created dummy variables indicating whether a given word appeared in the description or not, thereby encoding qualitative aspects of the properties into binary predictors. To refine the BoW representation and control the size of the feature space, we applied three frequency filters, retaining words that appeared in at least 1%, 5%, and 10% of the descriptions, which resulted in three datasets of varying sizes: light, medium, and large. This approach allowed us to explore the predictive power of nuanced textual features, capturing aspects of property descriptions that are not explicitly reflected in numerical variables. By encoding these textual features into structured predictors, the analysis aimed to complement the structured data with rich qualitative insights about the properties.

The high dimensionality of the BoW datasets, resulting from the inclusion of thousands of variables in each set, necessitated dimensionality reduction through Principal Component Analysis (PCA). The application of PCA allowed us to condense the information into a manageable number of components while retaining the most significant patterns and variation in the text. After applying PCA, the smallest BoW dataset (filtered at 10% frequency) resulted in 42 principal components, while the largest dataset (filtered at 1% frequency) contained nearly 400 components. This dimensionality reduction step was crucial to ensure computational efficiency in subsequent analyses, as well as to address issues of multicollinearity inherent in high-dimensional text data. Each of the three BoW-based datasets—light, medium, and large—was combined with the previously constructed variables from structured and spatial data, resulting in three comprehensive datasets of varying sizes. These datasets offered flexibility for modeling, enabling us to assess trade-offs between the richness of textual features and the computational cost, ultimately enhancing the predictive power of the framework.

2.4 Final Data

The final dataset was constructed by combining the textual variables derived from the Bag of Words (BoW) approach with the spatial variables and the original structured data. This integration resulted in three datasets of varying sizes (light, medium, and large), each containing a substantial number of variables. However, a critical challenge was addressing the missing values present in key variables such as bathrooms, surface area (m^2), and additional rooms, which had distinct patterns of missingness across the initial dataset and the textual features. To address this, the minimum value between the corresponding numerical variables from both sources was used as a conservative imputation method, significantly reducing the missingness. Despite this improvement, missing values remained in these and other numerical variables derived from the textual analysis, such as the number of parking spaces and floors. To further address these gaps, a hierarchical imputation strategy was applied. Missing values were first replaced with the median value calculated at the block level, as blocks capture

fine-grained local variation. If data were still missing, the imputation moved to broader geographic levels: UPZ, locality, and finally the entire dataset. The median was chosen over the mean to minimize the influence of outliers and provide a more robust central tendency measure, ensuring a realistic imputation in highly skewed distributions. This hierarchical imputation process ensured a systematic and contextually relevant filling of gaps in the dataset, maintaining consistency across the predictive features.

The integration of textual, spatial, and structured variables, coupled with a systematic imputation strategy, creates a robust dataset that captures the complex factors influencing property prices. By addressing missing values through hierarchical imputation and leveraging a wide array of predictors, the dataset effectively reflects the interplay between structural characteristics, location-specific effects, and qualitative descriptions. This comprehensive approach ensures the data is well-suited to support predictive modeling and address the objectives of understanding property price determinants in Bogotá.

2.5 Descriptive Statistics

The final dataset consists of 48,930 observations, with 38,644 belonging to the training set and the remaining 10,286 to the test set. For the analysis, the medium-sized dataset was selected as it provided a balance between predictive performance and computational feasibility, avoiding the challenges posed by the large dataset's dimensionality. The resulting dataset includes 118 variables: 6 numerical features capturing structural characteristics such as bedrooms and bathrooms, 83 principal components derived from textual data, a temporal variable indicating the month, and the remainder comprising spatial variables related to location and neighborhood effects.

In *Figure 1*, the histogram on the left illustrates the distribution of property prices in Bogotá, which exhibits a pronounced right-skewed shape. Most properties are concentrated below 1,000 million COP, with a steep decline in frequency as prices increase beyond this point. This skewness reflects the presence of a majority of moderately priced properties alongside a smaller number of high-priced outliers that extend the upper range of the distribution. On the right, the time series plot shows the evolution of mean and median property prices from early 2019 to mid-2021. Both metrics demonstrate a general upward trend, signaling an appreciation of property values over the analyzed period. The mean remains consistently higher than the median, indicating the influence of high-priced properties pulling the average upward. Notably, there are visible fluctuations in both measures, with more pronounced variability in the mean, which further suggests the impact of extreme values on average prices.

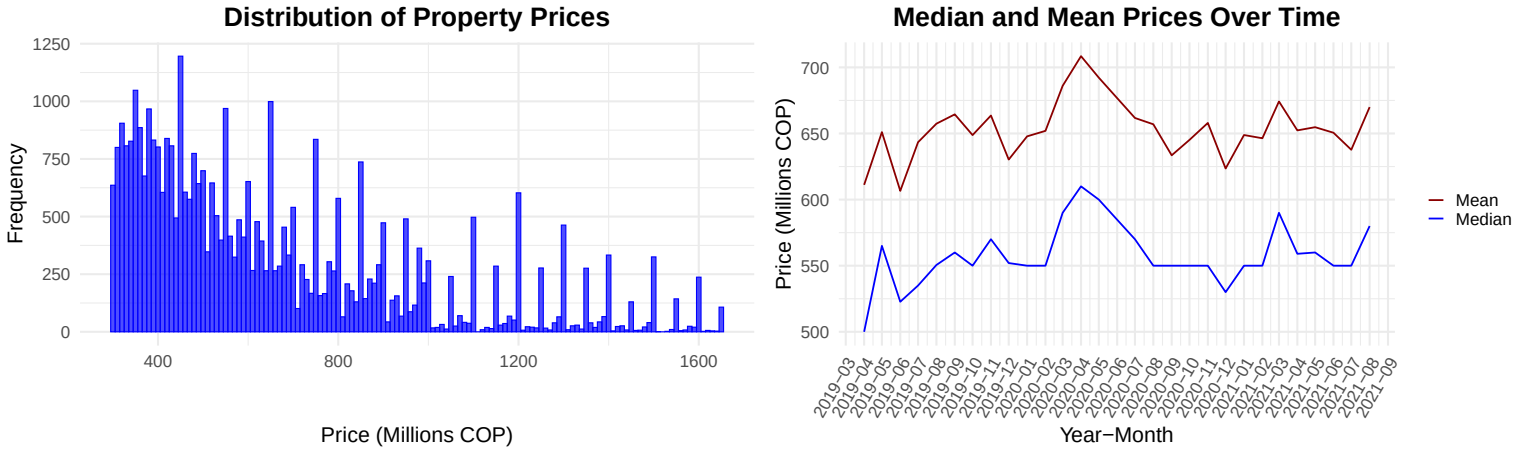

Figure 1. Distribution and Time Series (2019-2021) Property Prices

In *Figure 2*, the first map provides a spatial representation of median property prices across Bogotá. Higher median prices are concentrated in central areas, particularly in neighborhoods such as Chapinero, where the darkest shades of blue indicate the highest values. This spatial gradient

illustrates the significant role that location plays in determining property prices, with central zones benefiting from proximity to key urban amenities and infrastructure. In contrast, peripheral areas exhibit lower median prices, reflecting reduced demand and access to fewer services. This spatial variation underscores the importance of geographic location as a primary determinant of property values in the city

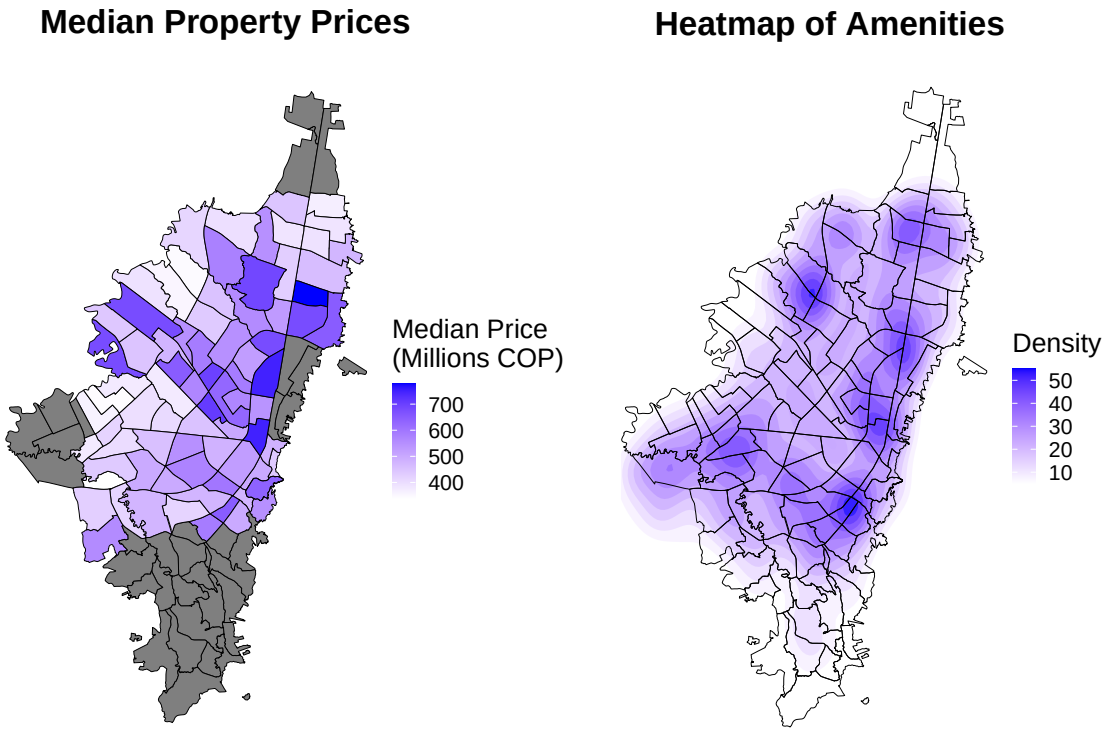

Figure 2. Spatial Distribution of Median Property Prices and Amenities

Table 1 provides a summary of the main numerical variables used in the analysis, including price, structural features, and spatial characteristics. The property price variable (in millions of COP) reflects significant variability, with a mean of 654.53 million COP and a standard deviation of 311.42 million COP. Prices range from a minimum of 300 million COP to a maximum of 1,650 million COP, with a median value of 559.99 million COP, indicating a skewed distribution. Structural features such as bedrooms, rooms, bathrooms, floors, parking spaces, and area reveal diverse property types within the dataset. On average, properties have 2.98 bedrooms and 2.61 bathrooms, while the number of rooms, which aggregates living spaces, averages at 2.76. The average number of floors is 3.64, with some properties having up to 50 floors, reflecting a mix of single-family homes and larger residential buildings. Area varies widely, with a mean of 125.80 square meters but ranging from 30 to 1,000 square meters, further emphasizing the dataset's heterogeneity. Parking spaces average at 1.83 per property, with some properties offering no parking and others up to five spaces.

Spatial and neighborhood-related variables add additional depth. The socioeconomic stratum (*estrato*) has an average value of 4.73, covering a range of Bogotá's socioeconomic diversity from strata 1 to 6. Block-level density is relatively low, with a mean of 0.03 persons per square meter, while the number of houses per block averages at 185.68, and persons per block average 402.37, though these figures display substantial variation, as shown by their high standard deviations. Cadastral values and commercial property values add an economic perspective to the data, with averages of 2.8 and 3.6 million COP, respectively, but significant variability suggests the coexistence of low and high-value areas within the city. Together, these variables provide a rich and detailed representation of both the structural and locational characteristics of the properties.

Statistic	N	Mean	St. Dev.	Min	Median	Max
Price (millions COP)	38,644	654.53	311.42	300.00	559.99	1,650.00
Bedrooms	48,930	2.98	1.47	0	3	11
Rooms	48,930	2.76	1.02	1.00	3.00	15.00
Bathrooms	48,930	2.61	1.04	1.00	2.00	15.00
Floors	48,930	3.64	3.99	0.00	3.00	50.00
Parkings	48,930	1.83	0.62	0.00	2.00	5.00
Area	48,930	125.80	83.82	1.00	107.00	1,000.00
Estrato	48,930	4.73	1.13	1	5	6
Density per Block	48,930	0.03	0.02	0.00	0.03	0.39
Houses per Block	48,930	185.68	229.20	0	111	2,951
Persons per Block	48,930	402.37	537.86	0	231	7,391
Cadastre Value	48,930	2.8	2.2	0	2.7	43.3
Commercial Value	48,930	3.6	2.8	0	3.5	56.1

Table 1. Descriptive Statistics of Selected Variables

Figure 3 explores the relationship between socioeconomic stratum (*estrato*) and property prices through boxplots for each stratum. A clear positive trend emerges, with median prices increasing consistently across strata. Properties in stratum 1 exhibit the lowest median prices, with a wider interquartile range and a higher density of lower-priced properties. In contrast, properties in stratum 6 show the highest median and maximum prices, with a greater concentration of higher-priced outliers. This pattern underscores the stratification of property values in Bogotá, where higher strata correspond to wealthier areas with more expensive properties.

Figure 4 examines the relationship between property area and price. The scatterplot on the left demonstrates a positive correlation between price and area, with larger properties generally commanding higher prices. However, a significant number of properties with areas below 500 m^2 cluster around a wide range of prices, highlighting the variability within this segment. The scatterplot on the right, which focuses on properties under 500 m^2 and transforms price into its logarithmic form, reveals a clearer linear relationship between log price and area. This transformation mitigates the influence of outliers and better captures the proportional increase in price relative to area, particularly for smaller properties. Both visualizations reinforce the importance of area as a primary determinant of property prices, though its effect varies depending on property size.

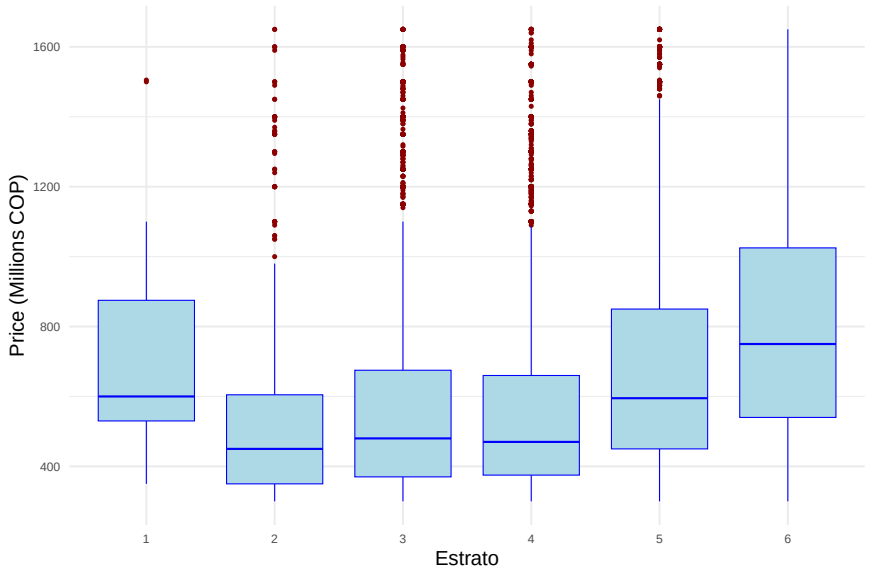

Figure 3. Distribution of Property Prices by Socioeconomic Stratum (*Estrato*)

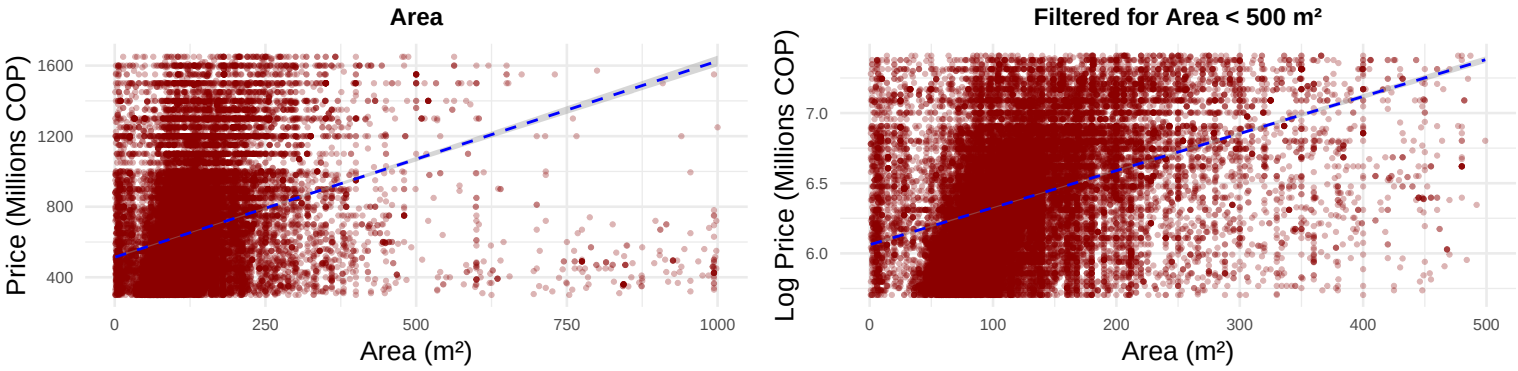

Figure 4. Relationship Between Property Area and Price

Figure 5 provides insights into the distances between properties and various urban amenities, such as banks, bus/train stations, parks, schools, and universities. The boxplots indicate substantial variability in these distances across the city, with median values generally below 1,000 meters for most amenities. However, certain amenities, like universities and hospitals, exhibit longer maximum distances, reflecting their more dispersed locations. Properties tend to be closer to amenities such as markets and police stations, suggesting a higher density of these facilities within residential neighborhoods. When considered alongside the median property price map in *Figure 2*, a pattern emerges wherein properties located closer to key amenities—particularly in central areas like Chapinero—tend to command higher prices. While the correlation between proximity to amenities and price is not universally clear across the city, the clustering of high-priced properties near densely amenitized areas in Chapinero suggests that access to amenities significantly influences property values in these premium zones. This highlights the critical role of urban accessibility in shaping the real estate market.

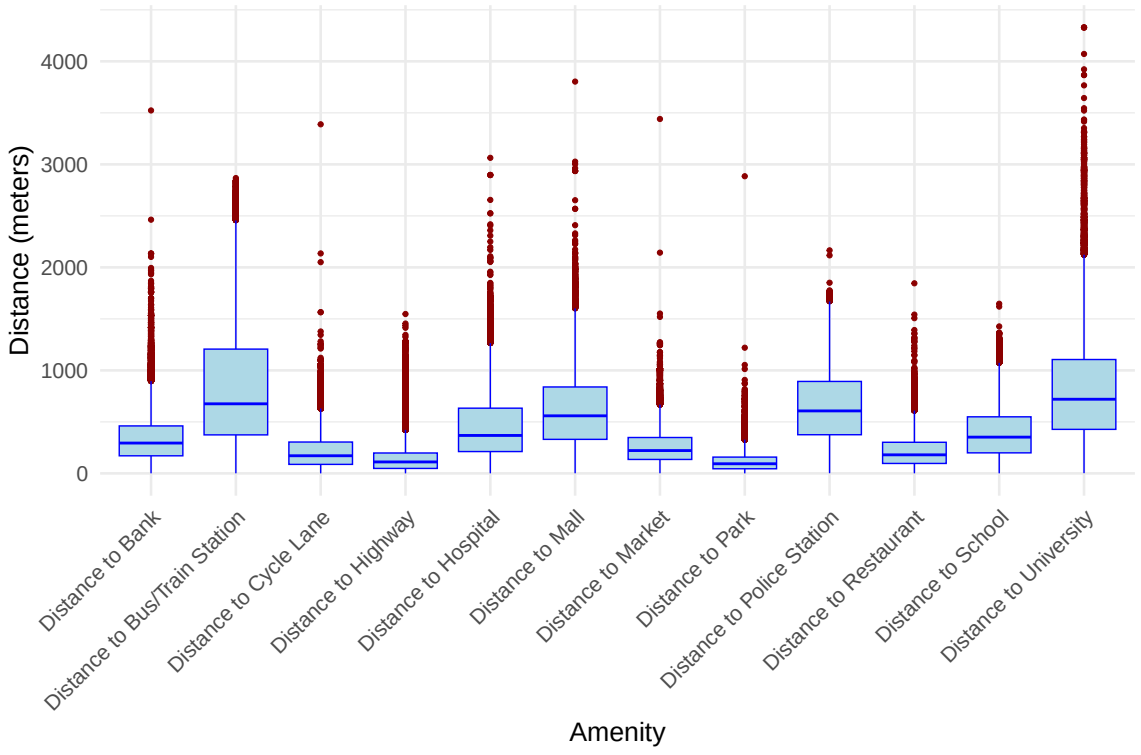

Figure 5. Distances to Urban Amenities Across Properties

The final dataset also includes 84 principal components derived from textual variables, capturing 90% of the variance in the descriptions. While these components are highly informative, their interpretability is limited, as they represent complex patterns in the textual data rather than specific, tangible property attributes. Despite this limitation, the combination of structured variables, spatial features, and principal components creates a robust framework for understanding property prices in Bogotá. This dataset aligns well with the problem's objective by integrating diverse predictors and addressing missing data effectively, enabling a comprehensive analysis of the structural, locational, and descriptive factors that influence property values.

3 Models

3.1 Predictive Strategy

This section focuses on developing predictive models for estimating property prices using the three datasets constructed in the previous section: small, medium, and large. The objective is to leverage a combination of structural, textual, and spatial variables to predict property prices effectively. As stated in the problem statement, the task involves modeling the price as a function of the property's characteristics, guided by the framework proposed by Rosen and Roback.

$$price = f(X)$$

The equation underscores the challenge of defining the functional form of the relationship between price and property attributes. However, with the extensive set of predictors available X —ranging from structural features like bedrooms and area to spatial factors such as distances to amenities—the datasets provide a robust foundation for tackling this challenge. The predictive power of these variables, as demonstrated in the previous sections, justifies their integration into multiple modeling approaches.

To address this task, a variety of predictive models will be implemented, each tailored to capture different aspects of the property price determinants. Linear Regression (LN) will serve as a baseline, providing a straightforward interpretation of the relationship between predictors and price under the assumption of linearity. Elastic Net (EN) builds on this by incorporating regularization, combining Lasso and Ridge penalties to handle high-dimensional datasets and improve model stability by selecting the most relevant variables. Tree-based methods, such as Random Forest (RF) and Gradient Boosting Machines (XGB and LGBM), will be used to capture non-linear relationships and complex interactions between predictors. RF excels in robustness and generalization by averaging multiple decision trees, while boosting algorithms like XGB and LGBM iteratively optimize prediction errors, making them highly effective for structured data. Neural Networks (NN) will be applied to exploit their strength in learning complex patterns, particularly useful for the high-dimensional and diverse nature of the dataset, including textual and spatial features. Finally, a Super Learner ensemble will combine the predictions of all these models, leveraging their individual strengths to maximize overall accuracy. This diverse modeling approach ensures flexibility and robustness in addressing the complexities of property price prediction.

The models will include normalization for numerical variables and appropriate encoding for categorical variables to ensure consistency and model performance. Moreover, spatial cross-validation will be employed to account for the geographic dependence of property prices, ensuring robust evaluation of model performance. For hyperparameter tuning, a systematic approach will be adopted, leveraging standard ranges commonly used in practice and conducting extensive experimentation to identify the optimal configurations for each model. This ensures that each model is fine-tuned to maximize its predictive accuracy, with the primary evaluation metric being the Mean Absolute Error (MAE). This metric, which measures the average magnitude of errors without penalizing large deviations disproportionately, is well-suited for the objectives of this analysis. Together, these strategies provide a comprehensive framework for addressing the problem of predicting property prices effectively.

3.2 Model Performance

After extensive testing, the medium-sized dataset was selected for modeling as it consistently outperformed the light dataset in predictive accuracy while remaining computationally feasible compared to the large dataset, which was too resource-intensive. Spatial cross-validation proved to be more effective than traditional cross-validation, as it better accounted for the geographic dependencies inherent in property prices. Ultimately, the best-performing models included Linear Regression (LN), Elastic Net (EN), LightGBM (LGBM), and the Super Learner ensemble. The hyperparameters tested for each model were chosen based on commonly used ranges and best practices in the literature, ensuring a systematic and informed tuning process. LN was used with normalized predictors to improve scale consistency, and the dependent variable was kept as price. For EN, hyperparameters included a lambda penalty ranging from 0 to 1 and an alpha regularization factor between 0 and 0.1, with 50 values tested for each parameter to fine-tune the model. For LGBM, a gradient boosting method, hyperparameters were tuned across multiple combinations, including 500 trees, tree depths of 3, 5, and 7, minimum node sizes of 10, 15, 20, and 30, and learning rates of 0.05, 0.1, and 0.2. Neural Networks were also extensively tested with architectures ranging from 1 to 2 layers, using layer sizes of 128, 64, and 28 nodes, ReLU activation functions, and dropout rates varied between 0.1 and 0.3. These configurations ensured a systematic exploration of the model space, allowing for the selection of robust and accurate predictors tailored to the complexity of the problem.

Model	**MAE Kaggle**	**Hyperparameters**
Linear Regression	200.6	Default settings
Elastic Net	203.2	Lambda: 1, Alpha: 0
LightGBM	188.8	Ntrees: 500, Min N: 30, Treedepth: 5, LR: 0.05
Neural Network	232.9	Layers: 2, Units: 128-64, Dropout: 0.2, 0.1
SuperLearner	183.5	Base Models: LN, EN, LGB, NN

Table 2. Models Performance: MAE Kaggle Score and Hyperparameters

Table 2 presents the Mean Absolute Error (MAE) scores and hyperparameters of the models evaluated on Kaggle. Linear Regression (LN), using default settings and normalized predictors, achieved an MAE of 200.6, establishing a baseline for comparison. Elastic Net (EN) performed similarly, with an MAE of 203.2, using a lambda of 1 and an alpha of 0, indicating the dominance of Ridge regularization in the model. LightGBM (LGBM) outperformed both LN and EN, achieving an MAE of 188.8 with 500 trees, a tree depth of 5, a minimum node size of 30, and a learning rate of 0.05. Neural Networks (NN), despite their flexibility, yielded a higher MAE of 232.9 when configured with two layers (128-64 nodes), ReLU activation, and dropout rates of 0.2 and 0.1. The Super Learner[1] ensemble achieved the best performance, with an MAE of 183.5, combining predictions from LN, EN, LGBM, and NN to leverage their complementary strengths.

The performance of Linear Regression (LN) slightly surpassed that of Elastic Net (EN), which is somewhat unexpected given EN's ability to regularize and handle high-dimensional datasets. This result suggests that the relationships between the predictors and property prices may exhibit a predominantly linear structure, allowing LN to capture these patterns effectively without the need for regularization. In contrast, EN's reliance on a high lambda and minimal alpha implies a heavy emphasis on Ridge regularization, which may have overly constrained the model, limiting its flexibility in fitting the data. This could explain why EN, despite its regularization capabilities, did not outperform LN in this case. The findings indicate that while regularization is useful for controlling overfitting, it might not always provide an advantage if the underlying data relationships are already well-aligned with linear assumptions.

[1]The Super Learner model was not submitted to the Kaggle competition due to time constraints but is included in this analysis for its methodological relevance and strong predictive performance.

On its part, LGBM outperformed all individual models, achieving a significantly lower MAE compared to both Linear Regression (LN) and Elastic Net (EN). This strong performance can be attributed to its ability to capture complex, non-linear interactions between variables while maintaining computational efficiency. The carefully tuned hyperparameters helped balance flexibility and regularization, enabling LGBM to generalize well without overfitting. These settings allowed LGBM to leverage the hierarchical structure in the data, particularly the spatial and categorical variables, which are challenging for purely linear models like LN and EN to model effectively. In contrast, Neural Networks (NN) performed poorly, with the highest MAE among the models tested. This result likely stems from the high dimensionality and diverse nature of the dataset, which require extensive fine-tuning and computational resources to optimize the numerous hyperparameters associated with NN architecture. While NN models can theoretically capture complex patterns better than tree-based models, the limited dataset size and the relatively shallow architecture tested (1-2 layers with 128-64 nodes) constrained their ability to fully exploit the data. Additionally, the dropout rates used to prevent overfitting (0.1-0.3) may have overly regularized the model, further limiting its capacity to learn. This underperformance highlights the challenges of effectively deploying NNs in smaller datasets or when computational resources restrict thorough hyperparameter tuning. Compared to LN, which leveraged the predominantly linear relationships in the data, and LGBM, which efficiently captured non-linear interactions, NNs appear less practical in this context despite their theoretical potential.

Finally, the Super Learner ensemble emerged as the best-performing model, achieving the lowest MAE of 183.5. By combining predictions from LN, EN, LGBM, and NN, the ensemble was able to fully exploit the diversity of variables, from the structural characteristics and spatial features to the textual principal components. The linear models, LN and EN, excelled at capturing the straightforward relationships between variables like bedrooms, bathrooms, and price, while LGBM efficiently handled the complex non-linear interactions present in features like distances to amenities and socioeconomic variables. Although NN struggled to perform on its own, its ability to recognize intricate patterns likely enhanced the ensemble by providing complementary insights from the high-dimensional textual data. This result highlights the value of the rich dataset constructed earlier, as the ensemble capitalized on the variety of predictors to address the multifaceted nature of property price prediction. The Super Learner's performance underscores that integrating diverse models tailored to different aspects of the data yields a more robust and accurate solution, demonstrating the importance of combining approaches to fully leverage the strengths of each variable type.

3.3 Variable Importance

Figure 6 displays the feature importance, measured by gain, for the LGBM model. Bathrooms stand out as the most influential predictor, contributing significantly more to the model's predictive power than any other variable. This highlights the importance of structural features in determining property prices, as buyers likely associate additional bathrooms with increased functionality and comfort. Area, the second most important variable, underscores the central role of property size in price determination, consistent with traditional real estate valuation. Parking spaces and bedrooms also rank highly, reflecting the value placed on convenience and livable space. Among the categorical and spatial variables, property type (e.g., apartment), socioeconomic strata (e.g., strata 6 and 3), and block-level features like density and cadastral valuation also show notable contributions, emphasizing the importance of location-specific characteristics in shaping property demand. Interestingly, several principal components (PCs) derived from the textual data also appear among the top predictors, such as PC4, PC2, and PC5, suggesting that the textual descriptions provided nuanced information about property attributes that complement the structured data. Spatial variables like distance to amenities (e.g., stations and cycle paths) have lower importance, which might indicate that proximity to amenities has less predictive power compared to structural or categorical features. This analysis underscores the multifaceted nature of property pricing, where a combination of structural, spatial, and textual features jointly influences model performance. The results highlight the value of incorporating diverse data sources to improve prediction accuracy.

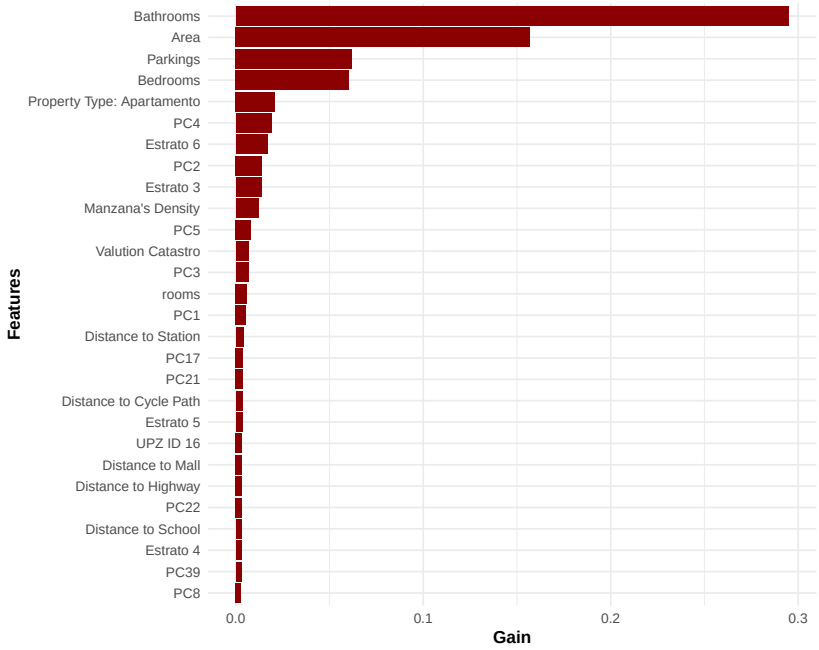

Figure 6. LightGBM Variable Importance (Top 30)

4 Conclution

The results of this study highlight the critical role that comprehensive data construction plays in developing effective predictive models for property pricing. By combining structural, textual, and spatial variables, the dataset captured a multidimensional representation of property characteristics. This integration allowed the models to leverage a wide variety of predictors, from traditional features like area and bedrooms to nuanced insights derived from textual descriptions and proximity to amenities. The hierarchical imputation strategy and the careful selection of principal components were essential for addressing data gaps and reducing dimensionality, ensuring a robust foundation for modeling. These steps underscore the importance of rigorous data preprocessing and feature engineering, particularly in contexts where data quality and completeness are challenging.

In terms of model performance, LightGBM stood out as the most effective standalone model, balancing computational efficiency and predictive accuracy. Its ability to capture non-linear interactions among variables proved crucial, especially given the complex relationships within the dataset. However, the Super Learner ensemble surpassed all individual models, demonstrating the power of integrating diverse modeling approaches. By combining linear models, tree-based methods, and neural networks, the ensemble successfully addressed the multifaceted nature of property pricing, leveraging the strengths of each model. The findings reaffirm the value of ensemble methods in predictive modeling, particularly in applications with diverse data sources and varying predictor types. Nevertheless, the performance of Neural Networks suggests room for improvement, as their computational requirements and dependence on extensive hyperparameter tuning made them less effective in this context.

While the study successfully addressed the primary goal of constructing a reliable predictive model for property prices in Chapinero, some limitations remain. The reliance on computationally intensive approaches limited the exploration of more complex neural network architectures and additional ensemble techniques. Furthermore, the relatively static nature of the dataset means that temporal changes in market conditions were not fully captured, which could affect model generalizability in dynamic markets. Future work could incorporate richer temporal data, such as macroeconomic indicators or policy changes, and explore alternative methods for integrating textual and spatial information. Despite these limitations, this study demonstrates the potential of machine learning for tackling complex real estate challenges and provides a scalable framework for enhancing decision-making in Colombia's real estate market.

References

[1] DANE. (2024). Comunicado de prensa. Recuperado de https://www.dane.gov.co/files/operaciones/PIB/cp-PIBItrim2024.pdf

[2] Jha, S. B., Babiceanu, R. F., Pandey, V., & Jha, R. K. (2020). Housing Market Prediction Problem Using Different Machine Learning Algorithms: A Case Study. *arXiv*. Recuperado de https://arxiv.org/abs/2006.10092

[3] Sharma, H., Harsora, H., & Ogunleye, B. (2024). An Optimal House Price Prediction Algorithm: XGBoost. *arXiv*. Recuperado de https://arxiv.org/abs/2402.04082

[4] Vargas-Calderón, V., & Camargo, J. E. (2019). A Model for Predicting Price Polarity of Real Estate Properties Using Information of Real Estate Market Websites. *arXiv*. Recuperado de https://arxiv.org/abs/1911.08382