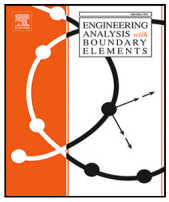
Engineering Analysis with Boundary Elements 169 (2024) 105988

Contents lists available at ScienceDirect

Engineering Analysis with Boundary Elements

journal homepage: www.elsevier.com/locate/enganabound



Machine learning applied to estate pricing for residential rentals in dynamic urban markets—The case of São Paulo city

Wesley F. Maia [a], Sergio A. David [b,a,*]

[a] Institute of Mathematics and Computer Sciences, University of São Paulo, São Carlos 13566-590, Brazil
[b] Department of Biosystems Engineering, University of São Paulo, Pirassununga 13635-900, Brazil

ARTICLE INFO

Keywords:
Rental pricing
Machine learning
Real estate analysis
Ensemble models
Multiple regression

ABSTRACT

This study conducts a comprehensive investigation into real estate rental pricing in São Paulo city, employing an innovative approach that combines advanced machine learning techniques with geospatial and natural language processing (NLP) analyses. The research analyzed a robust dataset comprising 47,243 rental listings, gathered through web scraping techniques. Following a rigorous data cleaning and preprocessing procedure, the study focused on 35,486 instances, incorporating a variety of variables that go beyond conventional metrics, including textual descriptions and geographic information, enriching the analysis and market understanding. Several regression models were implemented and compared, including linear approaches, Support Vector Machines, and ensemble methods such as Gradient Boosting, LightGBM, and XGBoost. The Blending model, which integrates multiple modeling techniques, stood out as the most accurate, achieving a Root Mean Squared Logarithmic Error (RMSLE) of 0.2923 on the test set. This result emphasizes the superiority of hybrid modeling strategies in complex pricing tasks. The findings of this study have significant practical implications. They provide landlords and tenants with a powerful data-driven tool for informed decision-making, reflecting the nuances and complexity of São Paulo's real estate market. The practical implementation of the model in an interactive web application not only demonstrates its utility in the real-world scenario but also serves as a model for future applications in real estate analysis. This work contributes to mitigating the waste of time and energy when it comes to searching for and pricing residential rentals in a large city, through the use of machine learning that shows its power and potential in accurately estimating rental prices in dynamic urban markets, allowing that more assertive and economical decisions can be taken within a social-sustainable-technological perspective.

1. Introduction

The decision to rent or buy a property presents a significant dilemma, particularly in complex urban markets like São Paulo. This choice is influenced by various factors, including personal financial conditions, life goals, desired location, and the anticipated duration of stay. Recent trends, especially among millennials, show a growing preference for renting over buying, driven by evolving social and economic factors. This study addresses the critical research gap in accurately predicting rental prices in such a dynamic market, leveraging advanced machine learning models to enhance decision-making for all stakeholders [1].

A FipeZap study comparing rental and purchase costs across fifty Brazilian cities highlights that in high-value real estate markets, renting may be more financially advantageous than buying. This finding is especially pertinent to São Paulo's diverse and challenging real estate market [2].

In this context, there arises a need for an effective methodology for the evaluation and prediction of property rental prices. Machine learning techniques emerge as powerful tools in this scenario, offering the possibility to surpass conventional methods of real estate analysis by considering a broader spectrum of variables, including location, property size, specific features, among others [3]. This study seeks to contribute to this research area by applying and comparing various machine learning models in the task of pricing rentals in the city of São Paulo.

The selection of suitable machine learning algorithms for this task is a critical aspect. While linear regression is often employed due to its simplicity and interpretability, studies such as those by Ghosalkar et al. highlight its limitations in more complex contexts [4]. Alternatively, approaches like decision trees and neural networks, suggested by Valenti et al. and Rai et al. respectively, demonstrate greater

* Correspondence to: Departamento de Engenharia de Biossistemas, Av. Duque de Caxias Norte, 225 - Jardim Elite - Campus da USP - Pirassununga, São Paulo, 13635-900, Brazil
E-mail addresses: wesley.ferreira.souza@usp.br (W.F. Maia), sergiodavid@usp.br (S.A. David).

https://doi.org/10.1016/j.enganabound.2024.105988
Received 8 June 2024; Received in revised form 29 August 2024; Accepted 1 October 2024
Available online 8 October 2024

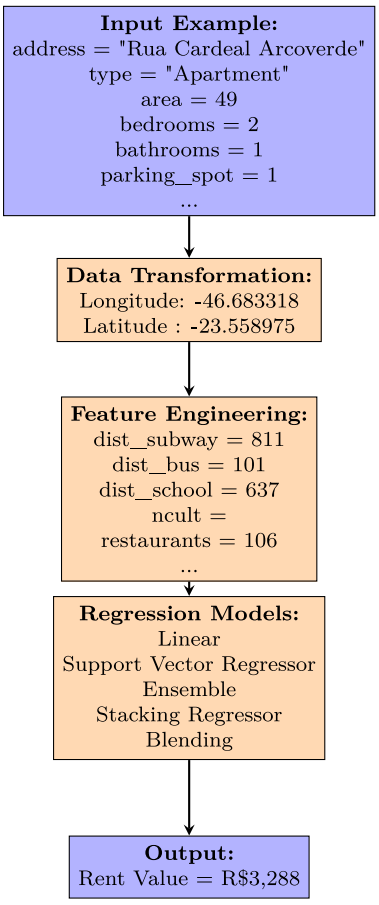


Fig. 1. Processing Flow for Rental Pricing in São Paulo: From Input to Estimated Value. The numbers in this flowchart represent specific features of a property that are used as inputs for the machine learning models to predict rental prices. For example, 'dist_subway = 811' indicates that the property is 811 meters away from the nearest subway station. These values are processed through feature engineering techniques and then fed into various regression models to generate an estimated rent value. In this example, the predicted rent value is R$3288. The accuracy of this prediction depends on the model used and the quality of the input data.

effectiveness in dealing with the complexity and high dimensionality of data typical in large real estate markets [5,6].

The challenge of predicting rental prices in São Paulo stems not from the difficulty of obtaining data but from the lack of publicly available comprehensive datasets. While there is no central repository for rental information, our study overcomes this limitation by leveraging advanced web scraping techniques to compile a large dataset of 47,243 rental listings. Additionally, we address the inherent variability in the data by enriching it with geospatial features and processing it to ensure accuracy and relevance. Although Natural Language Processing (NLP) was employed, its use was primarily focused on data pre-processing for the 'description' variable, ensuring that textual information was structured effectively for inclusion in our models. This approach allows us to utilize a data-driven methodology that is robust and well-suited to the complexities of São Paulo's real estate market.

Recent studies have expanded on the foundational work of hedonic pricing by incorporating advanced data analytics and machine learning techniques. For example, a 2022 study focused on predicting housing prices in the Chicago suburbs using machine learning demonstrated that XGBoost outperformed other models in the volatile post-pandemic market, underscoring the robustness of ensemble methods in capturing complex data patterns [7]. Similarly, another recent study focused on real estate price estimation in French cities, utilizing geocoding and machine learning to enhance prediction accuracy [8]. These advancements provide a contemporary context for this study, which applies a blending model to capture the intricacies of São Paulo's dynamic real estate market.

This study aims to fill a gap in the literature by employing a machine learning approach to analyze and predict rental prices in São Paulo. Using a substantial dataset, this research covers multiple aspects that influence rental prices, from basic property characteristics to their proximity to important urban infrastructure. Fig. 1 illustrates the adopted processing flow, from exemplary data input to the estimation of rental value, exploring a range of models ranging from linear regressions to more sophisticated ensemble techniques. The numbers shown in Fig. 1 represent key input features and their corresponding values for a specific property in São Paulo. These values are derived from the raw input data through a series of transformations and feature engineering steps. For instance, 'dist_subway = 811' is calculated based on the geographic coordinates of the property and the nearest subway station. Each of these engineered features is then fed into the regression models, which have been trained on a large dataset of rental properties to predict the rent value. The final output, in this case, R$3288, represents the estimated monthly rent for the property based on the input features. This estimation process is repeated for each property in the dataset, and the overall performance of the models is evaluated based on the accuracy of these predictions. The goal is to develop a robust and reliable predictive model that reflects the complexity of the São Paulo real estate market.[1]

This paper is organized as follows. Section 2 explores a Literature Review and presents a detailed analysis of previous works related to rental pricing and the use of machine learning in this domain. Section 3 describes the dataset used, the methodology for data collection and processing, and details the machine learning algorithms implemented, highlighting the systematic approach adopted from data input to generating rental estimates. In Section 4, we discuss the results achieved by the models, emphasizing key findings and interpretations. Section 5 explores the applicability of the model in the real world through an interactive website and compares model predictions with actual rental prices. Finally, Section 6 summarizes the study, reflects on its contributions, and point out future research directions, emphasizing the importance of continuously improving the model and exploring new approaches to better capture the complexities of the real estate market of a great urban center such as São Paulo city.

2. Literature review

The theory of hedonic demand, described by Besanko et al. [9], provides a fundamental basis for understanding real estate market pricing. This theory posits that each characteristic of a property — location, amenities, and other features — directly contributes to its overall value. This approach has been particularly relevant in urban contexts, where the diversity of properties and varied consumer preferences create a complex and multifaceted market.

Previous studies have employed various methodologies to predict real estate prices in urban areas. Yoshida et al. [10] conducted a comprehensive study on apartment rent prediction in Japan, comparing regression-based and machine learning-based approaches. They found that machine learning models like XGBoost and Random Forest significantly outperformed traditional methods like OLS (Ordinary Least Squares) and NNGP (Nearest Neighbor Gaussian Processes), especially when handling large datasets. Their study highlights the importance of spatial dependencies in predictive models, which are particularly relevant for large, heterogeneous cities.

Similarly, Ogundunmade et al. [11] applied machine learning models to predict housing rent prices in Ibadan City, Nigeria. They utilized

[1] https://github.com/WMaia9/rent-prediction

both OLS and XGBoost, revealing that certain property features, such as the number of toilets and detached house types, significantly influence rent prices. This study underscores the applicability of machine learning in different urban contexts and provides insights into the specific factors that drive rent prices in emerging markets.

Zhang et al. [12] also explored the use of machine learning for predicting house rent in India, utilizing a stacking ensemble method combining Gradient Boosting, LightGBM, and CatBoost. Their approach demonstrated improved predictive accuracy over traditional models, particularly in handling diverse datasets that include both numerical and categorical variables.

Hedonic models, which integrate both property characteristics and consumer preferences, have proven to be valuable tools in the analysis and valuation of property prices. Sartoris Neto and Fava [13] emphasize how these models can effectively establish the relationship between property size, location, and market value. These models are also crucial for understanding how external factors, such as infrastructure and accessibility, influence property prices [14,15].

Spatial heterogeneity is another important aspect in real estate valuation. Huang et al. [16] and other researchers [17,18] emphasize the need to consider the location and specific characteristics of different regions when assessing property prices. This approach is particularly relevant in large and diverse cities, where property appreciation can vary significantly from one neighborhood to another.

Property rental, as discussed by Shelton [19] and others [20,21], represents a viable and increasingly popular alternative in the real estate market, especially in urban contexts. This modality offers flexibility and aligns with mobility trends and changes in housing preferences. Promoting property rental can also be a driver for real estate market development and job generation, as highlighted in recent studies [22].

Mendonça [23] addresses the significant impact of housing credit policies in Brazil, highlighting their influence on property sales and rentals. This analysis underscores the importance of government and economic policies in shaping the real estate market. From a legal perspective, Diniz [24] emphasizes the role of lease contracts as crucial instruments in the rental market.

Sustainability has emerged as a central pillar in the real estate sector. The growing interest in sustainable practices is reflected in the demand for green, energy-efficient buildings constructed with eco-friendly materials. These attributes not only contribute to environmental protection but can also add significant value to properties, making them more attractive to a conscientious segment of the market [25,26].

Urban mobility and its influence on real estate valuation are also highly relevant topics. Proximity to efficient transportation modes such as subways and bus corridors can significantly increase the value of a property. The evolution of urban policies towards greater sustainability and the promotion of alternative transportation modes, such as bicycles and electric vehicles, are changing consumer preferences and, consequently, influencing property prices [27,28].

Urban markets like Mumbai, Jakarta, and Mexico City face significant socio-economic and infrastructural challenges that closely resemble those in São Paulo. Mumbai's real estate market, characterized by socio-economic segregation and high population density, shares similarities with São Paulo's complex urban landscape. Studies have highlighted the need for advanced models to handle such complexities, particularly in capturing the varied property characteristics and market dynamics [29,30]. Jakarta, undergoing rapid urbanization and infrastructural development, presents another parallel, where property values are heavily influenced by these factors [31]. These similarities suggest that the methodological approach developed in this study, particularly the use of blending models, could be effectively applied to these markets, providing a robust framework for real estate price prediction. Although Mexico City has not been directly studied in the same way, its urban dynamics and challenges indicate that similar methodologies could be successfully adapted to its context.

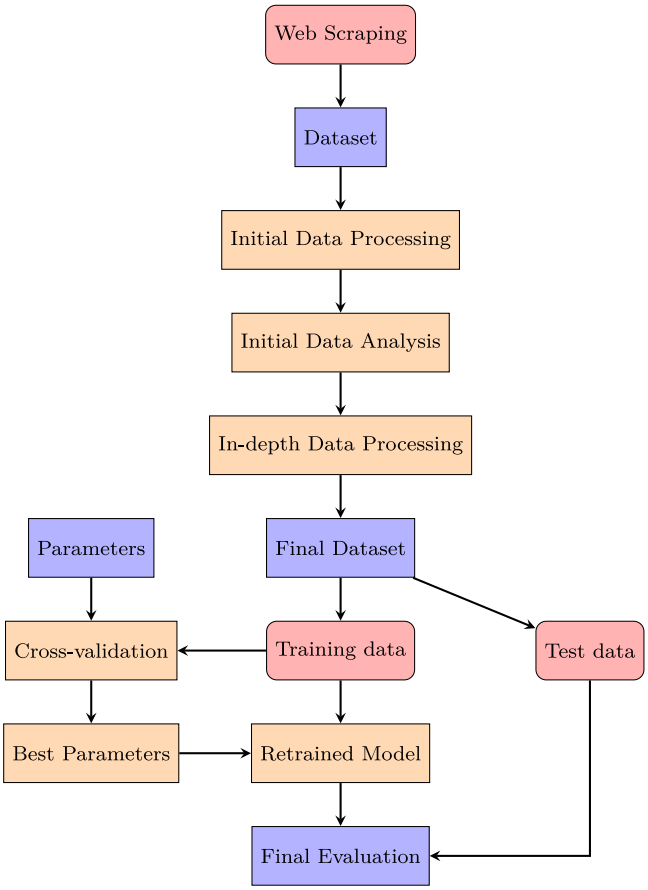


Fig. 2. Revised ETL process diagram reflecting the data flow and model training approach for real estate rental price prediction. This includes the selection of parameters, cross-validation for model tuning, and the identification of best parameters for retraining the model.

Finally, predictive analysis represents an important frontier in the real estate market field. The ability to use large volumes of data to anticipate market trends, identify regions with appreciation potential, and predict changes in real estate demand is crucial for investors and developers. These advanced techniques provide a deeper understanding of the market and enable more informed investment decisions [32–34].

These developments in the real estate market reflect a combination of technological advancements, environmental awareness, and changes in urban mobility policies, all converging to shape an ever-evolving sector. Understanding these factors is essential for accurately analyzing and pricing properties in the current landscape.

3. Data and methodology

In this study, we address the complex task of pricing rental properties in the city of São Paulo using a structured and detailed method, as illustrated in Fig. 2. The process begins with data collection through web scraping, followed by various stages of processing and analysis to prepare the data for modeling. The core of this work involves the application of machine learning models, carefully trained and evaluated using techniques such as cross-validation and hyperparameter tuning, aiming to develop an effective system for rental pricing. The presented diagram provides an overview of the adopted workflow, highlighting key steps from the initial data collection to the final evaluation of the models.

Table 1
Description of dataset variables.

Variable	Description
description	A brief textual description of the property, including key features and selling points.
price	Monthly rental price in Brazilian Reais (BRL). This is the target variable for the predictive model.
bedrooms	The total number of bedrooms in the property, which can influence the rental price significantly.
bathrooms	The number of bathrooms in the property, contributing to the property's overall amenities.
total_area	The total area of the property measured in square meters, indicating the size of the living space.
parking_spot	The number of parking spots available with the property, which is an important factor in urban areas.
address	The full address of the property, used for geocoding and spatial analysis.
link	The URL link to the property listing on the real estate portal, allowing for reference and verification of the data.

3.1. Construction and analysis of the dataset

The construction of the dataset for this study began with a web scraping [35] step on a renowned Brazilian real estate portal. This method enabled the retrieval of 47,243 records of properties available for rent in São Paulo in November 2023. After a meticulous cleaning and validation process, the dataset was refined to 35,486 valid instances.

During the collection, several essential variables were captured for each property, which are crucial for pricing analysis. These variables include information such as the price, number of bedrooms, bathrooms, total area, and address. A crucial processing step involved converting the geographical location of the properties, originally in string format, into precise latitude and longitude coordinates. This was achieved using the Geopy library [36], which allowed for the transformation of addresses into accurate geolocated data (see Table 1).

To ensure that the properties were effectively located within the city limits of São Paulo, official map data of the city provided by IBGE[2] was used. This ensured that the dataset accurately reflected the real estate market of the city.

Fig. 3 shows the geographical distribution of residences, where each point represents a property whose data was collected and validated. This geospatial analysis is crucial for understanding the distribution and concentration of rental listings in different regions of the city, vital information for accurate property pricing. Thus, this dataset serves as a solid foundation for the upcoming stages of modeling and predictive analysis.

3.2. Enrichment of the dataset with geospatial information

The data enrichment step was essential to enhance the rental pricing analysis. Obtaining the latitude and longitude of each property allowed us to integrate additional geospatial information, crucial for understanding the impact of location on property pricing.

With the geographic coordinates in hand, we developed a set of new variables, totaling 7 additional features. These variables reflect the proximity of each property to essential urban infrastructure such as schools, subway stations, and bus stops, obtained through the GeoSampa portal [37]. Furthermore, information about cultural establishments and food venues, collected from Overpass Turbo [38], was aggregated by calculating the number of these points within a 1 km radius of each property.

Potential biases in the dataset include the overrepresentation of certain neighborhoods due to higher online listing activity, as well as inaccuracies in property descriptions provided by landlords or real estate agents. Despite these challenges, the dataset provides a comprehensive overview of São Paulo's rental market during the specified timeframe, making it a solid foundation for subsequent modeling and predictive analysis (see Table 2).

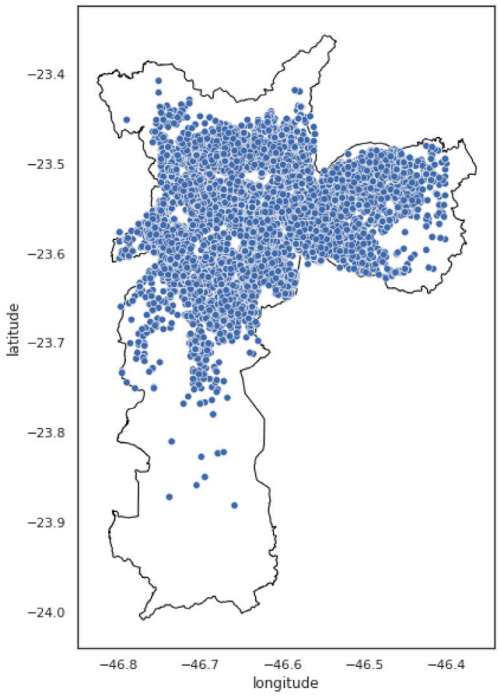


Fig. 3. Geographical distribution of available rental residences in the city of São Paulo. Each point represents a property included in the dataset, and the spatial distribution illustrates how rental listings are concentrated in various regions of the city. This distribution is crucial for understanding the geographical coverage of the dataset and ensuring that the model accounts for regional variations in rental prices.

Table 2
Geospatial variables added to the dataset.

Variable	Description
dist_subway	Distance (in meters) to the nearest subway station. This variable is crucial for understanding accessibility to public transportation, which can significantly influence rental prices.
dist_bus	Distance (in meters) to the nearest bus stop. Similar to dist_subway, this variable helps measure the property's accessibility to public transport.
school	Distance (in meters) to the nearest school. The proximity to educational institutions often impacts the attractiveness of a property, especially for families.
ncult	Number of cultural venues within a 1km radius. This variable includes cultural spaces such as libraries, museums, parks, concert venues, and other cultural attractions, reflecting the cultural vibrancy of the neighborhood, which can be a significant factor in rental pricing, especially in urban areas.
food	Number of food establishments within a 1km radius. This reflects the availability of dining options nearby, which is often a key consideration for renters.
latitude	Latitude of the property. This coordinate is used in conjunction with longitude to precisely locate the property on a map.
longitude	Longitude of the property. Together with latitude, this provides the exact geographical location of the property.

To calculate the distances between properties and points of interest, the haversine equation was employed [39]:

$$d = R \cdot \arccos\left(s_a s_b + c_a c_b \cos(\Delta\lambda)\right) \tag{1}$$

where:

[2] https://cidades.ibge.gov.br/brasil/sp/sao-paulo/panorama

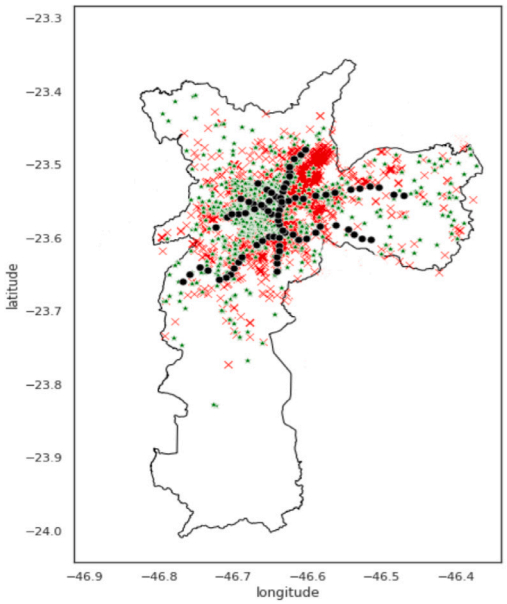


Fig. 4. Spatial distribution of schools (green), subway stations (black), and food establishments (red) in São Paulo. This figure demonstrates the proximity of rental properties to key urban amenities, which are important features influencing rental prices. The spatial relationship between these external variables and the properties helps in analyzing how accessibility and convenience impact rental values.

- $s_a = \sin(\text{lat}_a)$ and $s_b = \sin(\text{lat}_b)$ represent the sine of the latitudes of the property and the point of interest, respectively,
- $c_a = \cos(\text{lat}_a)$ and $c_b = \cos(\text{lat}_b)$ represent the cosine of the latitudes of the property and the point of interest, respectively,
- $\Delta\lambda = \text{lon}_b - \text{lon}_a$ is the difference in longitudes between the point of interest and the property,
- R is the average radius of the Earth, approximately 6371 km (see Fig. 4).

The inclusion of these geospatial variables significantly enriches the dataset, providing a robust foundation for predictive modeling and offering valuable insights into the relationship between property locations and their rental prices.

3.3. Natural language processing for textual data

Handling textual data in the context of machine learning involved Natural Language Processing (NLP) [40] techniques. This step was crucial to convert the initial variable 'description' (Property Description) - which contains unstructured data - into a structured numerical format suitable for analysis and predictive modeling.

3.3.1. Text preprocessing and transformation

Initially, tokenization was applied, represented by the following mathematical function:

$$T(W) = \{w_1, w_2, \ldots, w_n\} \tag{2}$$

where $T(W)$ is the resulting set of tokens, and W is the original sentence. Each w_i is an individual token extracted from the sentence.

Next, stop words and prepositions were removed, which are elements without relevant meaning for the model. Mathematically, this can be represented as the subtraction of the set of stop words (S) from the set of tokens:

$$R(T) = T(W) \setminus S \tag{3}$$

where $R(T)$ represents the remaining tokens after stop word removal.

Table 3
Description of textual variables in the dataset.

Variable	Description
suite	Presence of a suite in the property
furnished	Property is furnished
barbecue	Presence of a barbecue area
hall	Presence of a hall
balcony	Presence of a balcony
duplex	Property is duplex
townhouse	Presence of a townhouse
air conditioning	Presence of air conditioning
pool	Presence of a pool
gym	Presence of a gym
office	Presence of an office
elevator	Presence of an elevator

Lemmatization was the next step, where each token was reduced to its base form. This operation is represented as:

$$L(w_i) = \text{root}(w_i) \tag{4}$$

for each token w_i, where $\text{root}(w_i)$ is the root or lemmatized form of the token.

3.3.2. Vector representation: Bag of words

The Bag of Words (BoW) technique was used to transform lemmatized texts into numerical vectors. The BoW representation of a document d can be mathematically expressed as:

$$BoW(d) = [f(w_1, d), f(w_2, d), \ldots, f(w_m, d)] \tag{5}$$

where $f(w_i, d)$ is the frequency of token w_i in document d, and m is the total number of unique tokens in the corpus.

3.3.3. Feature and amenity extraction

From the preprocessed text, relevant property features and amenities were identified. This identification was performed by analyzing the presence or absence of specific keywords, resulting in binary variables for each feature, as presented in Table 3.

This approach allows for a detailed analysis and quantification of textual information, significantly enriching predictive modeling and providing valuable insights into property pricing. However, it is essential to consider the possibility of inaccuracies in the descriptions provided by landlords, which can impact the final data quality.

3.4. Machine learning models

In this study, we apply a range of well-established machine learning models to predict rental prices in São Paulo. We focus on comparing their performance in the specific context of this dataset, rather than introducing each model in detail. Readers are referred to standard references for comprehensive descriptions of these models [41,42,42–45,45].

The comparative analysis conducted in this study is essential for understanding the effectiveness of different models in capturing the complexity of rental pricing in São Paulo. The uniqueness of this study lies in its rigorous evaluation of these models in a real-world dataset, emphasizing the strengths and weaknesses of each approach.

The Linear Regression model, while straightforward and interpretable, struggled with the non-linear relationships present in the dataset. This led to higher prediction errors, indicating that the model could not adequately capture the complex interactions between features. Ridge and Lasso models offered slight improvements due to their regularization techniques, which reduced overfitting; however, these models still fell short in addressing the dataset's non-linearity.

Support Vector Machines (SVM) demonstrated better performance, particularly in handling the high-dimensional and complex nature of the data. SVM's ability to manage non-linear relationships through

kernel functions resulted in a lower RMSLE compared to linear models. This highlights SVM's robustness in scenarios where data complexity is a significant factor.

Among the ensemble methods, XGBoost showed the most promising results. Its iterative approach to improving weak learners, combined with its built-in regularization, allowed it to efficiently handle the diverse and heterogeneous data found in São Paulo's rental market. XGBoost's superior performance underscores its ability to capture intricate patterns and interactions within the dataset, making it particularly well-suited for this task.

Blending models, which combine the strengths of multiple models, achieved the best overall performance. The lowest RMSLE on the test set was observed with the Blending approach, suggesting that this technique effectively balances the interpretability of simpler models with the accuracy of more complex ones. However, the blending model also required careful tuning of weights to avoid overfitting, demonstrating a potential drawback in terms of complexity and the need for meticulous calibration.

The results of this comparative analysis provide valuable insights into which models are most effective for predicting rental prices in São Paulo. While no entirely new machine learning methods were introduced, the significance of this work lies in its systematic comparison of existing techniques in a challenging, real-world context. These findings can inform future studies and practical applications in similar urban markets, where data complexity and feature interactions pose significant challenges.

3.5. Model evaluation metric

The primary metric used to evaluate the performance of the models in this study is the Root Mean Squared Logarithmic Error (RMSLE). RMSLE is particularly effective in assessing the accuracy of predictions, especially when the goal is to penalize large errors more heavily. It is defined as:

$$\text{RMSLE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log(y_i + 1) - \log(\hat{y}_i + 1))^2} \tag{6}$$

where y_i are the actual values, $\hat{y}_i$ are the predicted values, and N is the total number of observations.

This metric is advantageous because it places greater emphasis on large errors due to the logarithmic transformation and the squared term. As a result, RMSLE is a robust measure for evaluating models where prediction accuracy, particularly for larger values, is critical. A lower RMSLE value indicates better model performance, signifying higher prediction accuracy.

While RMSLE is the main metric used, other indicators such as standard deviation and residuals were also considered to provide additional insights into model performance. Standard deviation quantifies the dispersion of the predictions around the mean, offering a sense of how spread out the predictions are. Residuals, calculated as the difference between observed and predicted values, help identify patterns in prediction errors, further guiding model refinement. However, these additional metrics are primarily supportive, with RMSLE being the key measure of model effectiveness.

3.6. Data splitting and model optimization

The methodology adopted for data splitting and optimization is essential to ensure the effectiveness and generalization capability of machine learning models (see Fig. 5).

3.6.1. Data splitting into training and testing sets

Splitting the data into training and testing sets allows for the validation of model performance on data not used during training. The split was done as follows:

- **Training Set (70% of the data):** Used for model training.
- **Testing Set (30% of the data):** Used to evaluate the model's performance post-training.

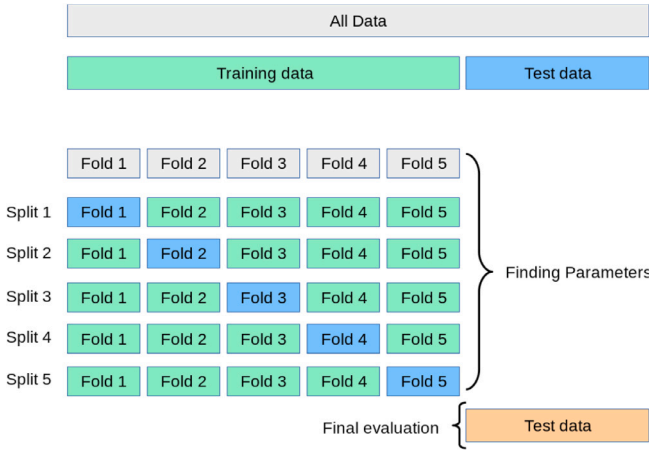


Fig. 5. K-fold cross-validation scheme.

3.6.2. Grid-search for hyperparameter optimization

Hyperparameters are settings external to the model and not learned from the data. Grid-Search [46] is a technique used to explore the best combination of hyperparameters, leading to a model with more accurate predictions. The objective function for Grid-Search can be defined as:

$$\min_{\theta} \text{Validation}_{\text{Error}}(\text{Model}, \theta) \tag{7}$$

where θ represents the set of hyperparameters, and $\text{Validation}_{\text{Error}}$ is the validation error for the model given hyperparameters θ.

3.6.3. K-fold cross-validation

K-fold cross-validation [47] is employed to ensure model robustness and prevent overfitting, especially relevant in the Grid-Search phase. This technique divides the training data into k subsets, following an iterative process for training and validation:

$$\text{CV}_{\text{Error}} = \frac{1}{k}\sum_{i=1}^{k}\text{Error}(\text{Model}, \text{Data}_{\text{Train}_i}, \text{Data}_{\text{Validation}_i}) \tag{8}$$

In this equation, Error is the error metric calculated for each cross-validation iteration, where $\text{Data}_{\text{Train}_i}$ and $\text{Data}_{\text{Validation}_i}$ are the subsets of data used for training and validation in the i-th iteration, respectively.

These data splitting and optimization techniques are fundamental for the development of reliable and effective machine learning models. More details on implementation and hyperparameter optimization results are presented in Appendix.

3.7. Development of web application

The practical implementation of the rental price prediction model was carried out through the development of an interactive web application using Streamlit. This open-source framework is remarkably suitable for creating web applications focused on data analysis in Python, thanks to its simplicity and efficiency.

Streamlit [48] stands out for its ease of use, with a simple and intuitive syntax that allows for the rapid creation of web applications. Streamlit's ease of integration with various Python libraries and frameworks, such as Pandas, Numpy, and Matplotlib, is a significant advantage, providing an efficient way to turn data analytics and machine learning models into interactive and informative applications.

The developed application offers a user-friendly interface where users can easily input data about properties and obtain rental price estimates based on the trained model. This tool demonstrates the practical applicability of the rental price prediction model, serving as

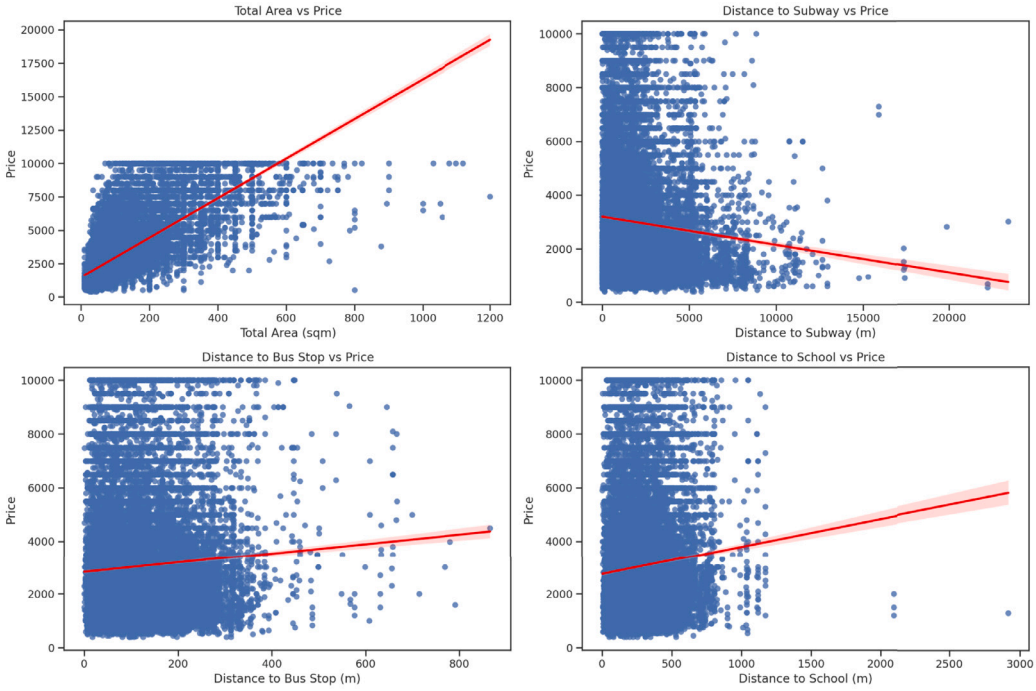


Fig. 6. Correlation analysis between total area, distance to infrastructure, and rental price. Each scatter plot includes a trend line to visualize the relationship. The top-left plot shows a strong positive correlation between property size and rental price, while the other plots illustrate weaker or inconsistent correlations between rental prices and proximity to urban infrastructure such as subway stations, bus stops, and schools. These visualizations provide important insights into how different features influence rental pricing in São Paulo.

a useful resource for landlords, tenants, and real estate professionals, facilitating informed decision-making in the real estate context of São Paulo.

4. Results and discussion

This section presents a comprehensive discussion of the results achieved in this study, starting with exploratory data analysis and feature engineering, followed by a detailed evaluation of machine learning models. The RMSLE metric is emphasized as a key indicator of model performance, both in the five-fold cross-validation (CV = 5) training process and in the testing phase. Initial analyses offer crucial insights into data characteristics and preparation required for effective modeling, while the subsequent model evaluations provide insights into the effectiveness of different algorithmic approaches in predicting rental prices in São Paulo.

4.1. Exploratory data analysis

Exploratory data analysis is a crucial step in the real estate pricing study, offering a deep understanding of the characteristics, patterns, and relationships present in the dataset. This section presents a detailed analysis of correlations between various variables and the rental price of properties. We explore both the distribution and skewness of the target variable, essential to guide effective modeling and predictive analysis.

In Fig. 6, we provide visual evidence of the correlations discussed in this section. The scatter plots, accompanied by trend lines, depict how different factors, such as the total area of a property and its distance to key urban infrastructure (subway stations, bus stops, and schools), relate to rental prices. The trend line in the upper-left plot shows a positive correlation between property area and price, confirming that larger properties generally command higher rents. Conversely, the other plots reveal weaker or inconsistent correlations between rental prices and proximity to infrastructure, suggesting that while location is important, other factors like property quality and market context may play a more significant role in determining rental values.

We evaluated how different factors, such as the total area of the property and the distance to key infrastructure (subway, bus stops, and schools), influence rental prices. Scatter plots with trend lines indicate a positive correlation between the total area of the property and the price, suggesting that larger properties tend to have higher rental prices. This trend is in line with general expectations in the real estate market.

In contrast, variables related to the proximity of urban infrastructure, such as subway stations, bus stops, and schools, did not show a strong or consistent correlation with rental prices. This observation implies that, while location and access to public services are important, they may not be the primary determinants of rental prices. Factors such as property quality, interior features, and the local market context are also relevant.

The graphs illustrating these correlations are shown in Fig. 6, highlighting the discussed relationships.

Complementing the scatter analysis, boxplots provided additional insights, especially regarding the distribution of rental prices in relation to the number of bathrooms, bedrooms, and parking spaces. We observed that there is a trend of increasing rental prices with an increase in the number of bathrooms and bedrooms, indicating a valuation of properties with more amenities. Additionally, the number of parking spaces also showed a positive relationship with the price, reinforcing the importance of these attributes in property pricing.

These findings are illustrated in the boxplot graphs in Fig. 7, which complement our understanding of the variables influencing property rental prices.

By combining both analytical approaches, we obtain a holistic and detailed view of the factors impacting property rental prices, essential for developing more accurate and effective predictive models.

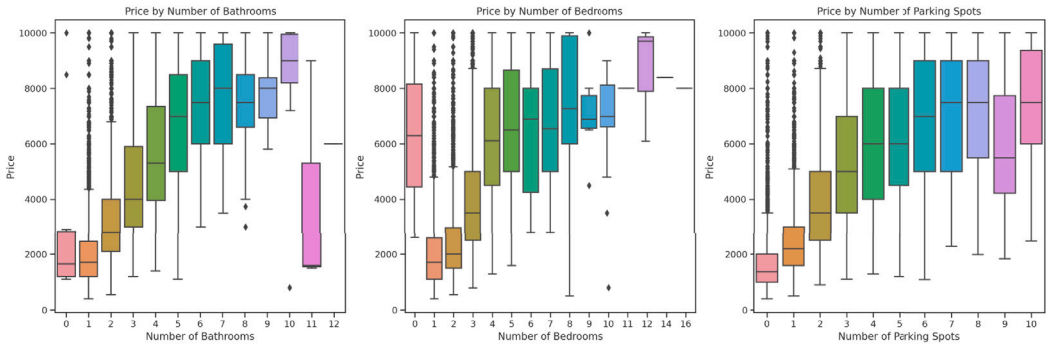


Fig. 7. Distribution of rental prices in relation to the number of bathrooms, bedrooms, and parking spaces.

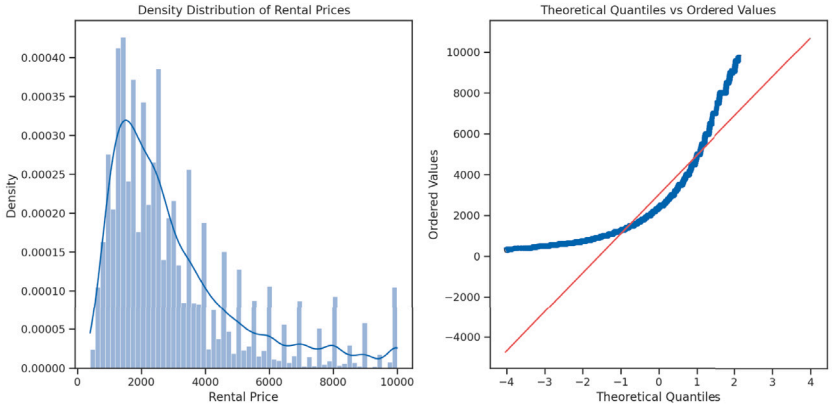


Fig. 8. Density distribution and Q-Q plot of rental prices.

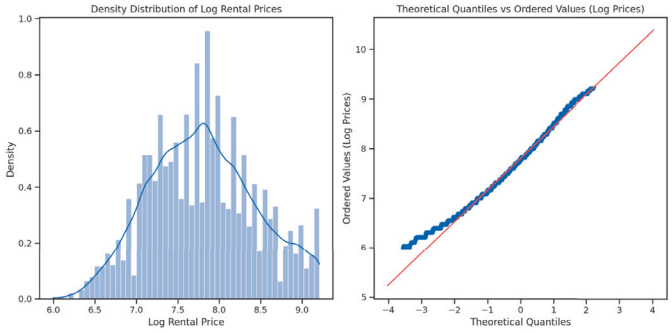


Fig. 9. Density distribution and Q-Q plot of rental prices after the logarithmic transformation.

4.2. Data normalization

Data normalization is a crucial step in preparing data for machine learning models, especially in contexts where the target variable exhibits skewness. In our study, we applied a logarithmic transformation to rental prices to normalize the distribution. This section discusses the statistical implications of this transformation.

Before the transformation, rental prices had a distribution with skewness of 1.48 and kurtosis of 1.82, indicating a distribution with a long right tail and a sharper peak than a normal distribution, as seen in Fig. 8.

The logarithmic transformation was applied as follows:

$$Y_{\log} = \log(Y) \tag{9}$$

where Y represents the original rental prices and $Y_{\log}$ are the transformed prices.

Table 4
Statistical measures of rental prices before and after the logarithmic transformation.

Statistical measure	Original	Transformation
Mean (μ)	3023.47	7.80
Standard Deviation (σ)	2093.32	0.64
Mode	2500	7.82
Skewness	1.48	0.15
Kurtosis	1.82	−0.51

Statistical measures before and after the transformation are presented in Table 4. We observed a significant reduction in skewness and kurtosis, bringing the data closer to a normal distribution, as evidenced in the density and Q-Q plot graphs (Fig. 9).

Normalization through logarithmic transformation is essential to ensure that machine learning techniques operate under the assumption of data normality, which can significantly improve predictive model performance.

4.3. Analysis and practical implications of results

The analysis of results obtained by machine learning models in the task of pricing rental properties in São Paulo reveals crucial insights into the nuances and complexities of the real estate market. In this section, we discuss in detail the performance of each model and its implications in the study's context.

The Linear Regression and Ridge models exhibited similar performances, with an RMSLE of 0.3829 ±0.0051 in training and 0.3798 in testing. This consistency between the training and testing sets indicates that both models were able to generalize well from the training data to unseen data, capturing the fundamental relationships in the dataset without overfitting or underfitting. The small difference between the

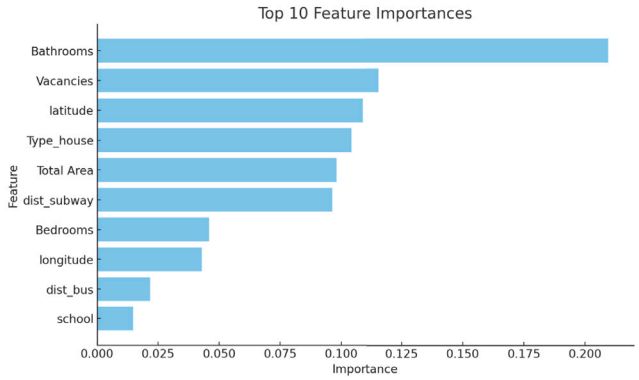


Fig. 10. Enter caption.

RMSLE values suggests that these models are reliable as baseline models, providing a robust starting point for comparison with more complex models.

In this research, stacking and blending techniques were employed to combine the strengths of multiple machine learning models, aiming to improve predictive accuracy. The selection of models for stacking and blending was based on a rigorous evaluation of their individual performances, as well as their complementary strengths. The models chosen for hybridization included linear models (e.g., Linear Regression, Ridge, and Lasso), Support Vector Machines, and ensemble methods (e.g., Gradient Boosting, LightGBM, XGBoost). Each of these models contributes differently to the prediction task, with linear models providing simplicity and interpretability, SVM handling non-linear relationships effectively, and ensemble methods excelling at capturing complex interactions within the data.

The Bayesian Ridge model, although conceptually similar to linear models, introduces an additional layer of Bayesian inference in regularization. However, this model showed a slightly higher RMSLE, both in training (0.4134 ± 0.0031) and testing (0.4145). This increase in RMSLE indicates that the additional complexity introduced by Bayesian regularization did not lead to better performance in this case. The model's higher RMSLE values suggest that, while it may offer theoretical advantages in terms of probabilistic interpretation, these did not translate into improved predictive accuracy for the São Paulo rental market dataset.

The Lasso and Elastic Net models, known for their stricter regularization approaches, showed slightly better results than Linear Regression and Ridge. For example, Lasso achieved an RMSLE of 0.3831 ± 0.0050 in training and 0.3801 in testing. These results suggest that the variable selection performed by Lasso and Elastic Net helps reduce overfitting, resulting in slightly better generalization on unseen data. The improvement in RMSLE values, though modest, highlights the importance of regularization in handling the potential collinearity of features and ensuring more stable predictions (see Fig. 10).

Support Vector Machine (SVM)-based models and Ensemble methods, including Gradient Boosting, LightGBM, and XGBoost, demonstrated substantial performance improvements. The SVM model, in particular, achieved an RMSLE of 0.3461 ±0.0064 in training and 0.3427 in testing, indicating its robustness in capturing complex, non-linear relationships in the data. The close alignment between training and testing RMSLE values reflects the model's ability to generalize well, making it a strong candidate for scenarios where data complexity is a significant challenge.

Ensemble models, such as Gradient Boosting, LightGBM, and XGBoost, further improved on these results, with XGBoost achieving an RMSLE of 0.3164 ±0.0040 in training and 0.3124 in testing. The low RMSLE values for both training and testing indicate that these models excel at capturing intricate patterns in the data without overfitting. Their performance underscores the effectiveness of ensemble techniques in aggregating the strengths of multiple weak learners to produce a powerful predictive model.

Table 5
Model results with RMSLE metrics for training and testing.

Model	Training (CV = 5)	Test
Linear	0.3829 ± 0.0051	0.3798
Bayesian	0.4134 ± 0.0031	0.4145
Ridge	0.3829 ± 0.0050	0.3798
Lasso	0.3831 ± 0.0050	0.3801
Elastic Net	0.4064 ± 0.0031	0.4035
SVM	0.3461 ± 0.0064	0.3427
Gradient Boosting	0.3148 ± 0.0050	0.3128
LightGBM	0.3190 ± 0.0026	0.3180
XGBoost	0.3164 ± 0.0040	0.3124
Stacking	0.3112 ± 0.0048	0.3198
Blending	**0.2973**	**0.2923**

In turn, stacking and blending models offer interesting approaches to integrating multiple machine learning models. The stacked model showed good performance on the training set but experienced a drop in effectiveness on the test set, suggesting a tendency to overfit. This is reflected in the RMSLE values, where the Stacking model achieved 0.3112 ±0.0048 in training but increased to 0.3198 in testing, indicating some level of overfitting. In contrast, the Blending model exhibited the best overall performance, with an RMSLE of 0.2973 in training and 0.2923 in testing. The Blending model's low RMSLE values across both training and testing sets suggest that it effectively balances the strengths of various models, leading to superior generalization and predictive accuracy. This performance demonstrates the value of combining diverse models to capture different aspects of the data, ultimately leading to more accurate predictions.

Further analysis was conducted to evaluate the impact of excluding geospatial variables from the feature set. The results are summarized in Table 6, which compares the RLMSE metrics for models trained on a reduced set of features, specifically focusing on property attributes without geospatial data. Notably, the models trained on these selected features exhibited higher RMSLE values compared to those incorporating geospatial variables, as shown in Table 5.

This comparison, illustrated in Fig. 11, reveals that the inclusion of geospatial features significantly enhances model performance, underscoring the importance of location-based variables in accurately predicting rental prices. The results demonstrate that while models trained solely on property attributes can provide reasonable estimates, the integration of geospatial data leads to more precise and reliable predictions, offering a distinct advantage in understanding market dynamics.

The results of these models have significant implications for the São Paulo real estate market. Highly effective models, such as Ensemble and Blending, demonstrate a remarkable ability to accurately price rental properties. The low RMSLE values achieved by these models on both training and testing sets highlight their potential as reliable tools for stakeholders in the real estate market. These advanced tools offer valuable support for landlords, tenants, and real estate professionals, enabling data-driven decisions, rental pricing optimization, and a deeper understanding of market dynamics.

The comparative analysis of the models trained on different feature sets clearly demonstrates that the inclusion of geospatial variables provides a significant performance boost, as evidenced by the lower RMSLE values in models incorporating these variables. This finding is consistent across various machine learning models, indicating that geospatial features are crucial for accurately capturing the factors influencing rental prices in São Paulo. By comparing the results in Tables 5 and 6, it is evident that models trained on comprehensive feature sets, including geospatial data, outperform those trained solely on property-specific attributes.

Table 6
Model results with RMSLE metrics for selected features before geospatial processing.

Model	Training (CV = 5)	Test
Linear	0.4539 ± 0.0037	0.4502
Bayesian	0.4539 ± 0.0037	0.4502
Lasso	0.4539 ± 0.0037	0.4502
Elastic Net	0.4539 ± 0.0037	0.4502
Ridge	0.4539 ± 0.0037	0.4502
SVM	0.4191 ± 0.0025	0.4160
Gradient Boosting	0.3814 ± 0.0019	0.3798
LightGBM	0.3841 ± 0.0029	0.3836
XGBoost	0.3980 ± 0.0028	0.3982
Stacking	0.3822 ± 0.0024	0.3857
Blending	**0.3742**	**0.3733**

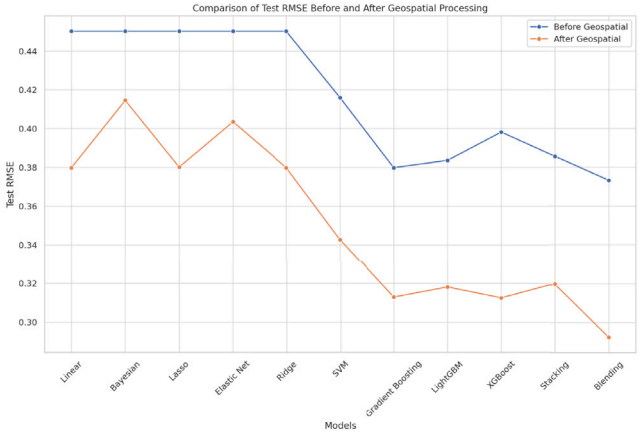


Fig. 11. Comparison of Test RMSLE Before and After Geospatial Processing Across Different Models. The plot illustrates the impact of geospatial features on model performance, with each line representing the RMSLE values for a given model before and after incorporating geospatial data.

This reinforces the importance of a holistic approach to feature selection in real estate price modeling, where the integration of location-based variables can significantly enhance predictive accuracy. These insights not only align with existing literature on the importance of location in real estate but also offer empirical evidence supporting the advanced variable selection strategy employed in this study. As such, the results presented provide a robust foundation for future research and practical applications in the field, highlighting the necessity of including geospatial data in predictive models for rental pricing.

5. Qualitative analysis of results

This section addresses the practical application of the machine learning model in a real-world context, using an interactive website to demonstrate the accuracy of the chosen model, i.e., the Blending model. Additionally, we discuss the comparison between actual rental prices and the model's estimates.

The Blending model, standing out for its superior performance in terms of RMSLE, was implemented in an interactive website,[3] offering users the ability to input property details and receive real-time rental price estimates. This integration demonstrates the practical applicability of the model in real market conditions, providing a valuable tool for both landlords and tenants.

The model's effectiveness in practice is evaluated through the comparison between actual rental prices, obtained from real estate websites, and the model's predictions. Table 7 illustrates this comparison (Real vs. Model) with selected examples, reflecting the model's accuracy in different real estate contexts:

[3] https://wmaia9-sprentweb-sprent-bvcrqd.streamlit.app/

Table 7
Comparison of real and model-predicted prices.

Property	Real (R$)	Model (R$)
Av. Campo Belo	2200	2341
R. Mandaqui	3500	3481
R. Correia Galvão	2300	2538
Av. Cursino	1100	2116
R. Bela Cintra	4500	3500
R. Doutor Jesuí	3800	3564
R. Nhu-Guaçu	2700	1200
R. Visconde de Inhom	2900	2905
Av. Carval de Frei	3490	3200
Av. Dolores Duran	4200	4731

In examples like Campo Belo and Mandaqui, the model demonstrated high accuracy, with predictions very close to actual market values. This reflects the model's ability to accurately capture the nuances and characteristics of the São Paulo real estate market in different types of properties, whether apartments or houses.

However, in other cases like R. Nhu-Guaçu, we observed a significant discrepancy between the actual and predicted price. Such differences can be attributed to various reasons. On one hand, there may be overpricing or underpricing by landlords, reflecting market expectations not aligned with general trends. On the other hand, such discrepancies may also indicate areas where the model requires additional adjustments, especially in cases where specific external factors or unique property features are not fully captured by the model.

This qualitative analysis of results reveals both the effectiveness and limitations of the Blending model, providing valuable insights for future improvements and adjustments. The ability to accurately predict rental prices in a variety of scenarios reinforces the model's potential as a useful tool for decision-makers in the real estate market.

6. Conclusion and future work

This study presented a detailed analysis of rental property prices in the city of São Paulo, using advanced machine learning techniques. The research focused on building and evaluating various regression models, with the Blending model standing out for its remarkable performance in terms of RMSLE. Initial exploratory analysis and subsequent normalization of the price variable were crucial to understanding real estate market dynamics and preparing the data for effective modeling.

The results obtained by the Blending model, demonstrated through an interactive website, showcased the model's effectiveness and practical applicability. This robust tool provides valuable insights for landlords, tenants, and real estate professionals, enabling more informed and data-driven decision-making.

Despite promising results, the study acknowledges the need for ongoing improvements and the exploration of new approaches. A promising avenue for future work is the incorporation of visual features of properties using Convolutional Neural Networks (CNNs) to analyze images. CNNs could be employed to capture details such as property condition, interior design, and lighting quality, which are difficult to quantify through structured data alone. By integrating these visual features with existing structured data, future models could offer a more holistic view of the factors influencing rental prices, enhancing predictive accuracy.

Another interesting direction involves integrating social media and news portal data using natural language processing techniques to capture market trends and consumer preferences. Social media platforms and news sources provide real-time data that reflect public sentiment, emerging trends, and economic factors influencing the real estate market. By incorporating this unstructured data, predictive models could be enhanced to detect and respond to market shifts more dynamically, offering a more responsive tool for stakeholders.

Moreover, future research could explore the use of transfer learning techniques with CNNs, where pre-trained models on similar tasks (e.g., real estate image classification) are fine-tuned for the specific context of São Paulo's rental market. This approach could reduce the need for large labeled datasets while still achieving high levels of accuracy in visual feature analysis.

In summary, this work not only contributes to the field of real estate analysis with technical rigor and practical applicability but also paves the way for future innovations that can further transform understanding and engagement in the real estate market.

CRediT authorship contribution statement

Wesley F. Maia: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Sergio A. David:** Writing – review & editing, Visualization, Validation, Supervision, Resources, Project administration, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix. Model hyperparameters

Table A.8 details the hyperparameters used in each regression model. These hyperparameters were carefully chosen to optimize the performance of the respective model.

This table provides a detailed summary of the hyperparameters for each model, offering clarity on the specific configuration adopted in the analysis.

Table A.8
Model hyperparameters.

Model	Hyperparameters
Bayesian Ridge	alpha_1 = 2.104e−05 alpha_2 = 8.871e−06 lambda_1 = 0.9517 lambda_2 = 0.01637 compute_score = False
Lasso	alpha = 0.0004225 max_iter = 1000000 tol = 0.001
Elastic Net	alpha = 0.0005033 l1_ratio = 0.8201 positive = True precompute = False selection = 'random' max_iter = 10000000 tol = 0.001
Ridge Regression	alpha = 12.7737
Support Vector	C = 46 epsilon = 0.009 gamma = 0.0003435
Gradient Boosting	n_estimators = 2501 learning_rate = 0.0322
LightGBM	objective = 'regression' num_leaves = 4 learning_rate = 0.01 n_estimators = 5000 max_bin = 200 bagging_seed = 7 feature_fraction_seed = 7 verbose = −1
XGBoost	learning_rate = 0.0092 n_estimators = 4492 max_depth = 4 min_child_weight = 0.0195 gamma = 0.0039 subsample = 0.308 colsample_bytree = 0.1605 scale_pos_weight = 3 reg_alpha = 6.89e−05 objective = 'reg:squarederror'
Stacking	meta_regressor = final use_features = True regressors = estimators
Blended	linear_reg = 0.005 svr_reg = 0.005 bayesian_ridge_reg = 0.005 ridge_reg = 0.05 lasso_reg = 0.1 elastic_net_reg = 0.1 gbr_reg = 0.1 lgbm_reg = 0.1 xgb_reg = 0.1 stacking_cv_reg = 0.435

References

[1] Hoffower H. Millennials aren't buying homes, and it might not be because they can't afford them: Some actually prefer to rent instead. Bus Insider 2019.
[2] Zylberstajn E. Índice fipezap de preços de imóveis anunciados. 2016, Online, Documento técnico.
[3] Choy L, Ho W. The use of machine learning in real estate research. 12, 2023, p. 740. http://dx.doi.org/10.3390/land12040740.
[4] Ghosalkar NN, Dhage SN. Real estate value prediction using linear regression. In: 2018 Fourth international conference on computing communication control and automation. ICCUBEA, IEEE; 2018, p. 1–5.
[5] Valenti A, Giuffrida S, Linguanti F. Decision trees analysis in a low tension real estate market: The case of troina (Italy). In: Computational science and its applications–ICCSA 2015: 15th international conference, banff, AB, Canada, June 22-25, 2015, proceedings, part III 15. Springer; 2015, p. 237–52.
[6] Rai H, Jagannathan M, Delhi VSK. Claim tenability assessment in Indian real estate projects using ANN and decision tree models. Built Environ Project Asset Manag 2021;11(3):468–87.
[7] Xu K, Nguyen H. Predicting housing prices and analyzing real estate market in the chicago suburbs using machine learning. 2022.
[8] Tchuente DN, Nyawa S. Real estate price estimation in french cities using geocoding and machine learning. Ann Oper Res 2022;1–38.
[9] Besanko D, Dranove D, Shanley M, Schaefer S. Economics of strategy. John Wiley & Sons; 2009.
[10] Yoshida T, Fujimoto Y. Spatial dependencies in machine learning models for apartment rent prediction. Real Estate Econ 2022;50:789–812.
[11] Ogundunmade TP, Abidoye M, Olunfunbi OM. Modelling residential housing rent price using machine learning models. Mod Econ Manag 2023;2(14):1–8.
[12] Zhang M, Chen Q, Wang L. Housing rent prediction in India using stacking ensemble machine learning techniques. Int J Hous Markets Anal 2019;13:123–45.
[13] Sartoris Neto A, Fava VL. Estimação de modelos de preços hedônicos: um estudo para residências na cidade de são paulo. 1996.
[14] Herath S, Maier G. The hedonic price method in real estate and housing market research: A review of the literature. 2010.
[15] Sirmans S, Macpherson D, Zietz E. The composition of hedonic pricing models. J Real Estate Literature 2005;13(1):1–44.
[16] Huang Z, Chen R, Xu D, Zhou W. Spatial and hedonic analysis of housing prices in Shanghai. Habitat Int 2017;67:69–78.
[17] Gilbukh S, Goldsmith-Pinkham P. Heterogeneous real estate agents and the housing cycle. tech. rep., National Bureau of Economic Research; 2023.
[18] Hu Y, Lu B, Ge Y, Dong G. Uncovering spatial heterogeneity in real estate prices via combined hierarchical linear model and geographically weighted regression. Environ. Plan. B: Urban Anal City Sci 2022;49(6):1715–40.

[19] Shelton JP. The cost of renting versus owning a home. Land Econom 1968;44(1):59–72.
[20] Mueller G. Real estate rental growth rates at different points in the physical market cycle. J Real Estate Res 1999;18(1):131–50.
[21] Ullah F, Sepasgozar SM. Key factors influencing purchase or rent decisions in smart real estate investments: A system dynamics approach using online forum thread data. Sustainability 2020;12(11):4382.
[22] Piazolo D, Dogan UC. Impacts of digitization on real estate sector jobs. J Prop Invest Finance 2021;39(2):47–83.
[23] Mendonça MJCd. O crédito imobiliário no Brasil e sua relação com a política monetária. Rev Brasileira Econ 2013;67:457–95.
[24] Diniz MH. Curso de Direito Civil Brasileiro: Responsabilidade Civil-v. 7. Saraiva Educação SA; 2010.
[25] Warren-Myers G. The value of sustainability in real estate: a review from a valuation perspective. J Prop Invest Finance 2012;30(2):115–44.
[26] Falkenbach H, Lindholm A-L, Schleich H. Review articles: environmental sustainability: drivers for the real estate investor. J Real Estate Lit 2010;18(2):201–23.
[27] Thériault M, Des Rosiers F. Modeling urban dynamics: Mobility, accessibility and real estate value. John Wiley & Sons; 2013.
[28] Stamm D, Riggs W. Real estate and new mobility. In: Disruptive transport: driverless cars, transport innovation and the sustainable city of tomorrow. Routledge 2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN; 2018, p. 66–75.
[29] Abdul Shaban ZA. Socio-spatial segregation and exclusion in mumbai. Int J Urban Reg Res 2016;153–70.
[30] van Ham M, Tammaru T, Ubarevičienė R, Janssen H. Urban socio-economic segregation and income inequality: A global perspective. Springer; 2023.
[31] Deden Rukmana DR. Income inequality and socioeconomic segregation in jakarta. Urban Stud 2023;135–52.
[32] Grybauskas A, Pilinkienė V, Stundžienė A. Predictive analytics using big data for the real estate market during the COVID-19 pandemic. J Big Data 2021;8(1):1–20.
[33] Khobragade AN, Maheswari N, Sivagami M. Analyzing the housing rate in a real estate informative system: A prediction analysis. Int J Civil Engine Technol 2018;9(5):1156–64.
[34] Lorenz F, Willwersch J, Cajias M, Fuerst F. Interpretable machine learning for real estate market analysis. Real Estate Econ 2023;51(5):1178–208.
[35] Zapimoveis-scraper. 2024, https://pypi.org/project/zapimoveis-scraper/, Descrição do projeto: zapimoveis-scraper is a Python package that works as a crawler and scraper using beautifulsoup4 to get data from zap imóveis.
[36] Geopy documentation. 2024, https://geopy.readthedocs.io/en/stable/.
[37] GeoSampa. 2024, https://geosampa.prefeitura.sp.gov.br/PaginasPublicas/_SBC.aspx. Acesso em 2024.
[38] Van Den Hoek J, Friedrich HK, Ballasiotes A, Peters LE, Wrathall D. Development after displacement: Evaluating the utility of OpenStreetMap data for monitoring sustainable development goal progress in refugee settlements. ISPRS Int J Geo-Inf 2021;10(3):153.
[39] Chopde NR, Nichat M. Landmark based shortest path detection by using A* and haversine formula. Int J Innov Res Comput Commun Eng 2013;1(2):298–302.
[40] Chowdhary K, Chowdhary K. Natural language processing. Fundam Artif Intell 2020;603–49.
[41] Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. IEEE Intell Syst Appl 1998;13(4):18–28.
[42] Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Stat 2001;1189–232.
[43] Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, et al. Xgboost: extreme gradient boosting. 1, (4):2015, p. 1–4, R package version 0.4-2.
[44] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. Lightgbm: A highly efficient gradient boosting decision tree. In: Advances in neural information processing systems, vol. 30, 2017.
[45] Pavlyshenko B. Using stacking approaches for machine learning models. In: 2018 IEEE second international conference on data stream mining & processing. DSMP, IEEE; 2018, p. 255–8.
[46] Liashchynskyi P, Liashchynskyi P. Grid search, random search, genetic algorithm: a big comparison for NAS. 2019, arXiv preprint arXiv:1912.06059.
[47] Anguita D, Ghelardoni L, Ghio A, Oneto L, Ridella S, et al. The'K'in K-fold cross validation. In: ESANN. 2012, p. 441–6.
[48] Streamlit. 2024, https://streamlit.io/, Streamlit é uma plataforma para construir aplicativos web com Python.