

The Investigation of Rental Market Rent Models Based on Machine Learning

Shiqi Tang *

Department of Architecture, Sichuan Agricultural University, Chengdu, Sichuan Province, China

* Corresponding Author Email: jotoole75316@student.napavalley.edu

Abstract. In terms of the renting field, landlords often set rental prices subjectively based on market trends, lacking standardized criteria. This can lead to arbitrary pricing and potential market instability. In contrast, machine learning, with its advancements across various industries, offers a solution to model these complex, non-linear relationships. Although numerous studies have investigated the factors that impact rental prices, there is a scarcity of research dedicated to predicting rental prices, in contrast to the extensive body of work centered on forecasting property prices within the real estate sector. Firstly, this study examines key factors influencing rental prices. Utilizing a Kaggle dataset containing statistical information on Italian housing rentals, including architectural and geographical features, this study processes the data, handles missing values, removes outliers, and analyzes feature correlations. Various machine learning models, including Backpropagation (BP) neural networks, decision trees, random forests, gradient boosting regression, and multiple linear regression, are employed for rental price prediction. The study concludes by comparing model performance and discussing their strengths and weaknesses. Key findings reveal that neural networks outperform other methods in rental price prediction, showing the lowest Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) errors, while decision trees perform the poorest. Furthermore, bathroom count exhibits the highest correlation with rental prices, followed by property size, with villa status ranking third.

Keywords: Machine Learning; Neural Network; Rent Model Prediction.

1. Introduction

Currently, there are challenges within the rental housing market. Landlords typically determine rental prices based on prevailing market rates and trends for similar properties. This method is highly subjective and lacks a standardized rental criterion, often resulting in arbitrary pricing that can adversely affect the overall health of the market. In contrast, machine learning, a field that has achieved significant advancements in theory and practical applications across various industries like manufacturing, information technology, and finance, is ideally suited for modeling these complex, nonlinear relationships.

In the research of the housing rental market, numerous studies have been investigated on the factors influencing rental prices. For instance, Zhang et al. [1] examined the influencing factors of residential rent in Guangzhou. Wang et al. [2] analyzed the factors contributing to residential rent distribution within the inner ring of Shanghai using the Geographically Weighted Feature Price method. Meng et al. [3] conducted research on the factors influencing residential rent in Wuhan using the Feature Price Model. Using common linear functions as the basic modeling form and analyzing the regression coefficients of various feature factors, they established feature price models for residential rent in their respective cities. Lu et al. [4] studied the impact of public service facilities on rental housing prices in Beijing using feature price models and spatial econometric models. Tian et al. [5] analyzed the spatial differentiation of rent and influencing factors in Beijing based on nine factors categorized as location features, neighborhood features, and overall features. Han et al. [6] used the Feature Price Model to analyze the impact of Chongqing's rail transit on property prices along the rail line.

However, research on rent prediction in the rental housing market is relatively scarce, with more focus on predicting property prices in the real estate market. Apart from traditional valuation methods, researchers have continually applied and developed new real estate evaluation methods, including the

use of machine learning methods, particularly artificial neural networks, in real estate evaluation. Do and Grudnitski [7] used Artificial Neural Networks (ANN) in real estate evaluation and demonstrated the superiority of neural networks. In addition to these methods, some researchers proposed the application of other machine learning methods like Support Vector Machines (SVM) and Random Forests in real estate evaluation. Antipov and Pokryshevskaya [8] employed the Random Forest model in real estate evaluation and found it outperformed other methods, including SVM and neural networks, in their experiments. Yang [9] and Chen [10] also demonstrated the superiority of Random Forest in predicting relevant real estate prices.

Although machine learning has found extensive use among researchers in forecasting property prices within the real estate sector, some studies have explored the application of machine learning to predict rental prices within the rental housing market. For instance, Guo [11] developed two optimized rent prediction models, Random Forest and XGBoost, and compared them comprehensively. The results indicated that both models exhibited excellent predictive performance for rental prices, with mean absolute errors and root mean square errors both less than 0.1. This demonstrates that machine learning can be applied not only to property price prediction but also to rental price prediction.

In summary, there are several key issues in rental market research: a lack of comprehensive understanding of the factors influencing rent, a need for detailed comparative research on different rent prediction algorithms, and relatively limited research on international rental markets compared to domestic markets. Therefore, the goal of this paper is to address these limitations by comparing the application of various methods e.g. decision trees, gradient boosting, Random Forest, multiple linear regression, and neural networks in rent prediction. It aims to identify the most effective method and conduct in-depth research on various rent-influencing features, providing valuable reference information for the real estate market.

2. Method

2.1. Data Source and Features

The data used in this study was sourced from Kaggle [12], which provides statistical information about the rental of houses in Italy. The dataset consists of 31 variables, categorized into two main parts: building features and location features. The building features encompass 27 variables, while the location features include three variables, which are the region where the building is located, the city where the building is situated, and the specific area within the city. The target variable for prediction in this study is the rent amount. The dataset includes categorical variables, numerical variables, and character variables. The detailed information related to original features and sample instances can be found in Table 1.

Table 1. Original Features and Sample Instances.

Feature Names		Variable Types	Sample Instances
Building Characteristics	parking_spaces	Categorical variable (Class 1: Yes; Class 2: No)	0
	bathrooms_per_room	Numerical variable	1
	bathrooms	Numerical variable	1
	rooms	Numerical variable	1
	top_floor	Categorical variable (Class 1: Yes; Class 2: No)	0
	state	(Class 1: Excellent / Renovated; Class 2: Good / Habitable; Class 3: New / Under Construction; Class 4: Needs Renovation)	buono / abitabile
	energy_rating	Categorical variable (Class 1: A; Class 2: B; Class 3: C; Class 4: D; Class 5: E; Class 6: F; Class 7: G).	D
	sea_view	Categorical variable (Class 1: Yes; Class 2: No)	0
	centralized_warming	Categorical variable (Class 1: Yes; Class 2: No)	0
	surface	Numerical variable	35
	furnished	Categorical variable (Class 1: Yes; Class 2: No)	1
	balcony	Categorical variable (Class 1: Yes; Class 2: No)	0
	tv_system	Categorical variable (Class 1: Yes; Class 2: No)	0
	external_exposure	Categorical variable (Class 1: Yes; Class 2: No)	0

	optic_fiber	Categorical variable (Class 1: Yes; Class 2: No)	0
	electric_gate	Categorical variable (Class 1: Yes; Class 2: No)	1
	cellar	Categorical variable (Class 1: Yes; Class 2: No)	0
	common_garden	Categorical variable (Class 1: Yes; Class 2: No)	0
	private_garden	Categorical variable (Class 1: Yes; Class 2: No)	0
	alarm_system	Categorical variable (Class 1: Yes; Class 2: No)	0
	doorman	Categorical variable (Class 1: Yes; Class 2: No)	0
	pool	Categorical variable (Class 1: Yes; Class 2: No)	0
	villa	Categorical variable (Class 1: Yes; Class 2: No)	0
	entire_property	Categorical variable (Class 1: Yes; Class 2: No)	1
	apartment	Categorical variable (Class 1: Yes; Class 2: No)	0
	attic	Categorical variable (Class 1: Yes; Class 2: No)	1
	lofts	Categorical variable (Class 1: Yes; Class 2: No)	0
	mansard	Categorical variable (Class 1: Yes; Class 2: No)	0
Geographical Features	region	Character variables	piemonte
	city	Character variables	Torino
	quarter	Character variables	Via della
	price	/	€500 per month

2.2. Data Preprocessing

2.2.1. Handling Missing Values in Features

There are 11 features with missing values in the data. Based on the statistical information, each feature with missing values is processed as follows. After model validation, the following methods for handling missing values have effectively improved the data:

The missing data is categorized into five types for processing: 1) Delete data with missing labels. 2) Delete data with identical variables, including TV system, alarm system, and doorman, entire_property. 3) fill in some missing data (such as region, area, room status) with "unknown." This is because regional data and room status are untraceable, and they are character-type data, so they are filled with "unknown." 4) For some missing data: 1. this study filled in the missing energy efficiency rating of buildings or equipment with ",". The original data was divided into six levels: A, B, C, D, E, and F. The reason for missing data may be that the building has not been rated, so it is filled with ",". 2. This study filled in the remaining missing numerical data with "0". The main reason these data were not filled is that the building did not undergo the corresponding data collection, or the relevant information is 0. The Table 2 below provide more details.

2.2.2. Data Standardization

Data was standardized using $(data - data.min()) / (data.max() - data.min())$ to scale feature values within the range [0, 1], facilitating improved model training. Standardization can expedite model convergence and enhance model performance.

Table 2. Handling Missing Values in Feature.

Features with Missing Values	Number of Missing Values	Proportion of Missing Values	Treatment Method.
price	4087	0.0975	Delete Samples with Missing Data
region	12	0.0002	Fill with "unknown"
quarter	4233	0.0632	Fill with "unknown"
state	4339	0.0647	Fill with "unknown"
energy_rating	13580	0.2026	Fill with ","
bathrooms_per_room	6121	0.0913	Fill with "0"
bathrooms	2037	0.0304	Fill with "0"
rooms	5101	0.0761	Fill with "0"
sea_view	55	0.0008	Fill with "0"
centralized_warming	11125	0.1660	Fill with "0"
surface	4935	0.0736	Fill with "0"

2.2.3. Removing Outliers in Unit Price

To more accurately analyze the price distribution of the majority group (regular houses) and focus on the main trends, this study removes data with monthly house rents exceeding 5000 euros. This decision is based on observations from the price distribution chart, where houses with prices exceeding 5000 euros per month represent a relatively small portion of the data. Price data that exhibit exceptionally high or low values exert minimal influence on the comprehensive analysis of price trends. Consequently, this study opts to exclude such less representative data to sharpen the focus on predicting price trends within the predominant cohort.

2.3. Analysis of the Processed Dataset

2.3.1. Correlation of Various Features

Comparing the relationship between various features and transaction prices horizontally is firstly implemented.

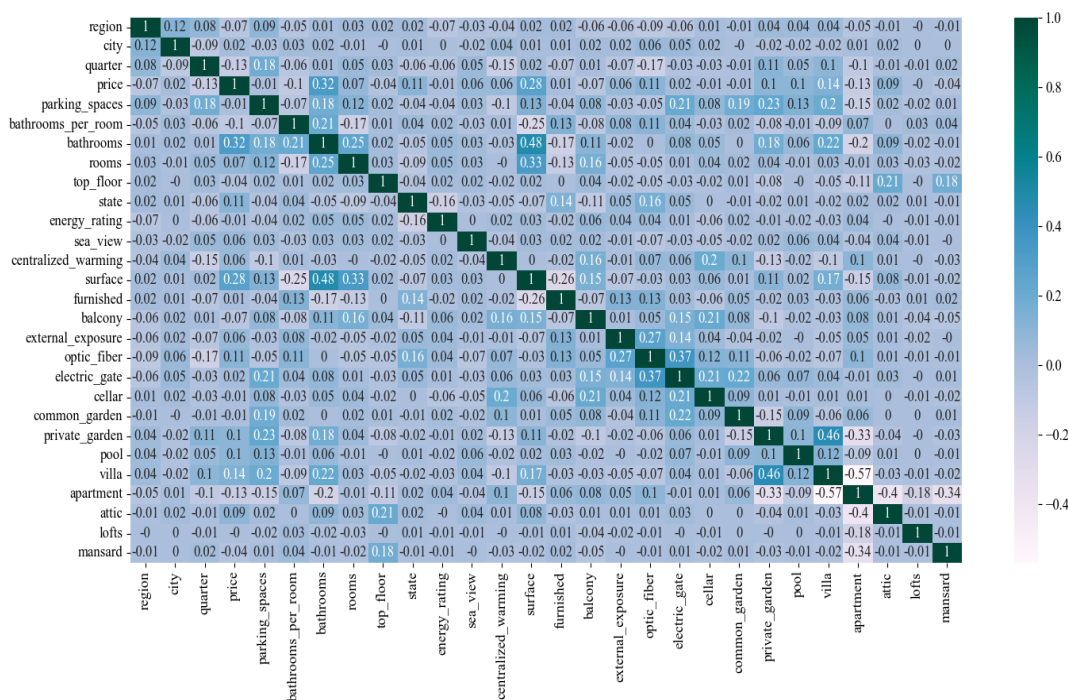


Fig 1. Feature Correlation (Photo/Picture credit: Original).

From the Fig. 1, it can be observed that the highest correlations between features are all below 0.5. Hence, the interconnections among the features exhibit relatively weak linearity, resulting in a low degree of linear commonality. As a consequence, there is no necessity for feature selection or feature engineering.

The three features with the highest correlation to rent are as follows: The number of bathrooms has the highest correlation with rent (0.32). The housing area (0.28) is the second most correlated feature with rent. Whether the property is a villa is the third most correlated feature with rent (0.14).

2.3.2. Relationship Between Building Features and Rent

For building features with fewer numerical variables and one or more categorical variables, this study uses box plots to visualize the data. Box plots effectively display the distribution of numerical variables under different categorical conditions, including statistics such as the median, quartiles, and outliers. When analyzing the relationship between a large number of numerical variables (e.g., area and average number of bathrooms per room), scatter plots are used. Scatter plots simultaneously display data points for two numerical variables, helping us observe their correlations, trends, and distribution characteristics.

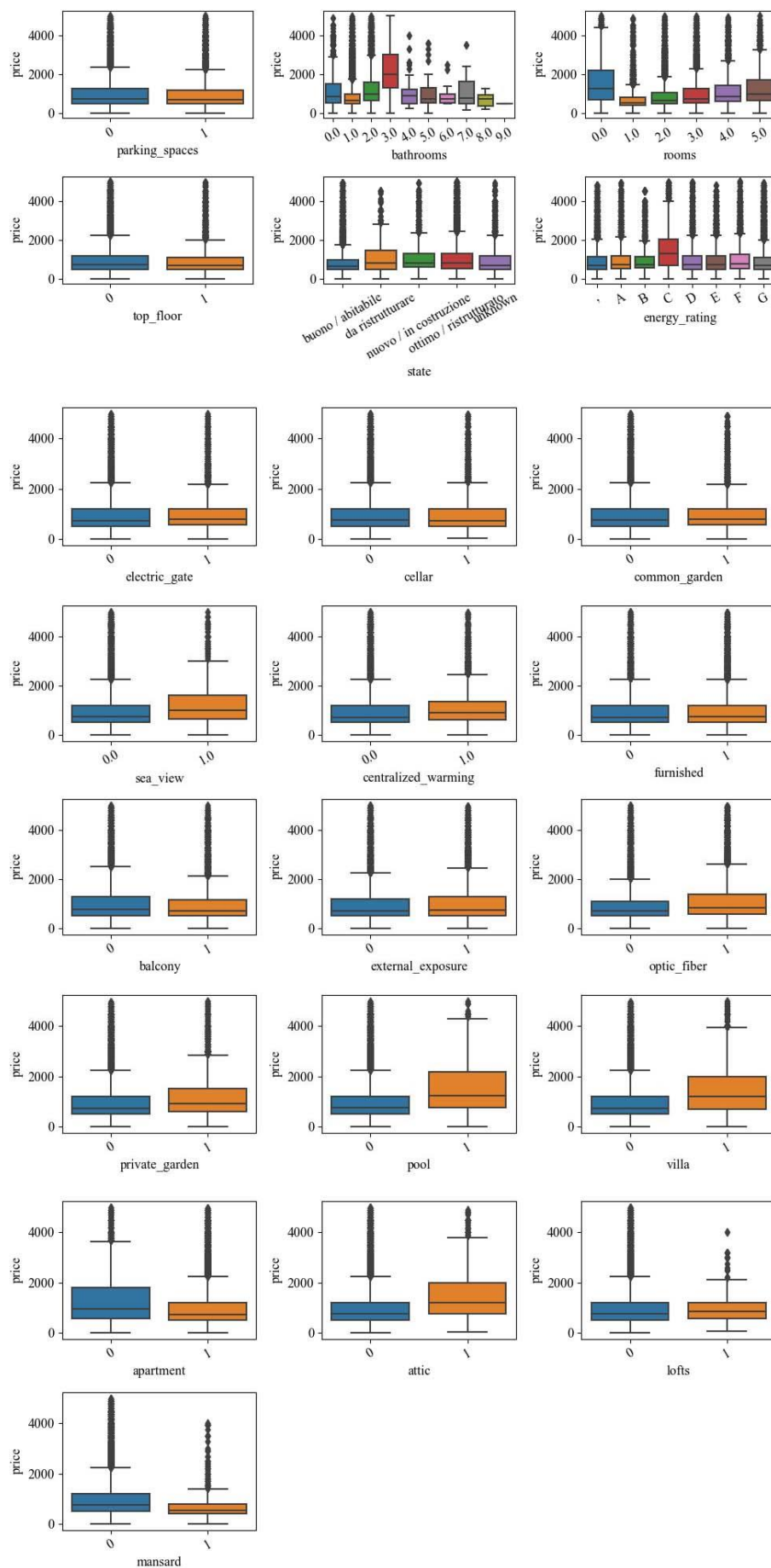


Fig 2. Boxplots Analyzing Building Features and Rent (Photo/Picture credit: Original).

The above graphs shown in Fig. 2 represent box plots of some features in relation to rent. From these graphs, the following observations can be made: 1) Features such as "sea view," "private garden," "swimming pool," and "villa" have an impact on rent. Properties with these features tend to have higher average rents compared to those without. 2) Houses with three bathrooms have significantly higher average rents than houses with other bathroom counts. Additionally, as the number of rooms increases, the average rent also tends to be higher, and room prices become more dispersed. Buildings with an energy efficiency rating of C have higher average rents. 3) For the other features shown in the graphs, their relationship with rent is not clearly evident.

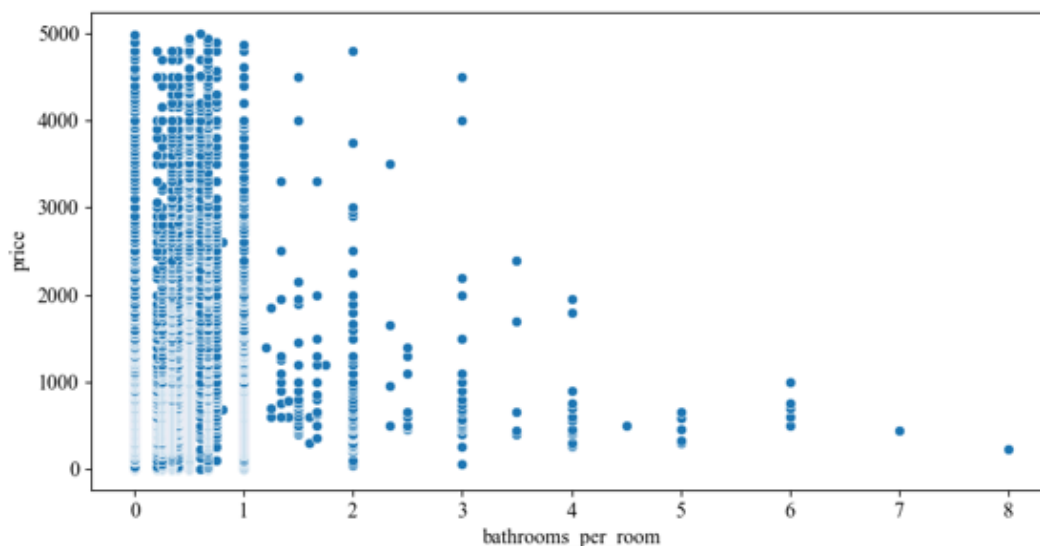


Fig 3. Scatterplot Analyzing the Feature of Average Bathrooms per Room and Rent (Photo/Picture credit: Original).

Fig. 3 illustrates the relationship between house prices and the average number of bathrooms per room. From the scatterplot, it is noticeable that the peak in house sales activity takes place when the average number of bathrooms per room falls within the range of 0 to 1. Furthermore, rent prices are distributed within the range of 0 to 5000 in this case. However, as the average number of bathrooms per room increases, the rent price range gradually narrows down and tends to approach lower values.

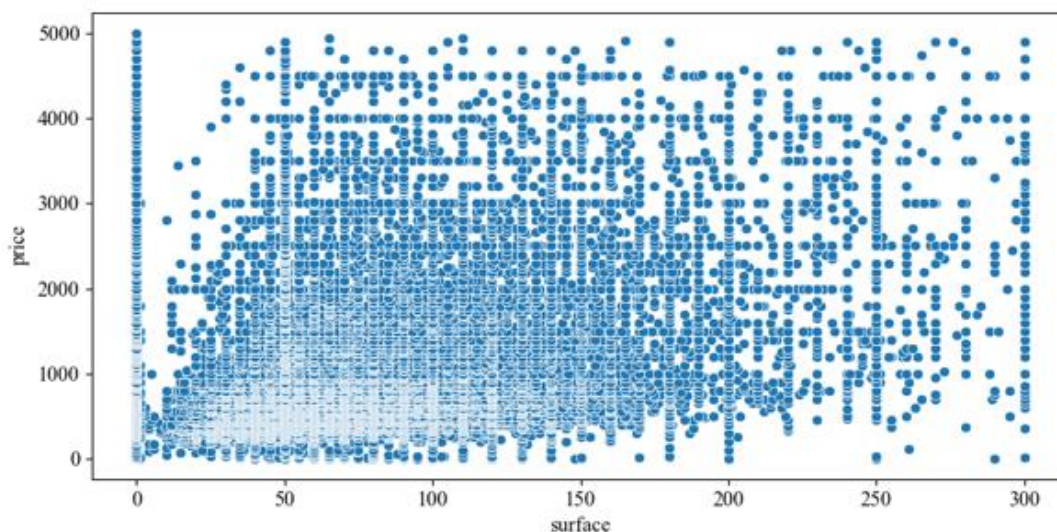


Fig 4. Scatterplot Analyzing the Feature of House Area and Rent (Photo/Picture credit: Original).

From the Fig. 4, it can be observed that house areas range from 0 to 300 square meters. As the size of the house decreases, the rent primarily clusters in lower price ranges. Conversely, with an increase in house size, the rent tends to exhibit greater dispersion.1. The Relationship between Location Features and Unit Price.

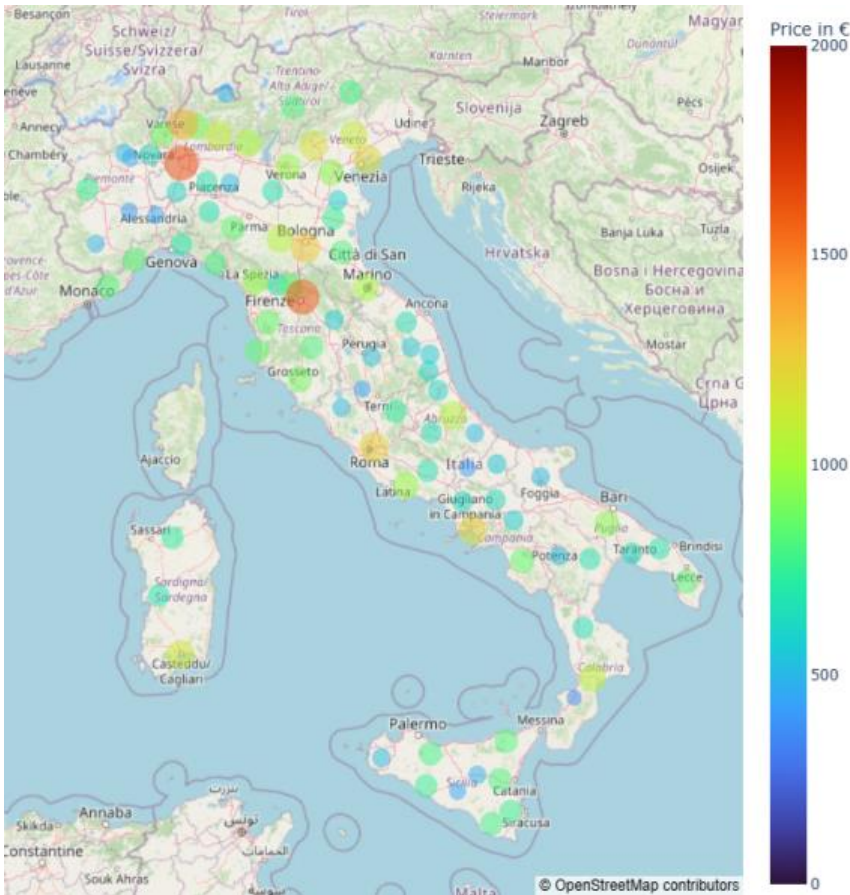


Fig 5. Scatterplot Map at the Province-City Level (Photo/Picture credit: Original).

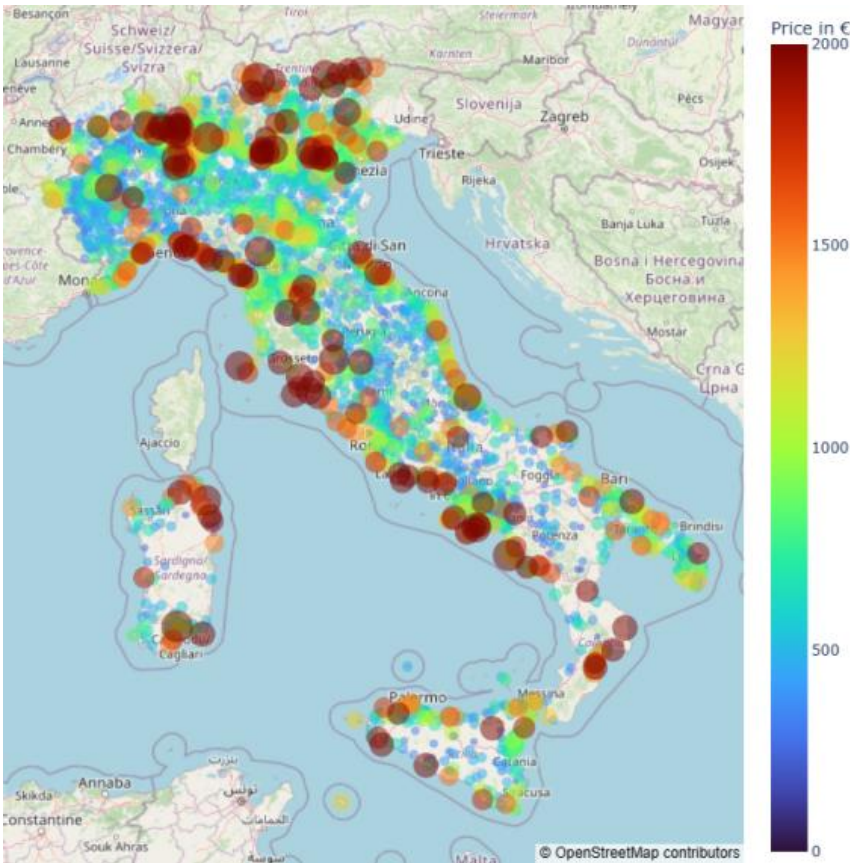


Fig 6. Scatterplot Map Grouped by Municipal Districts (Photo/Picture credit: Original).

The left graph groups data by provinces and cities, while the right graph group's data by city districts shown in Fig. 5 and Fig. 6. These scatter maps display the average price distribution in different cities, highlighting price differences through color and size. The specific steps include merging the original dataset with a dataset containing province coordinates, grouping by city, latitude, and longitude features, calculating the average house price for each city, sorting the data, and generating the scatter maps to emphasize price variations among different cities.

The maps illustrate that coastal regions exhibit higher house prices in contrast to inland areas. This observation is consistent with the prior analysis of the correlation between property attributes and rental rates, where the presence of a sea view was linked to elevated prices. Additionally, Milan, an inland city, and exhibits significantly higher rents compared to its surrounding areas. This can be attributed to Milan's status as a political and economic center, a hub for business, finance, and industry, as well as a center for fashion, culture, and the arts.

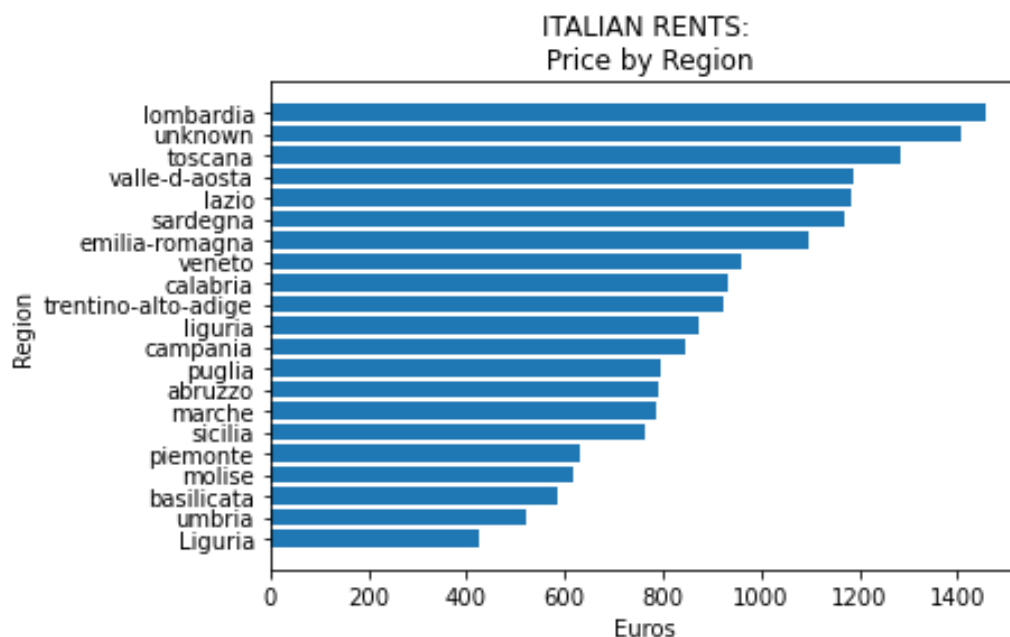


Fig 7. Scatterplot Map Grouped by Municipal Districts (Photo/Picture credit: Original).

In order to determine whether Milan, as the economic center of Italy, has higher average housing prices or if cities near the coast have the highest prices, this paper visualizes the calculated average rental prices for each region using a horizontal bar chart. The Fig. 7 shows that Lombardy, the region where Milan is located, has the highest average housing transaction price, excluding "unknown" values. The coastal region of Tuscany ranks second in average housing prices. This suggests that Milan's influence on rent, as an economic center, is generally greater than the influence of coastal areas.

3. Machine Learning Models

3.1. BP Neural Network Model

3.1.1. Principles of the Back Propagation (BP) Neural Network

Artificial Neural Networks (ANN) are based on the physiological study of the human brain, abstracting the neural networks of the brain to achieve complex information processing similar to the human brain. Neural networks consist of a large number of nodes (neurons) connected to each other, with neurons serving as the basic processing units of the network. The information processing function of neural networks is determined by the input-output characteristics of neurons, network topology, connection weights, and neuron thresholds, among other factors.

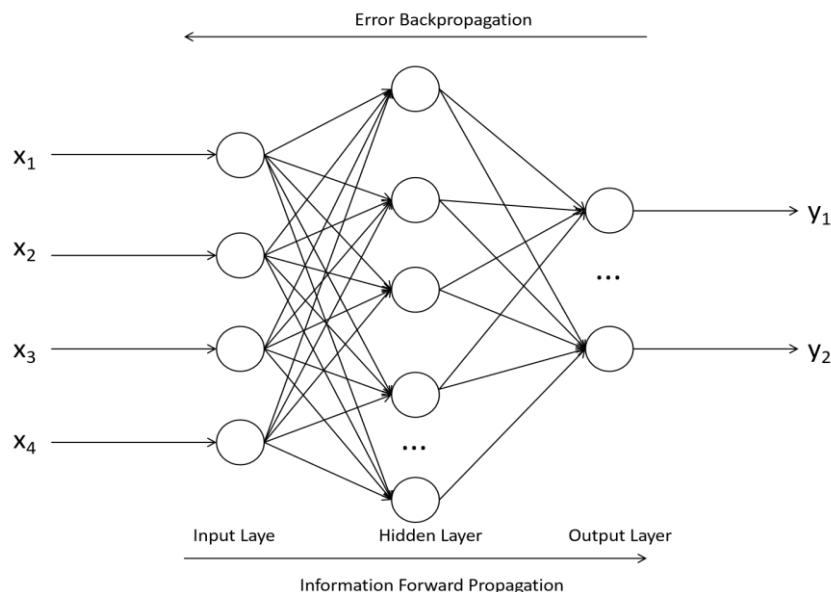


Fig 8. Neural Network Process (Photo/Picture credit: Original).

3.1.2. Artificial Neural Network (ANN) Models

Artificial neural networks encompass various models, and one widely used model is the BP neural network. The BP neural network consists of an input layer, one or more hidden layers, and an output layer, as shown in Fig. 8.

3.1.3. Building the BP Neural Network Model

This study constructed an ANN model for housing price prediction using the TensorFlow and Keras libraries. The following steps outline the process: 1) Creating the Model Object: The research created a Sequential model object, which is a simple model container for stacking neural network layers sequentially. 2) Adding the Input Layer the research added a fully connected layer as the input layer with 80 neurons, using Rectified Linear Unit (ReLU) as the activation function. The ``input_shape`` parameter specified the input data shape, i.e., the number of features. 3) Adding Hidden Layers: A hidden layer with 20 neurons was added, again using ReLU as the activation function. Hidden layers are responsible for extracting features and patterns from the data. 4) Adding the Output Layer: An output layer containing a single neuron with a linear activation function was added. In regression tasks, a linear activation function is typically used in the output layer as it produces real values as predictions. 5) Compiling the Model: the research selected the Adam optimizer as the optimization algorithm and Mean Squared Error (MSE) as the loss function. MSE is a commonly used loss function in regression tasks to measure the squared difference between predicted values and actual values. Additionally, the research monitored Mean Absolute Error (MAE). 6) Displaying the Model Summary: The summary information of the model, including the names of each layer, output shapes, and the number of trainable parameters, was printed. 7) Training the Model: the research trained the model using the training dataset and conducted 100 epochs of training, with each batch containing 250 samples. A validation dataset was also used to monitor the model's performance. The training process continuously adjusted the model's weights and biases to minimize the loss function, aiming for better fitting of the training data. 8) Making Predictions: After training, the trained model ``model`` was used to predict the test dataset ``test_x``, resulting in prediction values ``y_pred``. The prediction results were then transformed into a DataFrame with the same index as the test dataset for further analysis and visualization.

In summary, this study built a deep neural network model consisting of input, hidden, and output layers for housing price prediction. The model was trained using the backpropagation algorithm and the Adam optimizer for regression tasks. Model performance metrics, including the loss function Mean Squared Error (MSE) and Mean Absolute Error (MAE), were monitored during training. Ultimately, the model was used to generate housing price predictions.

3.1.4. Building the BP Neural Network Model

To visualize the changes in the loss function and MAE during the training process of the neural network model, a code snippet was created. It displays two curves: one for training loss and validation loss, and another for training MAE and validation MAE. These curves provide insights into the effectiveness of the model's training process and can help identify issues such as overfitting or underfitting.

3.2. Random Forest Model

The random forest algorithm builds multiple decision trees, often using Classification and Regression Trees (CART), and combines their outputs to generate the final prediction. It introduces random attribute selection during tree construction, making it an ensemble algorithm.

3.3. Gradient Boosting Regression Model

Gradient Boosting is a boosting algorithm that trains new weak learners based on the negative gradient information of the loss function. It combines these learners to improve the overall model's predictive performance.

3.4. Decision Tree Regression Model

Decision trees are machine learning algorithms that create a tree-like structure where each internal node represents a feature, and each leaf node represents an output value or class. Decision tree regression predicts the output value by following the tree structure based on input features.

3.5. Linear Regression Model

Linear regression is a widely-used statistical method for predicting a continuous numeric output variable based on one or more input features. It assumes a linear relationship between the inputs and the output, aiming to find the best-fitting line to minimize the difference between predicted and actual values.

4. Results and Discussion

4.1. Evaluate the Model's Indicators

Model Evaluation Metrics to objectively compare the performance of various models, the following three metrics were established to evaluate the model's indicators:

4.1.1. Mean Absolute Error (MAE)

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (1)$$

Where "n" represents the number of samples, " y_i " represents the predicted value, and " $\hat{y}_i$ " represents the actual value. This metric represents the average difference between predicted values and actual values, effectively reflecting the actual level of overall error. It is an intuitive measurement. If the Mean Absolute Error is large, it indicates a higher overall error, implying poorer model performance.

4.1.2. Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (2)$$

This metric is the square root of the average of the squared differences between predicted values and actual values. RMSE is sensitive to outliers and is more influenced by larger values while reflecting overall error. A larger RMSE indicates a higher likelihood of significant errors in individual samples, signifying poorer model performance.

4.2. Fitting Results for Various Model

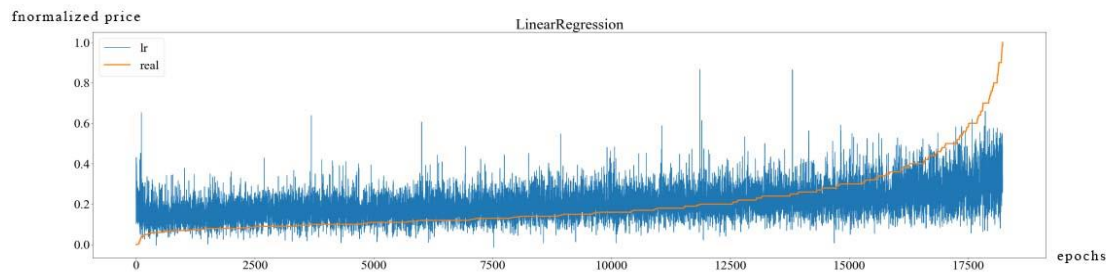


Fig 9. Fitting Results of the Linear Regression Model (Photo/Picture credit: Original).

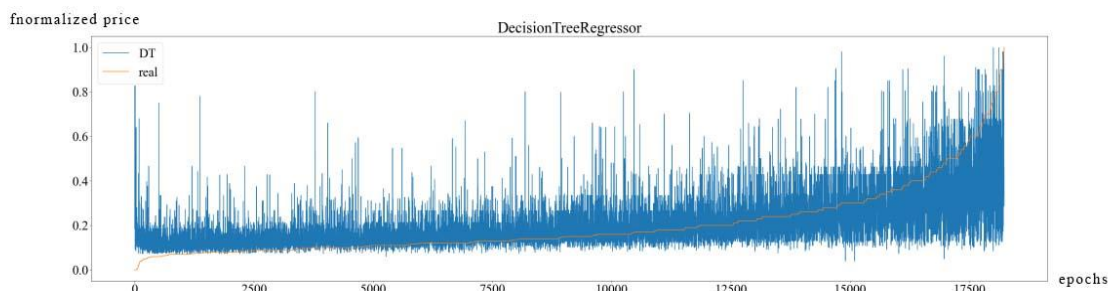


Fig 10. Fitting Results of the Decision Tree Model (Photo/Picture credit: Original).

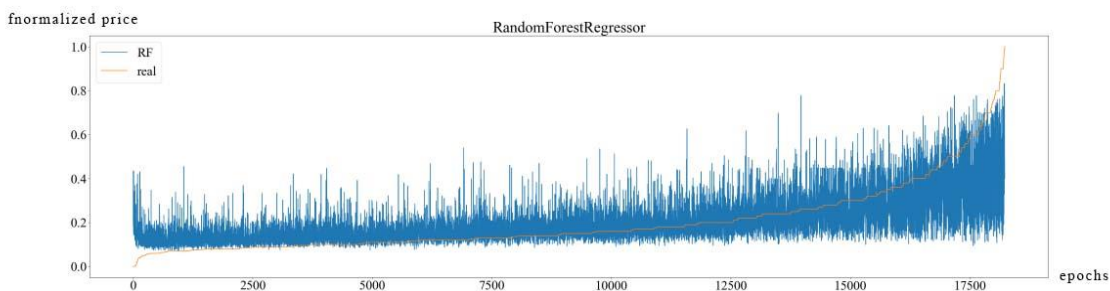


Fig 11. Fitting Results of the Random Forest Model (Photo/Picture credit: Original).

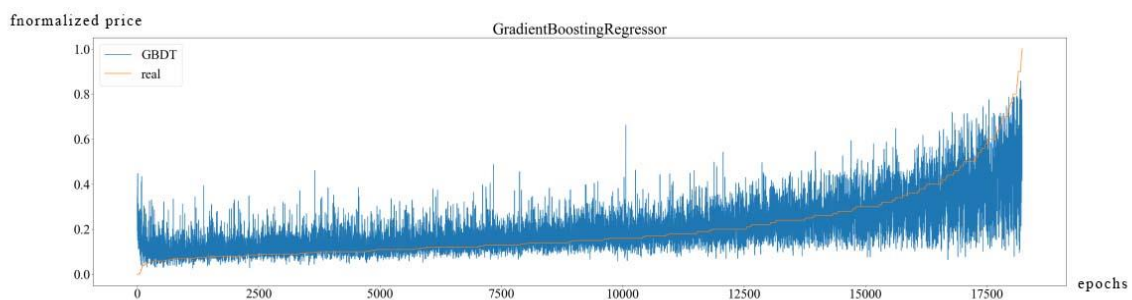


Fig 12. Fitting Results of the Gradient Boosting Regressor Model (Photo/Picture credit: Original).

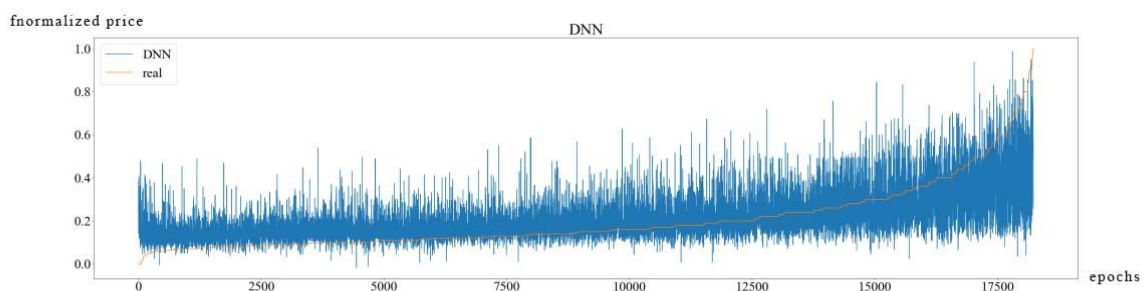


Fig 13. Fitting Results of the DNN Model (Photo/Picture credit: Original).

The experimental results are described from Fig. 9, Fig. 10, Fig. 11, Fig. 12 and Fig. 13, illustrating the performance of different models during the training process as a function of the number of training iterations. It is evident that the decision tree model exhibits the poorest predictive performance. These figures clearly depict significant fluctuations and instability in the predictions of the decision tree model. In contrast, the DNN demonstrates the best performance, characterized by relatively stable predictions with minimal fluctuations.

Upon analyzing the prediction trends, it becomes evident that the linear regression model displays lower performance. For instance, after a certain epoch (e.g., 17500), when the predictions of other models show an upward trend, the linear regression model fails to capture this upward trajectory. In contrast, the other four models adeptly capture and reflect this trend.

4.3. Comparison of Fitting Results for Various Model

From Table 3, it can be observed that the neural network has the lowest MAE and RMSE errors (i.e. 0.013, 0.076), indicating the best performance. Linear regression follows as the next best model, followed by random forest and gradient boosting. The decision tree's subpar performance is evident, as it exhibits a lower MAE but notably higher RMSE when compared to the top-performing neural network. The RMSE difference between them stands at 0.661, underscoring the disparity in their predictive accuracy.

Table 3. Comparison of Generalization Results for Five Models Based on the testing dataset.

	Linear Regression	Decision Tree	Random Forest	Gradient Boosting Regression	Neural Network
MAE	0.017	0.013	0.036	0.040	0.013
RMSE	0.091	0.737	0.130	0.128	0.076

Examining Fig. 10 and combining it with Table 3, it becomes evident that decision trees perform poorly in handling outliers, resulting in a lower MAE and significantly higher RMSE. This is likely because decision trees are prone to overfitting, a known drawback of this model. Overfitting implies that the model performs well on the training data but poorly on new, unseen data. Decision trees have additional shortcomings, including: 1) High Variance: Decision trees often exhibit high variance, meaning they are highly sensitive to minor variations in the data. This may lead to model instability, particularly in cases with small datasets or high noise. 2) Over-Splitting: Decision trees tend to over-split during training, generating overly complex tree structures. This can reduce the model's generalization ability on new data since it over-adapts to the training data's details.

In contrast, one reason for the superior performance of neural networks is their robust nonlinear modeling capability. Neural networks possess numerous tunable parameters and multi-layer structures, enabling them to capture and represent complex nonlinear relationships effectively. This makes neural networks well-suited for handling the intricacies in data, such as those encountered in predicting real estate prices. Additionally, random forests outperform decision trees due to their ensemble nature. Random forests are composed of multiple decision trees, offering robustness against outliers and generating relatively stable output results. This means that even when the input data contains outliers or noise, random forests can produce reliable predictions. However, linear regression performs poorly in empirical studies for several reasons. Firstly, linear regression necessitates pre-specifying a functional form, while the relationship between housing features and prices is typically nonlinear and complex. This highlights the inadequacy of linear regression models in effectively capturing the data. Additionally, linear regression demands that predictors remain free from collinearity, a condition that is frequently violated in housing data due to correlated variables like proximity to transportation and region. These factors collectively contribute to the suboptimal performance of linear regression models. However, this study still has some limitations required to be solved in the future:

1) The method used in this study to analyze the importance of model features relies on statistical prior knowledge, specifically the Pearson Correlation metric. However, it is important to

acknowledge that the perceived importance in statistics may not necessarily align with the importance observed during actual machine learning model training. Machine learning quantifies the importance of different features in influencing the final outcome based on the loss values between predicted results and ground truth. Furthermore, different machine learning models may yield varying assessments of feature importance. Therefore, it is imperative that future research directly analyzes the feature importance within the machine learning models employed in this study.

2) The dataset collected exhibits areas that require further clarification or may potentially contain anomalies. For instance, as depicted in Fig. 3, the observation that an increase in the number of bathrooms correlates with lower prices appears to be counterintuitive. To address this issue in future research, it is advisable to conduct a more thorough investigation, potentially through a literature review or the utilization of advanced data preprocessing techniques such as K-means clustering or the LOF (Local Outlier Factor) anomaly detection algorithm. These approaches can facilitate in rectifying potentially unreasonable or anomalous feature parameters, thereby enhancing the accuracy of the models during both the training and testing phases.

5. Conclusion

In this study, this research first analyzed the impact of various factors on rent prices. Next, this study created a neural network model and compared it with decision trees, gradient boosting, random forests, and multiple regression models to identify the best predictive model. The main findings of this research are summarized as follows: 1) the study reveals that among 30 features, the number of bathrooms exhibits the highest correlation with rent, followed by the house area. Additionally, whether the property is a villa ranks third in terms of its correlation with rent. 2) Neural networks not only find applications in the real estate sales market but also perform well in the housing rental market. When compared to decision trees, gradient boosting, random forests, and multiple regression models, neural networks emerge as the best predictive model.

In summary, this experiment successfully implements a rent prediction method based on deep learning and has achieved significant results. It can assist government authorities in managing the rental market, aid intermediaries in formulating more competitive rental price strategies, and provide renters with better-suited housing options.

Furthermore, the model can be further optimized through techniques such as five-fold cross-validation and hyperparameter tuning to enhance its precision. Regarding feature importance analysis, the current approach utilizes prior knowledge based on statistics, specifically the Pearson correlation coefficient. However, in practical applications, the significance of features during machine learning model training may not perfectly align with traditional statistical insights. Future research endeavors might delve into employing more direct machine learning techniques for the analysis of feature importance, aiming to provide a more precise understanding of how various features influence rental prices. Concerning data quality and handling of outliers, the study has identified certain unusual correlations in the dataset, such as the negative correlation between the number of bathrooms and rent. This necessitates further investigation and data cleansing. Subsequent work may involve comprehensive data quality control measures and the application of advanced data processing techniques to address outliers, thereby improving model performance.

References

- [1] Zhang Ruoxi, Jia Shijun. Study on the influencing factors of housing rent in Guangzhou. *Journal of Construction Management*, 2014, 28(6): 118-123. (in Chinese)
- [2] Wang Jiali, Ji Minhe, Deng Zhongwei. Analysis on the Causes of residential rent distribution in Shanghai's Outer Ring based on geographical weighted characteristic price method. *Areal Research and Development*, 2016, 35(5): 72-80. (in Chinese)
- [3] Meng Junchi, Zhang Shaojun. Study on Influencing Factors of Housing rent based on Characteristic Price Model: A case study of Wuhan City. *Value Engineering*, 2016, 27(3): 199-200. (in Chinese)

- [4] Jun Jun, Wang Zhiwen, Zhang Ji. The impact of public service facilities on rental housing rent in urban areas of Beijing. *Price Monthly*, 2017, 482(7): 7-12. (in Chinese)
- [5] Tian Kunrui. Spatial differentiation and influencing factors of rent in Beijing. Lanzhou Jiaotong University, 2017. (in Chinese)
- [6] Han Yongchao, Chen Chun, SHEN Haojing. Study on the impact of Chongqing Rail Transit on Housing prices based on Eigenprice model-a case study of Rail Line 3. *Price Monthly*, 2017, 476(1): 6-10. (in Chinese)
- [7] A. Quang Do. And Gary Grudnitski. A Neural Network Approach to Residential Property Appraisal. *Real Estate Appraiser*, 1992, 58:38-45.
- [8] Evgeny A. Antipov, Elena B. Pokryshevskaya. Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. *Expert Systems with Applications*, 2012,39(2):1772-1778
- [9] Yang Muxi. Research on second-hand house price evaluation based on Random forest Model. Changsha: Central South University, 2012 .(in Chinese)
- [10] Chen Yijia. Research on valuation model of second-hand houses in Beijing based on stochastic forest theory. Beijing Jiaotong University, 2015. (in Chinese)
- [11] Guo Rumeng. Rent Price forecast and analysis of Influencing factors in Beijing. Beijing University of Technology, 2019. (in Chinese)
- [12] Tommaso Ramella. Eda exploratory analysis, 2023, <https://www.kaggle.com/code/tommasoramella/eda-exploratory-analysis/notebook>.