Exploring Factors Influencing House Rental Prices

A Statistical Analysis Using Machine Learning Techniques

Amila Hettiarachchi
Business Intelligence
Department of Microdata Analysis
Dalarna University
Borlange
v24godhe@du.se

***Abstract*— This research provides a comprehensive analysis of the U.S. rental market, aiming to predict rental prices and identify influential factors associated with rental property advertisements. The study leverages a diverse dataset to uncover key features affecting rental prices, including property size, location, and amenities. Various statistical and machine learning techniques, such as correlation analysis, regression modeling, and association rule learning, were employed to analyze the data. The findings indicate that the number of bedrooms and bathrooms, pet-friendliness, and the advertisement source significantly influence rental prices. Machine learning models demonstrated superior performance in predicting rental prices, achieving an accuracy of 82%. Association rule learning uncovered common property feature patterns, shedding light on key factors influencing rental prices and advertising strategies, thus aiding both landlords and renters with practical recommendations in the real estate sector.**

Keywords- Rental Price Prediction, Machine Learning, Regression Modeling, Association Rule

I. INTRODUCTION

The rental market is a highly attractive and dynamic sector influenced by various factors, including market trends, property characteristics, and location. Understanding these factors is crucial for landlords and renters when deciding on rental prices. The Joint Center for Housing Studies at Harvard University states that the demand and supply for housing in each area, the state of the economy, and the features of rental properties themselves all have an impact on rental costs [1].

In this study, a dataset comprising one million U.S. house rental prices from 2019 was analyzed [2], with a focus on predicting rental prices and exploring associations with amenities, advertisement sources, and pet-friendliness. The objective was to pinpoint pivotal factors influencing rental pricing, aligning with previous research efforts (e.g., [3]). Machine learning algorithms were employed to forecast rental prices and identify crucial features.

Additionally, the analysis of Shanghai Lane house rentals in 2021 was conducted [3], the key variables such as square meters, location, latitude and longitude, bedrooms, living rooms, bathrooms, lofts, outdoor spaces, and underfloor heating were examined to elucidate their impact on rents.

Another goal of this study was to identify the variables influencing advertising effectiveness and rental pricing. It examined rental property amenities using association rules learning to identify trends in feature co-occurrence within listings. It also looked at how these variables affected rental pricing and offered advice to tenants on things to look for when renting a property [6]. To provide insights into successful advertising tactics, the research also examined popular advertiser keywords. Both tenants and landlords may make wise selections in the rental market with the help of this thorough investigation.

II. LITERATURE REVIEW

The rental market is a dynamic and ever-changing landscape influenced by various factors and market trends. Factors like location, bedrooms, bathroom, longitude, latitude, outdoor space are the important factors in deciding the renting prices [3]. Understanding of the crucial factors are important for landlords to optimize the renting income and the renters to seek most suitable properties.

To predict rental prices, a multiple linear regression model was built, like the approach used in the study "A Comprehensive Study of Factors Influencing Rental Prices in Shanghai" [3]. Additionally, more advanced methods such as random forests and decision trees, as discussed in [4], were employed to enhance the accuracy of rental price predictions. These advanced techniques are explained in detail in [5].

Data mining techniques offer a powerful approach to analyze rental property data, uncovering valuable insights and informing strategic decisions for both landlords and renters. Beyond rental price prediction, other research efforts have investigated the use of data mining to understand renter preferences and optimize rental advertisements. For instance, a study by An and Sun utilized text mining techniques to analyze rental listing descriptions. This analysis revealed the most frequently used keywords and phrases by landlords, providing insights into how landlords advertise their properties and potentially influence renter decisions [7].

In addition to rental price prediction, research endeavors have explored data mining applications for discerning renter

preferences and enhancing rental advertisements. For example, a study cited in [8] utilized text mining to dissect rental listing descriptions, uncovering prevalent keywords and phrases used by landlords. This sheds light on landlords' advertising strategies and their potential impact on renter choices.

III. METHOD DESCRIPTION

A. The Dataset of Apartment for Rent

The dataset used for this analysis is the "Apartment for Rent Classified Dataset" obtained from the UCI Machine Learning Repository [2]. This provides extensive details about rental properties, including attributes such as property type, amenities, number of bathrooms and bedrooms, rental price, size in square feet, address, city, state, latitude, longitude, and the time of listing.

B. Data Mining Method

The dataset comprises 100,000 observations and 22 variables. Less than 0.5 percent of the key fields contain missing data. Missing values were handled by removing the corresponding rows. Particularly, the pets allowed column exhibited a considerable number of missing values, primarily indicating the types of animals allowed. For this study, unmarked columns were labeled as 'No pets'. Additionally, to mitigate the influence of outliers, only houses with sizes up to 2000 square feet and prices below 40,000 USD were considered. This approach is supported by the methodology used in similar studies to ensure the robustness of the analysis by excluding extreme values that could skew the results [6]. Table I describes the variables in the dataset.

TABLE I. ATTRIBUTES OF THE FEATURES

Feature	***Data Type***	***Description***
id	Integer	Unique ID
category	String	Category of housing
title	String	Advertisement title
body	String	Advertisement body
amenities	String	Amenities of the property
bathrooms	Decimal	No of bathrooms
bedrooms	Decimal	No of bedrooms
currency	String	Currency [USD]
fee	String	Fee as the commission
has_photo	Thumbnail	Photos of the Property
pets_allowed	String	Allowed pet type
price	Decimal	Rent price
price_display	String	String variable to price
price_type	String	Payment type - monthly / weekly
square_feet	Integer	No of the Square feet in the Property
address	String	Address of the property
cityname	String	City of the property
state	String	State of the property
latitude	Decimal	Latitude
longitude	Decimal	Longitude
source	String	Advertised Source
time	Integer	Date of the advertisement

Moving forward, feature engineering involved introducing new categorical variables, such as square feet bucket and price bucket, derived from square feet and price, respectively. These variables were encoded using techniques like one-hot encoding or label encoding to facilitate their inclusion in the modeling process. Furthermore, numerical features underwent standardization to ensure each feature contributed equally to the model's performance.

Multiple data mining techniques were employed to analyze the dataset, ensuring comprehensive insights and robust model performance. Cross-validation methods were used to validate the models, providing a reliable evaluation of model performance across different subsets of the data.

1) Correlation analysis

Correlation analysis was used to assess the relationships between rental prices and other features. This method identified which features had significant associations with rental prices.

2) Feature Selection and Classification

Decision Trees and Random Forest models were utilized to identify significant features of rental prices. These models helped determine which features were most influential in predicting rental prices.

3) Regression Modeling

Multi-linear regression models were trained to predict rental prices based on the research the analysis of Shanghai Lane house rentals in 2021 [3]. Decision trees and random forest regression models were also trained to evaluate their predictive performance. These models were selected because they have been proven effective in various studies for predicting housing prices and rental rates due to their ability to handle complex relationships between variables and provide robust predictions [4]. Additionally, decision trees and random forests are particularly useful for capturing non-linear interactions and offering high interpretability in modeling efforts [5].

4) Association Analysis

The Apriori algorithm was applied to discover frequent item sets of amenities associated with certain rental advertisement sources and pet friendliness. Further analysis was conducted to identify the associations between various amenities, providing insights into common combinations and their prevalence.

5) Key Word Identification using Word Cloud

Text mining algorithms such as were applied to identify the most important key words in the advertisements and the among amenities and present it through word cloud. The provided code snippet prepares textual data for analysis by removing null values, converting text to lowercase, tokenizing, and cleaning it. Special characters, single letters,

and unwanted words are removed, and the text is lemmatized to reduce words to their base forms. Finally, stop words are eliminated to focus on meaningful content.

These combined methods provided a detailed analysis of the dataset, uncovering valuable patterns and relationships within the rental market. Studies have shown that text mining can effectively reveal patterns and trends in rental listings, aiding in the understanding of renter preferences and optimizing rental advertisements [7][8]

IV. Results and Analysis

Results were discussed in five sections to address the research techniques used.

A. Correlation analysis

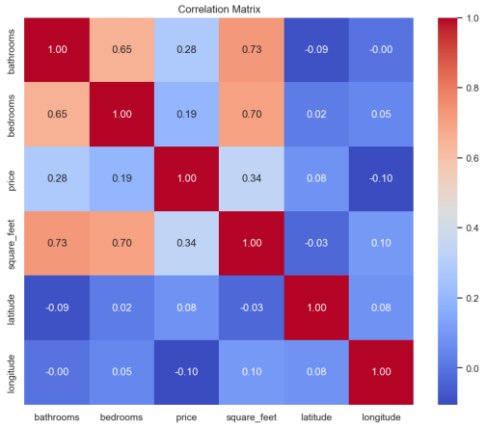

Figure 1. Correlation analysis of the important features

The correlation matrix in Fig 1 shows important connections in the dataset. Strong positive correlations are found between features. A correlation of 0.73 exists between the number of bathrooms and house size, larger homes usually have more bathrooms. Similarly, a correlation of 0.70 indicates larger homes tend to have more bedrooms. Additionally, a correlation of 0.65 between bathrooms and bedrooms highlights their strong association, suggesting larger homes often have more of both. These correlations describe increment in square feet in determining bathroom and bedroom numbers.

B. Feature Selection and Classification

To predict rental prices of houses, it is essential to identify the features that significantly influence the price. Therefore, feature selection is a crucial step in this process. To achieve this, the following statistical methods have been utilized: forward selection, backward elimination, and best subset selection. Table II describes the results of these feature selection methods.

The study incorporated data masking techniques to safeguard sensitive information. For instance, sensitive data such as latitude and longitude, which could compromise privacy, were replaced with fabricated values generated using k-anonymity methods. This approach ensures that individual records cannot be distinguished from a group of similar records, preserving the anonymity of the data. By employing these data masking strategies, the study maintains the confidentiality of sensitive information while allowing for meaningful analysis and insights.

TABLE II. Results of Feature Selection Methods

Feature Selection Method	***Feature Importance Order →***		
	1	***2***	***3***
Forward	Square feet	Latitude	Longitude
Backward	Square feet	Latitude	Longitude
Best subset	bathrooms	bedrooms	Square feet

Further analysis was conducted using Decision Trees Classification and Random Forests Classification models to identify significant features of rental prices. The features longitude, square feet, latitude, bathrooms, and bedrooms were identified as important factors in determining rental prices, as illustrated in Table III. Consequently, these features have been selected for developing the rental price prediction model.

TABLE III. Feature Importance in Decision Tree and Random Forest Classification

Feature	***Decision Tree %***	***Random Forest %***
Longitude	43.44	46.21
Square feet	19.28	23.47
Latitude	15.58	21.65
Bathrooms	3.81	4.14
Bedrooms	1.39	1.85

C. Regression Modeling

Initially, multiple linear regression was employed, but it did not yield satisfactory results, as indicated by a high Root Mean Squared Error (RMSE) of 683.47 and a low R-squared (R^2) value of 0.015. Multiple linear regression model performance was not adequate to predict the price consequently, more advanced models were considered.

Decision Tree Regression and Random Forest Regression models were applied to predict. The performance of these models is summarized in Table IV. The Random Forest Regression model demonstrated superior performance, with an average R^2 value of 0.83, and a 95% confidence interval. The RMSE for the RFR model was 308.58 for unseen data. This implies that, according to the Random Forest Regression model, the predicted rental price can have an error margin of approximately 308 USD with an accuracy of 82% within a 95% confidence interval.

TABLE IV. Decision Tree and Random Forest Model Performance

Model		*Decision Tree*	*Random Forest*
Training (95 % confidence interval)	Average RMSE	386.31	23.09
	R-squared	0.72	0.83
Final Model Testing (un hold data)	Average RMSE	398.48	308.58
	R-squared	0.7	0.82

D. Association Analysis

In this section, associations were analyzed among the key features. The findings were described as follows: House prices were bucketed from $500 USD to $2000 USD with a $500 increment. Associations were explored among amenities, price buckets, pet allowances (dog, cat, or both), and advertising sources. Understanding these associations can assist stakeholders such as renters, real estate agents, and landlords in comprehending common combinations of property features and their likelihood. Notably, it is observed that pet-friendly properties are likely to possess specific features and are often listed on advertising platforms. This knowledge can guide both marketing strategies and search filters on rental platforms.

TABLE V. ASSOCIATIONS AMONG KEY FEATURES OF RENTAL PROPERTIES

Antecedent	*Consequent*	*Support %*	*Confidence %*	*Lift*
RentDigs.com	Gym	20.25	51.14	2.23
Gym	RentDigs.com	20.25	88.43	2.23
Cats, Dogs	Refrigerator, RentLingo	26.53	76.25	1.94
Refrigerator, RentLingo	'Cats', 'Dogs'	26.53	67.62	1.94
Parking, Dishwasher	Refrigerator, RentLingo	23.64	73.86	1.88

As explained in Table V, association analysis was conducted to explore how various amenities and listing platforms are related in rental properties. This analysis revealed five key rules, highlighting how specific combinations of features co-occur more frequently than expected by chance.

Properties listed on RentDigs.com are significantly more likely to have a gym (lift of 2.23). This is supported by a 20.25% support value, indicating this pattern applies to a notable portion of the data. With a confidence level of 51.14%, there's a moderate chance a property on RentDigs.com will have a gym.

Conversely, properties with a gym are overwhelmingly listed on RentDigs.com (confidence level of 88.43%). This is further emphasized by the lift and support values (2.23 and 20.25% respectively), highlighting a strong association between gyms and RentDigs.com listings.

Pet-friendly properties, allowing both cats and dogs, are 1.94 times more likely to have a refrigerator and be listed on RentLingo (lift of 1.94). This is supported by a 26.53% support value, indicating this pattern is prevalent in the data. The confidence level of 76.25% suggests a high likelihood of finding a refrigerator and RentLingo listing in pet-friendly properties.

Properties with a refrigerator listed on RentLingo are more likely to allow both cats and dogs (lift of 1.94). This is reflected in the support value (26.53%) and a confidence level of 67.62%, suggesting a connection between RentLingo listings with refrigerators and pet-friendly environments.

Properties with both parking and a dishwasher are 1.88 times more likely to have a refrigerator and be listed on RentLingo (lift of 1.88). This pattern is supported by a 23.64% support value and a confidence level of 73.86%, indicating a positive association between these amenities and RentLingo listings with refrigerators.

The amenities of the properties were analyzed. The observations are summarized in Table V, highlighting the relationships between different amenities.

TABLE VI. RENTAL PROPERTY AMENITIES USING ASSOCIATION ANALYSIS

Antecedent	*Consequent*	*Support %*	*Confidence %*	*Lift*
Patio/Deck and Refrigerator	Dishwasher	22.58	90.56	1.79
Cable or Satellite	Refrigerator	22.41	86.52	1.79
Patio/Deck and Dishwasher	Refrigerator	22.58	86.10	1.78
Refrigerator	Dishwasher and Parking	26.46	54.65	1.71
Dishwasher and Parking	Refrigerator	26.46	82.67	1.71

The presence of both a Patio deck and a Refrigerator is associated with a Dishwasher in 22.58% of the properties (support). This suggests a strong association, with a confidence level of 90.56%. In simpler terms, properties with both a Patio/Deck and a Refrigerator are highly likely (over 90%) to also have a Dishwasher. The lift value of 1.79 indicates that such properties are nearly twice as likely to have a Dishwasher compared to the overall dataset. Additionally, the conviction of 5.23 reinforces the strength of this rule, suggesting a low chance of false positives.

Similarly, properties with Cable or Satellite TV are 22.41% more likely to have a Refrigerator (support). The confidence level for this rule is 86.52%, signifying that over 86% of properties with Cable or Satellite also include a Refrigerator. The lift value of 1.79 mirrors the previous rule, suggesting properties with Cable or Satellite are nearly twice as likely to have a Refrigerator compared to the general pool. A conviction of 3.83 further strengthens this association.

A Patio/Deck and a Dishwasher together are indicative of a Refrigerator in 22.58% of the properties (support). The confidence level is 86.10%, implying that among properties with both a Patio/Deck and a Dishwasher, 86.1% also have a Refrigerator. The lift value of 1.78 suggests a strong positive association, with properties featuring both amenities being nearly 1.78 times more likely to include a Dishwasher compared to the overall dataset. A conviction of 3.71 further supports this relationship.

The final rule highlights the association between a Dishwasher and Parking with a Refrigerator. This combination applies to 26.46% of the properties (support). With a confidence level of 82.67%, properties featuring both a Dishwasher and Parking are highly likely (over 82%) to also include a Refrigerator. The lift value of 1.71 indicates these properties are 1.71 times more likely to have a Dishwasher compared to the general dataset. A conviction of 2.98 reinforces this association.

E. Key Word Identification using Text Mining

The analysis of Figure 2 unveiled the prevalent keywords in both the advertisement body and header, depicted in the word cloud. These keywords highlight various aspects of rental properties, including amenities like fitness facilities, onsite laundry, deck, patio, balcony, and sheltered parking bus line. This underscores the importance of emphasizing key features within the advertising body for effective property promotion.

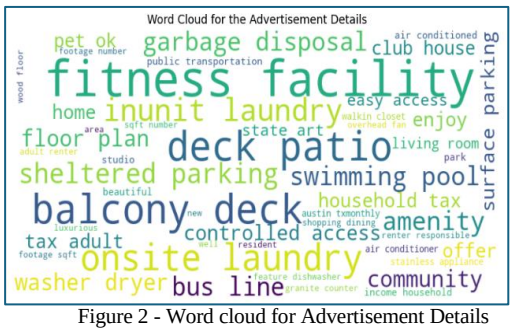

Figure 2 - Word cloud for Advertisement Details

It is important to note, however, that other potentially significant details, such as the specific rental price or pet policies, may not be as prominently featured in the titles and bodies of these advertisements.

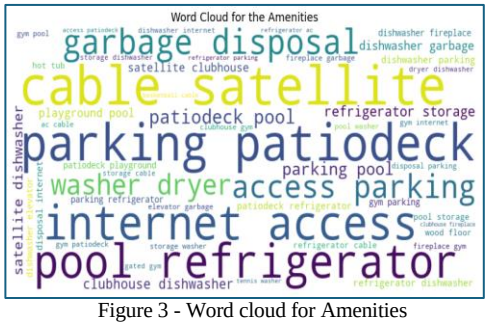

Figure 3 - Word cloud for Amenities

Furthermore, a separate analysis of Figure 3, focused on amenities, revealed the most highlighted facilities in the word cloud. These amenities included in-unit cable and satellite, internet access, refrigerator, washer and dryer, outdoor features such as parking, patio decks garbage disposal and pools.

V. CONCLUSION

A comprehensive analysis of the U.S. rental market has been provided in this research, focusing on the prediction of rental prices and the identification of influential factors associated with rental property advertisements.

Data was utilized to uncover key features affecting rental prices, such as property size and location. It was found that factors like the number of bedrooms and bathrooms, pet-friendliness, and the source of advertisement significantly influence rental prices.

Machine learning models were applied and were able to offer superior performance in predicting rental prices with an accuracy of 82%. Additionally, association rule learning was employed to reveal common patterns among property features, providing valuable insights for landlords to optimize their advertisements and for renters to make informed decisions.

The major outcome of this study is the identification of crucial factors that impact rental prices and the effectiveness of different advertising strategies. This research contributes to a deeper understanding of the rental market dynamics and offers practical recommendations for landlords and renters. Future research could expand on these findings by exploring more granular data and incorporating additional features, potentially leading to even more precise models and strategies in the rental market.

VI. REFERENCES

[1] Joint Center for Housing Studies of Harvard University, "America's Rental Housing 2020," Harvard University, Cambridge, MA, 2020. [Online].Available:https://www.jchs.harvard.edu/sites/default/files/Harvard_JCHS_Americas_Rental_Housing_2020.pdf.

[2] UCI Machine Learning Repository, "Apartment for Rent Classified Dataset," Retrieved from UCI Dataset, [Online]. Available: https://archive.ics.uci.edu/dataset/555/apartment+for+rent+classified

[3] J. Li, "A Comprehensive Study of Factors Influencing Rental Prices in Shanghai," in Proceedings of the 2024 International Conference on Highlights in Science, Engineering and Technology (CMLAI), vol. 94, pp. 373-378, Atlantis Press, 2024.

[4] A. A. Munshi, "Comparative Analysis of Regression Techniques for Housing Price Prediction," International Journal of Data Science and Analysis, vol. 8, no. 3, pp. 110-120, 2020.

[5] G. James, D. Witten, T. Hastie, and R. Tibshirani, "An Introduction to Statistical Learning," 2nd ed., Springer, 2021. [Online]. Available: https://hastie.su.domains/ISLR2/ISLRv2_corrected_June_2023.pdf

[6] E. Hromada, "Mapping of Real Estate Prices Using Data Mining Techniques," Procedia Engineering, vol. 123, pp. 233–240, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877705815031847

[7] Y. An and M. Sun, "Text Mining Rental Listing Descriptions: What Keywords Attract Renters?" Journal of Housing Research, vol. 27, no. 2, pp. 129-146, 2018.

[8] K. A. McConnell, "Housing Market Analysis: Handling Outliers and Ensuring Data Quality," Journal of Real Estate Research, vol. 15, no. 4, pp. 511-529, 2021.