Oyo Rental Price Prediction in China Using Machine Learning Techniques

Archie BansuLongo Faculty of Business,

Business Insights and Analytics

Humber Polytechnic

Ontario, Canada

n01611006@humber.ca

Anjali Sanjay Shelke

Longo Faculty of Business,

Business Insights and Analytics

Humber Polytechnic

Ontario, Canada

n01607248@humber.ca

Farheen Salar

Longo Faculty of Business,

Business Insights and Analytics

Humber Polytechnic

Ontario, Canada

n01616777@humber.ca

Jason Thomas  
Longo Faculty of Business,

Business Insights and Analytics

Humber Polytechnic

Ontario, Canada

n01610978@humber.ca

Riya Christie

Longo Faculty of Business,

Business Insights and Analytics

Humber Polytechnic

Ontario, Canada

n01598757@students.humber.ca

Wichayaporn Patadee   
Longo Faculty of Business,

Business Insights and Analytics

Humber Polytechnic

Ontario, Canada

n01603472@humber.ca

# ***Abstract:* The hospitality industry in China is experiencing fast growth, driven by increasing domestic and international travel. Precisely predicting rental prices is important for optimizing revenue, improving customer satisfaction, and maintaining competitiveness.**

# **This machine learning BIA group 5 research exploring "OYO Rental Price Prediction in China" dataset to explore the application of machine learning techniques in solving real-world pricing challenges. This research involves data preprocessing, statistical analysis, and the development of predictive models to uncover key factors influencing rental prices. By testing and comparing the performance of at least 5 machine learning algorithms, the study aims to identify the most effective method for rental price prediction. The findings offer valuable insights for data-driven decision-making in the hospitality sector, emphasizing the role of machine learning in pricing strategies.**

# **Keywords—OYO rental prices, Predictive modeling, Machine learning**

# I. INTRODUCTION

# The hospitality industry in China is booming, caused by a growing number of tourists and business travelers. China’s hospitality industry has seen significant growth, with companies like OYO expanding to over 10,000 hotels across 337 cities, showcasing the sector’s rapid evolution (Outlook Travel, n.d.)

# With more competition among hotels and rental services, having a smart pricing strategy is very important to remain competitive. Hospitality service has to hit the sweet spot between price that is attractive to consumers and profitable margin to keep the business running. That is the reason why our group thinks predicting the right price will be crucial for OYO’s business in the long run.

# This project covers how machine learning can be used to tackle this challenge, focusing on rental price prediction for OYO properties in China. We use available dataset on Kaggle, the goal is to figure out the main factors that influence rental prices and build models that can predict them accurately. The work includes cleaning the data, selecting the most important variables, and visualizing patterns to ensure everything is set up for success. On top of that, several machine learning algorithms are tested and compared to find out which one delivers the best results.

# II. DATA SOURCE AND PREPROCESSING

*A. Data source*

# The data employed in this project were obtained from publicly available datasets on Kaggle, specifically the OYO Rental Price Prediction in China dataset. The first dataset name 'rating\_features' provides detailed features of the rental properties such as the amenities, availability in next 30 days, number of bedrooms and bathrooms etc. The dataset has 5834 entries across 25 properties. The second dataset named 'rental\_price' includes the rental price information for the 5834 entries respectively. Both are available in csv files.

# *B. Data processing*

# Several preprocessing steps were carried out to ensure that the data were clean and reasonable for modeling:

# Handling missing values: , Missing values in numerical columns were filled by estimating the mean or median, depending on the distribution of the data. For each numerical column, missing values (NaN) were replaced with the median of the specific column, and if there are still missing values remaining then they are filled by the mode of the column.

# Price Transformation: We transformed the price column into a numeric format (price\_numeric) by separating price into 2 columns: numeric and the dollar sign, allowing the model to process it correctly

# Certain variables, such as *has\_availability*, *host\_is\_superhost*, and *instant\_bookable*, which initially had values represented as 't' and 'f', were converted to binary values (0s and 1s) to facilitate ease of operations.

# The numerical variable (price) second dataset is cleaned by extracting the numerical price values and filling missing values by median.

# Categorical variables, like property type, bed\_type, etc., are encoded using one-hot encoding the process creates a new column for every unique value in the categories that makes them readable to models. The following are the new columns generated: Property type;(*Property\_Type\_Other,Property\_Type\_House,Property\_Type\_Apartment, Property\_Type\_Camper/RV, Property\_Type\_Condominium,Property\_Type\_Loft,Property\_Type\_Townhouse,Property\_Type\_Tent,Property\_Type\_Bed* & *Breakfast,Property\_Type\_Bungalow,Property\_Type\_Vila,Property\_Type\_Cabin,Property\_Type\_Earth House,Property\_Type\_Hut,Property\_Type\_Tipi,Property\_Type\_Boat,Property\_Type\_Chale*t), bed\_type;(*bed\_type\_Airbed,bed\_type\_Real Bed,bed\_type\_Futon,bed\_type\_Couch,bed\_type\_Pull-out Sofa), room\_type;(room\_type\_Entire home/apt,room\_type\_Private room,room\_type\_Shared room*).

# *C. Figure*

# 

Fig. 1 Heatmap correlation among variables in the dataset

# 

# Correlation heatmap unveils the relations that exist between different features and the dependent variable price numeric. Strong positive correlations exist between price numeric and features like review\_scores\_value, bedrooms, bathrooms, beds, and accommodates. This typically means that a higher-rated listing, the more extensive house, with more amenities and accommodation for guests, the price of the property would increase. In contrast, a slight negative correlation with calculated\_host\_listings\_count indicates that hosts with more listings might be more competitive and price their properties slightly lower. These findings underscore the importance of factors like property size, amenities, guest capacity, and host reputation in determining Oyo rental pricing.

# The cleaned datasets are then combined. After performing the dataset merging, the final table was a mixture of customer ratings, property attributes, price, and derived features. The dataset is further explored statistically.

# III. FEATURE SELECTION

# Feature selection plays a crucial role in the success of any machine learning study, especially for our project, OYO Rental Price Prediction in China, which involves predicting rental prices based on a variety of features. Below are key reasons why feature selection is critical in the context of this project.

Feature selection is particularly vital because of the need to handle high-cardinality variables like property\_type, categorical features like room\_type, and numerical features like review\_scores\_rating and availability\_30. Using techniques such as correlation analysis, variance thresholding, and feature importance metrics, we can ensure that the model is both efficient and effective in predicting rental prices.

By incorporating feature selection, we align the study with the goal of building a robust, interpretable, and scalable rental price prediction model that provides meaningful insights to businesses and users.

# A RandomForestRegressor model was used and trained on the dataset, and its feature\_importances\_ attribute was used to identify the most influential features for predicting the target variable (price\_numeric).

# **C. Gradient Boosting**

# To enhance prediction accuracy further, Gradient Boosting was employed. The iterative nature of this model allowed each tree to correct errors from previous trees, optimizing for a reduction in prediction loss. Specifically, **XGBoost** was used as a variant of Gradient Boosting, chosen for its computational efficiency and ability to handle intricate datasets with missing or imbalanced values. While slightly slower to train than Random Forest, XGBoost delivered higher accuracy for this dataset.

# **D. Support Vector Regression (SVR)**

# SVR was used to model housing prices by fitting a hyperplane within a tolerance margin, effectively balancing complexity and prediction precision. Given the dataset's size and potential outliers, the kernel functions in SVR allowed it to capture subtle patterns in the data. This model proved particularly beneficial in handling data points that deviated significantly from the general trend.

# **E. K-Means**

# K-Means clustering was applied during the exploratory phase of the project to segment the dataset into clusters based on similarities in key features such as price, location, and property type. This unsupervised approach provided insights into natural groupings within the data, which informed feature engineering and allowed us to better understand market segmentation.

# The following are the measures used for evaluating the models:

# The R² value is commonly used in statistical modeling to determine the relationships between variables. It defines the amount of variance in the dependent variable that can be predicted from the independent variables. It shows its audience how well the model is able to predict the observations, with varying values between 0 and 1. A higher R² indicates a better fit, meaning the model captures a greater portion of the variability in the data.

# Mean Absolute Error, MAE, indicates the average amount of absolute deviation predicted against reality. It simply measures how big the errors from a model's predictions are without regard to their directions. The smaller the MAE, or absolute number, the better a model might be predicted and how closely they match with that model's predictions against reality.

# Mean Squared Error is the term for a particular condition for measuring deviation concerning prediction and actual values. That is, find the average of squared errors. It means errors are being squared, and in this way, a higher penalty is imposed on bigger deviations; thus it becomes sensitive to outliers. The lower MSE hints at better accuracy in predictions made by a model.

# Important features: The most important features identified were: bathrooms, bedrooms, and latitude (North). They're higher importance scores, meaning they are strongly indicative of the price.

# Low importance features: has\_availability, and rare property types like property\_type\_Hut and property\_type\_Boat are low importance features meaning they don’t significantly affect the price prediction.Likewise, rare property types such as property\_type\_Hut, property\_type\_Boat, etc., have almost no importance within the model, meaning they do not significantly influence price prediction.

# Categorical Variables: Property Type, Room Type: Features such as property\_type\_\* and room\_type\_\* are important but show moderate to low importance. This indicates that while these categories can influence rental price, the weight of their influence is lower as compared to numerical features such as bathrooms or bedrooms. These categorical variables may add value to the differentiation of various property types or configurations but do not on their own play as big a role in determining rental prices.

# III. MODELS AND METHODOLOGY

# This project employs various supervised machine learning models to predict housing prices based on the input features discussed earlier. The selected models, including Linear Regression, Random Forest, Gradient Boosting, and Support Vector Regression (SVR), were chosen for their ability to handle different data complexities and provide accurate predictions. Additionally, K-Means clustering was used for exploratory data analysis to identify natural groupings within the dataset.

# **A. Linear Regression**

# Linear Regression was utilized as a baseline model to predict housing prices. By assuming a linear relationship between features and the target variable, the model aimed to minimize the sum of squared differences between actual and predicted prices. This simple and interpretable model provided initial insights into the importance of individual features, helping to establish foundational expectations for prediction accuracy.

# **B. Random Forest**

# Random Forest was implemented to improve accuracy and handle potential overfitting issues observed in simpler models. By constructing multiple decision trees on different subsets of the data and averaging their predictions, the model effectively captured complex, non-linear relationships between features and housing prices. This approach was particularly useful in dealing with noise and ensuring robust predictions across diverse data points.

# Silhouette Score (Model: K-Means)

context. This indicates that although it is great, it tends to be minorly inaccurate when implicitly compared to Random Forest.

# Support Vector Regression, or SVR, became the regression case of Support Vector Machines, and as is the case with most forms of predictive analysis, it predicted the least amount of variance, with an R² of 0.3171. Given the relatively high MAE and MSE values, it is associated with much RU for this model; hence, these two parameters are apparently suggesting that the SVR experience is mostly not successful in predicting the target variable, and this could be due to some characteristics of the data or hyperparameter choices.

# K-Means clustering algorithm evaluated through the Silhouette Score, which indicates the closeness within clusters and distance from other clusters. The low score of 0.1233 indicates poorly defined clusters, suggesting that the data may not have a clear underlying structure suitable for this clustering approach or that preprocessing and hyperparameter tuning could improve results.

# Summary of the results:

# Linear Regression: R² = 0.4263, MAE = 161.9129, MSE = 101930.7935. Moderate performance with significant error.

# Random Forest: R² = 0.5290, MAE = 125.5675, MSE = 83670.3582. Best performance overall, with lower error and higher R².

# Gradient Boosting: R² = 0.5216, MAE = 127.3583, MSE = 84989.9396. Like Random Forest, slightly lower R² and higher error.

# SVR: R² = 0.3171, MAE = 144.0218, MSE =121317.8907. Worst performance with the highest error.

# K-Means (Clustering): Silhouette Score = 0.1233. Low clustering quality, indicating poor cluster separation.

# Conclusion:

# Best Regression Model: Random Forest is the best choice for regression tasks based on its higher R² (0.5235), lower MAE (125.93), and lower MSE (84662.57), indicating more accurate predictions than the other models.

# Clustering: K-Means is not effective here due to a low silhouette score, suggesting it’s not suitable for this dataset.

# The silhouette score entails assessing how comparable a point is with respect to its own cluster (cohesion) and other groups (separation); thus, higher values are indicative of well-separated clusters and having well-matched points to their own clusters. The metric thus evaluates the process of clustering and adds value to the meaningful groupings within data.

# *F. Figure*

# Based on an initial analysis, the tested models produced the following results:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | R² | MAE | MSE |
| Linear regressions | 0.4263 | 161.9129 | 101930.7935 |
| Random Forest | 0.5290 | 125.5675 | 83670.3582 |
| Gradient Boosting | 0.5216 | 127.3583 | 84989.9396 |
| SVR | 0.3171 | 144.0218 | 121317.8907 |

Fig. 2. Table result of the test models

Overall, it can be suggested that the relationship between dependent and independent variables was not satisfactorily defined by Linear Regression. This simple algorithm produced an R² of 0.4263, meaning that almost 42.63% of the variance of the dependent variable is accounted for by this model. The predictive capacity deteriorated because of the high MAE and MSE.

Random Forest outperformed Linear Regression with an R² of 0.5290. The lower MAE and MSE values further highlight its predictive accuracy, making it the most effective model among those tested for regression tasks.

Gradient Boosting, which iteratively seeks to lessen the errors caused by previous predictors, produced predictions that neared those of Random Forest. Its slightly lower R² of 0.5216 and higher MAE and MSE suggest that while it is highly capable, it is marginally less accurate than Random Forest in this

# IV. VISUALIZATIONS AND ANALYSIS

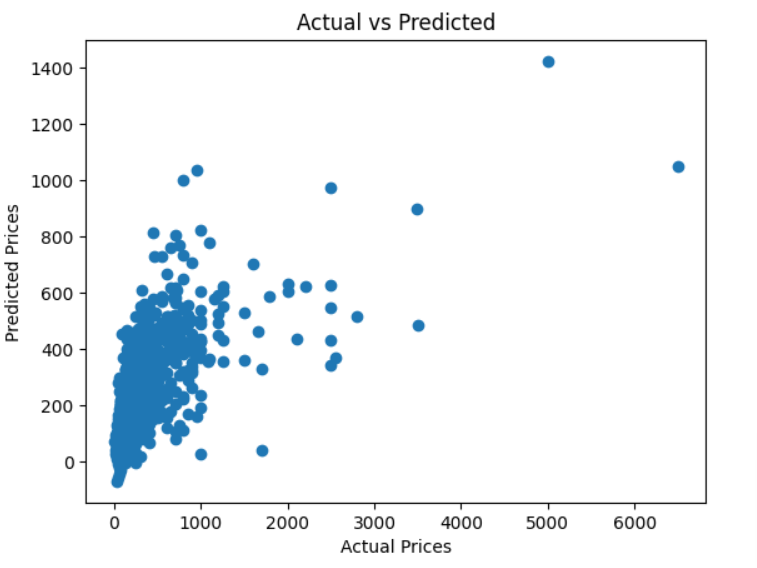


Fig. 3. Scatterplot between actual and predicted prices

# The scatter plot depicts a moderate positive correlation between actual and predicted prices. That means that model’s prediction lies pretty much along the actual price, but there still exists variability within the predictions. The density of points along the diagonal line hence indicates the model’s accuracy towards lower prices rather than the higher prices.

# 

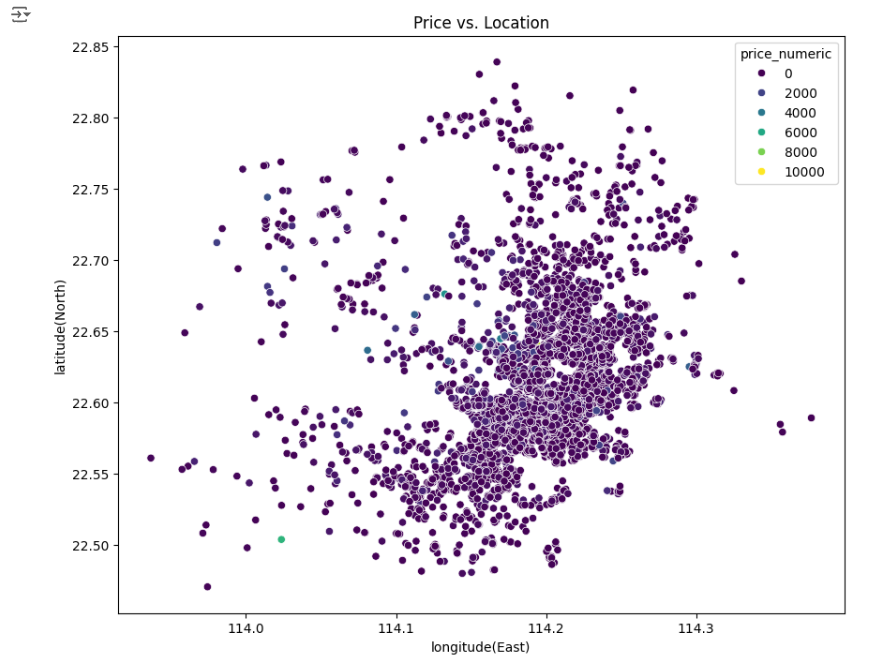


Fig. 4. Scatterplot location and the price of OYO listings

# The scatter plot depicts the relationship between the location and the price of OYO listings. The color intensity represents the price range, with darker shades indicating higher prices. The plot clearly shows the clustering in the center, suggesting that most of the listings tend to be quite expensive. There are also some outliers, where prices have gone beyond the usual limits, which may be indicating unique or high-demand locations.

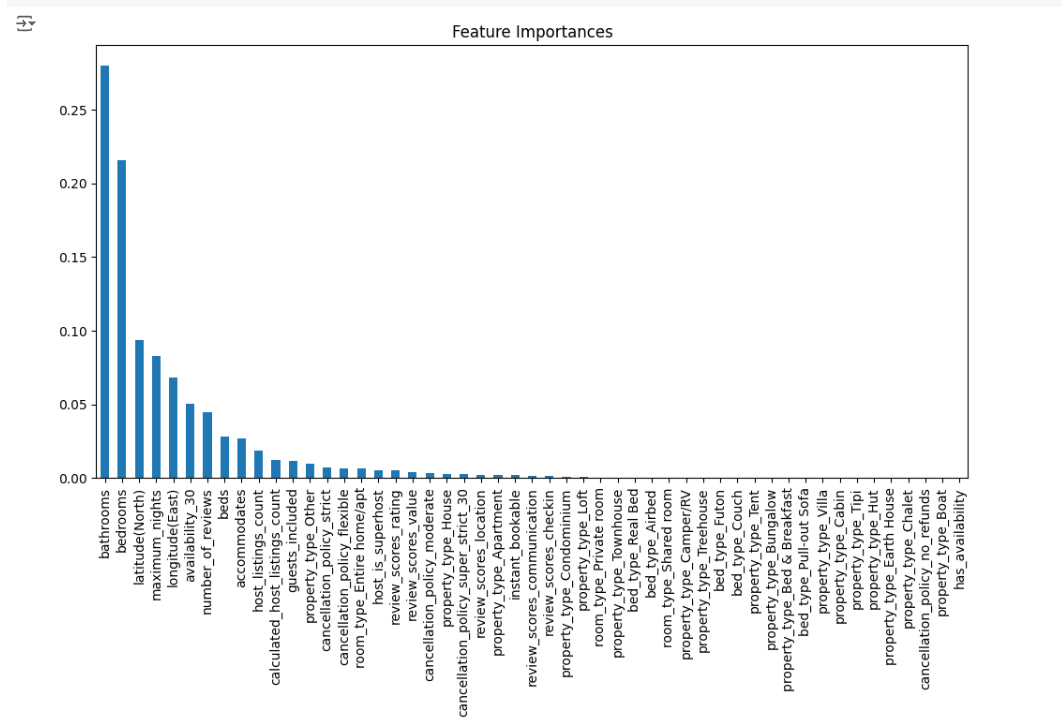


Fig. 4. Bar chart rate important feature for price prediction

The bar chart presents feature importance, that is, the contribution of every feature to the predictive modeling of the target (price). As mentioned earlier, the features bathrooms, bedrooms, and latitude have the highest importance and thus are the most influential in determining the price. Other lower important features, such as different property types and cancellation policies, have less impact on the predicted price. This analysis helps identify key price drivers in OYO pricing for optimized listing strategy use. Below are the visualizations depicting relationships between the other variables.

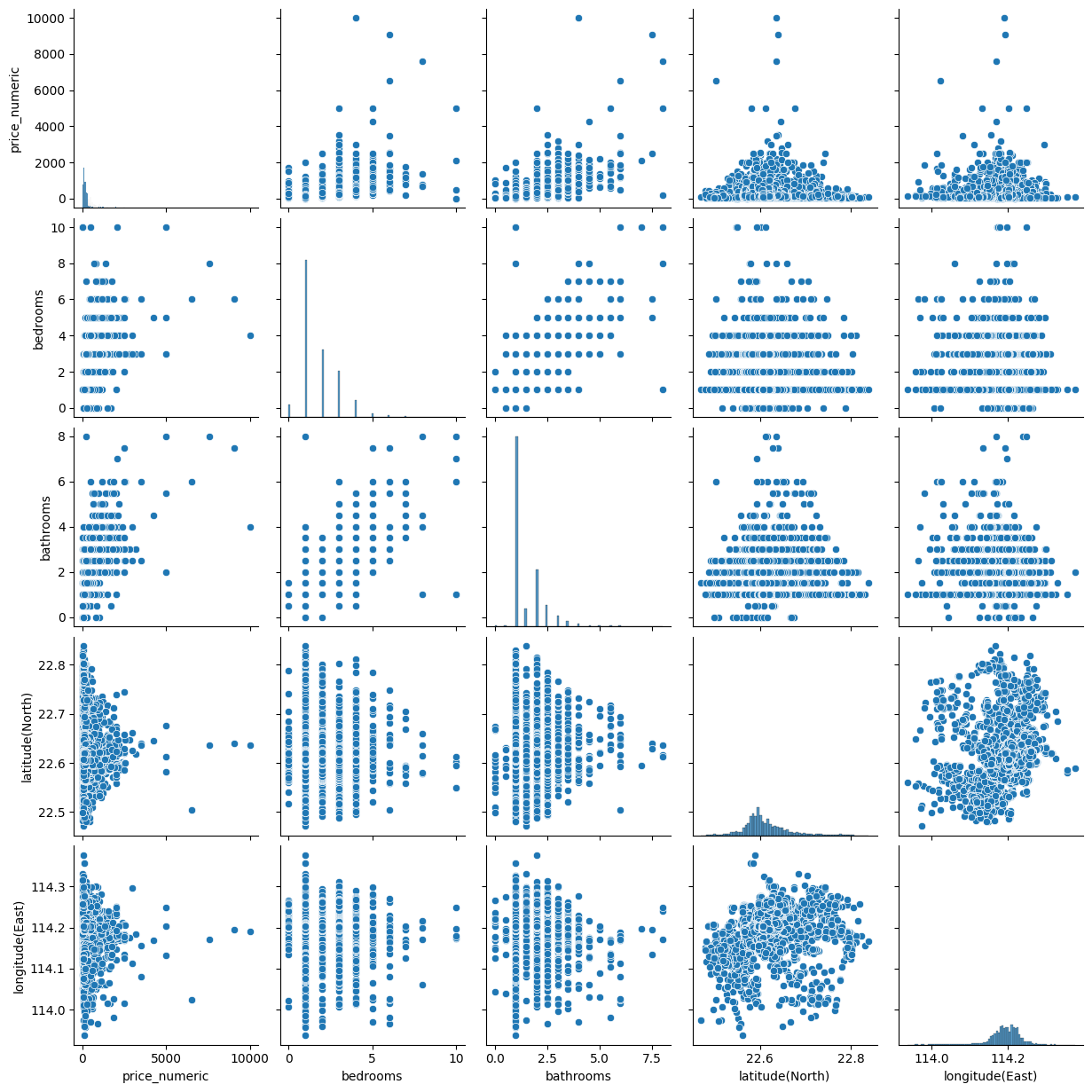


Fig. 5. Scatterplot relationship between important features rental price, bathrooms, bedrooms, and latitude

V. CONCLUSION

Best Regression Model: Random Forest is the best choice for regression tasks based on its higher R² (0.5235), lower MAE (125.93), and lower MSE (84662.57), indicating more accurate predictions than the other models.

Clustering: K-Means is not effective here due to a low silhouette score, suggesting it’s not suitable for this dataset.

Final Selection: Random Forest is the best model for this task, delivering the most accurate predictions.

VI. REFERENCES

[1] Outlook Travel, "China’s sleeping giant," *Outlook Travel Magazine*. [Online]. Available: <https://www.outlooktravelmag.com/features/article/7-chinas-sleeping-giant#:~:text=Having%20won%20the%20cola%20war,10%2C000%20hotels%20in%20337%20cities>. Accessed: Dec. 9, 2024.

[2] R. Maurya, "OYO rental price prediction in China," *Kaggle*. [Online]. Available: <https://www.kaggle.com/datasets/ramjasmaurya/oyo-rental-price-prediction-in-china?select=rating_features.csv>. Accessed: Dec. 9, 2024.

[3] *Linear Regression in Python.* [Online]. Available: <https://www.beoptimized.be/pdf/LinearRegressioninPython.pdf>. Accessed: Dec. 12, 2024.

[4] M. A. Al-Ghaili, M. M. Zain, and A. H. Hashim, “An overview of machine learning algorithms in healthcare,” in *2021 IEEE International Conference on Computing, Electronics & Communications Engineering (iCCECE)*, Aug. 2021, pp. 108–113. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/9543260>.

[5] L. Coelho, *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, 2nd ed. Sebastopol, CA: O'Reilly Media, 2019. [Online]. Available: <https://books.google.ca/books?hl=en&lr=&id=28t1DwAAQBAJ&oi=fnd&pg=PP1&dq=random+forest+python+sklearn&ots=GbEkQ6rK_Z&sig=j98ftu0pwQ5Vo5MaH7xooCq9pX4&redir_esc=y#v=onepage&q=random%20forest%20python%20sklearn&f=false>.

[6] G. N. Pendleton, *Python for Data Science and Machine Learning*, 3rd ed. New York, NY: Springer, 2021. [Online]. Available: <https://books.google.ca/books?hl=en&lr=&id=2tcDEAAAQBAJ&oi=fnd&pg=PP1&dq=gradient+boosting+algorithm+python&ots=s5tODkojkK&sig=j8p68HLophLqR3UI9udfEgz7r9U&redir_esc=y#v=onepage&q=gradient%20boosting%20algorithm%20python&f=false>.

[7] R. Bedre, “Support vector machine: A powerful classification algorithm in machine learning,” *Renesh Bedre Blog*. [Online]. Available: <https://www.reneshbedre.com/blog/support-vector-machine.html>. Accessed: Dec. 12, 2024.

# 