Analyzing Apartment Rental Classified Data: EDA and Modeling Report

Name: Tom Thomas
Student ID: 22008590

Introduction

This study compares and evaluates the efficiency of two machine learning techniques as they are applied to the modelling of apartment rental prices: random forest classification and linear regression. Key characteristics for rental apartments that are listed include square footage, location, number of bedrooms, and bathrooms. The dataset was compiled from apartments for rent classified 10K data. Evaluating these methods' ability to predict and categorize rental values according to various property features is the purpose.

Data Preprocessing

The apartment rent dataset undergoes several preprocessing steps to get it ready for efficient analysis and modelling. Initially, only columns that were relevant to the rental price modelling were kept, including square footage, number of bathrooms, bedrooms, and location features. Unwanted noise was eliminated from unnecessary data. To provide clean, complete data for models to be trained on, missing values were next found and eliminated. Since they would affect the distribution, outliers in the rental price column exceeding $5,100 per month were removed. To make categorical textual data understandable to machine learning models, one-hot encoding was used to encode them into numeric variables, such as the city name, state, and price type. In the end, the preprocessed data was divided 70-30 to form training and test sets.

Figure 1

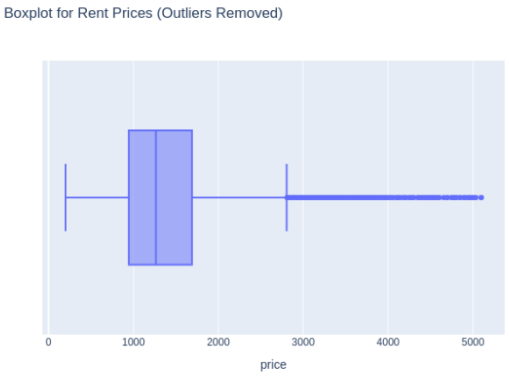

Figure 2

Correlation

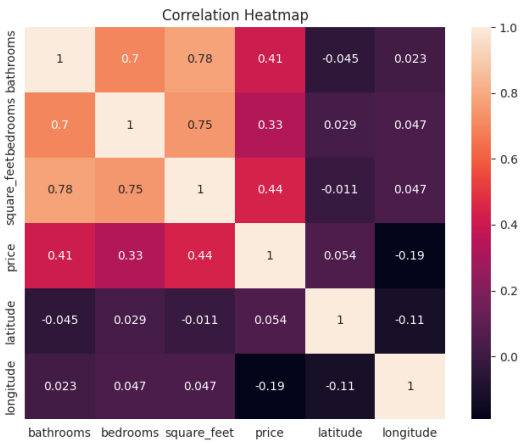

Figure 3

This heatmap was created to display how different features in the dataset on apartment rentals relate to one another, such as the number of bedrooms and bathrooms. This visual guide, acting like a map for studying the data, helped recognize patterns and understand the connections among different features more easily.

Linear Regression

For the linear Regression, Accuracy was evaluated using metrics including Mean Absolute Error = 22816531608895.54, Root Mean Squared Error = 109333964793262.19, and Mean Squared Error = 1.20. The high error scores show that the model struggled to calculate prices accurately. Plots comparing the actual and predicted prices and residuals (Figure 1) showed significant variations, indicating a poor fit. The model's poor accuracy is due to its oversimplified linear assumption, which fails to account for the complex, non-linear correlations present in the apartment pricing data. To accurately predict such intricacies, more complicated models are needed. So used the Support Vector Regression model for Accuracy and the values are Mean Absolute Error = 0.3955, Mean Squared Error = 0.3810 and Root Mean Squared Error = 0.6172.

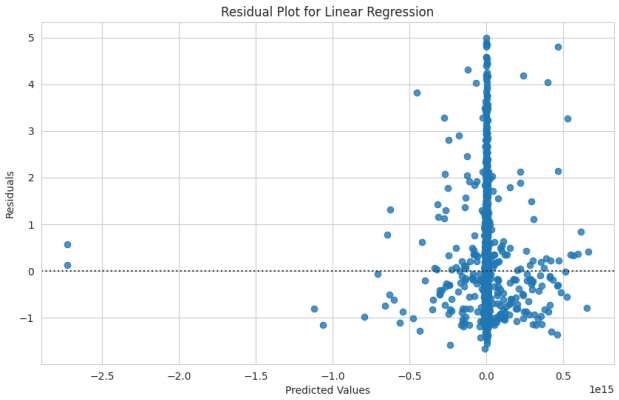

Figure 4

Random Forest Classification

The price was categorized as low, medium, or high based on value ranges. The random forest classifier was trained on features like bedrooms, bathrooms, location etc. to predict price category. A 5-fold cross-validation grid search was used to find the optimal model hyperparameters. The confusion matrix for classification (Figure 3) shows good performance in predicting the high and low-price categories. However, the medium-price category has poor metrics likely due to insufficient samples.

Accuracy: 0.85

	precision	recall	f1-score	support
High	0.89	0.91	0.90	1376
Low	0.00	0.00	0.00	16
Medium	0.75	0.72	0.73	568
accuracy			0.85	1960
macro avg	0.54	0.54	0.54	1960
weighted avg	0.84	0.85	0.84	1960

Table 1

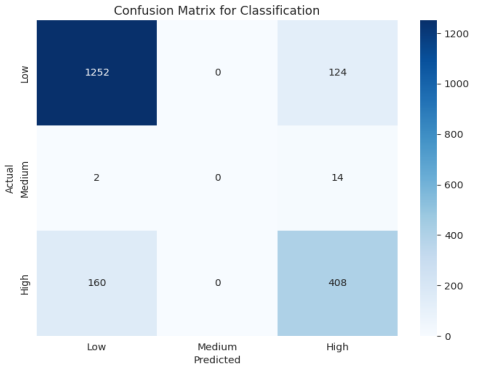

Figure 5

Conclusion

In conclusion, random forest classification and linear regression both performed well in their respective tasks. In addition to providing insightful information about the relationship between properties and rental costs, linear regression was exceptionally good at forecasting numerical rent values. However, random forest classification proved to be effective in classifying rent prices into different groups. The particular goals of the analysis and the unique features of the target variable will determine which of these approaches is best.

Reference

archive.ics.uci.edu. (n.d.). UCI Machine Learning Repository. [online] Available at: https://archive.ics.uci.edu/dataset/555/apartment+for+rent+classified.

Pang-Ning Tan (2020). Introduction to data mining: international edition. Harlow: Pearson Education.