PENTECOST UNIVERSITY

Student Name: Emmanuel Quartey

Student ID: PUIT/22210083

Course: Artificial Intelligence and Expert Systems

Date: 3rd January, 2026

Machine Learning Project Report

Exploratory Data Analysis and Machine Learning on the MinoAI Dataset

1. Introduction

The increasing availability of data in modern information systems has positioned machine learning as a critical tool for extracting insights and supporting data-driven decision-making. However, datasets obtained from real-world sources are often incomplete, inconsistent, and noisy. These characteristics necessitate careful data exploration, cleaning, and preprocessing before reliable analytical or predictive outcomes can be achieved.

This project applies exploratory data analysis (EDA), data cleaning, feature engineering, and supervised machine learning techniques to the MinoAI dataset. The primary aim is to demonstrate a structured and methodical machine learning workflow, beginning with raw data inspection and concluding with model evaluation. Emphasis is placed on justifying analytical decisions and clearly explaining the reasoning behind methodological choices.

2. Dataset Description

The MinoAI dataset consists of structured records representing multiple attributes related to listings. It contains a mixture of numerical and categorical variables, making it suitable for both exploratory analysis and predictive modeling. The dataset comprises approximately 48,000 observations and 16 original features.

As provided, the dataset reflects realistic data conditions, including missing values and potential outliers, thereby offering an appropriate foundation for real-world machine learning experimentation.

3. Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to gain an initial understanding of the dataset's structure, distributions, and variable relationships. Descriptive statistics such as the mean, median, standard deviation, and range were computed for numerical features.

Visual exploration using histograms and boxplots helped identify skewness and extreme values, while correlation matrices were used to examine relationships between numerical variables. Insights obtained during this phase informed subsequent data cleaning and modeling decisions.

4. Data Quality Issues and Data Cleaning

Several data quality issues were identified, including missing values, inconsistencies, duplicate records, and outliers. Logical imputation strategies were applied where appropriate, and features with limited analytical relevance were removed.

Outliers were detected using the Interquartile Range (IQR) method and treated to reduce their influence on model performance. Duplicate records were removed to improve data integrity and ensure unbiased learning.

5. Feature Engineering and Transformation

Feature engineering was applied to prepare the dataset for machine learning algorithms. Categorical variables were transformed using one-hot encoding to ensure numerical compatibility. Irrelevant or redundant features were removed to reduce noise and improve model efficiency.

These transformations resulted in a structured dataset suitable for supervised learning.

6. Machine Learning Approach

The analytical task was framed as a supervised regression problem, with price selected as the target variable. A Random Forest Regressor was selected due to its robustness, ability to model non-linear relationships, and resistance to overfitting compared to simpler linear models.

The dataset was divided into training and testing subsets to allow objective evaluation on unseen data.

7. Model Evaluation

Model performance was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the R^2 score. These metrics provided complementary perspectives on prediction accuracy, error magnitude, and explanatory power.

8. Results and Discussion

The trained model successfully captured important pricing patterns influenced by factors such as room type, location, and availability. However, some variance remained unexplained, suggesting that additional external factors such as seasonal demand or listing quality were not captured in the dataset.

These findings highlight both the strengths and limitations of the chosen modeling approach.

9. Conclusion

This project demonstrated a complete machine learning pipeline, from exploratory data analysis through data cleaning, feature engineering, model training, and evaluation. The structured approach ensured methodological transparency and reliable interpretation of results.

Future work could include incorporating additional features, performing hyperparameter tuning, or applying alternative machine learning models to improve predictive performance.

10. Tools and Technologies Used

- Python
- Pandas
- NumPy
- Matplotlib
- Seaborn
- Scikit-learn
- Jupyter Notebook.

11. Use of Artificial Intelligence Assistance

Portions of this project were completed with the assistance of an artificial intelligence tool (ChatGPT). The AI was used to provide guidance on code structure, clarification of machine learning concepts, and assistance in improving the academic clarity and presentation of the report. All analytical decisions, code execution, and interpretations were reviewed, validated, and finalized by the author.

12. Academic Integrity Statement

All work presented in this project is original and complies with academic integrity guidelines. All tools, libraries, and assistance used have been appropriately acknowledged.