DSC680-T301 Applied Data Science (2263-1)

Mahad Farah

Rent price prediction based on home features

1. Introduction / Background

Housing affordability has become one of the most important economic challenges in modern urban environments. Rent prices vary widely depending on multiple factors, including the size of a property, the number of bedrooms and bathrooms, the furnishing level, and especially location. While basic exploratory analysis can reveal correlations, it does not fully explain the complexity of rental pricing or provide actionable predictive insights.

The purpose of this project is to develop an end-to-end data science solution that predicts monthly rent prices based on housing characteristics. By applying both interpretable regression methods and advanced machine learning models, this analysis aims to quantify which features most strongly influence rent while capturing nonlinear patterns in housing markets.

Ultimately, this work is intended to support renters, housing analysts, and policymakers by providing a data-driven understanding of rental cost drivers.

2. Dataset Description

This project uses the House Rent Prediction Dataset from Kaggle, which contains approximately 4,700 property-level rental listings. Each listing includes information such as monthly rent, property size, number of bedrooms (BHK), bathrooms, furnishing status, floor details, tenant preference, and location attributes, including city and locality

The dataset provides a mix of numerical and categorical features, making it well-suited for predictive modeling. Because rent is influenced by both structural attributes and geographic context, this dataset allows for a comprehensive modeling approach that incorporates multiple dimensions of housing cost variation.

3. Problem Statement and Objectives

The main problem addressed in this project is:

Can rent prices be accurately predicted using property characteristics and location information, and what are the most important drivers of rental cost?

Key objectives include:

- Building predictive models that estimate monthly rent with strong accuracy
- Comparing baseline regression methods with more advanced ensemble approaches
- Evaluating nonlinear relationships between features and rent outcomes
- Interpreting model results to explain rent drivers in a clear, jargon-free way

4. Research Questions

- How accurately can rent be predicted from property and location features?
- Do rent drivers behave nonlinearly, such as diminishing returns to property size?
- How much predictive power comes from city and locality compared to structural housing features?
- Which modeling approach generalizes best under cross-validation and avoids overfitting?

5. Methodology and Modeling Approach

5.1 Baseline Models

- Baseline methods will provide interpretability and serve as benchmarks:
    - Linear Regression
    - Log-transformed Rent Regression

These models help establish whether rent can be explained through simple linear relationships.

5.2 Regularized Regression

- To improve stability and reduce multicollinearity, the project will incorporate:
    - Ridge Regression
    - Lasso Regression
    - Elastic Net Regression

These approaches support feature selection and improve generalization.

5.3 Advanced Ensemble Models

- To capture nonlinear patterns and feature interactions, advanced machine learning models will be applied:
    - Random Forest Regression
    - Gradient Boosting Models (XGBoost)

These models are expected to outperform linear baselines by learning complex relationships between housing features and rent price.

5.4 Evaluation Strategy

- Models will be evaluated using:
    - K-Fold Cross Validation
    - Hyperparameter tuning
    - Metrics such as MAE, RMSE, and R^2
    - Error analysis by city and property type

This ensures that models do not simply overfit the dataset but generalize well to unseen rental listings.

6. Planned Visualizations

- The final white paper will include at least 3–5 key illustrations, such as:
    - Distribution of rent prices (histogram/log scale)
    - Rent vs size scatterplot with nonlinear trend line
    - Average rent comparison across cities (bar chart)
    - Feature importance plot from Random Forest/XGBoost
    - SHAP summary plot showing top rent drivers

These visuals will support interpretation and communicate findings clearly to non-technical audiences.

7. Ethical Considerations

- Predictive rent modeling raises important ethical concerns.
    - Fairness risk: Location features may reflect income segregation and reinforce existing inequalities
    - Misuse risk: Such models could potentially be used to justify unnecessary rent increases, worsening affordability

To mitigate these risks, results will be framed as descriptive and analytical rather than prescriptive tools for rent inflation.

8. Target Audience

- This analysis is designed for:
    - Renters seeking transparency in housing costs
    - Real estate analysts evaluating pricing patterns
    - Policymakers concerned with affordability

The white paper will avoid unnecessary technical words and focus on interpretable insights.

9. Appendix

- The appendix will include:
    - Data dictionary of all variables
    - Preprocessing steps (encoding, missing value handling)
    - Model hyperparameters and tuning ranges
    - Full evaluation metrics table

10. Questions an Audience Would Ask

- These will be answered in Milestone 4:
    - How accurate is the best rent prediction model?
    - Which features matter most in determining rent?
    - Does furnishing status significantly increase rent?
    - How much does city/locality influence rent compared to size?
    - Can the model generalize to new cities not in the dataset?
    - What is the typical prediction error in currency terms?
    - Are expensive listings harder to predict?
    - Could landlords misuse this model to raise rent unfairly?
    - What limitations exist in the dataset?
    - How could the model be improved with additional data?

References (APA Style)

Banerjee, S. (2022). House Rent Prediction Dataset [Data set]. Kaggle. https://www.kaggle.com/datasets/iamsouravbanerjee/house-rent-prediction-dataset

Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.