Predicting Housing Prices using Machine Learning Algorithms

Ivan Martić (SV77/2020), Ivan Đukanović (SV79/2020)

1. Motivation

The motivation behind this project is to explore the application of machine learning algorithms in predicting housing prices. Accurate prediction of housing prices is crucial for various stakeholders such as real estate agents, buyers, and sellers. By developing a reliable predictive model, we aim to assist these stakeholders in making informed decisions regarding buying, selling, or investing in real estate properties.

2. Research questions

The specific problem we aim to address with our work is to predict housing prices based on various features such as location, size, amenities, and other relevant factors. We will utilise the "Ames Housing Dataset," which contains detailed information about residential properties in Ames, Iowa. This dataset includes features like the number of bedrooms, bathrooms, square footage, neighbourhood, and sale price.

3. Related work

Previous studies in the field of real estate prediction have employed various machine learning techniques to address similar problems. These techniques range from traditional linear regression models to more advanced ensemble methods like random forests and gradient boosting. Researchers have explored feature engineering, model selection, and hyperparameter tuning to improve predictive performance.

4. Methodology

Our approach involves preprocessing the dataset to handle missing values, encode categorical variables, and scale numerical features. We will then split the data into training and testing sets. We will experiment with different machine learning algorithms, including linear regression, random forest, gradient boosting and k-nearest neighbours (KNN). Hyperparameter optimization techniques such as grid search or randomised search will be employed to fine-tune the models.

5. Discussion

For experimentation, we will evaluate each model using metrics such as root mean squared error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2). We will analyse the results to identify the best-performing model and examine any patterns or trends in prediction errors. Additionally, we will conduct error analysis to understand the limitations and potential areas for improvement in our predictive models.

6. References

[1] Guo, Z., & Huang, G. B. (2019). Predicting Housing Prices with Gradient Boosting Machines. *Applied Sciences, 9*(21), 4624. Full article: House price prediction with gradient boosted trees under different loss functions (tandfonline.com)

[2] Kaggle: Ames Housing Dataset. Available online: Ames Iowa Housing Data (kaggle.com)

[3] Random Forests Documentation. Available online: RandomForestRegressor — scikit-learn 1.5.0 documentation