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. For experimentation, we evaluated each model using metrics such as root mean squared error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2). Below are the results for each model:

Linear Regression:

- RMSE: 457073.8943282675
- MAE: 346352.3614427174
- R^2: -25.057411402684856

Random Forest:

- RMSE: 26481.58090216987
- MAE: 16265.72561433447
- R^2: 0.9125325931168123

Gradient Boosting:

- RMSE: 27593.661856399176
- MAE: 15967.833827041599
- R^2: 0.9050320395703464

K-Nearest Neighbours:

- RMSE: 95871.24074218614
- MAE: 61710.76075085324
- R^2: -0.14639723774068636

The Random Forest model achieved the best performance with an RMSE of 26,481.58, MAE of 16,265.73, and R^2 of 0.9125. Gradient Boosting also performed well, with an RMSE of 27,593.66, MAE of 15,967.83, and R^2 of 0.9050. The K-Nearest Neighbors model did not perform as well, with an RMSE of 95,871.24, MAE of 61,710.76, and R^2 of -0.1464, indicating that it was not able to capture the variance in the data effectively.

Linear Regression model performed poorly with an RMSE of 457,073.89, MAE of 346,352.36, and a negative R^2 of -25.0574. This suggests that the model is a poor fit for this dataset, potentially due to its inability to handle the complexity and non-linearity present in the data.

From these results, it is evident that ensemble methods such as Random Forest and Gradient Boosting are more effective for predicting housing prices in this dataset. They can capture complex patterns and interactions among features better than simple linear models and K-Nearest Neighbors.

Error analysis revealed that the ensemble methods consistently outperformed the other models in terms of predictive accuracy. The patterns in prediction errors indicate that these models are better suited for handling the variability in housing prices.

In summary, the Random Forest model emerged as the best-performing model in our experiments. Future work could involve further tuning of hyperparameters, exploring additional features, or employing more advanced techniques such as stacking or blending to enhance predictive performance.

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