House Price Prediction: A Comparison of Linear Regression and K-Nearest Neighbors

Rediet Tadesse, Funmiso Adeniyi, McKenzie Young

ACM Reference Format:
Rediet Tadesse, Funmiso Adeniyi, McKenzie Young. 2024. House Price Prediction: A Comparison of Linear Regression and K-Nearest Neighbors. In . ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 OVERVIEW

The objective of this analysis was to predict house prices based on features from a dataset of housing information. We built two models: a **Linear Regression** model and a **K-Nearest Neighbors (KNN)** model and evaluated their performance on the test set using the Mean Squared Error (MSE) and R^2 score.

Understanding the relationships between features and their influence on house prices is critical for real estate market analysis. This study aims to provide insights into significant predictors of house prices and to evaluate the suitability of two popular predictive models for this task.

The dataset used in this analysis is available on Kaggle [2], and the presentation video can be accessed here: https://shorturl.at/GgNIB.

2 CORRELATION HEATMAP

A **correlation heatmap** was plotted to visualize the relationship between all numeric variables in the dataset. Strong correlations (values close to 1 or -1) indicate a stronger relationship between the variables.

From the heatmap (Figure 1), we observed that **area**, **bathrooms**, **stories**, and **parking** had positive relationships with **price**. This finding justified their inclusion as significant features in the model.

3 FEATURE SELECTION

To enhance model performance, we used **SelectKBest** with the `f_regression` scoring function to identify the top features contributing most to predicting house price. The selected features were:

- **Area**
- **Bathrooms**
- **Stories**
- **Airconditioning**
- **Parking**

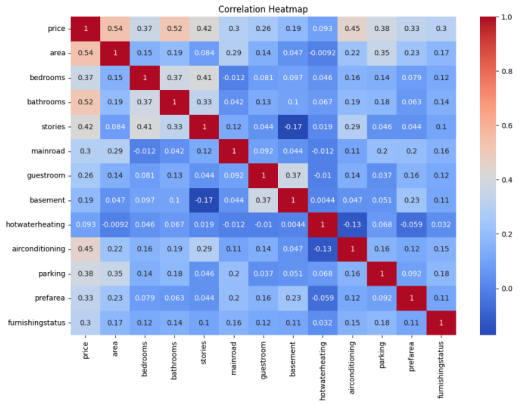

Figure 1: Correlation Heatmap of Features

These features were chosen based on their correlation with the target variable (**price**) and their importance in regression modeling. By reducing dimensionality, the model's performance was improved, and the risk of overfitting was minimized.

4 REGRESSION MODEL PERFORMANCE

4.1 Training Phase

Training Mean Squared Error (MSE): 1,217,211,760,217.0059

The training MSE indicates how well the model fits the training data. A lower value suggests a better fit.

4.2 Testing Phase

- **Test Mean Squared Error (MSE):** 2,104,079,659,093.142
- **Test R^2 Score:** 0.5837

The R^2 score, which ranges from 0 to 1, represents the proportion of variance in the target variable that is predictable from the independent variables. An R^2 score of 0.5837 indicates that the model explains approximately **58.37%** of the variance in house prices. This result demonstrates reasonable predictive power, although there is potential for improvement with more complex models or feature engineering.

5 KNN MODEL PERFORMANCE

For comparison, we also trained a **K-Nearest Neighbors (KNN)** regressor with k=5, predicting house prices based on the average of the nearest 5 neighbors:

- **Training Mean Squared Error (MSE):** 1,389,420,143,082.2935
- **Test Mean Squared Error (MSE):** 3,349,007,369,541.284
- **Test R^2 Score:** 0.3374

While the KNN model provided a simple approach to prediction, it struggled with larger datasets due to computational inefficiencies and performed poorly compared to the regression model.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Figure 2: Regression Line for Key Features

6 COMPARISON OF REGRESSION AND KNN MODELS

6.1 Key Differences

(1) **Performance Metrics:**
- The **Linear Regression model** outperformed the **KNN model** in terms of both training and test MSE, indicating better fitting and generalizability to unseen data.
- The $\mathbf{R^2}$ **score** for the regression model (0.5837) was higher than that of the KNN model (0.3374), suggesting that the regression model explained more of the variance in house prices.

(2) **Training Efficiency:**
- **Linear Regression** is computationally more efficient than **KNN**, as KNN requires finding the nearest neighbors for each prediction, which becomes computationally expensive for larger datasets.

7 CONCLUSION

The **Linear Regression model** provided more accurate and efficient predictions compared to the **KNN model**. The regression model's ability to model linear relationships in the data contributed to better performance metrics, making it the preferred choice for this dataset.

For future enhancements, exploring more complex models like **Decision Trees** or **Random Forests** could further improve predictive accuracy. Additionally, implementing feature engineering and optimizing hyperparameters for the models could yield even better results.

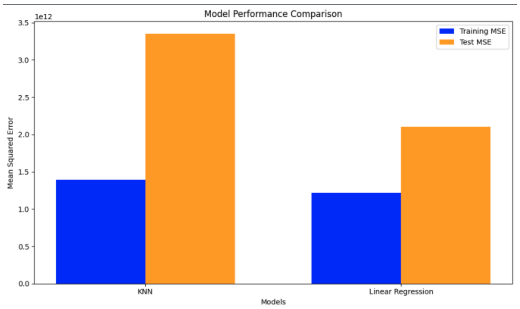

Figure 3: Comparison of Regression and KNN Performance

REFERENCES

[1] Housing Prices Dataset. https://www.kaggle.com/datasets/yasserh/housing-prices-dataset.
[2] Video Presentation. https://cdn-uploads.piazza.com/paste/kyu701o2foy4zg/c2b7dfff4f693be91913043a88ef390ee60a2befb18b14fe57e661b737dd9056/GMT20241210-224150_Recording_1920x1080.mp4.