PROJECT REPORT

House Price Prediction Using Machine Learning

Devtern

PATEL VISHRUTKUMAR DINESHBHAI

JANUARY 22, 2024

INDEX

Sr. No.	Title	Page
1	Introduction	2
2	Objective	2
3	Methods	2
3.1	Data Preparation	2
3.1.1	Data Collection	2
3.1.2	Data Preprocessing	2
3.1.3	Data Visualization of Dataset	3
3.2	Model Building	3
3.3	Model Evaluation	4
4	Model Development and Evaluation	4
4.1	Algorithm Selection	4
4.2	Training and Evaluation	4
4.3	Results	4
5	Conclusion	5
6	Future Scope	6
7	References	6

House Price Prediction Using Machine Learning

1. Introduction

In the current real estate market, determining the right price for a house is a complex task. The price of a house is influenced by a multitude of factors such as the type of dwelling, lot size, overall condition, year built, and more. Accurate price prediction is crucial for both buyers, who want to ensure they are not overpaying, and sellers, who aim to get the maximum possible price. It is also important for real estate agencies and online property listing platforms that need to provide reliable price estimates. However, due to the high variability and complexity of these factors, predicting house prices remains a challenging problem. The crucial part is the prediction of house prices based on various features to provide a stable and reliable estimate.

2. Objective

The primary goal of this project is to construct an effective machine learning model capable of accurately predicting house prices. This involves a multi-step process, including data collection, data preprocessing, model building using various machine learning algorithms, and thorough evaluation of the model's performance on a test set. By achieving this objective, we aim to contribute to the enhancement of the real estate industry and empower users with a more reliable tool for house price estimation.

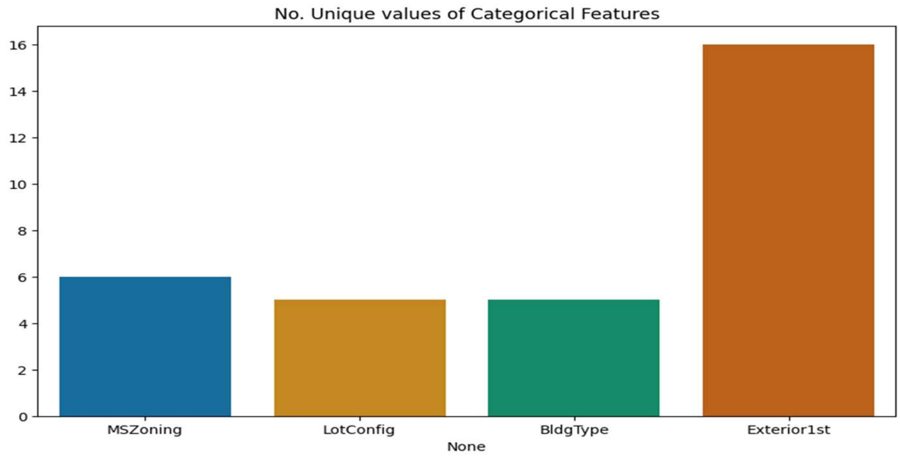
3. Methods

3.1 Data Preparation

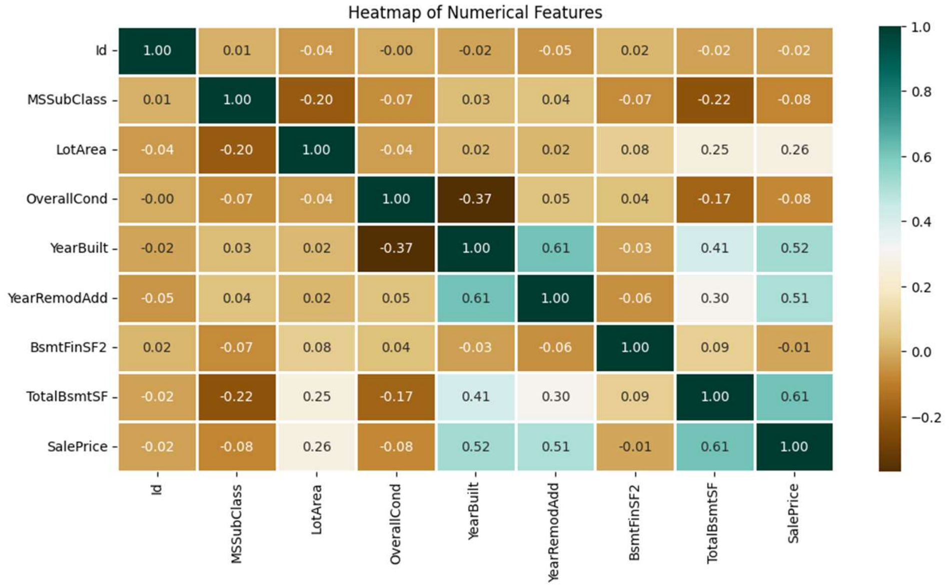
The dataset used for this study consists of historical data on house sales, including features like the type of dwelling, lot size, overall condition, year built, and other relevant factors. The following steps were undertaken to prepare the data for machine learning:

- **3.1.1 Data Collection:** Gather historical data on house sales.
- **3.1.2 Data Preprocessing:** Clean and preprocess the collected data to handle missing values, outliers, and inconsistencies. Feature engineering to extract relevant features from the data that might impact house prices.

- ### 3.1.3 Data Visualization of Dataset:



[Categorical Features of Dataset]



[Numerical Features of Dataset]

3.2 Model Building

Several machine learning algorithms are implemented including Random Forest Regressor, Decision Tree Regressor, Gradient Boosting Regressor, and K-Nearest Neighbors (KNN). These models are trained and tested on a split of the dataset to predict house prices based on historical patterns.

3.3 Model Evaluation

The model's performance is assessed using the **Mean Absolute Percentage Error (MAPE)** as the evaluation metric.

4. Model Development and Evaluation

4.1 Algorithm Selection

The following algorithms were used in the model:

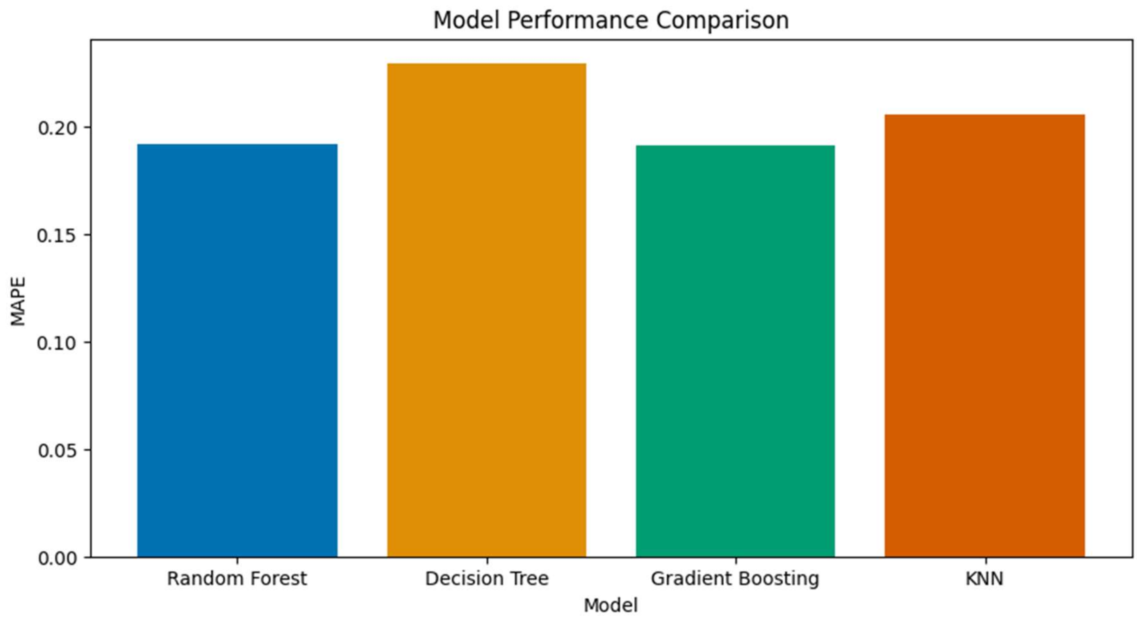
- **Random Forest Regressor**: This is a meta estimator that fits several classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.
- **Decision Tree Regressor**: This algorithm builds regression models in the form of a tree structure and breaks down our dataset into smaller subsets while at the same time an associated decision tree is incrementally developed.
- **Gradient Boosting Regressor**: This algorithm produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.
- **K-Nearest Neighbors (KNN)**: This algorithm assumes that similar things exist in close proximity and predicts the value of any given point in the dataset by averaging the values of the 'k' closest points.

4.2 Training and Evaluation

The dataset was split into a training set and a test set, with the training set used to train the model and the test set used to evaluate its performance. The models were trained to find the best parameters that minimize the difference between the predicted and actual house prices.

4.3 Results

The performance of each model was evaluated using the Mean Absolute Percentage Error (MAPE). The results were as follows:



- **Random Forest Regressor**: The Random Forest model achieved a MAPE of 19.20%.
- **Decision Tree Regressor**: The Decision Tree model achieved a MAPE of 22.96%.
- **Gradient Boosting Regressor**: The Gradient Boosting model achieved a MAPE of **19.16%.**
- **K-Nearest Neighbors (KNN)**: The KNN model achieved a MAPE of 20.58%.

5. Conclusion

In conclusion, the project ‘House Price Prediction using Machine Learning’ has demonstrated the effectiveness of various machine learning algorithms in predicting house prices. The **Gradient Boosting Regressor** model emerged as the best-performing model, with a **MAPE of 19.16%**, indicating a high level of accuracy. During the implementation of the project, several challenges were encountered, such as handling missing values and outliers in the data, and selecting the most relevant features for the prediction. However, these challenges were addressed through data preprocessing techniques and feature engineering.

6. Future Scope

The ‘House Price Prediction using Machine Learning’ project has a promising future scope. Here are some potential enhancements and expansions for the system:

- **Incorporating Additional Data Sources**: The model’s performance could potentially be improved by incorporating additional data sources. For instance, data related to the neighborhood such as proximity to schools, hospitals, and public transportation could be included. Also, macroeconomic factors like interest rates, inflation rates, and housing market trends could provide valuable context.
- **Optimizing the Algorithm for Better Performance**: There is always room for improvement in machine learning models. More advanced techniques such as hyperparameter tuning, ensemble methods, or deep learning could be explored to enhance the model’s performance.
- **Expanding the System to Cover Multiple Cities or Regions**: Currently, the model is trained on a specific dataset. However, house prices can vary greatly from one city or region to another. Expanding the system to include data from multiple cities or regions could make the model more robust and widely applicable.
- **Integration of Emerging Technologies**: Emerging technologies such as edge computing could be used to deploy the model closer to the data source, reducing latency and improving real-time prediction capabilities. Advanced machine learning techniques like reinforcement learning or neural networks could also be explored to improve the prediction accuracy.

7. References

1. Pedregosa et al., “Scikit-learn: Machine Learning in Python”, JMLR 12, pp. 2825-2830, 2011.
2. McKinney & others, “pandas: powerful Python data analysis toolkit”, 2010.
3. Hunter, J. D., “Matplotlib: Visualization with Python”, 2007.
4. Waskom M., “Seaborn: statistical data visualization”, 2021.

5. Jain, A., & Kumar, A. M., “House price prediction: a comparison of multiple linear regression and artificial neural networks”, Journal of AI and Data Mining, 2018.

6. Brownlee, J., “Machine Learning Mastery”, 2020.