Predicting House Prices - Documentation

Introduction

This documentation provides an overview of the "Predicting House Prices" project, outlining the project's objectives, data sources, methodology, and results. This project aims to create a machine learning model for predicting house prices based on various features, making it a valuable tool for real estate professionals and investors.

Project Objectives

The primary objectives of this project are as follows:

1. **Data Collection**: Obtain a dataset containing information about house prices and relevant features.

2. **Data Exploration**: Explore the dataset to understand its structure, check for missing values, and analyze the distribution of house prices and features.

3. D**ata Preprocessing**: Clean the data by handling missing values and outliers.

4. **Feature Selection**: Choose the most relevant features for predicting house prices.

5. **Model Selection**: Select an appropriate regression algorithm for building the prediction model.

6. **Model Training and Evaluation**: Train the selected model, evaluate its performance, and fine-tune hyperparameters if necessary.

7. **Comparison of Models**: Compare the performance of different regression models to identify the best-performing model.

Data Sources

For this project, we used the "Amsterdam House Price Prediction" dataset, which is available on Kaggle. The dataset contains information about houses in Amsterdam, including features such as area, longitude, latitude, and the number of rooms. House prices are provided as the target variable.

The dataset can be downloaded using the Kaggle API, as mentioned in the project README.

Methodology

Data Exploration and Preprocessing

- We performed data exploration to understand the dataset's structure and checked for any missing values.

- Missing values were handled by replacing them with the median value for each feature.

- Feature selection was based on the correlation matrix and feature importance techniques to choose relevant features for the model.

Model Selection

- The Random Forest Regression model was chosen for predicting house prices due to its robustness and accuracy in regression tasks.

Model Training and Evaluation

- The dataset was split into training and test sets.

- The Random Forest model was trained on the training data and evaluated using metrics such as Mean Squared Error (MSE) and R-squared (R2).

Comparison of Models

- The project also compared the performance of various regression models, including Linear Regression, Lasso Regression, Ridge Regression, Decision Tree, Random Forest, Gradient Boosting, and K-Nearest Neighbors.

- The best model was selected based on the R2 accuracy metric.

Results

The project achieved the following results:

- Successfully trained a Random Forest Regression model for house price prediction.

- Compared the performance of different regression models, identifying the best-performing model.

- Provided documentation to explain the entire project, data sources, data preprocessing, model selection, and evaluation metrics.

The best-performing model, which is the Random Forest Regression model, can be utilized for predicting house prices based on various input features, providing valuable insights for real estate professionals and investors.

Conclusion

The "Predicting House Prices" project demonstrates the process of building a machine learning model for predicting house prices. The project covers data collection, exploration, preprocessing, feature selection, model selection, training, and evaluation. It concludes with a comparison of different regression models to determine the most accurate model for predicting house prices. This work has practical applications in the real estate industry and can assist in making informed investment decisions.

For any further details or inquiries, please refer to the project's README or contact the project contributors.