Team Members:

Vittorio, Sheyma, Archie, Varvara and Najma

Topic:

Predicting house prices in the UK based on various features using machine learning models.

Dataset (link it if possible):

[UK House Price Index: data downloads April 2024 - GOV.UK (www.gov.uk)](https://www.gov.uk/government/statistical-data-sets/uk-house-price-index-data-downloads-april-2024)

**UK House Price Index Dataset**

* The dataset contains detailed house price information across different regions in the UK from January 1995 to April 2024.

1. Is this a regression or classification task?

This is a regression task since we are predicting a continuous variable (house prices).

1. What are your inputs? (features/columns/etc. that your model will train on)

Features will include:

* Date
* Region Name
* Sales Volume
* Detached Price
* Semi-Detached Price
* Terraced Price
* Flat Price
* New Price
* Old Price

1. What is your target? (the output that your model will predict)

The target variable is the Average Price.

1. What machine learning model(s) are you planning on using?

We are planning to use the following models:

* Linear Regression
* Random Forest Regressor
* Gradient Boosting Regressor

1. One of the criteria for the project is for you to optimize your model. What are some parameters or other things you could change about your model as you test its performance to optimize it?

Parameters to optimize:

* + - For Random Forest
      * Number of trees
      * max depth
    - For Gradient Boosting
      * Learning rate
      * number of boosting stages

We will use GridSearchCV for hyperparameter tuning and document the changes and performance in a CSV file.

1. What steps will you need to take to clean and prepare your data?

Steps include:

* Handling missing values
* Encoding categorical variables such as region names
* Normalizing numerical features
* Splitting the dataset into training and testing sets

1. Do you have any questions or concerns you think you will need help with?

Do we need to consider any specific domain knowledge in UK real estate to improve our model?

What are the best practices for dealing with highly correlated features?

Link to GitHub Repository