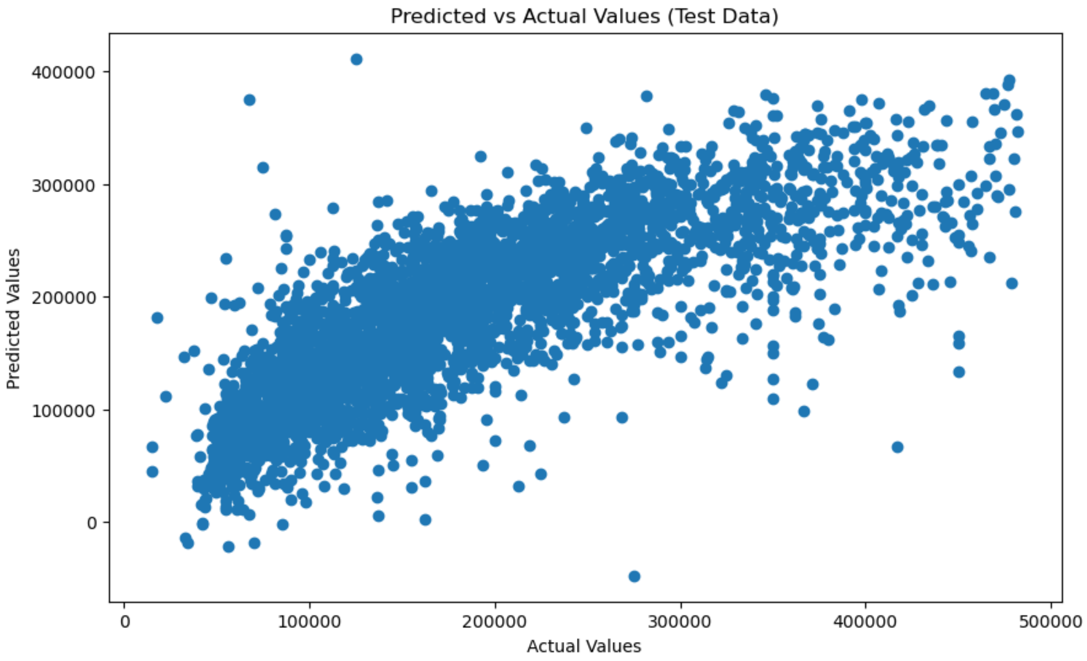
ML Project Documentation

Linear Regression

1. ***General Information on dataset:***
    a. *The name of dataset used: California Housing Data (1990).*
    b. *The number of samples used in training and testing:*
        i. Training : `14087`
        ii. Testing : `3522`
2. ***Implementation details:***
    a. *At feature extraction phase, how many features were extracted, their names, the dimension of resulted features.*
        i. **Number of features extracted:** After preprocessing and feature engineering, the dataset contains additional features.
        ii. **Names of resulted features:** `rooms_per_household`, `bedrooms_per_room` and `population_per_household`.
        iii. **Dimension of resulted features:** `(17609, 16)`
    b. *Is cross-validation is used in any of implemented models? Yes.*
        i. *The number of fold:* `3-fold cross-validation`.
        ii. *Ratio of training/validation.*
            1. Training set: 2/3 of the data
            2. Validation set: 1/3 of the data
3. ***Results details:***

    1.*Loss curve*



2. *Accuracy*

```
Average R squared score :  0.6231679206055769
```

```
Evaluation score on 3 cross-validation sets :  [0.6145668  0.62858995
0.62634702]
```

KNN

1. ***General Information on dataset:***
    a. *the name of dataset used:*
    b. *the number of samples used in training and testing.*
        i. Training: `14087`
        ii. Testing: `3522`
2. ***Implementation details:***

- *At feature extraction phase, how many features were extracted, their names, the dimension of resulted features.*
    - **Number of features extracted:** After preprocessing and feature engineering, the dataset contains additional features.

- **Names of resulted features:** `rooms_per_household`, `bedrooms_per_room` and `population_per_household`.
- **Dimension of resulted features:** `(17609, 16)`

- *Is cross-validation is used in any of implemented models? Yes.*

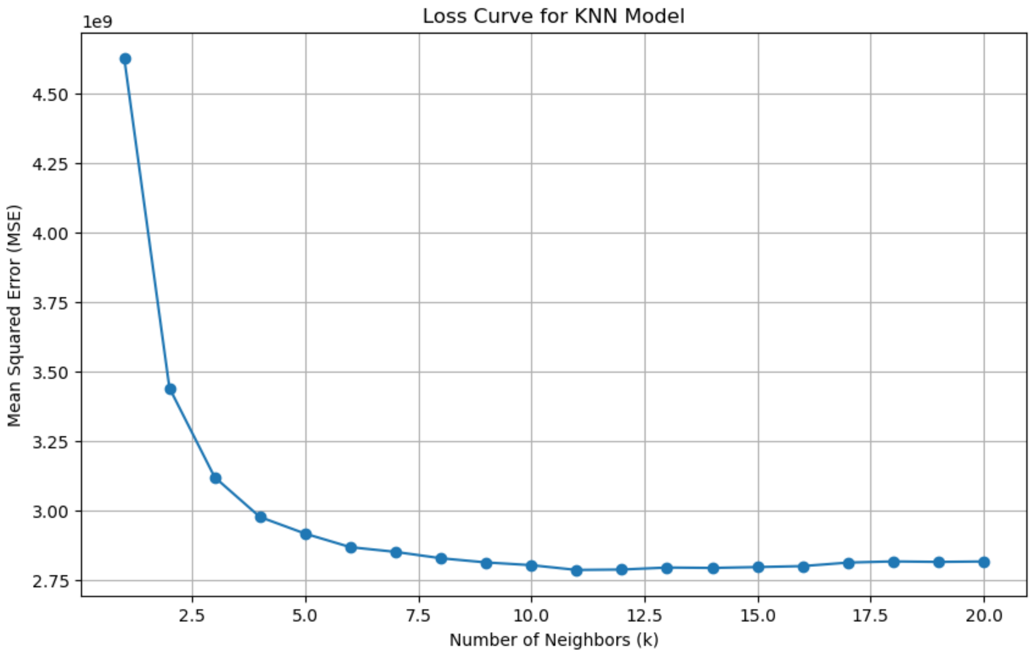
1. *The number of fold:* *`3-fold cross-validation`*.
2. *Ratio of training/validation.*
    a. Training set: 2/3 of the data
    b. Validation set: 1/3 of the data

- *Hyperparameters used in your model, as initial learning rate, optimizer, regularization, batch size, no. of epochs, etc…*

1. **K-Nearest Neighbors (KNN):**
    - **`n_neighbors`**: Number of neighbors to consider `18 in your example`.
2. **Gradient Boosting Regressor:**
    - **`n_estimators`**: Number of boosting stages `100`.
    - **`learning_rate`**: The step size shrinkage used to prevent overfitting `0.1`.
    - **`max_depth`**: Maximum depth of the individual trees `8 in your example`.
3. **XGBoost:**
    - **`max_depth`**: Maximum depth of a tree `8`.
    - **`n_estimators`**: Number of boosting rounds `100`.
    - **`objective`**: The learning task and corresponding objective, set to `reg:squarederror` for regression.
    - **`random_state`**: Seed for reproducibility `42`.

1. ***Results details:***

KNN Loss curve

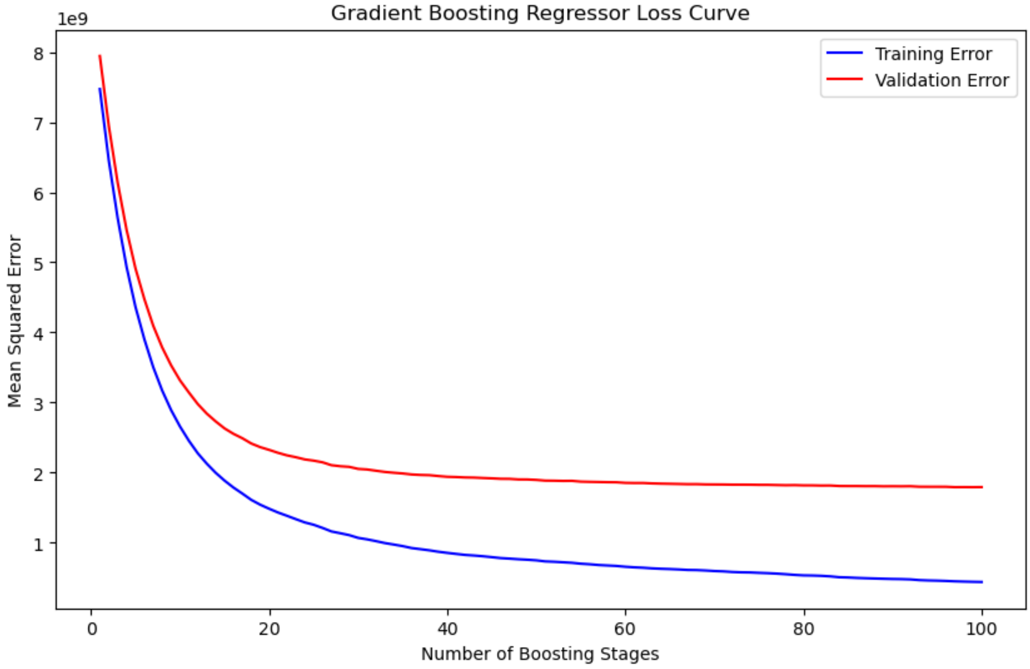


KNN Accuracy

```
Evaluation score on 3 cross-validation sets :  [0.67168832 0.68923142
0.68611015]

Average R squared score :  0.6823432959082879
```

Loss Curve For Gradient Boosting Regressor.

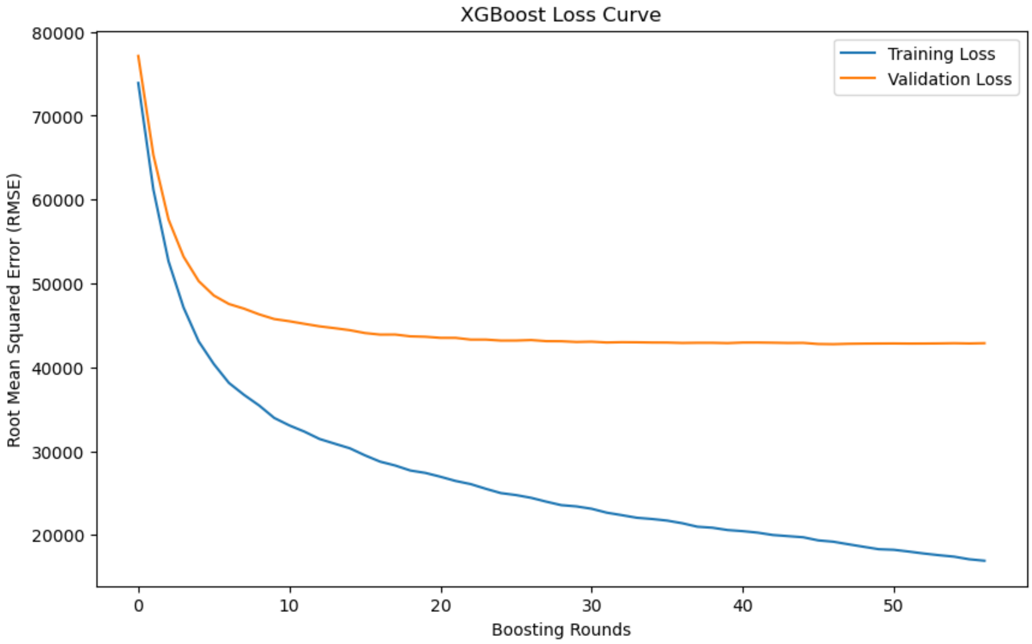


Gradient Boosting Regressor Accuracy.

```
Evaluation score on 3 cross-validation sets :  [0.78743338 0.80353852 0.80631629]
```

```
Average R squared score :  0.7990960650730976
```

Loss Curve For **eXtreme Gradient Boosting Regressor (XGBoost)**



eXtreme Gradient Boosting Regressor (XGBoost) ***Accuracy***

```
Average R squared score :  0.7906049245321869
```

```
Evaluation score on 3 cross-validation sets :  [0.78188369 0.79387968 0.79605141]
```

Logistic Regression

1. ***General Information on dataset:***
   a. *the name of dataset used: Food-101*
   b. *number of classes and their labels:*

        i. Five classes.

        ii. ['cheesecake', 'cup_cakes', 'donuts', 'hamburger', 'pizza']

    c. *the total number of samples in dataset and the size of each (in case of images):*

        i. number of samples: 3500 samples

        ii. Images size: (224 * 224)

    d. *the number of samples used in training, validation and testing:*

        i. training = `2800`

        ii. validation = `0.2` of training set (560 samples).

        iii. testing = `700`

2. ***<u>Implementation details:</u>***

- *At feature extraction phase, how many features were extracted, their names, the dimension of resulted features.*
    - *how many features were extracted:* 128 features were extracted for each image using the SIFT.
    - *the dimension of resulted features:* `(2800, 224, 224, 3)`
- *Is cross-validation is used in any of implemented models? No.*

1. ***<u>Results details:</u>***

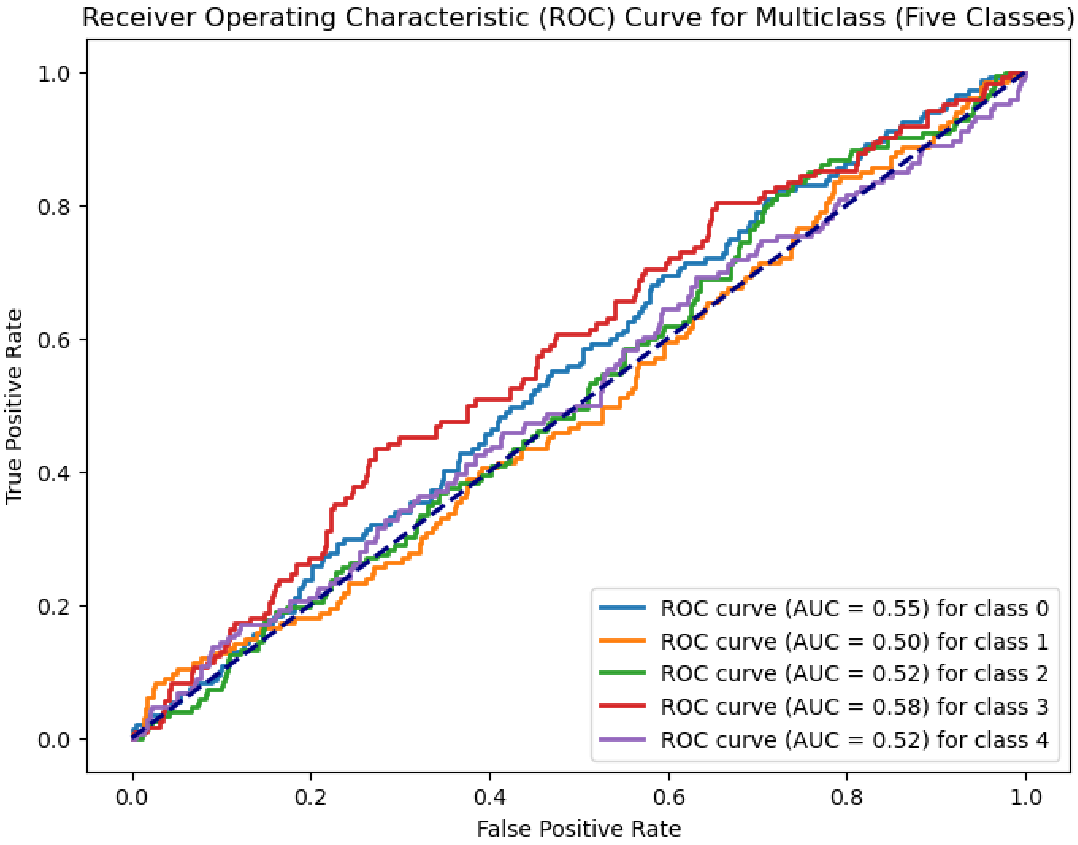
    Accuracy

    `Accuracy: 0.22714285714285715`

    Confusion matrix

```
Confusion Matrix:
[[38 31 22 26 30]
 [23 24 29 33 24]
 [30 32 29 34 27]
 [31 17 20 40 14]
 [29 25 27 37 28]]
```

    ROC curve


Receiver Operating Characteristic (ROC) Curve for Multiclass (Five Classes)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (AUC = 0.55) for class 0
ROC curve (AUC = 0.50) for class 1
ROC curve (AUC = 0.52) for class 2
ROC curve (AUC = 0.58) for class 3
ROC curve (AUC = 0.52) for class 4