Assignment 2 Report

Prepared by

SM Afiqur Rahman

&

Jubaer Ahmed Bhuiyan

Course: Introduction to Machine Learning

Instructor: Karteek Popuri

Answer to Question 1:

The cross_val_score function in scikit-learn by default employs 5-fold cross-validation, which divides the training data into 5 folds, trains the model on 4 folds, and assesses the model's success on the 5th fold. This procedure will be repeated five times, with each version using a different fold for evaluation, and the average error over all repetitions will be returned. The amount of the dataset and the available processing resources determine what amount folds are used for cross-validation.

In general, a smaller number of folds (for example, 5 or 10) is usually adequate for cross-validation when the dataset size is big, as it is in our instance with 800 observations. Even though using more folds may not substantially increase the model's precision, it will take longer to compute in order to run cross-validation.

We chose to fold the data in 5 parts. The dataset was divided into 5 equal portions, and the model was trained and evaluated 5 times, with each fold acting as the test set once. This is a popular option for the number of folds in cross-validation because it finds a balance between using just the right amount of data in each fold to obtain an accurate estimate of the model performance and avoiding using too many folds that would result in a high variance estimate.

Error rates: -

Validation approach error (MSE): 125.64721878839762

Validation approach RSE: 11.209247021472835

Validation approach R2: 0.5531427987011468

Cross-validation approach error (MSE): 103.36660292429413

Cross-validation approach RSE: 10.166936752252083

Cross-validation approach R2: 0.6356084343348203

Using various methods, the validation strategy and cross-validation (CV) approach provide estimates of model effectiveness. Using a validation dataset that wasn't used for training, the validation method gauges the model's success. By dividing the training dataset into k-folds and evaluating the model k times, with each fold serving as the validation dataset, the CV method, on the other hand, assesses the model's performance. We can see from our findings that the cross-validation method produces a mean squared error (MSE) of 103.37, which is lower than the MSE of 125.65 produced by the validation method. This indicates that the cross-validation method, which utilizes all of the available data for training and validation, offers a superior approximation of the model's performance.

Furthermore, the cross-validation RSE values (10.17) are also lower than the validation approach's RSE values. (11.21). This indicates that a better approximation of the model's capacity to adapt to new data is provided by the cross-validation method.

Last but not least, the R2 number obtained through cross-validation (0.64) is marginally better than the R2 obtained through the validation method. (0.55). This indicates that the cross-validation method offers a more accurate estimation of the percentage of variation in the target variable that the model accounts for.

Overall, compared to the validation method, which only uses one validation set, the cross-validation approach yields a more accurate estimate of model performance because it utilizes all the data that are accessible for both training and validation.

Answer to Question 2:

Error rates using Simple Linear Regression Model: -

Validation approach error (MSE): 125.64721878839762

Validation approach RSE: 11.209247021472835

Validation approach R2: 0.5531427987011468

Cross-validation approach error (MSE): 103.36660292429413

Cross-validation approach RSE: 10.166936752252083

Cross-validation approach R2: 0.6356084343348203

Error Rates using Ridge Regression model: -

MSE on test data: 91.04629491809004

RSE on test data: 9.541818218667238

R2 on test data: 0.5952908477478914

On the test data, the Ridge regression model works better than the basic linear regression model in terms of MSE and R2. On the test data, the Ridge regression model's RSE is also less than the basic linear regression model's, demonstrating a superior fit between the model and the data.

In addition, compared to the validation method used for the simple linear regression model, the cross-validation technique used to adjust the Ridge regression model's regularization parameter offers a more trustworthy approximation of the model's performance on unobserved data. This is because cross-validation reduces the bias in the estimate of the model's performance by providing an estimate of the model's performance on numerous iterations of the data.

Overall, compared to the basic linear regression model, the Ridge regression model seems to be a superior option for estimating the compressive strength of concrete based on the provided characteristics. It is important to keep in mind that alternative regression models or machine learning methods might work more effectively, so it is wise to test out a number of models and evaluate their performance before settling on one.

Answer to Question 3:

Error rates using Simple Linear Regression Model: -

Validation approach error (MSE): 125.64721878839762

Validation approach RSE: 11.209247021472835

Validation approach R2: 0.5531427987011468

Cross-validation approach error (MSE): 103.36660292429413

Cross-validation approach RSE: 10.166936752252083

Cross-validation approach R2: 0.6356084343348203

Error Rates using Ridge Regression model: -

MSE on test data: 91.04629491809004

RSE on test data: 9.541818218667238

R2 on test data: 0.5952908477478914

Error Rates using Lasso Regression model: -

Residual standard error (RSE): 9.541580680442047

R2 score: 0.5953109975128402

Here are a few findings:

- The lowest R2 score, greatest MSE and RSE values, and worst performance of the three models all point to the Simple Linear Regression model's performance.
- The reduced MSE and RSE numbers and higher R2 score on the test data show that the Ridge Regression model performs better than the Simple Linear Regression model. This indicates that when compared to the Simple Linear Regression model, the Ridge Regression model is more generalizable to novel data.
- With nearly equal RSE and R2 numbers, the Lasso Regression model performs similarly to the Ridge Regression model. The Ridge Regression model is marginally better at predicting the outcome variable than the Lasso Regression model, which has a slightly smaller RSE number.

The Ridge and Lasso Regression models appear to perform better than the Simple Linear Regression model overall, with the Lasso Regression model slightly outperforming the Ridge Regression model in terms of predictive performance.