Assignment 11 – Machine Learning

Types of Cross Validation methods in Machine Learning:

**1. Holdout Method**

**2. K-Fold Cross-Validation**

**3. Stratified K-Fold Cross-Validation**

**4. Leave-One-Out Cross-Validation (LOOCV)**:

**5. Leave-P-Out Cross-Validation**

**6. Time Series Cross-Validation**

**7. Repeated K-Fold Cross-Validation**

**8. Nested Cross-Validation**:

**1. Holdout Method:**

Removing a part of the training data set and sending that to a model that was trained on the rest of the data set to get the predictions

Calculate the error estimation which tells how our model is doing on unseen data sets.

Split the dataset into two parts: a training set and a validation set.

Typically, use a larger portion for training (e.g., 70-80%) and the rest for validation.

**2. K-Fold Cross-Validation**:

Divide the dataset into 'K' equal subsets or folds.

Train the model 'K' times, each time using 'K-1' folds for training and 1 fold for validation.

Calculate performance metrics for each validation fold.

Average the metrics to get the final performance estimate.

**3. Stratified K-Fold Cross-Validation**:

Similar to K-Fold, but ensures that each fold contains approximately the same distribution of class labels as the entire dataset.

Useful when dealing with imbalanced datasets.

**4. Leave-One-Out Cross-Validation (LOOCV)**:

Each data point serves as a validation set while the rest are used for training.

Extremely computationally expensive for large datasets.

**5. Leave-P-Out Cross-Validation**:

Similar to LOOCV, but leaves 'p' data points out for validation instead of just one.

**6. Time Series Cross-Validation**:

Suitable for time series data where the order of data points matters.

Splits the dataset sequentially, ensuring that validation sets come after training sets in time.

**7. Repeated K-Fold Cross-Validation**:

Repeats K-Fold CV multiple times with different random splits.

Provides more reliable performance estimates.

**8. Nested Cross-Validation**:

Combines cross-validation for model selection (e.g., hyper parameter tuning) with cross-validation for performance evaluation.

Helps prevent over fitting during hyper parameter tuning.

**Finding the Best Model : Using Grid search CV –Machine Learning -Supervised Learning-Regression :**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Support Vector Machine - Regression -GridSearchCV** | | | | | | | | | |
|
| **s.no.** | **C** | **gamma** | **kernel** | **split0\_test\_score** | **split1\_test\_score** | **split2\_test\_score** | **mean\_test\_score** | **std\_test\_score** | **rank\_test\_score** |
| 0 | 10 | auto | rbf | -0.122 | -0.072 | -0.116 | -0.103 | 0.022 | 20 |
| 1 | 10 | auto | poly | 0.838 | 0.835 | 0.792 | 0.822 | 0.021 | 1 |
| 2 | 10 | auto | sigmoid | -0.124 | -0.073 | -0.116 | -0.105 | 0.022 | 22 |
| 3 | 10 | auto | linear | -0.030 | 0.024 | -0.087 | -0.031 | 0.046 | 13 |
| 4 | 10 | scale | rbf | -0.121 | -0.062 | -0.127 | -0.103 | 0.029 | 21 |
| 5 | 10 | scale | poly | -0.117 | -0.084 | -0.180 | -0.127 | 0.040 | 28 |
| 6 | 10 | scale | sigmoid | -0.126 | -0.075 | -0.118 | -0.106 | 0.022 | 27 |
| 7 | 10 | scale | linear | -0.030 | 0.024 | -0.087 | -0.031 | 0.046 | 13 |
| 8 | 100 | auto | rbf | -0.113 | -0.064 | -0.112 | -0.096 | 0.023 | 19 |
| 9 | 100 | auto | poly | 0.765 | 0.732 | 0.686 | 0.727 | 0.033 | 2 |
| 10 | 100 | auto | sigmoid | -0.124 | -0.073 | -0.116 | -0.105 | 0.022 | 22 |
| 11 | 100 | auto | linear | 0.536 | 0.532 | 0.487 | 0.518 | 0.022 | 9 |
| 12 | 100 | scale | rbf | -0.148 | -0.095 | -0.201 | -0.148 | 0.043 | 34 |
| 13 | 100 | scale | poly | -0.122 | -0.101 | -0.192 | -0.138 | 0.039 | 31 |
| 14 | 100 | scale | sigmoid | -0.152 | -0.104 | -0.138 | -0.131 | 0.020 | 30 |
| 15 | 100 | scale | linear | 0.536 | 0.532 | 0.487 | 0.518 | 0.022 | 9 |
| 16 | 1000 | auto | rbf | -0.090 | -0.037 | -0.131 | -0.086 | 0.039 | 17 |
| 17 | 1000 | auto | poly | -21.199 | -7.258 | -14.434 | -14.297 | 5.692 | 38 |
| 18 | 1000 | auto | sigmoid | -0.124 | -0.073 | -0.116 | -0.105 | 0.022 | 22 |
| 19 | 1000 | auto | linear | 0.675 | 0.673 | 0.642 | 0.663 | 0.015 | 7 |
| 20 | 1000 | scale | rbf | -0.142 | -0.098 | -0.203 | -0.148 | 0.043 | 33 |
| 21 | 1000 | scale | poly | -0.075 | -0.057 | -0.149 | -0.094 | 0.040 | 18 |
| 22 | 1000 | scale | sigmoid | -1.636 | -1.642 | -1.479 | -1.586 | 0.076 | 35 |
| 23 | 1000 | scale | linear | 0.675 | 0.673 | 0.642 | 0.663 | 0.015 | 7 |
| 24 | 2000 | auto | rbf | -0.049 | -0.005 | -0.092 | -0.049 | 0.035 | 16 |
| 25 | 2000 | auto | poly | -62.062 | -24.203 | -44.428 | -43.565 | 15.468 | 39 |
| 26 | 2000 | auto | sigmoid | -0.124 | -0.073 | -0.116 | -0.105 | 0.022 | 22 |
| 27 | 2000 | auto | linear | 0.712 | 0.736 | 0.659 | 0.702 | 0.032 | 3 |
| 28 | 2000 | scale | rbf | -0.132 | -0.092 | -0.195 | -0.139 | 0.043 | 32 |
| 29 | 2000 | scale | poly | -0.022 | -0.004 | -0.102 | -0.043 | 0.043 | 15 |
| 30 | 2000 | scale | sigmoid | -5.572 | -5.741 | -5.162 | -5.492 | 0.243 | 36 |
| 31 | 2000 | scale | linear | 0.712 | 0.736 | 0.659 | 0.702 | 0.032 | 3 |
| 32 | 3000 | auto | rbf | -0.012 | 0.033 | -0.046 | -0.008 | 0.032 | 12 |
| 33 | 3000 | auto | poly | -117.498 | -151.358 | -45.855 | -104.904 | 43.983 | 40 |
| 34 | 3000 | auto | sigmoid | -0.124 | -0.073 | -0.116 | -0.105 | 0.022 | 22 |
| 35 | 3000 | auto | linear | 0.672 | 0.714 | 0.643 | 0.676 | 0.029 | 5 |
| 36 | 3000 | scale | rbf | -0.121 | -0.082 | -0.184 | -0.129 | 0.042 | 29 |
| 37 | 3000 | scale | poly | 0.028 | 0.049 | -0.056 | 0.007 | 0.046 | 11 |
| 38 | 3000 | scale | sigmoid | -12.159 | -12.654 | -11.192 | -12.001 | 0.607 | 37 |
| 39 | 3000 | scale | linear | 0.672 | 0.714 | 0.643 | 0.676 | 0.029 | 5 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Random Forest - Regression -GridSearchCV** | | | | | | | | | |
|
| **s.no** | **critereon** | **max features** | **n\_estimators** | **split0\_test score** | **split1\_test score** | **split2\_test score** | **mean\_test\_score** | **std\_test\_score** | **rank\_test\_score** |
| 0 | mse | auto | 10 | 0.813 | 0.776 | 0.772 | 0.787 | 0.018 | 23 |
| 1 | mse | auto | 100 | 0.823 | 0.800 | 0.790 | 0.804 | 0.014 | 16 |
| 2 | mse | auto | 1000 | 0.828 | 0.806 | 0.789 | 0.808 | 0.016 | 14 |
| 3 | mse | sqrt | 10 | 0.810 | 0.811 | 0.769 | 0.797 | 0.020 | 20 |
| 4 | mse | sqrt | 100 | 0.822 | 0.816 | 0.792 | 0.810 | 0.013 | 12 |
| 5 | mse | sqrt | 1000 | 0.830 | 0.819 | 0.792 | 0.814 | 0.016 | 3 |
| 6 | mse | log2 | 10 | 0.800 | 0.788 | 0.770 | 0.786 | 0.012 | 25 |
| 7 | mse | log2 | 100 | 0.831 | 0.817 | 0.789 | 0.812 | 0.017 | 6 |
| 8 | mse | log2 | 1000 | 0.831 | 0.818 | 0.793 | 0.814 | 0.016 | 1 |
| 9 | mae | auto | 10 | 0.802 | 0.789 | 0.776 | 0.789 | 0.010 | 22 |
| 10 | mae | auto | 100 | 0.817 | 0.802 | 0.783 | 0.801 | 0.014 | 18 |
| 11 | mae | auto | 1000 | 0.819 | 0.802 | 0.786 | 0.802 | 0.013 | 17 |
| 12 | mae | sqrt | 10 | 0.807 | 0.804 | 0.781 | 0.797 | 0.011 | 19 |
| 13 | mae | sqrt | 100 | 0.829 | 0.815 | 0.791 | 0.812 | 0.016 | 8 |
| 14 | mae | sqrt | 1000 | 0.829 | 0.816 | 0.791 | 0.812 | 0.016 | 9 |
| 15 | mae | log2 | 10 | 0.791 | 0.779 | 0.775 | 0.782 | 0.007 | 27 |
| 16 | mae | log2 | 100 | 0.828 | 0.817 | 0.788 | 0.811 | 0.017 | 11 |
| 17 | mae | log2 | 1000 | 0.829 | 0.818 | 0.792 | 0.813 | 0.016 | 4 |
| 18 | friedman\_mse | auto | 10 | 0.793 | 0.803 | 0.758 | 0.785 | 0.019 | 26 |
| 19 | friedman\_mse | auto | 100 | 0.827 | 0.800 | 0.791 | 0.806 | 0.015 | 15 |
| 20 | friedman\_mse | auto | 1000 | 0.828 | 0.806 | 0.790 | 0.808 | 0.016 | 13 |
| 21 | friedman\_mse | sqrt | 10 | 0.810 | 0.780 | 0.771 | 0.787 | 0.017 | 24 |
| 22 | friedman\_mse | sqrt | 100 | 0.829 | 0.814 | 0.793 | 0.812 | 0.015 | 10 |
| 23 | friedman\_mse | sqrt | 1000 | 0.830 | 0.817 | 0.793 | 0.813 | 0.015 | 5 |
| 24 | friedman\_mse | log2 | 10 | 0.807 | 0.796 | 0.783 | 0.796 | 0.010 | 21 |
| 25 | friedman\_mse | log2 | 100 | 0.831 | 0.811 | 0.793 | 0.812 | 0.016 | 7 |
| 26 | friedman\_mse | log2 | 1000 | 0.830 | 0.818 | 0.794 | 0.814 | 0.015 | 2 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Decision tree - Regression -GridSearchCV** | | | | | | | | | |
|
| **S.No.** | **criterion** | **max\_features** | **splitter** | **split0\_test\_score** | **split1\_test\_score** | **split2\_test\_score** | **mean\_test\_score** | **std\_test\_score** | **rank\_test\_score** |
| 0 | mse | auto | random | 0.688 | 0.690 | 0.703 | 0.694 | 0.007 | 4 |
| 1 | mse | auto | best | 0.704 | 0.653 | 0.709 | 0.688 | 0.025 | 8 |
| 2 | mse | sqrt | random | 0.683 | 0.651 | 0.587 | 0.640 | 0.040 | 14 |
| 3 | mse | sqrt | best | 0.721 | 0.677 | 0.635 | 0.677 | 0.035 | 10 |
| 4 | mse | log2 | random | 0.621 | 0.617 | 0.568 | 0.602 | 0.024 | 17 |
| 5 | mse | log2 | best | 0.681 | 0.659 | 0.685 | 0.675 | 0.011 | 11 |
| 6 | mae | auto | random | 0.705 | 0.771 | 0.713 | 0.730 | 0.030 | 1 |
| 7 | mae | auto | best | 0.692 | 0.719 | 0.658 | 0.690 | 0.025 | 7 |
| 8 | mae | sqrt | random | 0.729 | 0.742 | 0.655 | 0.709 | 0.038 | 2 |
| 9 | mae | sqrt | best | 0.685 | 0.670 | 0.723 | 0.693 | 0.023 | 5 |
| 10 | mae | log2 | random | 0.696 | 0.713 | 0.608 | 0.672 | 0.046 | 12 |
| 11 | mae | log2 | best | 0.691 | 0.680 | 0.730 | 0.700 | 0.021 | 3 |
| 12 | friedman\_mse | auto | random | 0.681 | 0.674 | 0.715 | 0.690 | 0.018 | 6 |
| 13 | friedman\_mse | auto | best | 0.703 | 0.657 | 0.696 | 0.685 | 0.020 | 9 |
| 14 | friedman\_mse | sqrt | random | 0.638 | 0.409 | 0.553 | 0.533 | 0.095 | 18 |
| 15 | friedman\_mse | sqrt | best | 0.686 | 0.654 | 0.620 | 0.653 | 0.027 | 13 |
| 16 | friedman\_mse | log2 | random | 0.650 | 0.622 | 0.616 | 0.629 | 0.015 | 15 |
| 17 | friedman\_mse | log2 | best | 0.698 | 0.482 | 0.662 | 0.614 | 0.094 | 16 |

***Result:* The best model is Support Vector Machine Regression –Grid Search CV – Mean r2\_score value = 0.822 .**