Nested K-Fold Cross-Validation

Luke Neuendorf (lneuendorf@wisc.edu)
Github Repo: https://github.com/lneuendorf/Fantasy-Footall-Rookie-RB-Model-Selction

I. Introduction

Over the past year, I have played around with predicting rookie NFL running back fantasy football success using their college production and draft capital. I used regression algorithms to predict total fantasy points over their first three seasons. However, I never had an unbiased way of comparing algorithm performance. The problem with regular k-fold cross validation is hyperparameters tuning and accuracy measurements are performed on the same train-test pair. This often leads to information leakage and bias [1]. Nested k-fold CV attempts to solve this problem by tuning hyperparameters on inner folds and testing accuracy on outer folds.

II. Algorithm

The training set is split into k outer folds. The number of rows per outer fold is equated with Equation 1.

$$rows_{per\_fold} = \lfloor rows_{in\_training\_set} / k \rfloor \quad \text{(Equation 1)}$$

From there, k-1 folds are passed to the inner loop, where gridsearch CV is used to tune hyperparameters. After this, the model accuracy is measured on the held out outer fold. The same process is repeated for a total of k iterations, such that each outer fold takes a turn measuring accuracy and each set of k-1 folds are used to train hyperparmeters. The overall accuracy is measured by averaging k accuracy outputs from each test set. The process is demonstrated below in Figure 1.

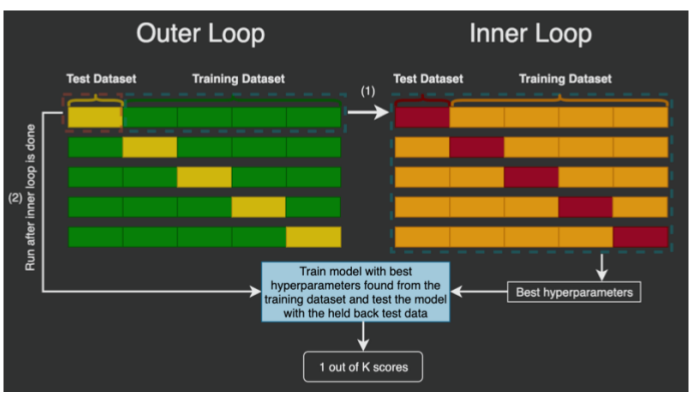

Figure 1: Nested K-Fold CV visualization [1]

III. Dataset

The dataset, which combines 142 college running back stats, sixteen NFL combine stats, draft age, draft pick, and NFL fantasy production, was compiled by Jerrick Backous [2]. I reduced the features to thirteen by removing columns with over 10% missing data and/or a correlation coefficient to the target variable below 0.7. The thirteen features and their correlations to the target variable (NFL Total Fantasy Points in Years 1-3) are shown in a correlation matrix in Figure 2. Note that “NFL Total Fantasy Points in Years 1-3” is named “Tot” in this figure.

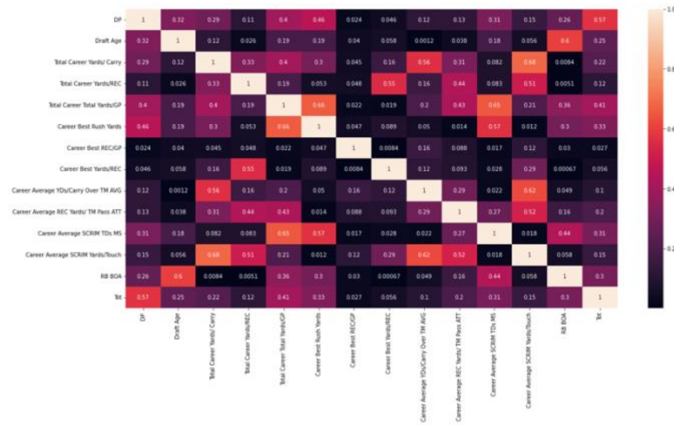

Figure 2: correlation matrix of dataset after feature elimination

Below in Figure 3 is the bucketed target variable distribution. As is visible in the chart, the data is heavily right-skewed. This skew is caused by the fact that most NFL running backs do not get much playing time, as there are only 32 starting running backs. Therefore, the average fantasy point total lies closer to zero. Additionally, the sample of running backs that produce large numbers in their first three seasons is relatively small as they have to be very skilled.

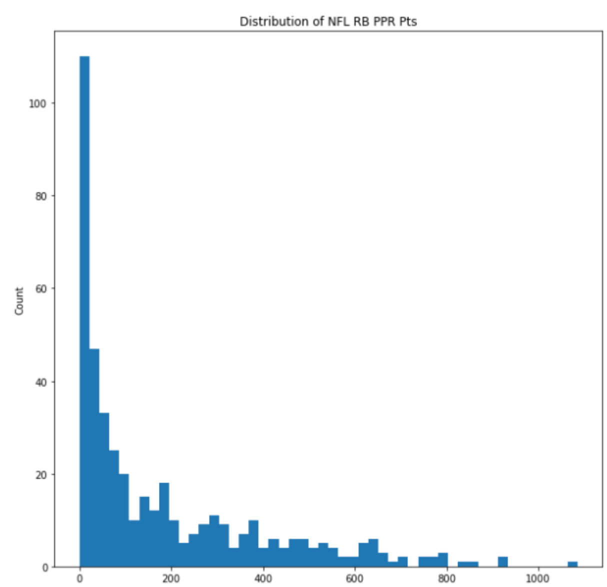

Figure 3: distribution of target variable

IV. Result

I started by making a baseline model to compare the other model accuracy scores to. Here, the baseline model is the mean value of the target variable, which ended up being 182.548 fantasy points. Then, I calculated the MAE of the baseline model using the Equation 2, where y is the prediction and ŷ is the true value.

$$MAE = (1/n)\ \Sigma\ |y_i - \hat{y}_i| \quad \text{(Equation 2)}$$

Next, I ran the program three times, generating the average MAE of each outer test fold for seven regression algorithms. Then, I averaged the three MAEs produced in each iteration to produce an average MAE for each algorithm. This is shown in Figure 4.

	Mean Absolute Error (MAE)			
Regression Algo	**Iter. 1**	**Iter. 2**	**Iter. 3**	**Average**
Baseline Model	-	-	-	167.287
Decision Tree	124.628	124.667	120.994	123.430
Elastic Net	169.415	169.330	169.330	169.358
XGBoost	118.914	117.934	117.140	117.996
Linear	168.625	178.954	178.066	175.215
RANSAC	166.214	166.182	163.270	165.222
Gaussian Process	169.090	169.147	169.139	169.125
Support Vector	157.548	155.808	154.619	155.992

Figure 4: results after running the program three times

As can be seen in Figure 4, XGBoost and Decision Trees far outperformed all other tested algorithms. Linear Regression, Gaussian Process Regression, and Elastic Net Regression had a worse average MAE than the baseline model, making these algorithms bad candidates to use in this scenario.

V. References

[1] C. Hansen, “Nested Cross-Validation Python Code,” ML From Scratch, 19-Dec-2019. [Online]. Available: https://mlfromscratch.com/nested-cross-validation-python-code/. [Accessed: 21-Aug-2022].

[2] J. Backous, “NFL Prospect Database,” Patreon. [Online]. Available: https://www.patreon.com/devydata. [Accessed: 21-Aug-2022].

[3] S. Raschka, “Nested CV for Algorithm Selection (L11 Model Eval. Part 4),” YouTube, 24-Nov-2020. [Online]. Available: https://www.youtube.com/watch?v=XXFLFWHP9Nc&t=717s. [Accessed: 21-Aug-2022].

Console Output:

```
In [1]: runfile('/Users/lukeneuendorf/Documents/UWMad/SM22/CS540/P6/p6.py', wdir='/
Users/lukeneuendorf/Documents/UWMad/SM22/CS540/P6')
Data Preprocessing:
    Num features before preprocessing: 160
    Num of features after deleting columns with ...
        more than 10.0% null values: 98
        correlation coefficient lower than 0.7: 13

Learning About the Data:
    Skew of target variable: 1.4075
    The positve skew value means the target distribution is right-skewed.
    A correlation heat map and target var distribution bar chart have also been
created.

Nested K-Fold CV:
    Baseline Model (mean of y-var): 182.5476

    Decision Tree
        Inner Fold 1
            MAE: 116.3742
            Best Hyperparameters:
                ccp_alpha: 0.01
                max_depth: 3
                min_samples_leaf: 20
        Inner Fold 2
            MAE: 140.8913
            Best Hyperparameters:
                ccp_alpha: 0.2
                max_depth: 3
                min_samples_leaf: 15
        Inner Fold 3
            MAE: 140.8799
            Best Hyperparameters:
                ccp_alpha: 0.25
                max_depth: 2
                min_samples_leaf: 10
        Inner Fold 4
            MAE: 110.6182
            Best Hyperparameters:
                ccp_alpha: 0.02
                max_depth: 4
                min_samples_leaf: 20
        Inner Fold 5
            MAE: 123.8414
            Best Hyperparameters:
                ccp_alpha: 0.0
                max_depth: 3
                min_samples_leaf: 15

        Average Outer Fold MAE's: 126.521

    Elastic Net
        Inner Fold 1
            MAE: 167.5419
            Best Hyperparameters:
                alpha: 1.0
                l1_ratio: 0.01
        Inner Fold 2
            MAE: 159.1441
            Best Hyperparameters:
                alpha: 1.0
                l1_ratio: 0.01
        Inner Fold 3
            MAE: 148.1466
            Best Hyperparameters:
                alpha: 100.0
                l1_ratio: 0.34
        Inner Fold 4
            MAE: 193.2677
            Best Hyperparameters:
                alpha: 100.0
                l1_ratio: 0.29000000000000004
        Inner Fold 5
            MAE: 180.7899
            Best Hyperparameters:
                alpha: 100.0
                l1_ratio: 0.3

        Average Outer Fold MAE's: 169.778

    XGBoost
        Inner Fold 1
            MAE: 118.0019
            Best Hyperparameters:
                gamma: 1
                learning_rate: 0.01
                max_depth: 2
                n_estimators: 300
                subsample: 0.7
        Inner Fold 2
            MAE: 126.7428
            Best Hyperparameters:
                gamma: 1
                learning_rate: 0.01
                max_depth: 2
                n_estimators: 300
                subsample: 0.7
        Inner Fold 3
            MAE: 110.7511
            Best Hyperparameters:
                gamma: 1
                learning_rate: 0.01
                max_depth: 2
                n_estimators: 200
                subsample: 0.7
        Inner Fold 4
            MAE: 105.3213
            Best Hyperparameters:
                gamma: 1
                learning_rate: 0.01
                max_depth: 2
                n_estimators: 200
                subsample: 0.7
        Inner Fold 5
            MAE: 127.3145
            Best Hyperparameters:
                gamma: 1
                learning_rate: 0.01
                max_depth: 2
                n_estimators: 200
                subsample: 0.7

        Average Outer Fold MAE's: 117.6263
```

```
    Linear Regression
        Inner Fold 1
            MAE: 163.8971
            Best Hyperparameters:
                positive: True
        Inner Fold 2
            MAE: 158.9179
            Best Hyperparameters:
                positive: True
        Inner Fold 3
            MAE: 142.197
            Best Hyperparameters:
                positive: True
        Inner Fold 4
            MAE: 196.4709
            Best Hyperparameters:
                positive: True
        Inner Fold 5
            MAE: 183.9059
            Best Hyperparameters:
                positive: False

        Average Outer Fold MAE's: 169.0778

    RANSAC Regression
        Inner Fold 1
            MAE: 166.2869
            Best Hyperparameters:
                max_trials: 200
                min_samples: 0.9
        Inner Fold 2
            MAE: 163.0368
            Best Hyperparameters:
                max_trials: 200
                min_samples: 1
        Inner Fold 3
            MAE: 144.6341
            Best Hyperparameters:
                max_trials: 200
                min_samples: 1
        Inner Fold 4
            MAE: 200.104
            Best Hyperparameters:
                max_trials: 200
                min_samples: 1
        Inner Fold 5
            MAE: 185.0041
            Best Hyperparameters:
                max_trials: 200
                min_samples: 1

        Average Outer Fold MAE's: 171.8132

    Gaussian Process Regressor
        Inner Fold 1
            MAE: 165.2084
            Best Hyperparameters:
        Inner Fold 2
            MAE: 158.8079
            Best Hyperparameters:
        Inner Fold 3
            MAE: 147.8186
            Best Hyperparameters:
        Inner Fold 4
            MAE: 193.1411
            Best Hyperparameters:
        Inner Fold 5
            MAE: 180.6391
            Best Hyperparameters:

        Average Outer Fold MAE's: 169.123

    Support Vector Regression
        Inner Fold 1
            MAE: 141.218
            Best Hyperparameters:
                C: 1
                epsilon: 0.1
                kernel: rbf
        Inner Fold 2
            MAE: 154.5787
            Best Hyperparameters:
                C: 1
                epsilon: 0.5
                kernel: rbf
        Inner Fold 3
            MAE: 116.0784
            Best Hyperparameters:
                C: 1
                epsilon: 0.1
                kernel: rbf
        Inner Fold 4
            MAE: 189.6404
            Best Hyperparameters:
                C: 1
                epsilon: 0.5
                kernel: poly
        Inner Fold 5
            MAE: 175.5905
            Best Hyperparameters:
                C: 10
                epsilon: 0.1
                kernel: rbf

        Average Outer Fold MAE's: 155.4212
```