Abstract

Real predictive maintenance datasets are generally difficult to obtain and in particular difficult to publish, we analyze a synthetic dataset that reflects real predictive maintenance encountered in the industry. This dataset was provided by the University of California Irvine Machine Learning Repository. The goal of our project was to test our synthetic dataset using four different methods of cross-validation: Validation set approach, Leave one out-cross validation(LOOCV), K-fold cross-Validation, Repeated K-fold cross-Validation. To see which method provides the best results.

Keywords: synthetic, dataset, K-fold, LOOCV,

Write Up Report

Introduction/Content:

- Predictive Maintenance Dataset is a synthetic dataset that reflects real predictive maintenance data encountered in industry.
- The dataset consists of 10,000 data points stored as rows with 14 features in columns (two columns in use)
- UID: unique identifier ranging from 1 to 10,000
- Air temperature [K]: generated using a random walk process later normalized to a standard deviation of 2 K around 300 K
- Process temperature [K]: generated using a random walk process normalized to a standard deviation of 1 K, added to the air temperature plus 10 K.
- Rotational speed [rpm]: calculated from power of 2860 W, overlaid with a normally distributed noise
- Torque [Nm]: torque values are normally distributed around 40 Nm with a Ïf = 10 Nm and no negative values.
- Our goal was to find the best Cross-Validation method to represent our model (we are trying to predict the difference in Root Mean squared Error (RMSE). We tested using four different methods of cross validation: i) Validation set approach ii) Leave one

out-cross validation(LOOCV) iii) K-fold cross-Validation iv) Repeated K-fold cross-Validation.

Implementation:

Our goal was to find the best Cross-Validation method to represent our model. We are trying to predict the different in Root Mean squared Error (RMSE). Root Mean squared Error (RMSE) is the square root of the averaged squared difference between the actual value and the predicted value of the target variable. It gives the average prediction error made by the model, thus decrease the RMSE value to increase the accuracy of the model. Along with Root Mean squared Error, our project will also predict the Mean Absolute Error (MAE), and R^2Error. However, for the purpose of this project, we will only talk about the prediction error measured by mean-square error. To get the prediction error measured by mean-square error or misclassification rate as minimized as we can, we put our model through four different methods of cross-validation.

Types of Cross-Validation:

During the process of partitioning the complete dataset into the training set and the validation set, there are chances of losing some important and crucial data points for the training purpose. Since those data are not included in the training set, the model has not got the chance to detect some patterns. This situation can lead to overfitting or underfitting of the model. To avoid this, there are different methods of cross-validation techniques that guarantee the random sampling of training and validation data set and maximize the accuracy of the model. Some of the most popular cross-validation methods are:

- Validation Set Approach
- Leave one out cross-validation(LOOCV)
- K-fold cross-Validation
- Repeated K-fold cross-validation

Validation Set Approach(or data split)

In this method, the dataset is divided randomly into training and testing sets, and then:

- A random sampling of the dataset
- Model is trained on the training data set, for our model we assumed different train data set.
- The resultant model is applied to the testing data set

Pros and cons:

- One of the most basic and simple techniques for evaluating a model.
- No complex steps for implementation.
- Predictions done by the model are highle dependent upon the subset of observations used for training and validation.
- Using only one subset of the data for training purposes can make the model biased.

Leave One Out Cross-Validation(LOOCV)

This method also splits the dataset into 2 parts but it overcomes the drawbacks of the Validation set approach, and then:

- Train the model on N-1 data points
- Testing the model against that one data point which was left in the previous step
- Calculate prediction error
- Repeat the above 3 steps until the model is not trained and tested on all data points
- Generate overall prediction error by taking the average of prediction errors in every case

Pros and cons:

- Less bias model as almost every data point is used for training.
- No randomness in the value of performance metrics because LOOCV runs multiple times on the dataset
- Training the model N times leads to expensive computation time if the dataset is large.

K-fold Cross-Validation

This cross-validation technique divides the data into K subsets(folds) of almost equal size. Out of these K folds, one subset is used as a validation set, and the rest others are involved in training the model.

Pros and cons:

- Fast computation speed.
- A very effective method to estimate the prediction error and the accuracy of a model.
- A lower value of K leads to a biased model and a higher value of K can lead to variability in the performance metrics of the model. Thus, it is very important to use the correct value of K for the model(generally K = 5 and K = 10 is desirable).

Repeated K-fold cross-validation:

As the name suggests, in this method the K-fold cross-validation algorithm is repeated a certain number of times.

Pros and cons:

- In each repetition, the data sample is shuffled which results in developing different splits of the sample data.
- With each repetition, the algorithm has to train the model from scratch which means the computation time to evaluate the model increases by the times of repetition.

Performance:

Here we will discuss what we found after running our model through all of the above methods of cross-validation. In summary, we found that if the data set has fewer than 3 predictor and the values are large, Validation Set Approach(or data split) had the least mean squared error. While Leave One Out Cross-Validation(LOOCV) had the highest mean squared error and also was the slowest of all the methods, as mentioned below in the results of x=Air Temperature K VS. y=Process Temperature K, and x=Process Temperature VS. y=Air Temperature.

On the other hand, with 3 or more predictors and small values of those predictors K-fold Cross-Validation had the least mean squared error, and Validation set or data split method had the highest mean squared error, as mentioned below in the results of x=Rotational Speed RPM vs the Model.

Results:

Air Temperature K VS. Process Temperature K(large value predictors/variables)

Validation Set Approach(or data split)

Resampling results:

$RMSE = 0.9025509 \quad R^2 = 0.7986079$

Leave One Out Cross-Validation(LOOCV)

Resampling results:

$RMSE = 0.9644706 \quad R^2 = 0.767486$

K-fold Cross-Validation

Resampling results:

$RMSE = 0.964253 \quad R^2 = 0.7677189$

Repeated K-fold cross-validation

Resampling results:

$RMSE = 0.9642442 \quad R^2 = 0.767614$

Process Temperature VS. Air Temperature (large value predictors/variables)

Validation Set Approach(or data split)

$RMSE = 0.7220549 \quad R^2 = 0.7638512$

Leave One Out Cross-Validation(LOOCV)

Resampling results:

$RMSE = 0.7154205 \quad R^2 = 0.7674834$

K-fold Cross-Validation

Resampling results:

$RMSE = 0.7152523 \quad R^2 = 0.7676246$

Repeated K-fold cross-validation

Resampling results:

$RMSE = 00.00 \quad R^2 = 00.00$

Rotational Speed RPM vs the Model(Small value predictors/variables)

Validation Set Approach(or data split)

Resamlping Results:

$RMSE = 82.79853 \quad R^2 = 0.7714102$

Leave One Out Cross-Validation(LOOCV)

Resampling Results:

$RMSE = 86.81852 \quad R^2 = 0.7654775$

K-fold Cross-Validation

Resampling results:

$RMSE = 86.71066 \quad R^2 = 0.7668473$

Repeated K-fold cross-validation

Resampling results:

$RMSE = 86.65665 \quad R^2 = 0.7669639$

Conclusion:

Using various methods of cross-validation with different predictors, we were successfully able to achieve the outcomes of our desire. We saw some big changes in the mean square values with various predictors with small values and contrary there were also some minor increases and decreases with predictors of the same values. All in all, we can conclude that if running a large data set with 3 or more predictors leave-one-out-cross-validation(LOOCV) worked the best with roughly about 3 minutes of processing time(sample size of 10,000) and the lowest mean squared value. However, if running a large data set with 3 or fewer predictors and values of those predictors about the same(large), the validation set or data split approach worked the best with little to no time processing and least mean squared value.

References:

AI4I 2020 Predictive Maintenance Dataset Data Set. UCI Machine Learning Repository: Ai4i 2020 predictive maintenance dataset data set. (n.d.). Retrieved November 10, 2021, from https://archive.ics.uci.edu/ml/datasets/AI4I+2020+Predictive+Maintenance+Dataset#.

Cross-validation in R programming. GeeksforGeeks. (2021, September 15). Retrieved November 11, 2021, from https://www.geeksforgeeks.org/cross-validation-in-r-programming/.