Volume 9, Issue 03, March 2024

ISSN (online): 2456-0006

International Journal of Science Technology Management and Research

Available online at: www.ijstmr.com

Cross Validation A Better Approach For Improving Performance of Machine Learning Model

Kuldeep Nargave
M Tech (CSE) 4th semester
Computer Science and Engineering Department
L.N.C.T. (Bhopal) Indore Campus Indore M.P.

Mr Nilesh Avinash Joshi
Assistant Professor
Computer Science and Engineering Department
L.N.C.T. (Bhopal) Indore Campus Indore M.P.

Abstract: ***Cross-Validation is just a method that simply reserves a part of data from the dataset and uses it for testing the model (Validation set), and the remaining data other than the reserved one is used to train the model. Leave one out cross validation. There are several reasons that why we are motivated towards K-fold Cross Validation. We have a lot of readily available data that can explain a lot of things and help us identify hidden patterns. However, if we only have a small dataset, splitting it into a training and test dataset at a 80:20 ratio respectively doesn't seem to do much for us. However, when using K-fold cross validation, all parts of the data will be able to be used as part of the testing data. k-fold cross validation, Stratified k-fold cross validation, Time Series cross validation, Leave one out cross validation. In the proposed work we used K fold cross validation and split the dataset into training data and test data. Evaluating the model and determining its accuracy. We want to determine whether the model is over- or under-fitted. Evaluation finally determining the quality of the model. By using k-fold cross-validation, we are able to "test" the model on k different data sets, which helps to ensure that the model is generalizable. Finally, it lets us choose the model which had the best performance.***

Keywords: Cross-Validation, Training data, Test data, Accuracy, Performance

I. Predictive Modeling

Machine learning model performance assessment is just like assessing the scores, how we used to evaluate our sores in high schools and colleges for the meeting the eligibility criteria for getting the best courses or getting selected in the campus interviews for companies for the job and clearing cut-off scores for many more competition exams for getting selected. So apparently, the GOOD score recognizes the fact that the candidate is always good. The same is been expected in the machine learning model, and that should achieve the expected results in predictions/forecasting/calcification problem statements. Even in the ML world, the model has been trained in the context of data, model, and code. As we are aware that there are multiple ways to evaluate the performance of the model. As a team always has the responsibility to build a generalized model and certain performance is expected in the production, at the same time we have to convince the customer, stakeholders of the same and add the key business benefits based on the advisory committee by SME's guidance to meet the goals[1,11].

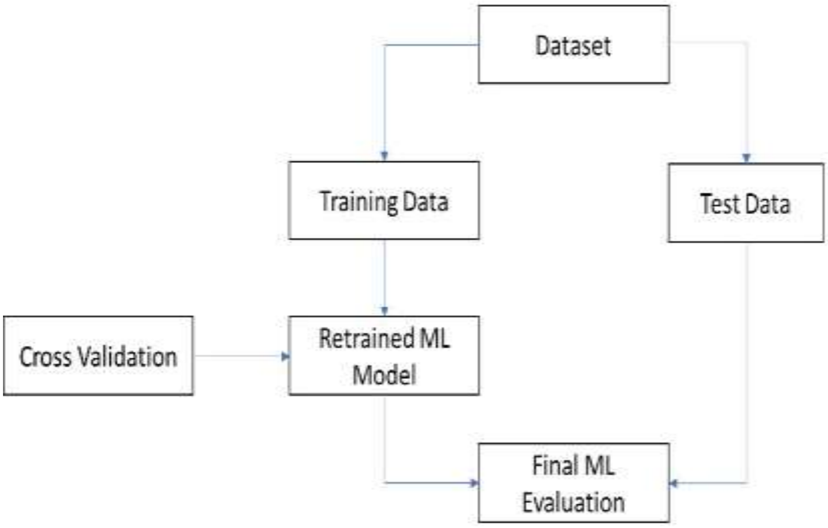


Figure 1 Cross validation data flow

As we are an ML engineer team, we must provide the performance of the model in the numeric range. Let's say the performance of the model would be 85-90%. Sometimes the performance of the model in training and testing will not behave the same in production, in many cases, Overfitting or Underfitting will be experienced during the production environment. Accuracy is the just number, for getting a better understanding of a prediction-based problem that corrects the predictions which are made by the model built by the team with the available number of records. So we need to train the model across different combinations of data.

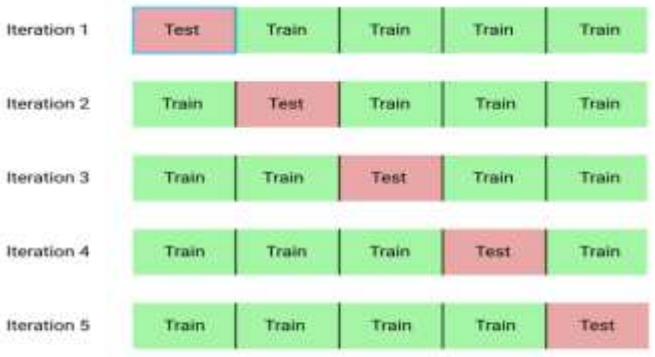
II. Motivation

There are several reasons that why we are motivated towards K-fold Cross Validation. We have a lot of readily available data that can explain a lot of things and help us identify hidden patterns. However, if we only have a small dataset, splitting it into a training and test dataset at a 80:20 ratio respectively doesn't seem to do much for us. However, when using K-fold cross validation, all parts of the data will be able to be used as part of the testing data. This way, all of our data from our small data set can be used for both training and testing, allowing us to better evaluate the performance of our model. Splitting dataset into training and testing as a 80:20 ratio or a train-test split provides us with only one result to refer to in our evaluation of the model. We are not sure about the accuracy of this result, if it's due to chance, the level of bias, or if it actually performed well. However, when using k-fold cross validation, we have more models that will be producing more results. For example, if we chose our k value at 10, we would have 10 results to use in our evaluation of the model's performance. If we were using accuracy as our measurement; having 10 different accuracy results where all of the data was used in the test phase is always going to be better and more reliable than using one accuracy result that was produced by a train-test split - where all the data wasn't used in the test phase. You would have more trust and confidence in model's performance if the accuracy outputs were 94.0, 92.8, 93.0, 97.0, and 94.5 in a 5-fold cross validation than a 93.0 accuracy in a train-test split. This proves to us that our algorithm is generalizing and actively learning and is providing consistent reliable outputs[10].

In a random train-test split, we assume that the data inputs are independent. Let's further expand on this. Let's say we are using a random train-test split on a speech recognition dataset for British English speakers who reside from London. There are 5 speakers, with 250 recording each. Once the model performs a random train-test split, both the training and testing dataset will learn from the same speaker, which will be saying the same dialogue. Using K-fold cross validation will allow you to train 5 different models, where in each model you are using one of the speakers for the testing dataset and the remaining for the training dataset. This way, not only can we evaluate the performance of our model, but our model will be able to perform better on new speakers and can be deployed in production to produce similar performance for other tasks[8].

III. K fold Cross Validation

K-fold cross-validation is defined as a method for estimating the performance of a model on unseen data. This technique is recommended to be used **when the data is scarce** and there is an ask to get a good estimate of training and generalization error thereby understanding the aspects such as underfitting and overfitting. This technique is used for hyperparameter tuning such that the model with the most optimal value of hyperparameters can be trained. It is a **resampling technique without replacement.** The **advantage** of this approach is that each example is used for training and validation (as part of a test fold) exactly once. This yields a **lower variance estimate of the model performance** than the holdout method. This technique is used because it helps to avoid overfitting, which can occur when a model is trained using all of the data. By using k-fold cross-validation, we are able to "test" the model on k different data sets, which helps to ensure that the model is generalizable[9].



Iteration 1	Test	Train	Train	Train	Train
Iteration 2	Train	Test	Train	Train	Train
Iteration 3	Train	Train	Test	Train	Train
Iteration 4	Train	Train	Train	Test	Train
Iteration 5	Train	Train	Train	Train	Test

Figure 2 Working K-fold cross-validation**.**

IV. Literature Survey

In 2018 Yoonsuh Jung proposed **"Multiple predicting K-fold cross-validation for model selection"**. They proposed a new CV method is proposed within the framework of K-fold CV. The proposed method uses (K − 1) folds of the data for model validation, while the other fold is for model construction. This provides (K − 1) predicted values for each observation. The traditional K-fold CV can be improved by the suggested MPCV by reducing the variation in the validation error. In turn, this elevates the variation in the model construction. They provide some guidance to select an appropriate value of K. However, it is not easy to provide a universal rule for the choice of K, since it depends on the sample size, number of parameters, structure of data, and so on. A naive rule is to choose K such that

$K \approx \log(n)$ and $n/K > 3d$. The second condition means that we need certain amount of data for the model construction to capture the complexity of the data structure reasonably well [1].

In 2019 Nagadevi Darapureddy, Nagaprakash Karatapu, proposed **"Research of Machine Learning Algorithms using K-Fold Cross Validation"**. In machine learning, Classification is one of the most important research area. Classification allocates the given input to a known category. In this paper different machine algorithms like Logistic LR), Decision tree (DT), Support vector machine (SVM), K nearest neighbors (KNN) were implemented on UCI breast cancer dataset with preprocessing. The models were trained and tested with k-fold cross validation data. Accuracy and run time execution of each classifier are implemented in python it can be observed that the accuracy for classifying the dataset using decision tree classifier is more when compared with other classifiers using k-fold cross validation. Other cross validation techniques like repeated kfold, leave one out (LOO) can be implemented. Further the remaining learning algorithms like Gradient Descent, Nesterov accelerated gradient, Adagrad, Adam, Adadelta, momentum-based gradient descent can be applied to deep networks to compare the accuracy [2].

In 2019 Andrius VabalasID, Emma Gowen, Ellen Poliakoff , proposed **"Machine learning algorithm validation with a limited sample size".** High dimensional data with a small number of samples is of critical importance for identifying biomarkers and conducting feasibility and pilot work, however it can lead to biased machine learning (ML) performance estimates. They applied ML to predict autistic from non-autistic individuals showed that small sample size is associated with higher reported classification accuracy. They investigated whether this bias could be caused by the use of validation methods which do not sufficiently control overfitting. Simulations show that K-fold Cross-Validation (CV) produces strongly biased performance estimates with small sample sizes, and the bias is still evident with sample size of 1000. Nested CV and train/test split approaches produce robust and unbiased performance estimates regardless of sample size. The results suggest how to design robust testing methodologies when working with small datasets and how to interpret the results of other studies based on what validation method was used[3].

In 2020 Sitefanus Hulu , Poltak Sihombing , Sutarman proposed **"Analysis of Performance Cross Validation Method and K-Nearest Neighbor in Classification Data"**. To produce data classifications that have data accuracy or similarity in proximity of a measurement result to the actual numbers or data, testing can be done based on accuracy with test data parameters and training data specified by Cross Validation. The data sharing with Cross Validation has better data recognition with a percentage of 100%. The results of the K-NN test results in the classification of data using iris data sets using variation test values 3, 4, 5, 6, 7, 8, 9, have 100% percentage accuracy with 75 true amount of data and 0 incorrect amount of data.. The data sharing with Cross Validation has better data recognition with a percentage of 100%. K-NN test results in data classification using iris data sets using variation test values of 3, 4, 5, 6, 7, 8, 9, have a percentage accuracy of 100% with 75 correct amount of data and 0 incorrect amount of data. Percentage of variation in the value of K K-Nearest Neighbor 3,4,5,6,7,8,9. and variations in the number of K-Fold 1,2,3,4,5,6,7,8,9,10. has a percentage of 100% on K-Fold 4 and 7[4].

In 2021 Muhammad Asrola, Petir Papilob, Fergyanto E Gunawan proposed **"Support Vector Machine with K-fold Validation to Improve the Industry's Sustainability Performance Classification"**. The SVM model was enriched by the model tuning and k-fold validation to enhance the model performance. Our previous research in bioenergy industry inspired us to develop an accurate model for sustainability performance classification and improved Multi-Dimensional Scaling (MDS) model which were commonly applied. The result showed that in the model training stage, SVM with polynomial model had the highest accuracy to classify sustainability performance. Ten folds validation with cost (4), gamma (0.25) and coef0 (16) as tuning parameter performed 98.32% of accuracy in data testing. This result had proof that SVM with polynomial kernel model was able to classify sustainability performance accurately. This model is potentially substituted previous common models in industry's sustainability. performed 98.32% of accuracy in data testing. This result had proof that SVM with polynomial kernel model was able to classify sustainability performance accurately. This model is potentially substituted previous common models in industry's sustainability assessment which were not adaptive and less accurate[5].

In 2021 Kwanele Phinzi , Dávid Abriha and Szilárd Szabó proposed "**Classification Efficacy Using K-Fold Cross-Validation and Bootstrapping Resampling Techniques on the Example of Mapping Complex Gully Systems".** Their main objective of this work was to investigate the performance of support vector machines (SVM) and random forest (RF) algorithms in extracting gullies based on two resampling methods: bootstrapping and k-fold cross-validation (CV). They used Planet Scope data, acquired during the wet and dry seasons. Results revealed that gullies had significantly different ($p < 0.001$) spectral profiles from any other land cover class regarding all bands of the Planet Scope image, both in the wet and dry seasons. However, NDVI was not efficient in gully discrimination. Based on the overall accuracies, RF's performance was better with CV, particularly in the dry season, where its performance was up to 4% better than the SVM's. Nevertheless, class level metrics showed that SVM combined with CV was more successful in gully extraction in the wet season. They found the following outcomes. • Gullies were spectrally different in all bands of the Planet Scope images, both in the dry and the wet seasons. NDVI values did not differ from all land cover classes regarding the reflectance values; thus, it was not involved in gully classification [6].

In 2022 Zeyang Lin , Jun Lai , Xiliang Chen , Lei Cao and Jun Wang proposed **"Curriculum Reinforcement Learning Based on K-Fold Cross Validation".** They proposed a curriculum reinforcement learning method based on K-Fold Cross Validation that can estimate the relativity score of task curriculum difficulty. Drawing lessons from the human concept of curriculum learning from easy to difficult, this method divides automatic curriculum learning into a curriculum difficulty assessment stage and a curriculum sorting stage. To solve the problems of complex curriculum sorting and slow convergence in curriculum reinforcement learning, they proposes a curriculum reinforcement learning method based on K-Fold Cross Validation, which can automatically sort curriculums through curriculum difficulty assessment and curriculum sorting without relying on expert experience, and it can be applied in multi-agent deep reinforcement learning algorithm based on replay buffer space. Through simulations in the cooperative environment and the adversarial environment, the usability and superiority of the algorithm was proven, which is of solid research and practical significance[7].

In 2022 Kaiyu Suzuki , Yasushi Kambayashi and Tomofumi Matsuzawa proposed **"CrossSiam: k-Fold Cross Representation Learning" .** They applied k-fold cross validation to the task of classifying images using deep learning, which is a method that compares and evaluates models appropriately model of a given problem; this technique is easy to understand and easy to implement, and it produces results in lower bias estimates. However, k-fold cross validation reduces the amount of data per neural network, which reduces the accuracy. The approach can be very important in the field where reliability is required, such as automated vehicles and drones in disaster situations. In this study, we propose a method to improve the reliability of multi-agents such as unmanned robots and drones by making autonomous decisions based on camera information. They apply k-fold cross validation to representation learning as a baseline. [8].

In 2022 Sashikanta Prusty, Srikanta Patnaik and Sujit Kumar Dash proposed **"SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer".** Cancer is the unregulated development of abnormal cells in the human body system. Cervical cancer, also known as cervix cancer, develops on the cervix's surface. This causes an overabundance of cells to build up, eventually forming a lump or tumour. As a result, early detection is essential to determine what effective treatment we can take to overcome it. Therefore, the novel Machine Learning (ML) techniques come to a place that predicts cervical cancer before it becomes too serious. The proposed system used stratified k-fold analysis and CV techniques to provide medical practitioners with the data they needed to make better diagnostic decisions. The best performing of four supervised machine learning algorithms was implemented on the proposed framework. They discovered that the model RF6 scored 98.10 percent for Hinselmann, 95.80 percent for Schiller, RF8 scored 97.49 percent for Cytology, and RF9 scored 97.95 percent for Biopsy. The most significant contribution of this work is the integration of a more robust cervical cancer [9].

V. PRAPOSED APPROACH

We should train the model on a large portion of the dataset. Otherwise, we'll fail to read and recognise the underlying trend in the data. This will eventually result in a higher bias. We also need a good ratio of testing data points. As we have seen above, less amount of data points can lead to a variance error while testing the effectiveness of the model.

We should iterate on the training and testing process multiple times. We should change the train and test dataset distribution. This helps in validating the model effectiveness properly Do we have a method which takes care of all these 3 requirements. That method is known as "k-fold cross validation". It's easy to follow and implement. Below are the steps for it:

1. Randomly split your entire dataset into k"folds"
2. For each k-fold in your dataset, build your model on k – 1 folds of the dataset. Then, test the model to check the effectiveness for kth fold
3. Record the error you see on each of the predictions.
4. Repeat this until each of the k-folds has served as the test set!
5. The average of your k recorded errors is called the cross-validation error and will serve as your performance metric for the model.

Now, one of most commonly asked questions is, "How to choose the right value of k

VI. COMPARATIVE ANALYSIS

Comparing Accuracy scores for all 5 Folds

Table 1 Comparing Accuracy scores for all 5 folds.

Fold. No.	Accuracy
1	0.88122807
2	0.82105263
3	0.81736842
4	0.84736842
5	0.83690265

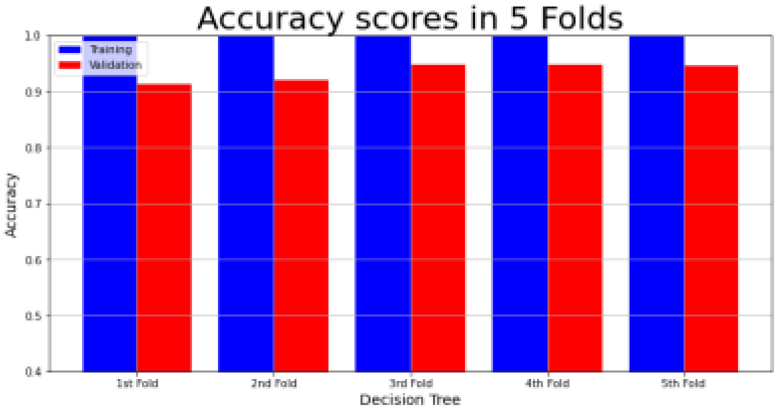


Figure 3 Accuracy scores for 5 Folds

CONCLUSION

Creating datasets to train and validate our model from data collection is the most common machine learning approach to increase the model's performance. The split ratio of the dataset could be 70 : 30 or 80 : 20. The holdout approach is the most common cross-validation approach. The issue with this approach is that we are unsure whether a good validation accuracy score of the model denotes a good model. Cross-Validation is just a method that simply reserves a part of data from the dataset and uses it for testing the model (Validation set), and the remaining data other than the reserved one is used to train the model, Leave one out cross validation, k-fold cross validation ,Stratified k-fold cross validation, Time Series cross validation K-fold cross-validation is defined as a method for estimating the performance of a model on unseen data. This technique is recommended to be used when the data is scarce and there is an ask to get a good estimate of training and generalization error thereby understanding the aspects such as underfitting and overfitting

REFERENCE

1. Yoonsuh Jung “Multiple predicting K-fold cross-validation for model selection” Journal of Nonparametric StatisticsISSN: 1048-5252 (Print) 1029-0311 (Online) Journal homepage: https://www.tandfonline.com/loi/gnst20.
2. Nagadevi Darapureddy, Nagaprakash Karatapu, “Research of Machine Learning Algorithms using K-Fold Cross Validation International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-8 Issue-6S, August 2019
3. Andrius VabalasID, Emma Gowen , Ellen Poliakoff „Machine learning algorithm validation with a limited sample size” https://doi.org/10.1371/journal.pone.0224365 November 7, 2019.
4. Sitefanus Hulu , Poltak Sihombing , Sutarman Analysis of Performance Cross Validation Method and K-Nearest Neighbor in Classification Data International Journal of Research and Review Vol.7; Issue: 4; April 2020 Website: www.ijrrjournal.com Research Paper E-ISSN: 2349-9788; P-ISSN: 2454-2237.
5. Muhammad Asrola,, Petir Papilob, Fergyanto E Gunawan Support Vector Machine with K-fold Validation to Improve the Industry’s Sustainability Performance Classification 5th International Conference on Computer Science and Computational Intelligence 2020 ScienceDirect Available online at www.sciencedirect.com Procedia Computer Science 179 (2021) 854–862
6. Kwanele Phinzi , Dávid Abriha and Szilárd Szabó Classification Efficacy Using K-Fold Cross-Validation and Bootstrapping Resampling Techniques on the Example of Mapping Complex Gully Systems Bootstrapping Resampling Techniques on the Example of Mapping Complex Gully Systems. Remote Sens. 2021, 13, 2980. https:// doi.org/10.3390/rs13152980.
7. Zeyang Lin , Jun Lai , Xiliang Chen , Lei Cao and Jun Wang Curriculum Reinforcement Learning Based on K-Fold Cross Validation © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons
8. Kaiyu Suzuki, Yasushi Kambayashi and Tomofumi Matsuzawa CrossSiam: k-Fold Cross Representation Learning Suzuki, K., Kambayashi, Y. and Matsuzawa, T. CrossSiam: k-Fold Cross Representation Learning. DOI: 10.5220/0010972500003116 In Proceedings of the 14th International Conference on Agents and Artificial Intelligence (ICAART 2022) - Volume 1, pages 541-547 ISBN: 978-989-758-547-0; ISSN: 2184-433X Copyright c 2022 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved.
9. Sashikanta Prusty, Srikanta Patnaik and Sujit Kumar Dash SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer rusty S, Patnaik S and Dash SK (2022), SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer. Front. Nanotechnol. 4:972421. doi: 10.3389/fnano.2022.972421.
10. Jerzy Wieczorek | Cole Guerin | Thomas McMahon K-fold cross-validation for complex sample surveys Received: 30 July 2021 Accepted: 9 January 2022 DOI: 10.1002/sta4.454 2022 The Authors. Stat published by John Wiley & Sons Ltd.
11. Isaac Kofi Nti and Owusu Nyarko-Boateng Performance of Machine Learning Algorithms with Different K Values in K-fold CrossValidation I.J. Information Technology and Computer Science, 2021, 6, 61-71 Published Online December 2021 in MECS (http://www.mecs-press.org/) DOI: 10.5815/ijitcs.2021.06.05.