Practical Machine Learning

Auto-sklearn fail

Bimal Pandey

Introduction: Auto-sklearn is an automated machine learning (AutoML) tool that aims to simplify the process of building machine learning models. Auto-sklearn automates the selection of the best machine learning model, data preprocessing, and hyperparameter tuning. This automation saves time and effort, particularly for non-machine learning professionals. Auto-sklearn uses Bayesian optimization to efficiently search the hyperparameter space incorporates meta-learning to learn from previous experiments and apply this knowledge to new datasets. Meta-learning helps auto-sklearn to initialize the search space efficiently and guide the optimization process.

Analysis of the code: In the given code the the performance of Autosklearn in terms of accuracy is 0.63 which is outperformed by the Random Forest with its default hyperparameters gaining accuracy of 0.67. I found that the code in the fail.py uses default resampling strategy and search space is not specified. However, Autosklearn uses large space by default which might not be the actual reason behind the poor performance. Then, I implemented the resampling strategy using cross-validation which increased the performance of model. In this problem, the use of cross validation proved to be beneficial. The use of cross validation decreases the overfitting and helps in better performance.

Furthermore, the number of folds in cross validation is varied to examine the accuracy level of the Autosklearn to validate whether cross validation is really impacting the performance. After some trails, I found that the performance of model was changed and all of these results were better than previous and they also outperformed the RF with default hyperparameters.

3- fold= 0.66

5- fold= 0.698

6 -fold= 0.68

10- fold= 0. 6789

Additionally, I tried measuring the performance using standard Scaler and one hot encoder but it didn't prove to be beneficial or had minimal impact on performance. From my experiment and knowledge, I can conclude that the dataset do not have scaling problem. Also the hyperparameter search need to be adjusted according to the dataset and their nature so that better performance can be achieved. Similarly, I noticed time also changes the performance of this model.