



4222-SURYAGROUPOF INSTITUTION

VIKRAVANDI-605652

NAANMUDHALVANPROJECT

EARTHQUAKEPREDICTIONMODELUSINGPYTHON

## AI\_PHASE2:

Consideradvancedtechniquessuchashyperparametertuningandfeature

Engineeringtoimprovethepredictionmodel’sperformance.

PREPAREDBY

J. DINESHKUMAR REGNO:422221106302 DEP:BE.ECE

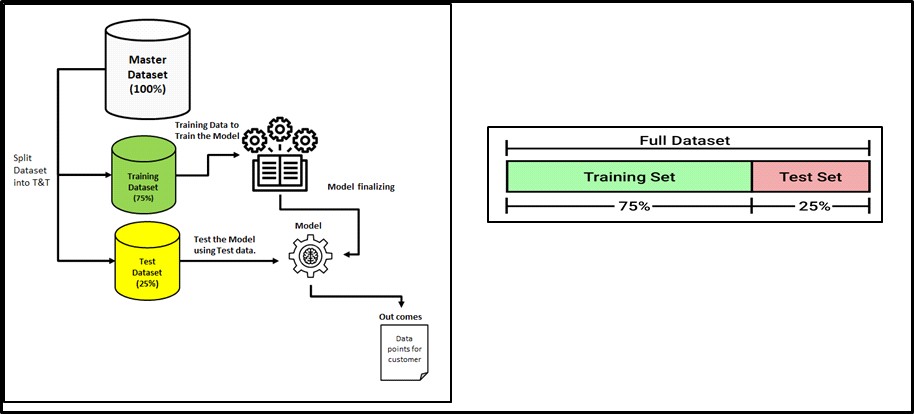
## INTRODECTION:

EveryML Engineer andDataScientist mustunderstand thesignificanceof **“HyperparameterTuning.** whileselecting your rightmachine/deep learning model and improving the performance of the model(s).Make it simple, for every single machine learningmodelselection isamajorexerciseanditispurelydependentonselectingtheequivalent set of hyperparameters, and all these are indispensable to train a model. It is always referring to the parameters of the selected model and be remember it cannot be learnt from the data, and it needs to be provided before the model gets into the training stage, ultimately the performance of themachinelearningmodelimproveswithamoreacceptablechoiceofhyperparametertuningand selectiontechniques.Themainintentionofthisarticleistomakeyouallawareofhyperparameter tuning.

HYPERPARAMETERTUNING:

Asweknowthatthereareparametersthat areinternally learnedfrom the given dataset andderivedfrom thedataset,they arerepresented in making predictions, classification and etc., These are so-called **Model Parameters,** and they are varyingwithrespecttothenatureofthedatawecouldn’tcontrolthissinceitdependsonthedata. Like ‘**m**‘ and ‘**C**‘ in linear equation, which is the value of coefficients learned from the given dataset.Some set of parameters that are used to control the behaviour of the model/algorithm and adjustable in order to obtain an improvised model with optimal performance is so-called **Hyperparamer.**

**HYPERPARAMETERLIFECYCLE:**



# Hyperparameter Space

Asweknowthatthereis alistofHPsforanyselectedalgorithm(s)andourjobis tofigureoutthe bestcombinationofHPsandtogettheoptimalresultsbytweakingthemstrategically,thisprocess willbeproviding us with theplatformfor **HyperparameterSpace** and this combination leads to providethebestoptimalresults,nodoubtinthatbutfindingthiscomboisnotsoeasy,wehaveto search throughout the space. Here every combination of selected HP value is said to be the “**MODEL**” and have to evaluate the same on the spot. For this reason, there are two generic approaches to search effectively in the HP space are **GridSearch CV** and **RandomSearch CV.** Here CV denotes **Cross-Validation.**

### DATA LEAKAGE:

Well!NowquicklywillunderstandwhatisDataleakageinML,thisismainlyduetonotfollowing someoftherecommendedbestpracticesduringtheDataScience/MachineLearninglifecycle.

Theresulting

is Data Leakage, that’s fine, what is the issue here, after successful testing with perfect accuracy followed by training themodelthen themodelhas been planned to moveinto production. Atthis moment ALL Is Well.

Still,theactual/real-timedataisappliedtothismodelintheproductionenvironment,youwillget poor scores. By this time, you may think that why did this happen and how to fix this. This is all because of the data that we split data into training and testing subsets. During the training the model has the knowledge of data, which the model is trying to predict, this results in inaccurate and bad prediction outcomes after the model is deployed into production.

### Causesof DataLeakage

DataPre-processing

ThemajorrootcauseisdoingallEDAprocessesbeforesplittingthedatasetintotestand train

<!--[endif]-->Doingstraightforwardnormalizingorrescalingon agivendataset

<!--[endif]-->PerformingMin/Maxvaluesofafeature

<!--[endif]-->Handlingmissingvalueswithoutreservingthetestandtrain

<!--[endif]-->RemovingoutliersandAnomalyonagiven dataset

<!--[endif]-->Applyingstandardscaler,scalin

# Data Leakage

Well!NowquicklywillunderstandwhatisDataleakageinML,thisismainlyduetonotfollowing someoftherecommendedbestpracticesduringtheDataScience/MachineLearninglifecycle.

Theresulting

is Data Leakage, that’s fine, what is the issue here, after successful testing with perfect accuracy followed by training themodelthen themodelhas been planned to moveinto production. Atthis moment ALL Is Well.

Still,theactual/real-timedataisappliedtothismodelintheproductionenvironment,youwillget poor scores. By this time, you may think that why did this happen and how to fix this. This is all because of the data that we split data into training and testing subsets. During the training the model has the knowledge of data, which the model is trying to predict, this results in inaccurate and bad prediction outcomes after the model is deployed into production.

### Causesof DataLeakage

DataPre-processing

ThemajorrootcauseisdoingallEDAprocessesbeforesplittingthedatasetintotestand train

<!--[endif]-->Doingstraightforwardnormalizingorrescalingon agivendataset

<!--[endif]-->PerformingMin/Maxvaluesofafeature

<!--[endif]-->Handlingmissingvalueswithoutreservingthetestandtrain

<!--[endif]-->RemovingoutliersandAnomalyonagiven dataset

<!--[endif]-->Applyingstandardscaler,scaling,assertnormaldistributiononthefulldataset

Bottomlineis,weshouldavoiddoinganythingtoourtrainingdatasetthatinvolveshaving knowledgeofthetestdataset.Sothatourmodelwillperforminproductionasageneralisedmodel.

will go through the available Hyperparameters across the various algorithms and how we could implement all these factors and impact the model.

StepstoPerformHyperparameterTuning

Selecttherighttypeofmodel.

<!--[endif]-->ReviewthelistofparametersofthemodelandbuildtheHP space

<!--[endif]-->Findingthemethodsforsearchingthehyperparameterspace

<!--[endif]-->Applyingthecross-validationschemeapproach

## GRIDESEARCH:

The**GridSearch**onethatwehavediscussedabove

usuallyincreasesthecomplexityintermsofthecomputationflow,SosometimesGSisconsidered inefficient since it attempts all the combinations of given hyperparameters. But the **Randomized Search**isusedtotrainthemodelsbasedonrandomhyperparametersandcombinations.obviously, the number of training models are small column than grid search.

### Insimpleterms,InRandomSearch,inagivengrid,thelistofhyperparametersare trained and test our model on a random combination of given hyperparameters.

GettingRandomForestClassifierobjectformyoperation.

fromsklearn.model\_selectionimportRandomizedSearchCV from sklearn.ensemble import RandomForestClassifier

fromscipy.statsimportrandintassp\_randint

AssigningmyTrainandTestspilttomyRandomForestClassifier object

#builda RandomForestClassifier

clf= RandomForestClassifier(n\_estimators=50)

Specifyingthelistofparametersand distributions

param\_dist={"max\_depth":[3,None],

"max\_features":sp\_randint(1,11),

"min\_samples\_split":sp\_randint(2,11),

"min\_samples\_leaf":sp\_randint(1,11), "bootstrap": [True, False], "criterion": ["gini", "entropy"]}

Definingthesample,distributionsandcross-validation

samples=8#numberofrandomsamples

randomCV = RandomizedSearchCV(clf, param\_distributions=param\_dist, n\_iter=samples,cv=3)

Allparameters aresetand,let’s dothefitmodel

randomCV.fit(X, y) print(randomCV.best\_params\_)

Output

{'bootstrap': False, 'criterion': 'gini', 'max\_depth': 3, 'max\_features': 3, 'min\_samples\_leaf': 7, 'min\_samples\_split': 8}

AspertheCross-Validationprocess,willfigureoutthemeanandgettheresults

randomCV.cv\_results\_['mean\_test\_score']

Output

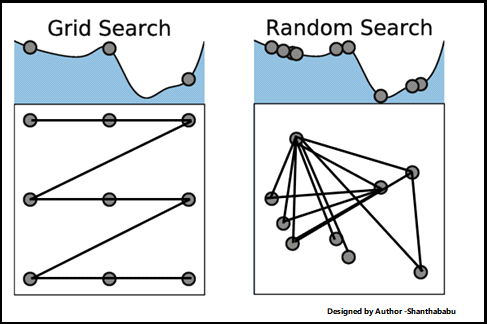
array([0.73828125,0.69010417,0.7578125,0.75911458,0.73828125,

nan, nan,0.7421875]) Best accuracy from training print(randomCV.score(X\_test,y\_test)) Output

0.8744588744588745

Youmayhaveaquestion,nowwhichtechniqueisbesttogo.Thestraightansweris RandomSearshCV, let’s

Theblowpictorialrepresentationwouldgiveyouthebestunderstandingof GridSearchCV and RandomSearshCV



# Conclusion

Sofar wehavediscussed in adetailed study of Hyperparameter visions with respectto the Machine Learning point of view, please remember a few things.

Eachmodelhasasetofhyperparameters,sowehavecarefullychosenthemandtweakedthem during hyperparameter tuning. I mean building the HP space.

AllhyperparametersareNOTequallyimportantandnodefinedrulesforthis.trytouse continuous values instead of discrete values.

MakesuretouseK-FoldwhileusingHyperparametertuningtoimproviseyour hyperparameter tuning and coverage of hyperparameter space.

Gowithabettercombinationforhyperparametersandbuildstrongresults.