*Recap:*

*Why we do Feature Selection:*

*1.) Higher Dimension mai bht saare ML algos ka performance khrbh hojata*

*2.) Complexity*

*3.) Interpretebility*

*3.) Inferece mai prblm krna , Inference mtlb ( Input & output ka relation smjhne )*

*Filter Method:*

*1 baar mai 1 Single feature uthate hai ,*

*& statistics use krke uska importance nikalne ki koshish krte hai.*

*Problem with this method is:*

*1 baar mai 1 single feature/column utharahe hai*

*Like feature 1 ko study krrhe hai & feature 2 ko study krrhe hai*

*,but hum feature 1 & feature 2 dono ko sath mai study nahi krrhe hai*

***E.g*** *jaise latitude & longitude column hai*

*Dono ko akele akele study krenge tou imp nahi lagega*

*But dono ko sath mai*

*Study krenge tou Importance smjhega.*

***Wrapper Method.***

*F1, F2, F3,……….Fn tak Feature hai*

*What we do yeah n-feature se Subset Generate krte hai.*

*F1,F2,F3, y*

*Har Subset ke upar ML algo apply krte hai.*

*R2 Score measure krte hai*

*Then*

*Jish subset ka accuracy ya matrix Highest hota hai*

*Vo combination & feature ko select krte hai.*

*Step 1 Subset Generation*

*Step 2 Subset ko Evaluate krrhe hai ML Algo lagake & Result nikalrahe hai, Then We compare*

*Steps:*

1. *Subset Generation:*

*First a subset of feature is generated. This can be done in a variety of ways.*

*E.g:*

*You might start with one feature and gradually add more, or start with all feature & gradually remove them, or generate subsets of features randomly.*

*The subset generation method depends on the specific*

*Type of wrapper method being used.*

1. *Subset Evaluation:*

*After a subset of features has been generated, a model is trained on this subset of features and the model’s performance is evaluated, usually through cross-validation. The performance of the model gives as estimate of the quality of feature in subset.*

1. *Stopping Criterion:*

*This process is repeated , generating and evaluated different subset of features until some stopping criterion is met. This could be certain number of subset evaluated.*

*( Kuch jageh hie use hota hai yeah )*

*There are 4 Techniques:*

1. *Exhaustive Feature Selection or Best Subset Selection*
2. *Forward Selection*
3. *Backward Elimination*
4. *Recurssive Feature Elimination*
5. *Exhaustive Feature Selection or Best Subset Selection*

*In filter method -> inidvidully cols ko study krskte hai*

*& combination ka prblm tha ki relations of 2 or more colms together vo ignore horaha hai.*

***Kam number of columns pe guranttee ache results deta hai***

*I have 2 colms*

*F1,f2,y*

*Mujhe best input cols pata krna hai*

*F1, y F2,y F1,F2,y*

*Teeno mai ML apply kiya :*

*0.6L 0.51 0.72*

*Maanlo 3 Columns hai or 3 features hai*

*f1 f2 f3 y*

*f1 y f2 y f3y ek model*

*f1 f2 y f2f3y f1f3y ek model aur*

*fir*

*f1 f2 f3 y se Ek aur Model*

*Sbh subsets pe 1 hie koisabhi ML Algo lagayenge*

*Then jiska best score hoga vohi subset ko rkhenge baki sbhko hatadenge.*

*Disadvantages:*

1. *Computational Complexity*

*N columns hai tou N number of Models Train karne padenge yeahisbhse bada disadvantage hai.*

*E.g jo last filter method sikhte time data liya tha*

*Usme 563 columns theah , so you can imagine the numbers of Model, it’s insane*

*This is the Biggest Flaw.*

1. *Risk of Overfitting:*

*By Checking all possible combinations of features, there’s a risk of overfitting the model, to the training data.*

*The feature combination that performs best on training data may not necessarily perform well on unseen data.*

1. *Requires a Good Evaluation Metric:*

*The effectiveness of Exhaustive metric used to assess the goodness of a feature depends on the quality of the evaluation metric used to assess the goodness of a feature subset.*

*If a poor metric is used, the feature selection may not produce optimal result.*

***Jaise feature badhte jayega generally mean r2 score aur zyaada rhega , it is not a good thing, in that scenario***

***We must use adjusted r2 score.***

*As data jo hai vo alag alag scale pe hai so , Linear Regression mai ese kehte hai ki 1 scale pe data laane*

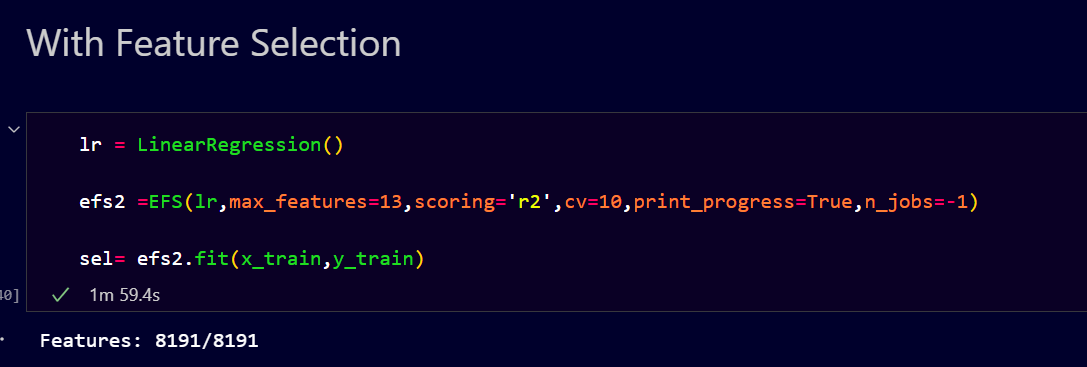
*so using Standard Scalar , saare data ko 1 hie scale pe leeayenge.*

from sklearn.preprocessing import StandardScaler

sc = StandardScaler()

x\_train = sc.fit\_transform(x\_train)

x\_test = sc.transform(x\_test)

**

*n-jobs jo hai vo basically cores ka kaam krta hai kindoff cores mai parallely execution karata hai bade task ko ishse time thoda kamm lagta hai*

***Don’t Trust r2 Score with High Number of Columns***

***Better to apply***

***Filter 1st & then Wrapper Method.***

***Bad Part:***

***Bade dataset pe apply nahi hoga , 20 Cols kebaad roone lagega .***

*Yeahi Problem Solve karneke liye hum use krte hai*

*Backward Elimination & Forward elimination*

*Kahi jageh pe backward elimination ko hie Sequential backward elimination bollte hai.*

***E.g***

***4 feature hai***

***Chaaro ka combination use krke model ko train krte hai***

***& we get accuracy of 0.89***

***Jaise iska naam hai Backward Elimination***

***& fir hum columns hatate jaate hai***

***Backward Elimination or Sequential backward elimination***

***F1 F2 F3 F4 model training 🡪 0.89***

***Then what we do ki***

***4 Models train krte hai***

*F1* ***F2 F3 F4 model training 🡪 0.81***

***F1*** *F2* ***F3 F4 model training 🡪 0.71***

***F1 F2*** *F3* ***F4******model training 🡪 0.91***

***F1 F2 F3*** *F4* ***model training 🡪 0.65***

***Which is giving best result ?***

***As we can see F1, F2, & F4 giving us best result & f3 ko remove krna sbhse safe hai.***

***As it is iterative process will repeat this again….***

*F1*  ***F2 F4 model training 🡪 0.79***

***F1****F2* ***F4 model training -> 0.81***

***F1 F2*** *F4* ***model training -> 0.83***

***Now we can see best result is f1 & f2 so will remove f4***

***F1*** *F2 model training -> 0.53*

*F1*  ***F2 model training -> 0.63***

***F2 giving Best Result***

***Iteration utne hie honge jitne Columns honge***

***Abh Hum kya krrhe hai ki***

***Har iteration ka Best Result likhrahe hai***

***F1 F2 F3 F4 🡪 0.89***

***F1 F2 F4 🡪 0.91***

***F1 F2 -> 0.83***

***F2 🡪 0.63***

*Now what we will do is yeah chaaro mai Comparision karenge*

*& jiska best result hoga ushe Select karenge*

*e.g above*

*f1,f2,f4 is best feature subset*

*so we will train our model on this subset.*

*This is how sequential Backward Elimination Works.*

***Benefits:***

***This is Much Faster.***

***Disadvantages:***

***We can miss the best combination.***

***Local Best miljayega global best miss hoga.***

***As process hie iska hai ki esa hoga hie***

***It is the only problem with this approach***

***Forward Selection***

***F1 F2 F3 F4 -> yaha hum zero 0 se start krte hai***

***Y ka mean lete hai***

***Add karna chalu krte hai***

***F1 -> 0.63***

***F2 -> 0.51***

***F3 -> 0.43***

***F4 -> 0.49***

***Best F1 hai ismese***

***Abh aur Column add krna hai***

***F1 Kesath 1-1 krke add krenge***

***F1 F2 -> 0.63***

***F1 F3 -> 0.71***

***F1 F4 -> 0.80***

***Yaha best f1 & f4 mila***

***F1 F4 F2 🡪 0.81***

***F1 F4 F3 -> 0.85***

***F1, F4 & F3 best mila ismese***

***Then now 1 hie add krna hai***

***F1, F4 ,F3 ,F2 -> 0.83***

*Now we will check all iteration & see which is giving best result*

***F1 -> 0.63***

***F1 F4 -> 0.80***

***F1 F4 F3 -> 0.85***

***F1, F4 ,F3 ,F2 -> 0.83***

***Abh ismese sbhse best***

***F1 , F4 & F3 hai then***

***We will go with f1,f4 & f3 for our Model.***

***Both benefits & disadvantages are same as backward for Forward also.***

***n-features hai***

***Exhaustive - > 2n-1 Model form honge***

***Backward -> n (n+1) /2***

***Forward -> n(n+1)/2***

***This is the summary.***

***When to use Forward & when to use Backward***

***E.g***

***Suppose***

***563 Columns hai***

***& you have to decide best 550***

***& then you will use backward eliminate***

***Coz yeahi fast hogga***

***Coz eliminate krte aarhe hai hum***

***&***

***Suppose aapko 50 best select krna hai***

***Then hum use krenge Forward***

***& if bss Best feature chaye honge then koibhi use krlenge tou chalega***

***Update !***

***Scikit Learn ka bhi 1 Library SequentialFeatureSelctor Library hai***

***Mlextend easily hojata ushse kaam pr***

***Sklearn Iska fayda kya hai basically ki Pipelines vgera banana easy hota hai.***

***Stopping Criteria:***

***Tolerance Level suppose 0.5 hai***

*Yeah basically jo best feature milte hai humme Har ek iteration mai usme difference check krta hai*

*Ki purane iteration ke best Feature accuracy & naye iteration ke best feature accuracy mai jo specify hai E.g 0.5 ka difference hai ki nahi, if hai tou aageh badhte rhega*

*If difference nahi hai then*

*It will not go further it will stop their only .*

*Optimize karneka tarika hai bss yeah.*