*Feature Selection*

*What is feature in ML ?*

*Simple terms mai yeah Input Column hai.*

*Suppose 10 column hai so ismese top 5 columns hum select krte hai & on basis of that*

*We predict future event.*

*Need of Feature Selection ?*

1. *Curse of Dimensionality or features*

*Ek certain number of features kesath hie aapko Optimum Result milte hai.*

*& ushse zyaada feature add krnese result khrbh hojata hai*

*Due to Sparsity.*

*& What Sparsity says is*

*2D mai 2 points ko loo & & 100D mai vahi 2 points loo ,*

*Jaise jaise Dimension badhrha vaise vaise khaalipan create horaha*

1. *Computational Complexity*
2. *Interpretability*
3. *Prediction b.) inference ( e.g loan kyu nahi millrha ishe )*

*E.g 500 ke factors pe yeah bata nahi skte ki ushe loan kyu nahi millraha.*

*Types of Feature Selections.*

1. *Filter Based Technique.*
2. *Wrapper Technique.*
3. *Embedded Technique.*
4. *Hybrid Technique.*

*Filter Based Feature Selection.*

*Filter based feature*

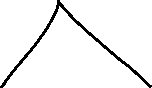
*Yeah Score assign krta hai har ek ek column ko & then based on score*

*It filters out the Columns , jo ache columns hai unhe rkhta hai.*

1. *Variance Threshold.*
2. *Correlation*
3. *ANOVA*
4. *Chi Square.*
5. *Mutual Information*

*Step 1 jo sbhmai mandatory hai kindoff.*

1. *Duplicate Features or Column agar hai tou unko Drop krdo.*
2. *Variance Threshold*



1. *Constant 2.) Quasi constant*
2. *Constant 🡪*

*A b c*

*1 1 y*

*2 1 s*

*3 1 x*

*B jo hai uska variance kuch hai hie nahi 0 hai vo bilkul spread out hai hie nahi,*

*But when we talk about A , A ka variance is spread out*

*We simply drop such constant columns jo Spreadout hie na hoo.*

1. *Quasi Constant Columns*

*1000 Rows*

*995 Rows mai b ka value hai: 1*

*5 Rows mai b ka value is : 0*

*The variance of b is very much close to 0*

*Yeah 2 Type ka agar feature millrha hai humme*

*Then we use a technique called*

*Variance Threshold.*

*Points to Consider:*

1. *Ignores Target Variables:*

*Variance threshold, it is univariate method, 1 column 1 baar mai,*

*Means it evaluates each feature independently & doesn’t even consider the relationship between each feature and the target variable,*

*This means it may keep irrelevant feature that have a high variance.*

***Basically Relationship capture nahi krrpaata with our Target variable.***

***So yeah dhyan rkhna hai***

***Lyk kuch columns ese hote hai jo threshold above rehte hai but they don’t have any relationship with Y or target variable.***

***& kahi baar iska ulta hota hai***

***We drop column whose threshold is below the mark but it is forming a good relationship with y or target variables.***

1. *Ignore Feature Extractions.*

*Basically vahi jo Upar bola maine ,*

*Variance Threshold doesn’t account for interactions between features. A Feature with low variance may become very informative when combined with another feature.*

1. *Sensitive to Data Scaling.*

*It is important to standardize the features before applying variance threshold.*

*There is 1 feature jaha 100000 mai value hai & there is one feature where 0.2,0.9 mai value hai*

*& humne threshold liya 0.1*

*Then yeah Threshold kehisabse vo kabhi drop hoga hie nahi.*

1. *Arbitary Threshold Value,*

***Generally bolte hai Scale krke 0.1 se 0.01 ke beech mai rkhlo***

***But esa koi rule nahi hai , So yeah soch vgera ke rkhna hota hai.***

*Correlation ( Pearson Co relation Coefficient )*

*Spearman bhi lagaskte hai even yaha.*

*2 approach hai*

*Suppose*

*F1,f2,f3,f4 4 columns hai & y is output*

*Approach 1:*

*Then hum f1 ka co-relation y se nikale*

*Then hum f2 ka co-relation coefficient y se nikale*

*Then hum f3 ka co-relation coefficient y se nikale*

*Then hum f4 ka co-relation coefficient y se nikale*

*Ese hum relation nikale*

*& esehie hum 1 Cuttof Decide karle E.g : 0.3 or -0.3*

*Agr yeah cutoff clear krega jo jo columns vo available rahenge &*

*Jo nahi karega vo available nahi rahenge.*

*Approach 2:*

*We can try to know about all the columns co-relations*

*Like:*

*F1 co-relation with F2,*

*F1 co-relation with f3,*

*F2 corelation with f3*

*& so on.*

*We can do this & we will get to know whether is their any Multicollinearity is their or not.*

*Most of the used approach is:*

*Multi collinearity reduce karte hai, & yeah reduce krneke vajese*

*Columns reduce hojate hai.*

*Correlation ke Disadvantages*

1. *Linearity Assumption:*

*Correlation measures the linear relationship between two variables.* ***It does not capture non-linear relationship****. If a relationships is non-linear , the correlation coefficient can be misleading.*

1. *It doesn’t capture Complex Relationship*

*E.g f1, f2 ka relation tou nikalrhe hai*

*But esa bhi possible hoskta ki*

*F1,f2 & f3 teeno ka relationship bhaut important tha Y ko Predict karnekeliye*

1. *Threshold Determination:*

*Just like variance threshold,*

*Defining what level of correlation is considered “high” can be subjective and may vary depending on the specific problem or dataset.*

1. *Sensitive to Outliers:*

*Correlation is sensitive to outliers.*

*Anova*

*Anova hum tbh use krte hai*

*Jbh Input mai Numerical Columns hai & output is Categorical.*

*There is also another variation.*

*& Inputs & Outputs Dono Numerical hai tou ish case mai bhi hum Anova lagaskte hai.*

*1-Way Anova*

*Disadvantages of Anova*

1. *Assumption of Normality*

*Anova assumes that the data for each group follow a normal distribution. This assumption may not hold true for all datasets.*

*Especially those with skewed distributions.*

***Jo saara group hai yeah maanke chalo ki Normal Distribution follow krega.***

1. *Assumption of Homogenity of Variance:*

*Har group ka variance same hona chaye*

*Anova assumes that the variances of the different groups are equal. This is the assumption of homogeneity of variance (also known as homoscedasticity). If this assumption is violated, it may lead to incorrect results.*

1. *Independence of Observations.*

*Anova assumes that the observations are independent to each other. This might not be the case in dataset.*

*Where observation are related.*

*E.g ( Time Series Data, Nested Data )*

*That’s why Time Series Data mai hum Anova use nahi krte hai.*

1. *Effect of Outliers:*

*Anova is sensitive to outliers. A single Outlier can significantly affect the F-statistic leading to a potentially erroneous conclusion.*

1. *Doesn’t Account for interaction:*

*Yaha bhi hum 1 baar mai 1 hie chiz ko dekhte hai not more than that,*

*Just like other univariate feature selection methods, ANOVA does not consider interactions between features.*

*Chi-Square*

*Ch-Square keliye dono data categorical chaye.*

*Chi Square basically 1 input Column & 1 output column ke beech relation batata hai*

*& bss categorical data pr kaam krta hai.*

*Numerical Data pe kaam nahi krta hai.*

*Step ->*

1. *Contigency Table*

***Agar Difference bhaut hie alag hai***

***then yeah bataraha hai ki***

***1 mai Changes laanese Dusre mai farak padraha hai***

***Then it means ki Dono categorical data mai Relation hai.***

*Disadvantages:*

*Categorical Data.*

*/|\/|\/*

***Mutual Information***

*Mutual Information (MI) is a measure of the dependency between 2 variables. It quantifies the amount of information obtained about 1 random variable through observing the other random variable. It is fundamental quantity in information theory.*

*It is a measure of 2 columns unka ekdusrepr dependency kitna hai vo batata hai Mutual information.*

*e.g*

*f1 ko dekhke y ko kitna bataskte hai*

***Disadvnatages:***

