Page No.
Date

Session - 5.4

- **Feature Selection -**
  - feature → i/p column
    target → o/p "

- We don't use all the features rather only use some. this is feature selection

- Why to do feature selection -
  1) Curse of dimensionality -
     ↳ In model only a certain number of feature will give optimum result.
     If we increase # features then due to sparsity ~~result~~ good result will not come

  2) Computational complexity
     ↳ reduce column reduce space & time complexity

  3) Interpretability -
     ↳ if less columns then we can interpret easily

- Technique/types of F.S
  ↳ filter based technique
  ↳ wrapper method
  ↳ Embedded technique
  ↳ hybrid technique

Page No.
Date

1) Filter based technique -
↳ we do some statistical measure in each column individually and filter the columns.
In this we have -
→ Variance Threshold
→ Correlation
→ Anova
→ Chisquare
→ mutual info

Step-1

→

f1	f2	f3	o/p
1	7	1	
2	6	2	
3	3	3	

→ Here we see that f1 & f3 are duplicates So we remove them

Step-2 . Variance Threshold
↳ It is applied on two types of features

1) constant

A	B	C
2	1	Y
7	1	N
2	1	N

Here we see that B column is constant so it has no variance hence remove it.

2) Quasi constant feature
→ assume a feature in which 99% out of 100 rows in 99% rows value is 1 and in remaining 5 rows value is 0. so var is 0 so remove it.

Page No.
Date

Step → • decide threshold (like 0.1)
• for all feature calculate threshold variance
• drop if var. is less than threshold

→ There are some points to be considered -

• we only focus on individual column. Suppose f1 has high high var so we kept it but what if it doesn't have any relat" with o/p and vice-versa means f2 has low var. So we remove it but it has high relat" with o/p

• we don't see about relat" between f1 & f2 what if f1 has low var. but is related to f2

• must do scaling
• challenging to find threshold.

→ Correlation (Pearson corr coef)
↳ we find correlat" b/w i/p & o/p
↳ it is b/w −1 to 1
↑ strong inverse linear relat"
↑ strong positive linear relat"

→ 0 mean no linear relat" (may be non-linear)

↳ we code -

f1 → y
f2 → y
|
→ y

find corr b/w f & y
if outside this (0-0.3 to 0.5) then remove take

or we can find corr b/w f_1, f_2 etc and if having more correlⁿ then keep one of them

· Disadvantages -
- It's possible that there is non-linear relatⁿ but we still dropit
- It doesn't capture complex relationship - we only see relatⁿ b/w f_1 & f_2 but what if f_1, f_2, f_3 has relatⁿ
- How to know threshold?
- Sensitive to outlier.

→ ANOVA - (Hypothesis test)

i/p	o/p
Numerical	categorical (>2 class)

f → o/p (find relatⁿ b/w them)
if strong then keep it.
→ strength of relatⁿ is find using f- statics

· Disadvantages -
- All groups that are formed should be normally distributed
- Variance of groups must be homogeneous
- Effect of outliers
- Doesn't take account in features relatⁿ

5) Chisquare test
is Already done

Page No.
Date

Advantage of filtered based feature select^{n}

- Simplicity
- Speed
- Scalability
- pre-processing steps

- Disadvantage.
  - doesn't consider interact^{n} b/w features
  - model agnostic
  - Threshold determination