Handling Multicollinearity in Data Science

A quick guide to detect and fix correlated predictors

What is Multicollinearity?

Definition

When predictors are highly correlated, the model struggles to separate their effects.

Impact

Unstable coefficients, unreliable p-values, poor interpretability.

How to Detect

Correlation Matrix

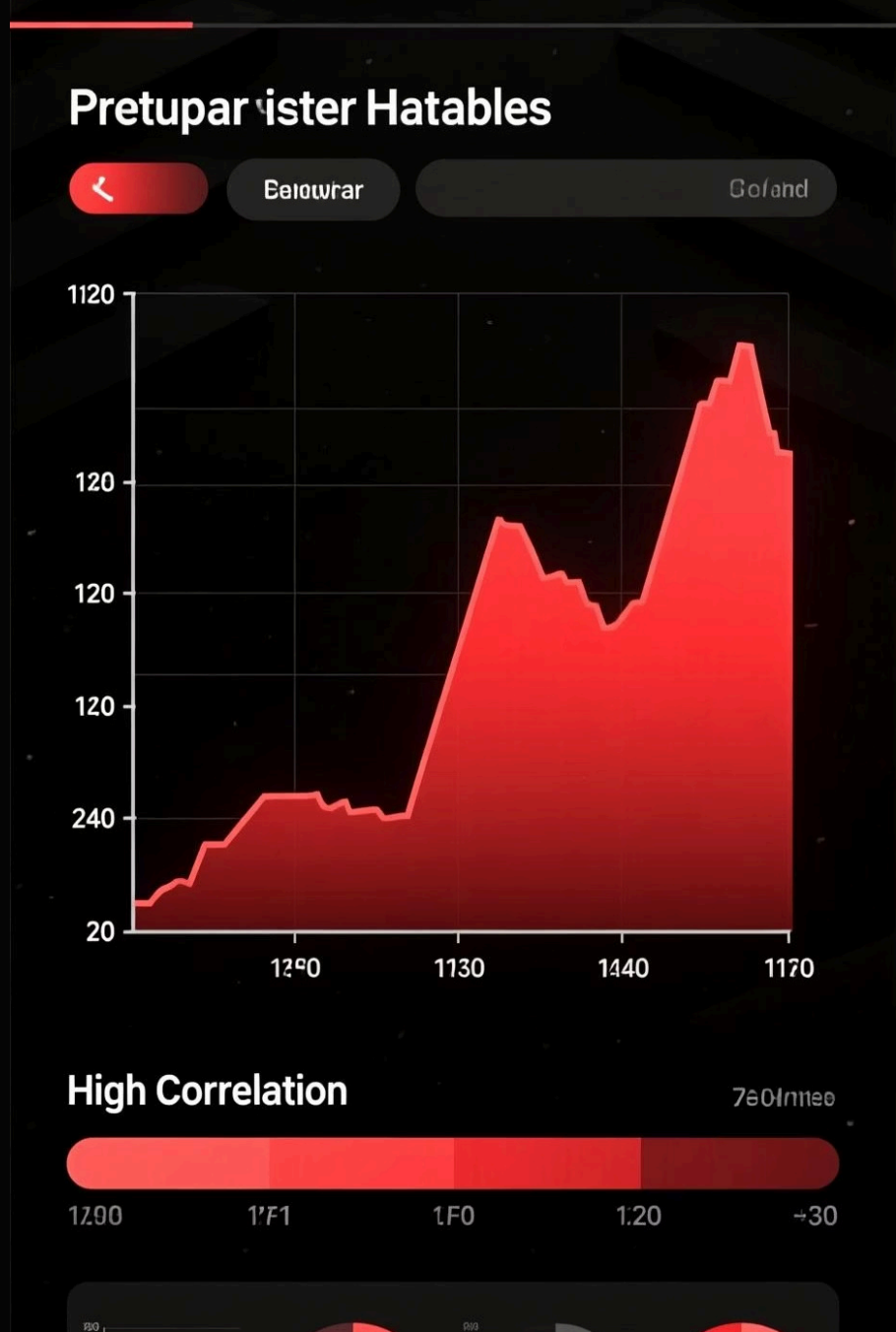
Look for strong correlation (>0.8) between predictors.

Variance Inflation Factor (VIF)

VIF > 10 is problematic.

Condition Number

>30 = strong multicollinearity.



How to Handle

Method	Pros	Cons	Best Use Cases
Drop a Variable	Simple, quick	May lose info	Redundant vars
Combine Variables	Keeps meaning, interpretable	Needs domain knowledge	Similar features
Ridge Regression	Keeps all vars, stabilizes model	Coefficients less interpretable	Focus on prediction
Lasso Regression	Auto feature selection	Can drop useful vars	High-dim data
PCA	Removes correlation, reduces dims	Harder to interpret components	High-dim, low need for interpretability
Domain Knowledge	Business relevance, compliance	Requires SME input	Banking, Healthcare

Rules of Thumb

VIF > 10 = serious multicollinearity

When the Variance Inflation Factor exceeds 10, you should take action to address the issue.

Always check correlation before regression

Make this a standard part of your data preprocessing workflow.

Balance interpretability vs prediction accuracy

Your approach should depend on your project goals and stakeholder needs.

Key Takeaway

If goal = explainability

Drop/merge features, apply domain knowledge.

If goal = prediction accuracy

Use Ridge, Lasso, or PCA.

Clean, modern infographic style with Blue/Teal + Grey color palette.