Notes

Weight of Evidence (WoE)

* Weight of Evidence is a measure used in statistics and machine learning, particularly in credit scoring, to evaluate predictive power of an independent variable in relation to a binary target variable.
* Use Case:
  + Helps transform categorical variables into numerical representations while preserving the predictive power.
  + Commonly used in logistic regression for credit risk modeling.
* Interpretation:
  + WoE > 0: Indicates a higher proportion of good outcomes.
  + WoE < 0: Indicates a higher proportion of bad outcomes.
  + WoE = 0 No predictive power.

Information Value (IV)

* Information Value quantifies the predictive strength of an independent variable in relation to the dependent variable.
* Use Case:
  + Determines the relevance of variables in predictive modeling.
  + Helps in feature selection.
* Interpretation:
  + IV < 0.02: Weak Predictor.
  + 0.02 < IV < 0.1: Moderate Predictor.
  + 0.1 < IV < 0.3: Strong Predictor.
  + IV > 0.3: Very Strong Predictor.

Collinearity

* Collinearity occurs when two or more independent variables in a regression model are highly correlated, leading to redundancy and instability in model estimates.
* Types:
  + Perfect Collinearity: When one variable is an exact linear function of another.
  + High Collinearity: When two variables have a strong correlation but are not exact duplicates.
* Detection Methods:
  + Variance Inflation Factor (VIF): Measures the extent of collinearity in regression.
  + Correlation Matrix: Checks the correlation between independent variables.
* Solutions:
  + Remove one of the correlated variables.
  + Use Principal Component Analysis (PCA) to reduce dimensionality.
  + Apply regularization techniques (e.g., Lasso Regression).

Residual Analysis

* Residual Analysis involves evaluating the difference between observed values and predicted values in a regression model to check model assumptions and performance.
* Use Case:
  + Detects heteroscedasticity (non-constant variance in residuals).
  + Identifies model misspecification.
  + Checks normality of residuals.
* Key Assumptions for Regression:
  + Residuals should be normally distributed.
  + Residuals should have constant variance (homoscedasticity).
  + No patterns should be present (Indicates a good model fir).
* Visualization Techniques:
  + Residual Plot: Checks for randomness.
  + Q-Q Plot: Assesses normality.
  + Histogram of residuals: Evaluates the distribution.

Heteroscedasticity (Sedasticity)

* Heteroscedasticity refers to the condition in which the variance of residuals in a regression model is not constant across all levels of an independent variable.
* Types:
  + Homoscedasticity (Constant Variance): Desirable property in linear regression.
  + Heteroscedasticity (Non-Constant Variance): Can lead to inefficient estimates and misleading statistical inference.
* Detection Methods:
  + Residual Plot: If residual variance increases with the independent variable, heteroscedasticity is present.
  + Breusch-Pagan Test: A statistical test for heteroscedasticity.
* Solutions:
  + Apply log transformation to stabilize variance.
  + Use weighted least squares regression.
  + Apply robust standard errors to adjust for heteroscedasticity.

Summary

* These concepts are fundamental for data preprocessing, feature selection, and regression modeling.
* Proper handling of these aspects ensures better model accuracy, interpretability, and robustness.