|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Assumptions | Benefits | Limitations | Handling of Imbalanced Data |
| Logistic Regression (Regularized) | * Linear relationship between features and log-odds of outcome * Independence of observations * No multicollinearity | * Simple and interpretable * Provides feature importance * Efficient for large datasets | * May underperform with non-Linear relationships * Sensitive to outliers * Assumes linearity of independent variables and log odds | * Can be improved with class weighting * Regularization helps with rare events |
| LOGISTIC REGRESSION (MULTILEVEL) | * Correct model specification * Independence of residuals within and across levels * Homoscedasticity between levels * Normally distributed random effects | * Accounts for hierarchical/nested data structures * Provides both cluster-specific and population-average estimates | * Computationally intensive * Requires sufficient sample size at both levels for reliable estimates | * Can handle imbalanced data by incorporating group-level random effects * Allows flexibility in modeling cluster-specific variations |
| Random Forests | * Features are independent * Trees are uncorrelated | * Handles non-linear relationships well * Robust to outliers and noise * Provides feature importance | * Can overfit on noisy data * Less interpretable than linear models * Computationally intensive for large datasets | * Can handle imbalanced data well through bootstrapping and random feature selection |
| XGBoost | * Monotonic relationship between features and target | * High predictive accuracy * Handles missing data * Efficient implementation | * Prone to overfitting * Sensitive to outliers * Requires careful tuning | * Built-in support for class weighting * Can adjust the scale\_pos\_weight parameter |
| LSTM Neural Networks | * Sequential data * Temporal dependencies in data | * Captures long-term dependencies * Handles variable-length sequences * Effective for time-series data | * Requires large amounts of data * Computationally intensive * Prone to overfitting with small datasets | * Can be improved with oversampling techniques * Careful batch construction needed |