SELECTING A PREDICTIVE MODELING ALGORITHM

Regression

Linear Regression

+ Easy to interpret
+ Quick runtime
- Requires linear relationships
- Poor for imbalanced datasets
- Sensitive to outliers

Count Regression

+ Specializes in target variables with Poisson distributions
- Highly specialized

Gamma Regression

+ Specializes in target variables with Gamma distributions
- Highly specialized

Classification or Regression

Neural Network

+ Resistant to collinearity
+ Quick to output results
+ Good with highly non-linear data
- Very difficult to explain outcomes
- Cannot handle missing values
- Sensitive to outliers
- Time consuming to create
- Requires user input to tune well

Boosted Model

+ Resistant to overfitting
+ Robust for imbalanced datasets
- Difficult to configure hyperparameters
- Slower to train than Forest Model

Spline Model

+ Easy to interpret
+ Self-selects most important predictors
+ Robust for highly non-linear relationships
- Computationally expensive

Support Vector Machine

+ Resistant to overfitting
+ Very flexible
- Trial & Error for settings on large datasets
- Computationally expensive

Classification

Decision Tree

+ Easy to interpret
+ Does not require scaling or normalization
+ Not sensitive to missing values or outliers
+ Self-selective for features
- Prone to overfitting

Forest Model

+ Same benefits as Decision Tree
+ Less prone to overfitting than Decisoin Tree
+ Usually require minimal hyperparemeter adjustments
- More difficult to explain than Decision Trees

Logistic Regression

+ Easy to interpret
+ Quick runtime
- Underperforms on complex relationships
- Limited to 2 category output

Naive Bayes Classifier

+ Easy to interpret
+ Easy to implement
- Can underperform compared to other properly tuned models
- Assumes independence between predictors
- Training data must include all possible outcomes