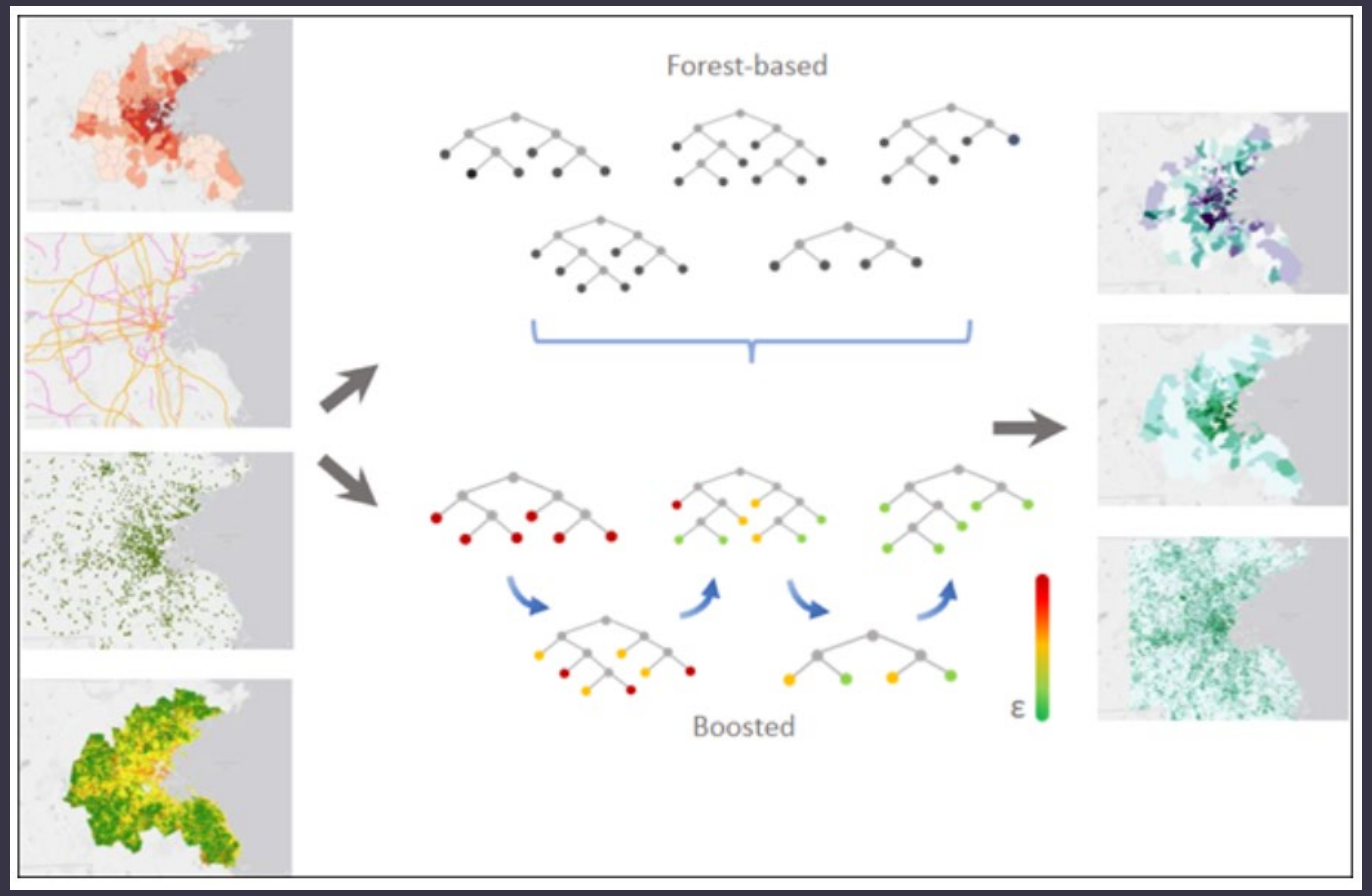
MODELING SPATIAL RELATIONSHIPS

Forest-based and Boosted Classification and Regression

RANDOM FOREST



https://pro.arcgis.com

FOREST BASED & GRADIENT BOOSTED MODELS

Forest based

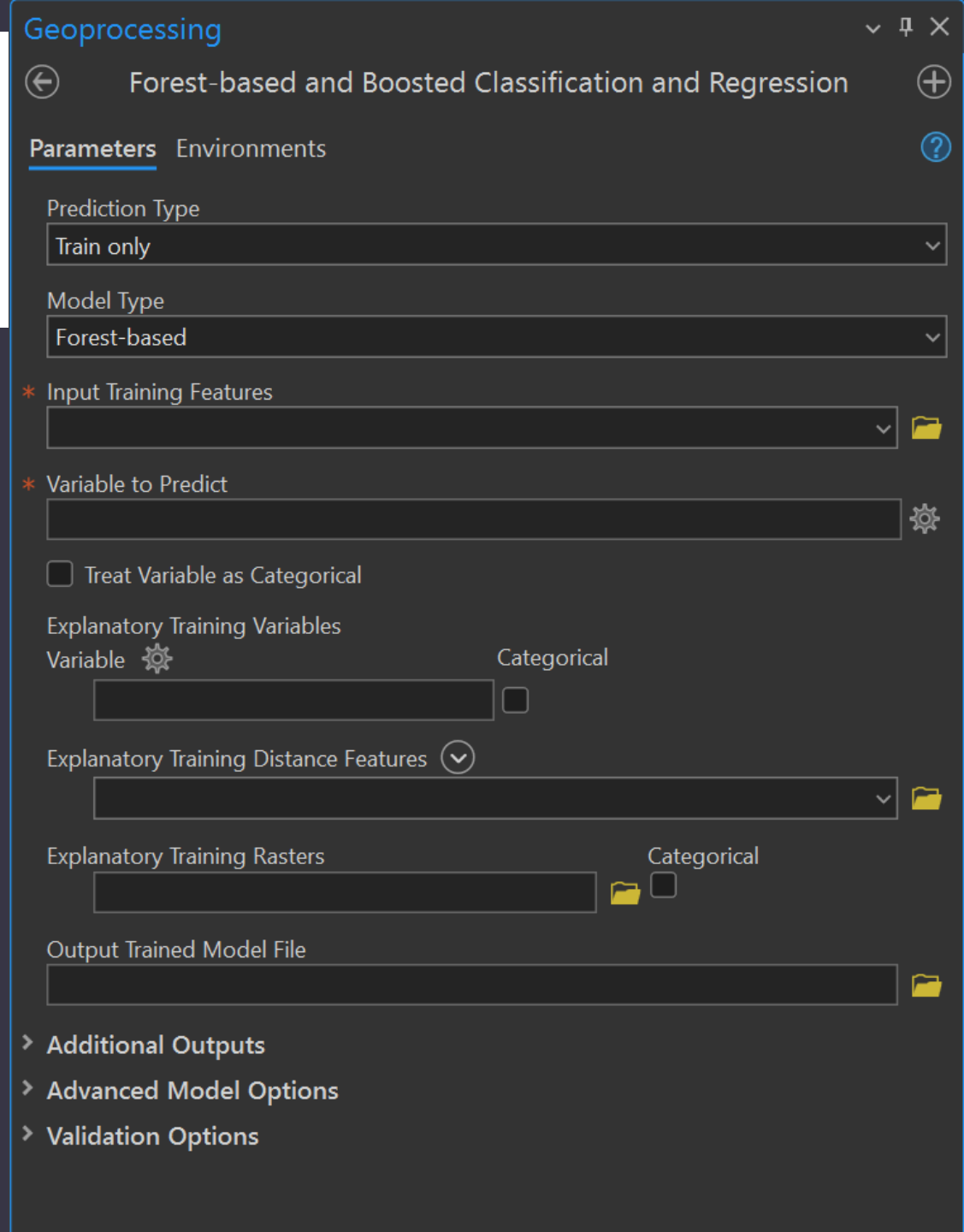
- Many independent decision trees
- Each decision tree is created from a random subset of training data and explanatory variables
- Each tree generates its own predictions
- Final prediction is based on an average of all decision trees in the entire forest

Gradient boosted

- Creates a series of sequential decision trees
- Each tree is built to minimize the error (bias) of the previous tree
- Combines weak learners to become a strong prediction model

HOW TO TRAIN A MODEL

- Prediction Type
  - Train only
  - Predict to features
  - Predict to raster
- Explanatory Training Variables
- Explanatory Training Distance Features
- Explanatory Training Rasters



VALIDATION AND TRAINING DATA DIAGNOSTICS

- R-squared – proportion of variance in the dependent variable that can be explained by the independent variable
- Goodness of fit
- 0.0 to 1.0 (higher the better)
- While training, the tool reserves 10% of data for validation – it only uses 90% of the training data and uses validation data to determine how well the model performed preventing Overfitting.

Training Data: Regression Diagnostics

R-Squared	0.844
Mean Absolute Error (MAE)	33329.720
Mean Absolute Percentage Error (MAPE)	0.194
Root Mean Square Error (RMSE)	45848.739
p-value	0.000
Standard Error	0.003

*Predictions for the data used to train the model compared to the observed categories for those features

Validation Data: Regression Diagnostics

R-Squared	0.705
Mean Absolute Error (MAE)	44251.616
Mean Absolute Percentage Error (MAPE)	0.250
Root Mean Square Error (RMSE)	62644.450
p-value	0.000
Standard Error	0.010

*Predictions for the test data (excluded from model training) compared to the observed values for those test features

THINGS TO KEEP IN MIND…

- These tools can perform well within the range of explanatory variables used to train the model.
- Forest-based and boosted models do not extrapolate.
- Number of trees (100 by default) can be increased with the complexity of relationships between variables, size of dataset, and variable to predict.
- Tool runtime is sensitive to the number of variables used per tree.
- The tool may generate slightly different results each time due to the randomness introduced in the algorithm.

THINGS TO KEEP IN MIND…

- A “good” model is subjective and varies based on the data.
- Some industries may have a r-square value (threshold) to consider if the model is satisfactory for making a prediction.
- Refining a Model – includes only most important variables. The variables with low importance that did not contribute to the model performance were excluded.
- Increasing number of trees in a model could result in a better model.
- Model Out of Bag Errors – Determines how much the model performance has improved by increasing the number of trees in the model.

https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/forestbasedclassificationregression.htm