







**cross-validation**

- **nfolds**: number of folds for N-fold cross-validation (default: 0, disabled)
- **keep\_cross\_validation\_predictions**: keep the predictions of the cross-validation models.
- **keep\_cross\_validation\_fold\_assignment**: keep the cross-validation fold assignment.
- **fold\_assignment**: cross-validation fold assignment scheme, if fold column is not specified. The "Stratified" option will stratify the folds based on the response variable, for classification problems. Must be one of: "AUTO", "Random", "Modulo", "Stratified".
- **fold\_column**: column with cross-validation fold index assignment per observation.

# Platt Scaling

- Many ML algorithms introduce biases when it comes to class probability

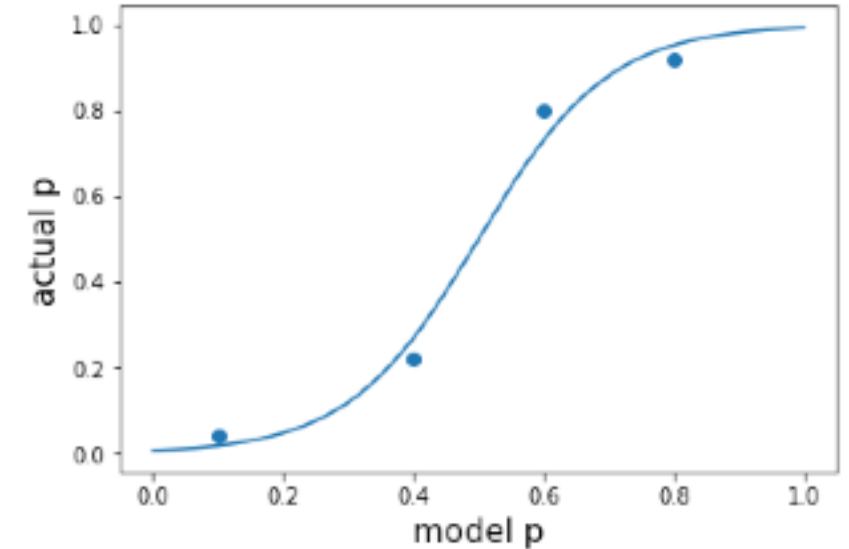
SVMs, GBMs { underpredict high-prob classes  
overpredict low-prob classes

Naïve Bayes { overpredict high-prob classes  
underpredict low-prob classes

- Correct for this by fitting a sigmoid to the model output:

$$P(\mathbf{x}) = \frac{1}{1 + \exp(Af(\mathbf{x}) + B)}$$

- Split data into two frames:
  - 1) Training dataframe: used to generate  $f(\mathbf{x})$
  - 2) Platt calibration frame: used to fit  $A$  and  $B$



# Cross-Validation

- **nfolds**: number of folds for N-fold cross-validation (default: 0, disabled)
- **keep\_cross\_validation\_predictions**: keep the predictions of the cross-validation models.
- **keep\_cross\_validation\_fold\_assignment**: keep the cross-validation fold assignment.
- **fold\_assignment**: cross-validation fold assignment scheme, if fold column is not specified. The "Stratified" option will stratify the folds based on the response variable, for classification problems. Must be one of: "AUTO", "Random", "Modulo", "Stratified".
- **fold\_column**: column with cross-validation fold index assignment per observation.