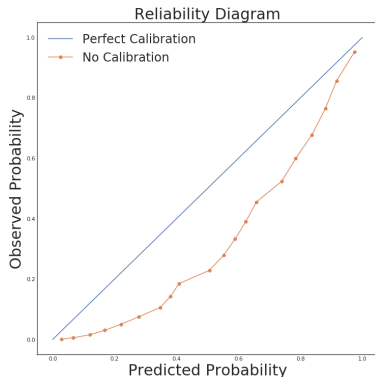




Applied Machine Learning

Performance Evaluation: Calibration versus Discrimination



Learning goals

- Understand difference between calibration and discrimination
- How to diagnose calibration with plots and metrics
- When calibration matters and types of miscalibration

PERFORMANCE METRICS AND USE CASE NEEDS



- **Match metric to task:** Optimize what reflects success in your application
- **Robustness to outliers:** Use *MAE* (instead of *MSE*) when occasional large errors should not dominate the metric; use *MSE* if they should
- **Class imbalance:** Use F1-score instead of accuracy
- **Unequal error costs:** Apply cost-weighted metrics for asymmetric penalties (i.e., weighting FP and FN differently)
- **Prediction type matters:**
 - *Discrete predictions:* Use confusion-matrix metrics (e.g., *F1 score*)
→ Focus on **discrimination** (how well classes are separated)
 - *Probabilistic predictions:* Use proper scoring rules (e.g., *Brier score*)
→ Consider **calibration** (matching predicted to actual probabilities)

DISCRIMINATION

Consider the true positive $Y = 1$, true negative $Y = 0$, and predicted label \hat{Y} .

Discrimination: Ability of \hat{Y} to well-separate positive/negative instances.

⇒ Confusion matrix-based measures are discrimination measures, e.g.,

$FPR = P(\hat{Y} = 1 | Y = 0)$, $TPR = P(\hat{Y} = 1 | Y = 1)$, AUC.



		Predicted Values		
		Positive	Negative	
True Values	Positive	TP	FN	<i>Sensitivity</i> $\frac{TP}{TP + FN}$
	Negative	FP	TN	<i>Specificity</i> $\frac{TN}{TN + FP}$
		<i>Precision</i> = $\frac{TP}{TP + FP}$	<i>Negative Predictive Value</i> = $\frac{TN}{TN + FN}$	<i>Accuracy</i> $\frac{TP + TN}{TP + FP + TN + FN}$







CALIBRATION

Consider the true positive $Y = 1$, true negative $Y = 0$, and predicted label \hat{Y} .

Calibration: When the predicted probabilities \hat{p} closely agree with the observed proportion for $Y = 1$ (for any reasonable grouping).

- **Calibration in the large:** Observed vs. predicted prob. in *full sample*.

Predicted prob. of survival (1 = survived, 0 = not) for 6 Titanic passengers of different gender

	Probability	Actual
	0.2	0
	0.8	0
	0.9	1
	0.1	0
	0.7	1
	0.3	1

Calibration in the large:
property of full sample

Mean predicted
probability of survival:
0.5

Observed probability
of survival in sample:
0.5

Calibration in the large:
 $0.5 = 0.5$ ✓



CALIBRATION









Consider the true positive $Y = 1$, true negative $Y = 0$, and predicted label \hat{Y} .

Calibration: When the predicted probabilities \hat{p} closely agree with the observed proportion for $Y = 1$ (for any reasonable grouping).

- **Calibration in the large:** Observed vs. predicted prob. in *full sample*.
- **Calibration in the small:** Observed vs. predicted prob. in *subsets*.

Predicted prob. of survival (1 = survived, 0 = not) for 6 Titanic passengers of different gender

	Probability	Actual
	0.2	0
	0.8	0
	0.9	1
	0.1	0
	0.7	1
	0.3	1

Calibration in the large:
property of full sample

Mean predicted
probability of survival:

0.5

Observed probability
of survival in sample:

0.5

Calibration in the large:

0.5 = 0.5 ✓

Calibration in the small:
property of subsets

Mean predicted
probability of survival:

male: 0.20

female: 0.80

Observed probability
of survival in sample:

male: 0.33

female: 0.67

Calibration in the small:

male: 0.20 \neq 0.33 ✗

female: 0.80 \neq 0.67 ✗

CALIBRATION AND DISCRIMINATION

A well-calibrated classifier can be poorly discriminating, e.g.

Obs. Nr.	true Y	\hat{f}_1	\hat{f}_2
1	1	1	0
2	1	1	0
3	0	0	1
4	0	0	1
Avg Prob	50%	50%	50%

- Both models (\hat{f}_1 and \hat{f}_2) result in the same calibration in the large (50%).
- However, \hat{f}_1 is better than \hat{f}_2 as it correctly classifies the real outcome Y .



CALIBRATION AND DISCRIMINATION



A well-discriminating classifier can have a bad calibration, e.g.

Obs. Nr.	truth Y	\hat{f}_1	\hat{f}_2
1	1	0.7	0.9
2	1	0.7	0.9
3	0	0.3	0.5
4	0	0.3	0.5
Avg Prob	50%	50%	70%

- Both models are well discriminating, i.e., setting thresholds $c_1 \in]0.3, 0.7[$ for \hat{f}_1 and $c_2 \in]0.5, 0.9[$ for \hat{f}_2 perfectly separates positive and negative observations (and will result, e.g., in a perfect AUC = 1).
- \hat{f}_2 is poorly calibrated as the averaged probabilities are 70% and do not match the truth proportion of positive observations (which is 50%).

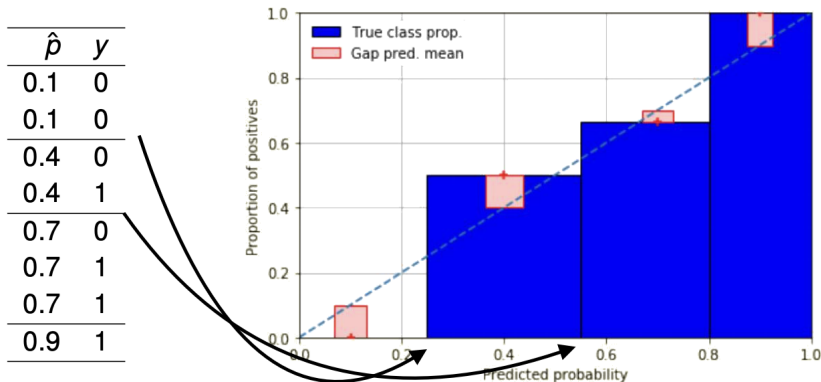
⇒ **How can we measure calibration quality?**

CALIBRATION PLOT / RELIABILITY DIAGRAM



To assess calibration visually, we can plot on the

- x-axis: average predicted probability (e.g., grouped by quantiles)
- y-axis: observed proportion of positive instances in each group



Gap pred. mean: Shows deviation between observed proportion of positives (red points) and average predicted probability within each group (red bar).

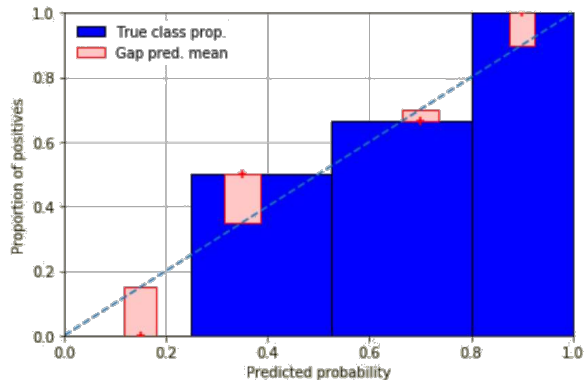
CALIBRATION PLOT / RELIABILITY DIAGRAM



Changing predictions will change the position of the red points on the x-axis

- The closer the red points to the diagonal line, the better calibration
- **Question:** Did we improve or worsen calibration?

\hat{p}	y
0.1	0
0.2	0
0.3	0
0.4	1
0.6	0
0.7	1
0.8	1
0.9	1

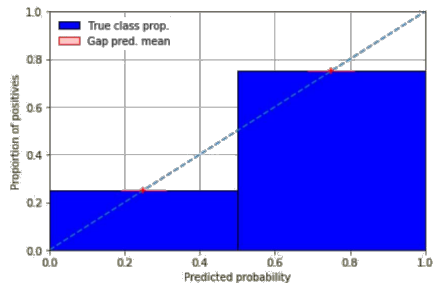


CALIBRATION PLOT / RELIABILITY DIAGRAM

Different groupings lead to a completely different picture

2 Groups

\hat{p}	y
0.1	0
0.2	0
0.3	0
0.4	1
0.6	0
0.7	1
0.8	1
0.9	1



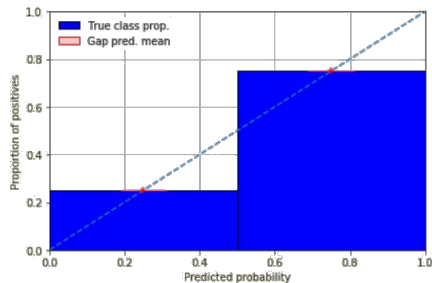
CALIBRATION PLOT / RELIABILITY DIAGRAM

Different groupings lead to a completely different picture



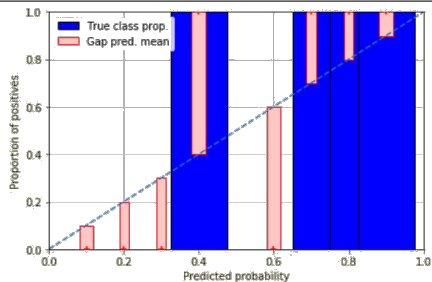
2 Groups

\hat{p}	y
0.1	0
0.2	0
0.3	0
0.4	1
0.6	0
0.7	1
0.8	1
0.9	1



8 Groups

\hat{p}	y
0.1	0
0.2	0
0.3	0
0.4	1
0.6	0
0.7	1
0.8	1
0.9	1



IMPORTANCE OF GROUPING PREDICTIONS

Evaluating calibration with bins/groups:

- We need enough predictions in each group for reliable estimates.
- Trade-off:
 - large groups give better empirical estimates
 - small groups allow a more fine-grained assessment of calibration



IMPORTANCE OF GROUPING PREDICTIONS



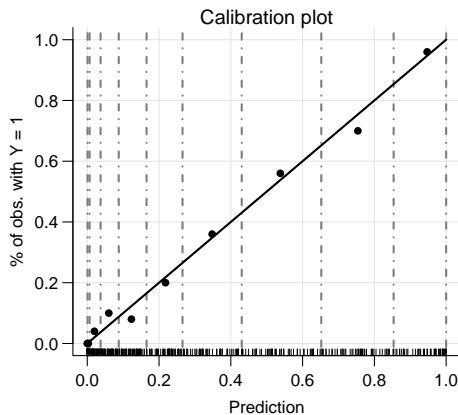
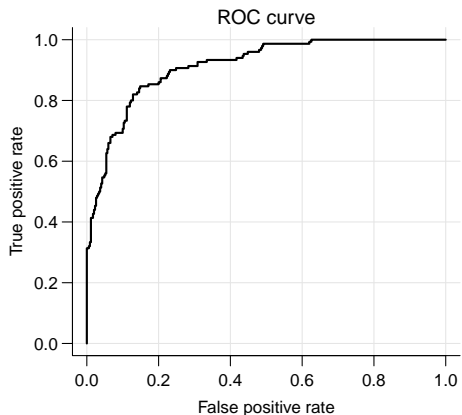
Evaluating calibration with bins/groups:

- We need enough predictions in each group for reliable estimates.
- Trade-off:
 - large groups give better empirical estimates
 - small groups allow a more fine-grained assessment of calibration

How groups/bins can be created:

- Equal-Width: Splits predicted probabilities into equally sized intervals.
⇒ Can lead to bins with few or no observations.
- Equal-Frequency: Each bin holds a similar number of observations.
⇒ Bin width can vary notably.

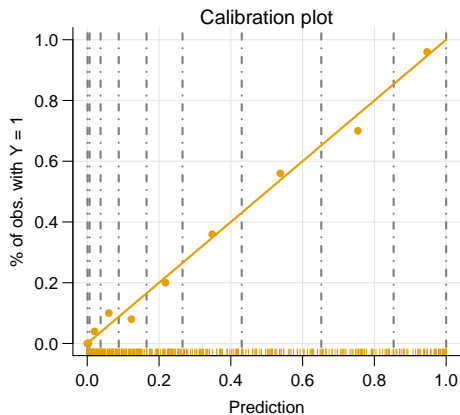
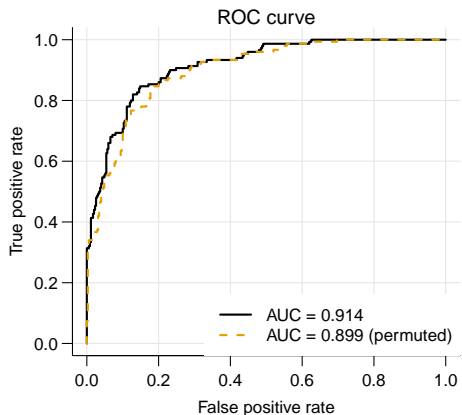
ROC CURVE VS. CALIBRATION PLOT



- **ROC:** Measures only discrimination as it is based on TPR and FPR and assesses only the ranking of predicted probabilities (not their magnitude).
- **Calibration plot:** Measures (for reasonable groups, here by deciles) how well the predicted probabilities match the proportion of positives.



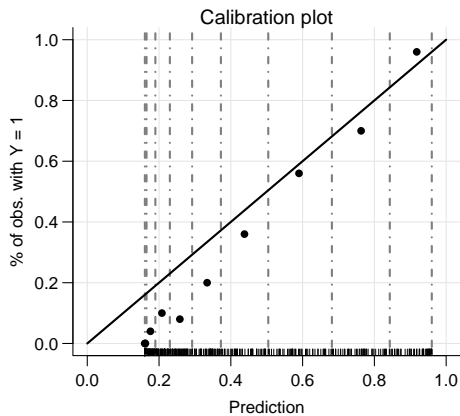
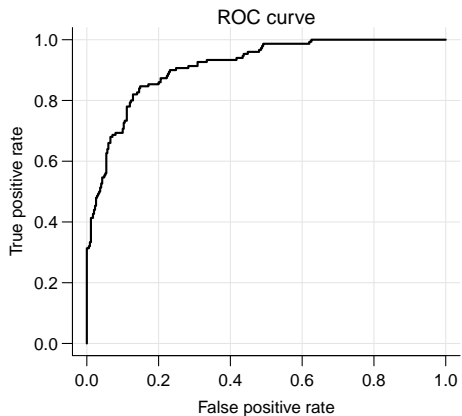
ROC CURVE VS. CALIBRATION PLOT



Permuting predictions within each group (e.g., randomly assigning different predictions to observations within each decile)

- worsens the ROC curve / AUC (ranking within each decile changes).
- does not affect the calibration plot (as we only look at averages).

ROC CURVE VS. CALIBRATION PLOT



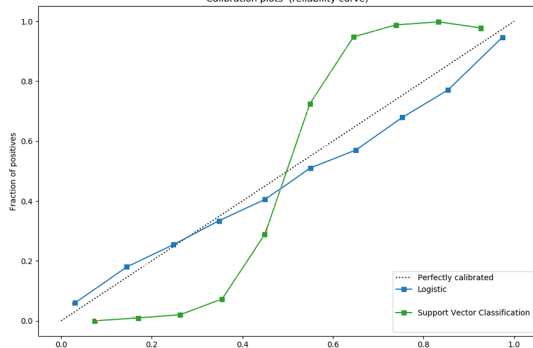
Monotonic transformations of the predicted probabilities (e.g., $\frac{\hat{p}+0.2}{1.2}$)

- do not affect the ROC curve as the ranking of \hat{p} will not change.
- affect the calibration plot (here it looks worse).

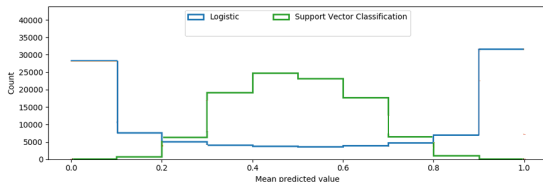
EXAMPLE: CALIBRATION PLOTS (EQUAL-WIDTH)



Calibration plots (reliability curve)



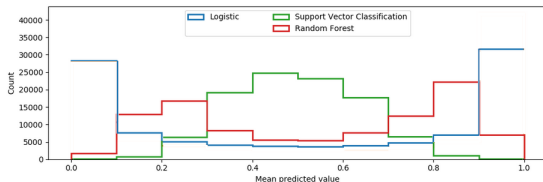
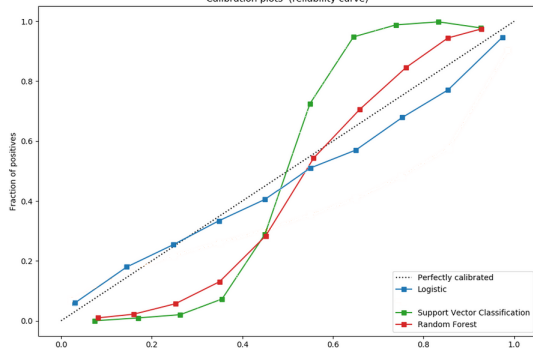
- Many algorithms (except logistic regression) return biased probabilities.
- Linear SVC: badly calibrated, as predictions refer to a distance (to decision boundary) and not to probabilities.



EXAMPLE: CALIBRATION PLOTS (EQUAL-WIDTH)



Calibration plots (reliability curve)

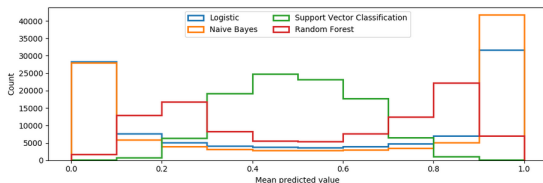
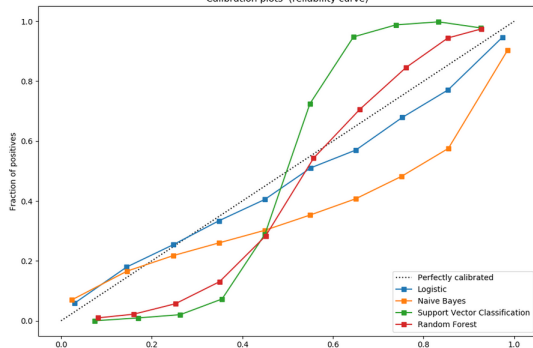


- Many algorithms (except logistic regression) return biased probabilities.
- Linear SVC: badly calibrated, as predictions refer to a distance (to decision boundary) and not to probabilities.
- Random Forest: probabilities close to 0 or 1 are very rare.

EXAMPLE: CALIBRATION PLOTS (EQUAL-WIDTH)



Calibration plots (reliability curve)



- Many algorithms (except logistic regression) return biased probabilities.
- Linear SVC: badly calibrated, as predictions refer to a distance (to decision boundary) and not to probabilities.
- Random Forest: probabilities close to 0 or 1 are very rare.
- Naive Bayes: pushes probabilities to 0 or 1 (see histograms).

METRICS FOR CALIBRATION

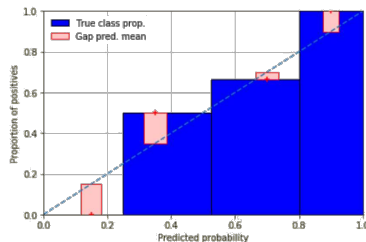


Goal: Measure agreement between pred. probabilities and observed freq.

- N instances; $\hat{p} \in [0, 1]$ predicted probabilities; $y \in \{0, 1\}$ true labels
- Predictions are grouped into M disjoint bins B_1, \dots, B_M
- \bar{p}_m : average predicted probability in bin B_m
- \bar{y}_m : average true label in bin B_m

Calibration-only Metrics: Measure how well predicted probabilities match observed frequencies, independent of class separation.

- **Expected Calibration Error (ECE):**
Average bin-wise gap between predicted probabilities and actual frequencies
$$\text{ECE} = \sum_{m=1}^M \frac{|B_m|}{N} |\bar{y}_m - \bar{p}_m|$$
- **Maximum Calibration Error (MCE):**
Largest bin-wise gap between predicted probabilities and actual frequencies
$$\text{MCE} = \max_{1 \leq m \leq M} |\bar{y}_m - \bar{p}_m|$$



HYPOTHESIS TEST FOR CALIBRATION



- **Goal:** Test if a classifier is really uncalibrated on the given data set
- Group instances into M bins based on percentiles of the predicted probabilities for the positive class (equal-frequency binning)
- Compute the test statistics of the Hosmer-Lemeshow (HL) test

$$H = \sum_{m=1}^M \left(\frac{(O_m^+ - E_m^+)^2}{E_m^+} + \frac{(O_m^- - E_m^-)^2}{E_m^-} \right)$$

- M number of bins
- O_m^+ observed positives or O_m^- observed negatives in bin m
- $E_m^+ = N_m \times \bar{p}_m$ expected positives
- $E_m^- = N_m \times (1 - \bar{p}_m)$ expected negatives in bin m
- N_m number of obs. and \bar{p}_m average predictions in bin m
- If H is small \Rightarrow observed outcomes match predicted probabilities (well-calibrated model)
- Follows a Chi-Squared distribution with $M - 2$ degrees of freedom

HL TEST: EXAMPLE



Obs. Nr.	truth Y	\hat{f}_1	\hat{f}_2
1	1	0.9	0.9
2	1	0.9	0.9
3	0	0.1	0.7
4	0	0.1	0.7

Group obs. by their predicted probability, creating $M = 2$ bins for each model.

Model \hat{f}_1 (Well-calibrated)				
Bin (\hat{p})	O^+	E^+	O^-	E^-
0.9	2	1.8	0	0.2
0.1	0	0.2	2	1.8

$$H = \left(\frac{(0-0.2)^2}{0.2} + \frac{(2-1.8)^2}{1.8} \right) + \left(\frac{(2-1.8)^2}{1.8} + \frac{(0-0.2)^2}{0.2} \right) = 0.44$$

Model \hat{f}_2 (Poorly calibrated)				
Bin (\hat{p})	O^+	E^+	O^-	E^-
0.9	2	1.8	0	0.2
0.7	0	1.4	2	0.6

$$H = \left(\frac{(0-1.4)^2}{1.4} + \frac{(2-0.6)^2}{0.6} \right) + \left(\frac{(2-1.8)^2}{1.8} + \frac{(0-0.2)^2}{0.2} \right) = 4.89$$

Conclusion: H-L statistic for \hat{f}_2 is over 10 times larger due to a large discrepancy for the $\hat{p} = 0.7$ bin ($O^+ = 0$ vs. $E^+ = 1.4$).

CALIBRATION: OVER- AND UNDER-ESTIMATES



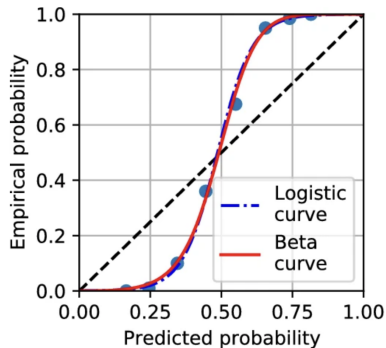
Example: Weather model predicting daily rain probabilities

- Calibration implies: predicted probabilities should (on average) match their observed frequencies
 - ⇒ E.g., '70% chance of rain' → should rain $\approx 70\%$ of the time
 - ⇒ Ensures predictions can be interpreted as actual risk/probabilities
- Consider the following predictions of a (good performing) classifier:

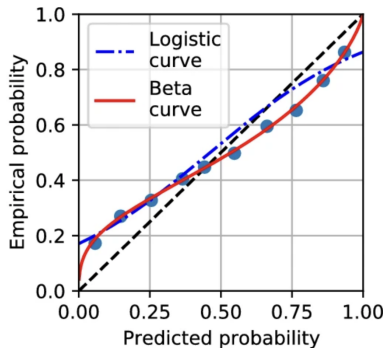
\hat{p}	y
0.1	0
0.1	0
0.4	0
0.4	1
0.7	0
0.7	1
0.7	1
0.9	1

- '10% chance of rain' was a slight over-estimate ($\bar{y} = 0/2 = 0\%$).
- '40% chance of rain' was a slight under-estimate ($\bar{y} = 1/2 = 50\%$).
- '70% chance of rain' was a slight over-estimate ($\bar{y} = 2/3 = 67\%$).
- '90% chance of rain' was a slight under-estimate ($\bar{y} = 1/1 = 100\%$).

CALIBRATION: OVER- AND UNDER-ESTIMATES



(a) Underconfidence



(b) Overconfidence

► "Filho et al" 2023

- **Underconfidence:** Predicted probabilities are too close to 0.5 & need to be expanded (underestimates near 1, overestimates near 0)
- **Overconfidence:** Predicted probabilities are too close to 0 or 1 & should be pulled toward 0.5 (overestimates near 1, underestimates near 0)