Stacking Calibrated Odds Against Floods

Emmanuel Rassou

Harvard Department of Statistics

HARVARD Faculty of Arts and Sciences DEPARTMENT OF STATISTICS

Problem Setup

Binary Classification of whether it will rain or not in Bangladesh

1. **Prioritize Recall over Precision.** Predicting no rain when there is a flood is less ideal compared to rain predicted when there is no flood.
2. **Interpretable Probabilities.** A prediction score of 0.8 for rain should represent a confidence of 80%.

Stacking as a more expressive form of Ensembling

We aim to improve upon Hussain et al.[2], who ensemble the predictions from 5 different classifiers using hard voting. A two-layer stacking scheme can weight the different methods more effectively using the final estimator to aggregate predictions.

ML Method	F1 Score(%)	Recall(%)	Precision(%)
Logistic Regression	**61.65**	54.49	71.00
KNN	42.72	31.23	67.63
Decision Tree	53.99	**56.15**	52.00
Support Vector Classifier	61.63	52.82	73.95
Random Forest	61.45	50.83	**77.66**
Hard Ensembling (theirs)	61.85	51.16	**78.17**
Soft Ensembling	60.32	50.50	74.88
Stacking (Log Reg)	**62.35**	**52.82**	76.08
Stacking (SVC)	61.93	52.16	76.21

Table 1. Stacking experiments using the same 5 base models.

Passive-Aggressive Classifier as a Final Estimator

Passive-Aggressive Classifier[1] is an online learning method that updates the parameters only when there is a clear mistake (aggressive), else it leaves the parameters alone (passive). This approach is a sample-efficient way for learning the weight combination of base learners to boost the F1-score and recall even more by trading for precision.

Stacking achieves an F1-score of **66.88%**, a recall of **69.10 %** and a precision of 64.80%.

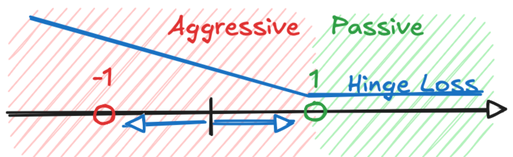
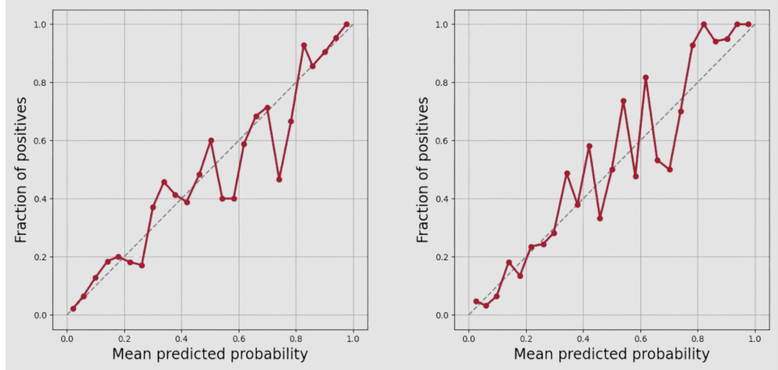


Figure: The hinge loss maximizes the margin to correct a mistake, but leaves the parameters alone when there is none.

Exploring Calibration

Accuracy depends on whether the score is ≥ 0.5 or not. A **calibrated** classifier ensures the score is a proxy for its confidence that it will rain.



Logistic Regression Reliability Curve | Support Vector Classifier Reliability Curve

We state our intuition as the **Reliability Curve Rule**:

- **Below** the perfect line indicates predictions are overconfident, and calibration trades off recall to improve precision
- **Above** the perfect line indicates predictions are underconfident, and calibration trades off precision to improve recall

Background: Methods of Calibration

1. **Platt Scaling:** apply a sigmoid function to raw uncalibrated predictions f_i

$$P(y_i = 1|f_i) = \sigma(Af_i + B) = \frac{1}{1 + \exp(Af_i + B)}$$

2. **Isotonic Regression:** Learn a piecewise non-decreasing mapping which generalizes the sigmoid correction

$$min_{\hat{f}} \sum_{i=1}^{n} (y_i - \hat{f}_i)^2 \text{ such that } \hat{f}_i \geq \hat{f}_j \iff f_i \geq f_j$$

3. **Angular Calibration:** Li and Sur[3] use sin and cos to interpolate between informative $w^T x_{new}$ and non-informative noise $Z \sim N(0, 1)$

$$\hat{f}(\hat{w}^T x_{new}; \hat{\theta}) = E_Z \left[\sigma \left(cos(\hat{\theta}) \frac{w^T x_{new}}{||\hat{w}||_\Sigma} + sin(\hat{\theta}) Z \right) \right]$$

Bregman Divergence(Σ) $\implies$ generality of method

Method	Log Reg	KNN	Decision Tree	SVC	Random Forest
Uncalibrated	61.65	42.72	**53.99**	61.63	61.45
Sigmoid	60.31	46.41	38.66	60.20	**62.18**
Isotonic	61.83	**48.40**	38.66	60.63	61.54
Angular	**63.87**	–	–	**63.28**	–

Table 2. Comparing F1 Scores of different calibration methods. Red indicates a tradeoff favoring precision over recall; Green indicates a tradeoff favoring recall over precision.

Experimentally, the Reliability Curve Rule holds for sigmoid and isotonic methods, but for angular, it trades precision for recall in both Log Reg and SVC. More general structure of angular allows the fitting of more complex miscalibrated patterns.

Combining Stacking and Calibration

Our goal is to retain the scores where (i) recall is prioritized while (ii) ensuring an interpretable confidence probability.

Final Estimator	Log Reg	SVC	Passive-Aggressive
Uncalibrated	62.35	61.93	**66.88**
Sigmoid (Δ)	+0	-0.37	-6.43
Isotonic(Δ)	-0.33	-1.60	-5.99
Angular(Δ)	+0	+0	+0

Table 3. Retention of F1 score after calibration applied to stacking final estimators

Stacking with a **Passive-Aggressive** final estimator and **angular calibration** best achieves both orthogonal goals.

Future Directions: We recommend quantitatively analyzing the interpretability of scores post-calibration and applying calibration to the base learners of the ensemble.

References

[1] Crammer et al.
Online passive-aggressive algorithms.
Journal of Machine Learning Research, 7:551–585, 2006.

[2] Hussain et al.
Weather forecasting using machine learning techniques: Rainfall and temperature analysis.
Journal of Advances in Information Technology, 15(12):1329–1338, 2024.

[3] Li et al.
Optimal and provable calibration in high-dimensional binary classification: Angular calibration and platt scaling.
https://arxiv.org/abs/2502.15131, 2025.
arXiv:2502.15131.