Equation Appendix

Notation

- TP, FP, TN, FN – true/false positive/negative counts.
- $N = \mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}$ – total number of instances.
- $y_i \in \{0, 1\}$ – ground-truth label for instance i.
- $\hat{p}_i \in [0, 1]$ – model-predicted probability of class 1.

1 Binary Classification Metrics

1.1 Accuracy, Precision, Recall

$$\text{Accuracy (ACC)} = \frac{\mathrm{TP} + \mathrm{TN}}{N}, \tag{1}$$

$$\text{Precision (P)} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{2}$$

$$\text{Recall (R)} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{3}$$

1.2 F_1-score

$$\mathrm{F}_1 \;=\; 2\,\frac{\mathrm{P} \cdot \mathrm{R}}{\mathrm{P} + \mathrm{R}}. \tag{4}$$

1.3 Matthews Correlation Coefficient (MCC)

$$\mathrm{MCC} = \frac{\mathrm{TP}\,\mathrm{TN} - \mathrm{FP}\,\mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}. \tag{5}$$

1.4 Area Under the ROC Curve (AUC)

The threshold-independent metric

$$\mathrm{AUC} = \int_0^1 \mathrm{TPR}(\mathrm{FPR})\,\mathrm{d}(\mathrm{FPR}), \tag{6}$$

where $\mathrm{TPR} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$ and $\mathrm{FPR} = \mathrm{FP}/(\mathrm{FP} + \mathrm{TN})$.

2 Loss Functions

2.1 Binary Cross-Entropy (Neural Network)

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\big[y_i \,\log \hat{p}_i + (1-y_i)\,\log(1-\hat{p}_i)\big]. \tag{7}$$

2.2 CatBoost Ordered Boosting Objective

CatBoost minimises the additive loss

$$\min_{f\in\mathcal{F}} \sum_{i=1}^{N} \ell\big(y_i, f\big(\boldsymbol{x}_i\big)\big) + \lambda\,\|f\|^2_{\text{model}}, \tag{8}$$

where ℓ is the log-loss, $\|\cdot\|_{\text{model}}$ the model-specific regulariser and $\lambda > 0$ the L2 term tuned over $[1, 10]$.

3 Log_2 MIC Transformation

$$\text{log2\_mic} = \log_2\big(\text{MIC} \,/\, \text{mg}\,\text{L}^{-1}\big). \tag{9}$$

4 Bootstrapped Confidence Intervals

For an estimator $\hat{\theta}$ and B bootstrap resamples $\{\hat{\theta}^{*(b)}\}_{b=1}^{B}$:

$$\text{CI}_{95\%} = \big[\,\hat{\theta}^{*(0.025)},\ \hat{\theta}^{*(0.975)}\,\big]. \tag{10}$$

5 DeLong Test for Paired ROC Curves

Given two vectors of AUC contributions V_1 and V_2:

$$\Delta = \text{AUC}_1 - \text{AUC}_2, \tag{11}$$

$$\sigma^2_{\Delta} = \frac{\text{Var}(V_1 - V_2)}{N}, \tag{12}$$

$$z = \frac{\Delta}{\sigma_{\Delta}}, \qquad p = 2\big(1 - \Phi(|z|)\big), \tag{13}$$

where $\Phi(\cdot)$ is the standard normal CDF.

6 Feed-Forward Neural Network Forward Pass

$$\boldsymbol{h}_1 = \mathrm{ReLU}\big(\boldsymbol{W}_1 \boldsymbol{x} + \boldsymbol{b}_1\big), \tag{14}$$

$$\boldsymbol{h}_2 = \mathrm{ReLU}\big(\boldsymbol{W}_2 \boldsymbol{h}_1 + \boldsymbol{b}_2\big), \tag{15}$$

$$\hat{y} = \sigma\big(\boldsymbol{w}_3^{\top} \boldsymbol{h}_2 + b_3\big), \tag{16}$$

with $\sigma(z) = 1/(1 + e^{-z})$.

7 Bayesian Optimisation Search Space

$$\text{depth} \sim \mathcal{U}\{4,\, 10\}, \tag{17}$$

$$\eta\,(\text{learning rate}) \sim \mathcal{U}(0.005,\, 0.3), \tag{18}$$

$$\lambda_{\mathrm{L2}} \sim \mathcal{U}(1,\, 10). \tag{19}$$

8 Stratified k-Fold Cross-Validation

Let $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ and strata $S_1, \ldots, S_K$ with class proportions preserved; then

$$\mathrm{CV}(\theta) = \frac{1}{K} \sum_{k=1}^{K} \mathcal{M}\big(\theta;\, \mathcal{D}_{\mathrm{train}}^{(k)}, \mathcal{D}_{\mathrm{test}}^{(k)}\big), \tag{20}$$

where $\mathcal{M}$ is a chosen metric (e.g. AUC).

End of Appendix