Right-Censored Regression for IBD with 9-Fold CatBoost and XGBoost Stacking: A Tobit-Corrected Ensemble

Abstract

The IBD target is right-censored at $c = 20$: values above 20 are recorded as 20. Ordinary MSE training on the observed label $Y = \min(Y^*, c)$ is biased near the upper tail. This report specifies a censor-aware stacking pipeline that (i) trains CatBoost only on uncensored cases ($Y < c$) using $K = 9$-fold out-of-fold (OOF) predictions, (ii) estimates a robust residual scale σ on uncensored residuals, (iii) applies a Tobit-style tail expectation correction to all base and meta predictions, and (iv) fits an XGBoost meta-learner on the corrected OOF signal. The final predictor refits CatBoost on *all* uncensored data and uses a double Tobit correction (base and meta) for coherent inference under censoring. NaN-labeled rows are treated as a pure scoring set and never used for fitting.

"Data rarely lie, but they are often gagged by measurement."

1. Problem and Notation

Let $Y_i^* \in \mathbb{R}$ denote the latent continuous IBD and $Y_i = \min(Y_i^*, c)$ the observed label with censoring threshold $c = 20$. Let $\delta_i = \mathbb{I}\{Y_i < c\}$ indicate uncensored observations. Let $X_i \in \mathbb{R}^p$ be features and $\mu(X) = \mathbb{E}[Y^* \mid X]$ the conditional mean of the latent outcome. The goal is to learn a predictor $\hat{\mu}(X)$ that is well calibrated despite censoring and to output predictions consistent with the right-censoring mechanism.

2. Theoretical Frame and Rival Models

Definition 1 (Tobit model). *Assume $Y^* \mid X \sim \mathcal{N}(\mu(X), \sigma^2)$ and observe $Y = \min(Y^*, c)$. The likelihood factors into density terms for $\delta = 1$ and survival terms for $\delta = 0$.*

Lemma 1 (Truncated normal tail mean). *If $Z \sim \mathcal{N}(0, 1)$, then $\mathbb{E}[Z \mid Z > z] = \varphi(z)/\{1 - \Phi(z)\}$, where φ and Φ are the standard normal pdf and cdf.*

Proposition 1 (Tobit tail correction). *Under homoscedastic normal errors, the conditional expectation for censored cases satisfies*

$$\mathbb{E}[Y^* \mid Y = c,\, X] = c + \sigma \frac{\varphi\left(\frac{c-\mu(X)}{\sigma}\right)}{1 - \Phi\left(\frac{c-\mu(X)}{\sigma}\right)}.$$

Rival approaches. *(i) Tobit MLE* directly optimizes the censored likelihood; *(ii) censored quantile/CLAD* relaxes normality; *(iii) custom censored losses for trees*. The present design instead decouples mean learning (trees) and tail reconstruction (Tobit correction).

3. Mechanism with Equations

Training set partition. Let $U = \{i : Y_i \text{ observed and } Y_i < c\}$ be the uncensored subset. Only U is used for fitting base learners and estimating σ.

Base learner. A K-fold ($K = 9$) OOF scheme trains CatBoost models $\{\text{CatBoost}_k\}_{k=1}^{K}$ on $U \setminus U_k$ and predicts U_k to obtain an OOF vector $\hat{\mu}^{\text{OOF}}_{\text{CatBoost}}$. A final $\text{CatBoost}_{\text{full}}$ is then refit on all of U.

Robust scale estimation. Let residuals on U be $r_i = Y_i - \hat{\mu}_{\text{CatBoost}_{\text{full}}}(X_i)$. A robust scale is

$$\hat{\sigma} \;=\; 1.4826 \cdot \text{MAD}(r) \;=\; 1.4826 \cdot \text{median}_i(|r_i - \text{median}_j\, r_j|)\,,$$

with a small lower bound to prevent degeneracy.

Tobit correction operator. For any prediction $\tilde{\mu}$ define

$$\text{TobitCorr}(\tilde{\mu};\, c, \hat{\sigma}) \;=\; \begin{cases} \tilde{\mu}, & \tilde{\mu} < c, \\ c + \hat{\sigma}\, \dfrac{\varphi\big(\frac{c-\tilde{\mu}}{\hat{\sigma}}\big)}{1 - \Phi\big(\frac{c-\tilde{\mu}}{\hat{\sigma}}\big)}, & \tilde{\mu} \geq c. \end{cases}$$

Meta-learner. Construct the meta-feature $s^{\text{OOF}} = \text{TobitCorr}\big(\hat{\mu}^{\text{OOF}}_{\text{CatBoost}};\, c, \hat{\sigma}\big)$. Train an XGBoost regressor on (s^{OOF}, Y) for $i \in U$, producing $\widehat{g}(\cdot)$. Let residuals of $\widehat{g}$ on U define a meta-scale $\hat{\sigma}_{\text{meta}}$ via MAD.

Final predictor. For a new x:

$$\hat{\mu}_{\text{CatBoost}}(x) \;\xrightarrow{\text{corr}}\; s(x) = \text{TobitCorr}\big(\hat{\mu}_{\text{CatBoost}}(x);\, c, \hat{\sigma}\big) \;\xrightarrow{\text{XGBoost}}\; \hat{y}_{\text{raw}}(x) = \widehat{g}\big(s(x)\big)$$

$$haty_{\text{raw}}(x) = \widehat{g}\big(s(x)\big) \;\xrightarrow{\text{corr}}\; \hat{y}_{\text{final}}(x) = \text{TobitCorr}\big(\hat{y}_{\text{raw}}(x);\, c, \hat{\sigma}_{\text{meta}}\big).$$

4. Cross-Disciplinary Map

- **Econometrics**: Tobit and limited dependent variable models for measurement-limited outcomes.
- **Survival**: right-censoring uses survival term $1 - \Phi(\cdot)$, analogous to non-observed event times.
- **Machine Learning**: stacking reduces variance and specification error; robust scale via MAD handles heavy-tailed residuals.

5. Thesis–Antithesis–Synthesis

Thesis. Train a regressor on Y with MSE. Simple but biased under censoring.
Antithesis. Full Tobit MLE handles censoring but may be brittle with many features and nonlinearity.
Synthesis. Learn $\mu(X)$ with strong trees on *uncensored* data, then restore tail mass with Tobit correction; combine via stacking for stability.

6. Operational Algorithm (Pseudocode)

Algorithm 1 Censored Stacking with $K = 9$

Require: dataset $\{(X_i, Y_i)\}$, censor c, folds $K = 9$
$U \leftarrow \{i : Y_i \text{ observed and } Y_i < c\}$ ▷ train only on uncensored
OOF for CatBoost
for $k = 1$ to K **do**
  Fit CatBoost_k on $U \setminus U_k$
  $\hat{\mu}^{\text{OOF}}_{\text{CatBoost}}[U_k] \leftarrow \text{CatBoost}_k(X_{U_k})$
end for
Fit $\text{CatBoost}_{\text{full}}$ on all U and compute residuals $r_i = Y_i - \text{CatBoost}_{\text{full}}(X_i)$
$\hat{\sigma} \leftarrow 1.4826 \cdot \text{MAD}(r)$
$s^{\text{OOF}} \leftarrow \text{TobitCorr}\big(\hat{\mu}^{\text{OOF}}_{\text{CatBoost}};\, c, \hat{\sigma}\big)$
Fit meta XGBoost on (s^{OOF}, Y) for $i \in U$; compute $\hat{\sigma}_{\text{meta}}$ via MAD on meta residuals
Prediction for new x**:**
$\hat{\mu} \leftarrow \text{CatBoost}_{\text{full}}(x)$; $s \leftarrow \text{TobitCorr}(\hat{\mu}; c, \hat{\sigma})$
$\hat{y}_{\text{raw}} \leftarrow \text{XGBoost}(s)$; $\hat{y}_{\text{final}} \leftarrow \text{TobitCorr}(\hat{y}_{\text{raw}}; c, \hat{\sigma}_{\text{meta}})$

7. Assumptions, Limits, and Testable Predictions

Assumptions. (i) Residuals on U are roughly symmetric; (ii) a single scale σ is adequate; (iii) the relation learned on U extrapolates toward the censored region.
Limits. If censoring probability depends strongly on X and alters the response mechanism in the upper tail, training only on U can shift $\hat{\mu}$. Heteroscedasticity degrades a single σ.
Mitigations. Model $\sigma(X)$ (auxiliary regressor for $|r|$), use censored quantile/CLAD, or implement a custom censored loss.
Predictions to test.

- Tail calibration: the predicted fraction with $\hat{y}_{\text{final}} \geq c$ should align with observed censoring rates conditional on X-strata.
- Improvement from correction: OOF RMSE/MAE decrease when replacing raw with Tobit-corrected signals.
- Stability: OOF metrics across folds are consistent; meta-learner reduces variance against single-model baselines.

8. Practical Notes

- **NaN targets** act as an external scoring/validation set and never participate in fitting.
- After OOF is built, the base CatBoost is refit on *all* uncensored data to maximize information usage.
- Checkpoints: per-fold CatBoost models, a final CatBoost, an XGBoost booster, the meta ridge (if used), and CSV artifacts for OOF and scored predictions.

9. Symbolic Recap

$$\boxed{\hat{Y}_{\text{final}}(X) \;=\; \text{TobitCorr}\Big(\underbrace{\hat{g}\big(\text{TobitCorr}(\hat{\mu}_{\textsc{CatBoost}}(X);\ c,\ \hat{\sigma})\big)}_{\text{meta on corrected base}};\ c,\ \hat{\sigma}_{\text{meta}}\Big),}$$

$$\boxed{\hat{\mu}_{\textsc{CatBoost}} = \text{9-fold OOF/final on } \{Y < c\}.}$$

References

- J. Tobin (1958). Estimation of relationships for limited dependent variables.
- T. Amemiya (1984). Tobit models: A survey.
- J. L. Powell (1984, 1986). Censored regression quantiles; Symmetrically trimmed least squares.
- D. H. Wolpert (1992). Stacked generalization.
- W. H. Greene. *Econometric Analysis*. J. M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*.