¿Cómo se deben tratar los datos 'censurados a la derecha' en encuestas de empleo cuando se utilizan algoritmos de regresión como XGBoost?¿Existe literatura científica que respalde la asignación de un valor máximo de observación ($T=max+1$) para individuos que no han experimentado el evento de inserción laboral en estudios de cohorte?¿Qué técnicas de preprocesamiento se recomiendan para modelos de regresión cuando la variable objetivo presenta una alta concentración en el valor cero?

Right censoring and many zeros require specialized methods; treating censored times as (T=\text{max}+1) is not scientifically standard and can bias XGBoost-type regresiones.

1. ¿Cómo tratar datos censurados a la derecha con XGBoost?

En estudios de tiempo hasta inserción laboral, los individuos aún desempleados al final del seguimiento tienen **tiempos censurados**, no “tiempos largos conocidos”.
La literatura sugiere tres familias de enfoques:

- **Modelos específicos de censura**: supervivencia paramétrica/no paramétrica, Cox, modelos con predictor o respuesta censurada, o regresión lineal con datos censurados (Koul et al., 1981; Matsouaka & Atem, 2020).
- **Ponderación por probabilidad inversa de censura (IPCW)**: adaptar modelos de regresión estándar ponderando observaciones según su probabilidad de no estar censuradas (Blanche et al., 2022; Matsouaka & Atem, 2020).
- **Imputación de valores censurados**: imputar la cola superior con modelos Tobit, cuantílicos o de distribución de cola (Pareto, GB2) para luego usar cualquier regresor, incluido XGBoost, pero con gran cuidado por el sesgo (Drechsler & Ludsteck, 2025; Beckmannshagen et al., 2025).

Ninguna de estas líneas recomienda simplemente fijar un valor único máximo para todos los censurados.

2. ¿Es válido asignar (T=max+1) a no insertados?

En la literatura de imputación para datos censurados (salarios top-coded, etc.) se muestran grandes sesgos cuando se usan reglas simplistas como truncar o fijar todos los censurados en el umbral o en una constante (Drechsler & Ludsteck, 2025; Beckmannshagen et al., 2025).
Los autores enfatizan que tales estrategias son **“simplistic imputation”** y “obviously problematic” para análisis de regresión, atenuando coeficientes y distorsionando la distribución (Drechsler & Ludsteck, 2025; Beckmannshagen et al., 2025).
Por analogía, asignar (T=max+1) a no insertados se alinea con estas prácticas desaconsejadas: ignora la naturaleza de la censura y produce estimadores sesgados.

3. Variable objetivo con muchos ceros: técnicas recomendadas

Cuando la variable de resultado es continua, no negativa y con **exceso de ceros**, se recomiendan enfoques específicos:

- **Modelos Tobit y dos-partes (hurdle)**: primera parte modela la probabilidad de ser >0, segunda parte el valor continuo condicional en >0 (Boulton & Williford, 2018).
- **Modelos zero-inflated (ZIP/ZINB)** y extensiones con boosting para datos de conteo con muchos ceros (Lee, 2020; Sidumo et al., 2023; Kim et al., 2024).
- Para regresión general con fuerte desbalance, pueden usarse técnicas de **re-muestreo y ponderación focalizadas en regiones de interés** (p.ej. WERCS, oversampling con ruido gaussiano) que mejoran el rendimiento de regresores estándar (Branco et al., 2019).

Propuestas prácticas de preprocesamiento

Problema	Enfoque sugerido	Citaciones
Censura a la derecha en tiempo	Supervivencia / IPCW / imputación basada en modelo, no (T=max+1)	(Koul et al., 1981; Blanche et al., 2022; Matsouaka & Atem, 2020; Drechsler & Ludsteck, 2025; Beckmannshagen et al., 2025)
Exceso de ceros continuos	Tobit, modelos de dos partes, zero-inflated, técnicas de re-muestreo en regresión	(Boulton & Williford, 2018; Lee, 2020; Sidumo et al., 2023; Branco et al., 2019; Kim et al., 2024)

FIGURE 1 Métodos recomendados para censura y exceso de ceros

Conclusión

Para cohortes de inserción laboral con censura a la derecha, lo metodológicamente sólido es usar modelos de supervivencia o, si se mantiene XGBoost, incorporar IPCW o imputación basada en modelos, evitando reglas ad-hoc como (T=max+1). Cuando la variable objetivo se concentra en cero, son preferibles modelos Tobit / dos-partes o técnicas de re-muestreo y zero-inflated frente a la simple regresión estándar.

References

Lee, S. (2020). ADDRESSING IMBALANCED INSURANCE DATA THROUGH ZERO-INFLATED POISSON REGRESSION WITH BOOSTING. *ASTIN Bulletin*, 51, 27 - 55. https://doi.org/10.1017/asb.2020.40

Koul, H., Susarla, V., & Ryzin, J. (1981). Regression Analysis with Randomly Right-Censored Data. *Annals of Statistics*, 9, 1276-1288. https://doi.org/10.1214/aos/1176345644

Boulton, A., & Williford, A. (2018). Analyzing Skewed Continuous Outcomes With Many Zeros: A Tutorial for Social Work and Youth Prevention Science Researchers. *Journal of the Society for Social Work and Research*, 9, 721 - 740. https://doi.org/10.1086/701235

Blanche, P., Holt, A., & Scheike, T. (2022). On logistic regression with right censored data, with or without competing risks, and its use for estimating treatment effects. *Lifetime Data Analysis*, 29, 441-482. https://doi.org/10.1007/s10985-022-09564-6

Sidumo, B., Sonono, E., & Takaidza, I. (2023). Count Regression and Machine Learning Techniques for Zero-Inflated Overdispersed Count Data: Application to Ecological Data. *Annals of Data Science*, 1-15. https://doi.org/10.1007/s40745-023-00464-6

Drechsler, J., & Ludsteck, J. (2025). Imputation strategies for rightcensored wages in longitudinal datasets. *Journal for Labour Market Research*, 59. https://doi.org/10.1186/s12651-025-00410-4

Branco, P., Torgo, L., & Ribeiro, R. (2019). Pre-processing approaches for imbalanced distributions in regression. *Neurocomputing*, 343, 76-99. https://doi.org/10.1016/j.neucom.2018.11.100

Matsouaka, R., & Atem, F. (2020). Regression with a right-censored predictor using inverse probability weighting methods. *Statistics in Medicine*, 39, 4001 - 4015. https://doi.org/10.1002/sim.8704

Kim, J., Ha, I., & Kim, S. (2024). Copula deep learning control chart for multivariate zero inflated count response variables. *Statistics*, 58, 749 - 769. https://doi.org/10.1080/02331888.2024.2364688

Beckmannshagen, M., König, J., Retter, I., Schluter, C., Schröder, C., & Tchokni, Y. (2025). Dealing with Censored Earnings in Register Data. *Jahrbücher für Nationalökonomie und Statistik*, 0. https://doi.org/10.1515/jbnst-2024-0037