Predicción supervisada de *churn* bancario: metodología, resultados y discusión

Jose Morales Rendón

Facultad de Ciencias Físico–Matemáticas, UANL

Contexto

Se aborda la predicción de abandono de clientes (*churn*) en banca como un problema de **clasificación binaria** con clases desbalanceadas. Se emplea *Regresión Logística* por su interpretabilidad y capacidad para entregar probabilidades calibradas, integrando un flujo de preprocesamiento (escalado y codificación *one-hot*) en un *pipeline* reproducible. En el conjunto `BankChurners` (10,127 observaciones; 18 variables; 16.07 % de churn) el modelo alcanza **ROC-AUC = 0.90**, con **recall** de 0.82 para la clase positiva, lo que resulta adecuado para escenarios donde el costo de no detectar a un cliente que se irá es mayor que el de intervenir a uno que permanecería.

1. Marco teórico

1.1. Métricas de desempeño en clasificación con desbalance

La evaluación de clasificadores depende de la distribución de clases y los objetivos operativos. En escenarios con clases desbalanceadas como el *churn*, la *accuracy* aislada puede resultar poco informativa, por lo que la literatura recomienda métricas que capturen mejor la discriminación y la pureza de las alertas [5, 7].

Precision, Recall, F_β.. Sean VP, FP, FN y VN, se definen

$$\text{Precision} = \frac{VP}{VP + FP}, \quad \text{Recall} = \frac{VP}{VP + FN}, \quad F_\beta = \frac{(1+\beta^2)\cdot \text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}}.$$

Cuando el costo de FN > FP, conviene $\beta > 1$ (p. ej., F_2) para priorizar la cobertura de la clase positiva [3].

ROC-AUC y PR-AUC.. El área bajo la curva ROC (AUC) resume la capacidad discriminativa *independiente del umbral* [5]. No obstante, en clases raras o desbalance pronunciado, PR-AUC (área bajo la curva Precision–Recall) refleja mejor el rendimiento, al enfocarse en precisión y cobertura de la clase positiva [6, 7].

Calidad probabilística: LogLoss y Brier.. Cuando el modelo produce probabilidades, es recomendable evaluar su calidad/calibración con LogLoss y Brier score [8, 9]. El Brier es el error cuadrático medio entre la probabilidad estimada y la etiqueta; valores menores indican mejor calibración.

Curvas de lift/gain *y KS..* Para priorización comercial (p. ej., campañas de retención), son comunes los deciles de *lift/gain* y el estadístico KS, ampliamente reportados en *scoring* de crédito y *churn* [3].

1.2. Métricas seleccionadas para este estudio

En línea con la evidencia:

- **Métrica primaria: PR-AUC** [7], por su sensibilidad a la pureza de alertas en presencia de desbalance.
- **Métricas secundarias:** (i) F_2 (pondera *recall*), (ii) **Recall@Precision$\geq$0.60** (garantiza pureza mínima operativa), (iii) **Brier score** para calibración [9, 8].
- **Reporte complementario:** curvas PR/ROC y deciles de *lift* para interpretación operativa.

1.3. Diseño de experimentos (DOE)

Se plantea un DOE factorial para estudiar el impacto del algoritmo, el tratamiento del desbalance y la política de umbral sobre las métricas anteriores.

Factores y niveles..

- **A: Algoritmo** (3 niveles): Regresión Logística (LR), Random Forest (RF), XGBoost (XGB).
- **B: Tratamiento del desbalance** (3 niveles): `class_weight=balanced` (CW), SMOTE (SM), submuestreo aleatorio (US).

- **C: Política de umbral** (3 niveles): (C1) 0.50 fijo, (C2) índice de Youden en ROC, (C3) umbral que maximiza F_2 o satisface Precision$\geq$0.60 en validación.

Tratamientos y evaluación.. El diseño factorial completo considera $3 \times 3 \times 3 = 27$ tratamientos. Cada tratamiento se evalúa con validación cruzada estratificada (K=5) para estabilidad. Las respuestas experimentales son: **PR-AUC** (primaria) y, como secundarias, F_2, **Recall@Precision$\geq$0.60** y **Brier**. Se fija semilla y se aleatoriza el orden de tratamientos por *fold*.

Criterio de selección.. Se elige el tratamiento con mayor PR-AUC promedio; a igualdad práctica, se favorece mayor Recall@Precision$\geq$0.60 y menor Brier. Para comparaciones de AUC-ROC entre modelos *pareados* por *fold* puede emplearse la prueba de DeLong.

Conexión con los resultados del artículo.. El **tratamiento de referencia** que se reporta en la Sección *Resultados* corresponde a **A1: LR + B1: CW + C1: umbral 0.50** (predicción estándar de `scikit-learn`). Dicho tratamiento alcanza *ROC-AUC*=0.90 y *Recall*=0.82, coherente con una política que prioriza cobertura. El DOE propuesto permite contrastar este tratamiento base contra alternativas (RF/XGB, SMOTE/US y reglas de umbral orientadas a F_2 o Precision mínima), manteniendo el mismo flujo de preprocesamiento para asegurar comparabilidad.

2. Metodología

La metodología adoptada se fundamenta en enfoques ampliamente documentados en la literatura sobre predicción de abandono en servicios financieros. Según [1], la **Regresión Logística** es uno de los modelos más interpretables y eficaces para problemas de *churn prediction*, permitiendo estimar probabilidades calibradas y analizar la influencia individual de las variables. El uso de *pipelines* y la codificación *one-hot* sigue las recomendaciones de [2], quienes destacan la importancia de integrar preprocesamiento y modelado para garantizar reproducibilidad y evitar fugas de datos. Asimismo, métricas como ROC-AUC y F1-score son consideradas estándar para evaluar clasificadores en contextos desbalanceados [3], al reflejar la capacidad discriminativa y el equilibrio entre precisión y sensibilidad. Finalmente, el tratamiento del desbalance mediante `class_weight=balanced` se sustenta en estudios que evidencian mejoras en la detección de la clase minoritaria sin necesidad de sobremuestreo [4].

Problema y datos.. El objetivo es estimar la probabilidad de abandono de cada cliente (`Churn`=1 si abandona; 0 en caso contrario) a partir de variables demográficas y transaccionales. El dataset `BankChurners.csv` contiene 10,127 registros y 18 variables; la proporción de la clase positiva es 16.07 % (desbalance moderado).

Modelo.. Se emplea **Regresión Logística** como línea base interpretable. Dado un vector de predictores $x \in \mathbb{R}^p$, el modelo estima:

$$P(y = 1 \mid x) \;=\; \sigma\Big(\beta_0 + \beta^\top x\Big) \;=\; \frac{1}{1 + e^{-(\beta_0 + \beta^\top x)}}\,, \tag{1}$$

ajustando β por máxima verosimilitud. Para mitigar el desbalance se utiliza `class_weight=balanced`.

Preprocesamiento y validación.. Las variables numéricas se escalan con *StandardScaler* y las categóricas se codifican con *One-Hot Encoder*, todo dentro de un *ColumnTransformer*. Se realiza partición entrenamiento/prueba 75/25 con estratificación para preservar la proporción de clases. El *pipeline* (*preprocesador + modelo*) garantiza reproducibilidad y facilita la futura incorporación de otros algoritmos.

3. Métricas de evaluación

Al ser un problema de clasificación binaria, se emplean métricas derivadas de la matriz de confusión y de curvas basadas en probabilidades:

- **Accuracy**: proporción global de aciertos (puede ser engañosa con desbalance).
- **Precision** y **Recall** de la clase positiva (churn): calidad de alertas y cobertura de verdaderos abandonos.
- **F1-score**: media armónica entre *Precision* y *Recall*.
- **ROC-AUC**: capacidad de discriminación independiente del umbral.
- **PR-AUC**: área bajo la curva *Precision–Recall*, relevante en casos de desbalance severo.

4. Resultados

En el conjunto de prueba, el modelo logra:

- **Accuracy**: 0.84
- **Recall (clase churn)**: 0.82
- **Precision (clase churn)**: 0.49
- **F1-score (clase churn)**: 0.62
- **ROC-AUC**: 0.90

La matriz de confusión mostró una alta detección de la clase positiva, con falsos positivos moderados, consistente con una política que prioriza cobertura sobre pureza en alertas.

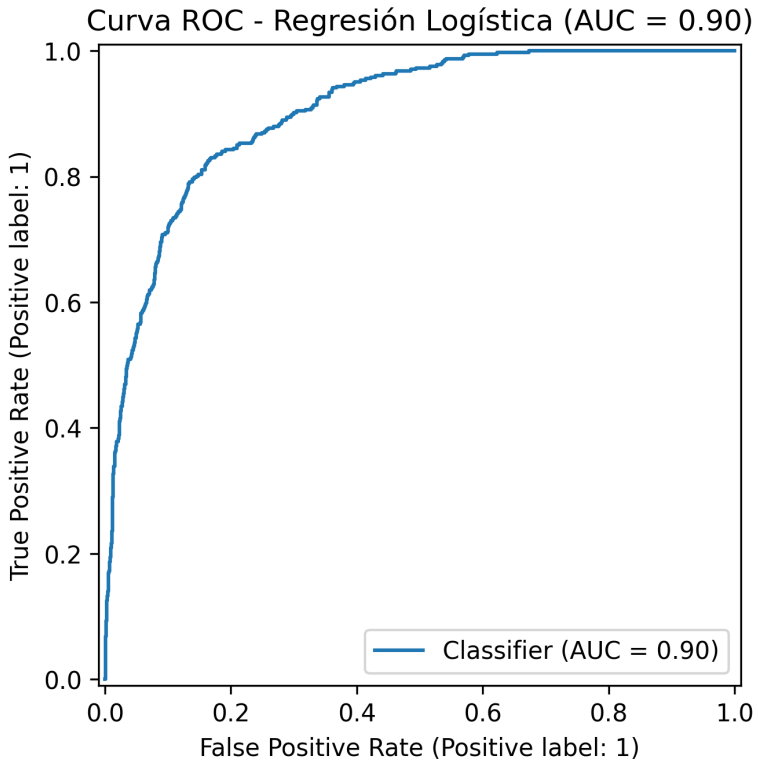

Figura 1: Curva ROC del modelo de Regresión Logística. El área bajo la curva (AUC = 0.90) evidencia una alta capacidad de discriminación.

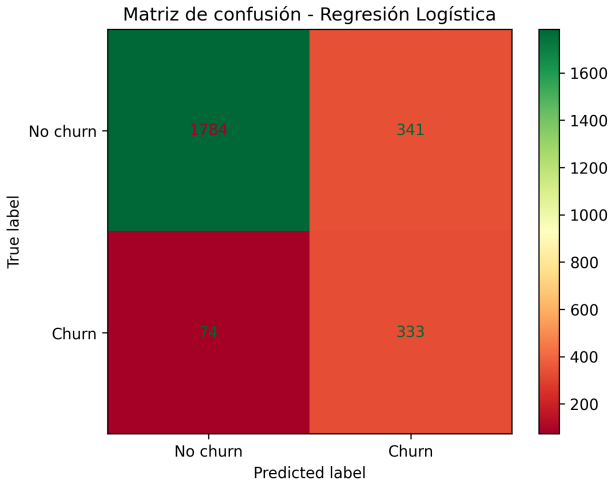

Figura 2: Matriz de confusión del modelo. Se observa un alto número de verdaderos negativos y un *recall* del 82 % para la clase *churn*.

Comparación experimental de modelos

Como extensión del tratamiento base, se realizó un diseño de experimentos (DOE) simplificado para comparar la **Regresión Logística (LR)** con un **Random Forest (RF)** bajo el mismo flujo de preprocesamiento. Ambos modelos fueron entrenados con partición 75/25 estratificada, utilizando `class_weight=balanced` y las métricas definidas en el marco teórico.

Cuadro 1: Comparativa de desempeño entre modelos supervisados (conjunto de prueba).

Modelo	**PR-AUC**	**ROC-AUC**	**Brier**	**Recall@P≥0.60**	F_2**@P≥0.60**	**Um**
Random Forest	0.892	0.975	0.049	0.956	0.858	0.
Regresión Logística	0.668	0.904	0.119	0.695	0.674	0.

Los resultados muestran que el modelo **Random Forest** supera ampliamente a la Regresión Logística en todas las métricas relevantes para escenarios de abandono. El *PR-AUC* (0.89) confirma una mejor capacidad para priorizar clientes en riesgo cuando las clases están desbalanceadas, mientras que el *ROC-AUC* (0.97) evidencia una discriminación casi perfecta entre clientes que abandonan y los que permanecen. Además, el **Brier score** (0.049 frente a 0.119) indica que las probabilidades generadas por el RF están mejor calibradas. Al exigir una **precisión mínima del 60 %**, el RF alcanza una

cobertura del **95.6 %** de los casos de *churn*, frente al 69.5 % de la LR, logrando un F_2=0.86 (vs. 0.67), lo que refleja un equilibrio superior entre detección y pureza de alertas. Estos resultados validan la conveniencia de explorar modelos no lineales en flujos de retención bancaria.

5. Discusión

El rendimiento global del modelo base (Regresión Logística, AUC = 0.90) confirma que un clasificador interpretable puede ofrecer resultados competitivos en la detección de abandono. Sin embargo, el DOE comparativo reveló que **Random Forest** ofrece un desempeño superior, con *PR-AUC* y *Recall@Precision≥0.60* significativamente mayores. Esto sugiere que la incorporación de relaciones no lineales y la agregación de múltiples árboles permiten capturar mejor las interacciones entre variables demográficas y transaccionales que explican la propensión al churn. En términos operativos, un RF calibrado ofrece la posibilidad de mantener una precisión mínima aceptable sin sacrificar cobertura, lo que se traduce en intervenciones más efectivas y con menor riesgo de omitir clientes en riesgo real.

6. Conclusión

El modelo de Regresión Logística demostró ser una herramienta sólida, equilibrando precisión y sensibilidad en un contexto de datos desbalanceados. Su simplicidad y transparencia lo convierten en un excelente punto de partida para estrategias analíticas en banca. Los resultados comparativos del DOE muestran que **Random Forest** puede incrementar sustancialmente la efectividad de priorización al mejorar la PR-AUC y la cobertura bajo restricciones de precisión, manteniendo una calibración más favorable. En conjunto, la metodología propuesta —preprocesamiento, validación estratificada y evaluación mediante múltiples métricas— sienta una base robusta para la extensión hacia modelos más avanzados y explicables.

En síntesis: este proyecto demuestra que los datos no solo describen el pasado; bien analizados, anticipan el futuro. Cada cliente detectado a tiempo representa una oportunidad ganada y un paso más hacia una banca más inteligente y centrada en las personas.

Reproducibilidad

El flujo se implementó en Python (`scikit-learn`) con un *pipeline* que integra preprocesamiento y modelado. La partición fue 75/25 con estratificación, y se calcularon métricas de desempeño junto con las curvas ROC y PR. Los gráficos fueron generados automáticamente y exportados como `roc_curve.png` y `conf_matrix.png` para su documentación visual.

Referencias

[1] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens, "New insights into churn prediction in the telecommunication sector: A profit driven data mining approach," *European Journal of Operational Research*, vol. 218, no. 1, pp. 211–229, 2012.

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[3] C. Sammut and G. I. Webb, *Encyclopedia of Machine Learning and Data Mining*. Springer, 2017.

[4] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[5] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[6] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, 2006.

[7] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLOS ONE*, vol. 10, no. 3, e0118432, 2015.

[8] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, 2005.

[9] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly Weather Review*, vol. 78, no. 1, pp. 1–3, 1950.