

Machine Learning para predecir cancelaciones y mejorar la retención en seguros

Luis Carlos Ocaña Hoeber

Universitat Oberta de Catalunya
Máster Universitario en Ciencia de Datos
Data Analytics in Industrial and Business Environments

3 de junio de 2025

- 1 Contexto y motivación
- 2 Objetivos del proyecto
- 3 Metodología
- 4 Datos
- 5 Métricas de evaluación
- 6 Modelos evaluados
- 7 Modelo resultante
- 8 Conclusiones y trabajo futuro

¿Por qué abordar el Churn?

- ⚠ **Riesgo real:** La fuga de clientes impacta en los beneficios y la sostenibilidad.
- 👤+ **Retener es 5-7x más barato** que captar un nuevo cliente.

Nuestra oportunidad

- 📈 El Machine Learning permite anticiparse y personalizar estrategias de retención.
- 💡 Convertir datos en acciones: ¡predecir, intervenir y fidelizar!

“Más datos, mejores decisiones, clientes más felices.”

Objetivos del proyecto

© Objetivo principal

Predecir la cancelación de pólizas con Machine Learning para mejorar la retención de clientes.

☰ Objetivos secundarios

- 🔍 Analizar y visualizar datos de clientes.
- ⚙️ Comparar y optimizar distintos modelos de ML (*Random Forest*, *XGBoost*, Redes Neuronales...).
- 📊 Evaluar rendimiento con métricas claras (*F1-score*, *AUC*).
 - 💡 Interpretar resultados y extraer factores clave de abandono.
 - 👍 Proponer estrategias prácticas para aumentar la fidelización.

Metodología CRISP-DM

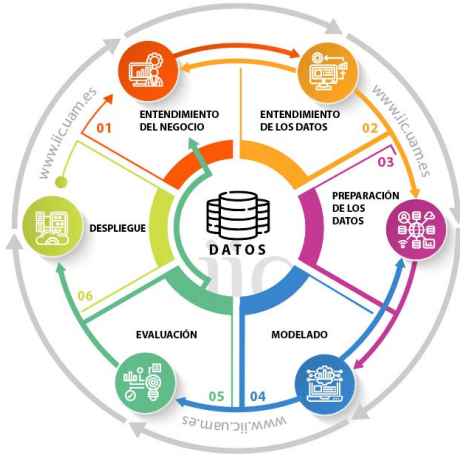


Figura: Diagrama del proceso CRISP-DM

Origen de los datos

- Dataset: **Auto Insurance Churn Analysis** (Kaggle)
- 1,6 millones de registros sintéticos (Texas, EE.UU.)
- Fusión de archivos: cliente, dirección, demografía y cancelaciones

Variables clave del análisis

- Variables personales: edad, antigüedad, hijos, estado civil
- Variables financieras: ingresos, valor vivienda, monto anual pagado
- Otras: propietario, nivel educativo, crédito

Preprocesamiento de datos

Limpieza

Eliminar outliers, registros erróneos y variables sin valor predictivo.



Transformación

Estandarización de numéricas (z-score) y codificación de categóricas (one-hot, binaria).



Balanceo





Oversampling, undersampling y SMOTENC para ajustar clases.



División

70 % entrenamiento / 30 % prueba manteniendo el balance de churn.

Métricas principales

-  **AUC-ROC**: Capacidad de distinguir entre clientes que hacen o no churn.
-  **F1-score**: Equilibrio entre precisión y recall en clase minoritaria.
-  **Precisión**: ¿De los predichos como churn, cuántos lo son realmente?
-  **Recall (Sensibilidad)**: ¿Cuántos churn reales detecta el modelo?

Matriz de confusión

- Analiza aciertos y errores: *TP*, *TN*, *FP*, *FN* para entender los fallos del modelo.

Modelo	Mejor técnica (según AUC)	AUC	F1-score
Regresión Logística	RandomUnderSampler	0.6876	0.2734
Naive Bayes	RandomOverSampler	0.6701	0.2795
KNN	RandomUnderSampler	0.6490	0.2692
Árbol de Decisión	SMOTENC	0.5775	0.2492
Bosque Aleatorio	RandomUnderSampler	0.6943	0.3571
AdaBoost	RandomUnderSampler	0.6943	0.4110
XGBoost	RandomUnderSampler	0.6943	0.4427
Red Neuronal	RandomOverSampler	0.6961	0.4516

Cuadro: Mejor técnica de balanceo por modelo según AUC y su F1-score correspondiente

Modelo resultante

Random Forest con SMOTENC

Comparación de métricas clave

Métrica	Antes	Después	Mejora
AUC	0.6889	0.6937	↑ 0.70 %
F1-score (Churn)	0.3961	0.4533	↑ 14.44 %
Recall (Churn)	0.38	0.45	↑ 18.42 %
Precisión (Churn)	0.41	0.46	↑ 12.20 %

Interpretación

Optimizar hiperparámetros permite al modelo detectar y clasificar mejor a los clientes con riesgo de churn, aunque el margen de mejora en recall y precisión deja claro que sigue habiendo margen para investigar y perfeccionar la estrategia.

Interpretación del modelo

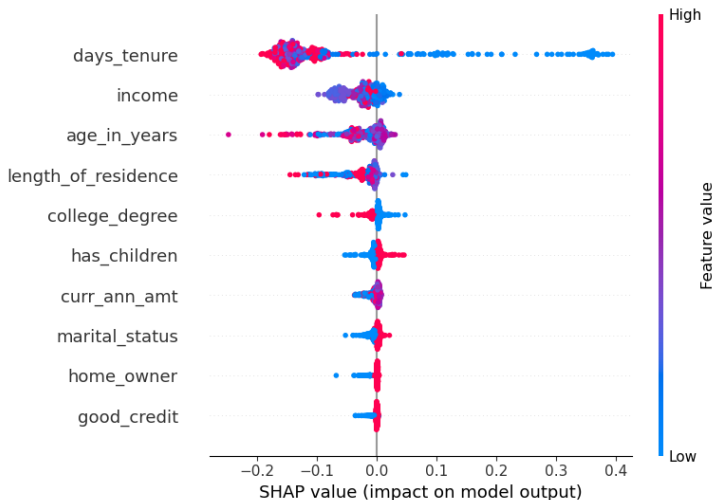


Figura: Interpretación del modelo utilizando valores SHAP

Limitaciones del modelo

Recall limitado para churn

- El modelo sólo identifica correctamente el **45 %** de los clientes que efectivamente hacen churn.

Muchos falsos positivos

- De todos los clientes predichos como churn, sólo el **46 %** realmente abandonan.

Impacto práctico

- Supone un coste operativo contactar a muchos clientes que no iban a irse.
- Aun así, el modelo permite priorizar esfuerzos frente a una intervención totalmente aleatoria.

Recomendaciones estratégicas de retención

Enfoque en clientes nuevos

- Priorizar la fidelización en los primeros 12-18 meses (mayor riesgo de churn).

Incentivos y fidelización

- Ofrecer bonificaciones o descuentos en la primera renovación.
- Desarrollar programas de puntos o beneficios a corto plazo.

Comunicación proactiva

- Explicar claramente beneficios y coberturas.

Estrategias personalizadas

- Clientes jóvenes: apps móviles, comunicación digital, recompensas por buen comportamiento.
- Clientes mayores: atención personalizada y procesos sencillos.

✓ Principales logros

- Modelo Random Forest + SMOTENC: F1-score **0.45** (churn), AUC **0.69**, exactitud global **87.6 %**
- Variables más influyentes: **antigüedad**, ingreso y edad

A Líneas futuras

- Incorporar variables de interacción y datos longitudinales
- Añadir datos de atención al cliente e historial de siniestros
- Analizar el impacto económico de las estrategias de retención
- Seguir mejorando el recall y la precisión del modelo

Machine Learning para predecir cancelaciones y mejorar la retención en seguros

Luis Carlos Ocaña Hoeber

Universitat Oberta de Catalunya
Máster Universitario en Ciencia de Datos
Data Analytics in Industrial and Business Environments

3 de junio de 2025