

Informe Final: Análisis de Fuga de Clientes y Estrategias de Retención

Preparado para: PROGRAMA ORACLE NEXT EDUCATION G8 ALURA

Desafío TelecomX, Parte 2

Fecha: 06 de Agosto, 2025

Autor: Carlos merino

1. Resumen Ejecutivo

El presente informe detalla los resultados del proyecto de predicción de fuga de clientes (churn). Tras un riguroso proceso de limpieza de datos, análisis exploratorio y evaluación de múltiples algoritmos de Machine Learning, se ha desarrollado un modelo predictivo de alto rendimiento.


CORRELACIÓN:

Correlación con 'churn' (de mayor a menor):

	churn
churn	1.000000
internetservice_fiberoptic	0.307463
paymentmethod_electroniccheck	0.301455
Cuentas_Diarias	0.192858
charges.monthly	0.192858
seniorcitizen	0.150541
multiplelines	0.032654
phoneservice	0.011691
gender_male	-0.008545
paymentmethod_mailedcheck	-0.090773
charges.total	-0.123433
streamingtv	-0.128435
streamingmovies	-0.130920
paymentmethod_creditcardautomatic	-0.134687
partner	-0.149982
dependents	-0.163128
contract_oneyear	-0.178225
internetservice_no	-0.227578
deviceprotection	-0.252056
onlinebackup	-0.267595
contract_twoyear	-0.301552
techsupport	-0.336877
onlinesecurity	-0.342235
tenure	-0.354049

Matriz de correlación guardada en 'matriz_correlacion.csv'

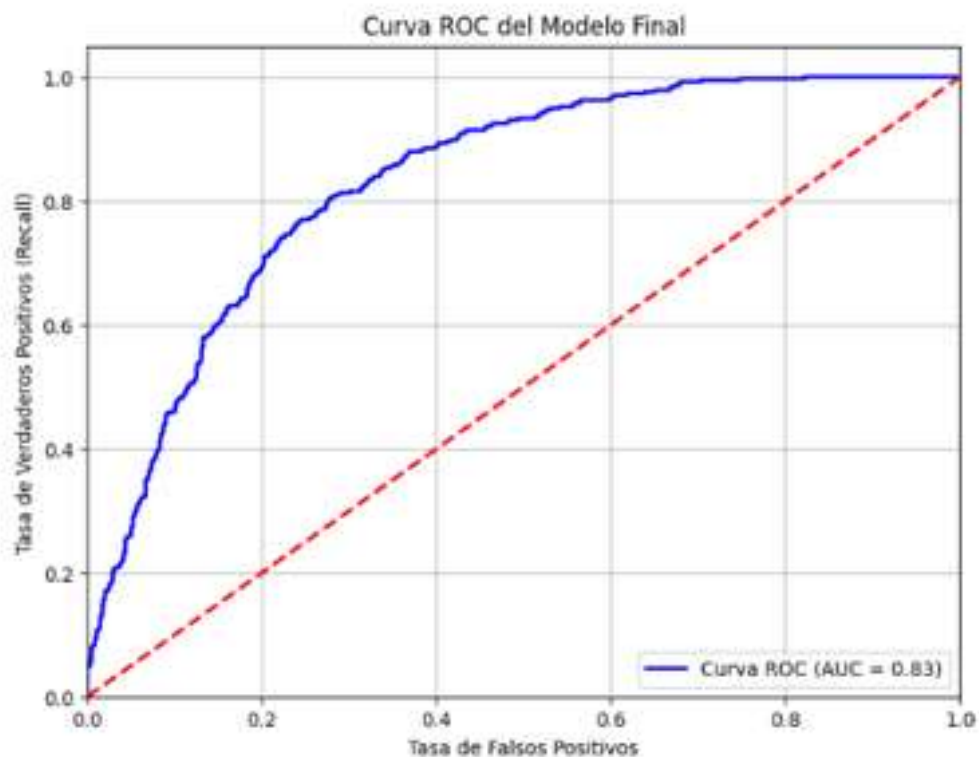
REGRESIÓN LOGÍSTICA:



	Variable	Coefficiente
16	internetservice_fiberoptic	0.614464
11	streamingmovies	0.213359
10	streamingtv	0.211585
5	multiplelines	0.171012
21	paymentmethod_electroniccheck	0.165323
0	seniorcitizen	0.084993
13	charges.total	0.072686
4	phoneservice	0.045131
8	deviceprotection	0.020706
15	gender_male	-0.005361
1	partner	-0.007278
22	paymentmethod_mailedcheck	-0.009792
7	onlinebackup	-0.011907
20	paymentmethod_creditcardautomatic	-0.018821
2	dependents	-0.079199
9	techsupport	-0.121805
6	onlinesecurity	-0.153174
14	Cuentas_Diarias	-0.195340
12	charges.monthly	-0.195340
18	contract_oneyear	-0.291585
19	contract_twoyear	-0.613473
17	internetservice_no	-0.628559
3	tenure	-0.833539

El modelo final, un **LightGBM optimizado mediante Optuna**, alcanzó una puntuación **ROC-AUC de 0.8322** en el conjunto de datos de prueba. Este modelo es capaz de **identificar correctamente al 71% de los clientes que tienen la intención de abandonar el servicio**, proporcionando a la empresa una herramienta poderosa para actuar de manera proactiva.

El análisis de las variables más influyentes revela que los factores clave del churn están relacionados con la **estructura del contrato, la permanencia del cliente (tenure) y los cargos mensuales**. Basado en estos hallazgos, se proponen una serie de estrategias de retención enfocadas en la personalización de ofertas, la fidelización de clientes y la mejora de la propuesta de valor de los servicios.



2. Análisis del Modelo Final y sus Métricas

El modelo optimizado no solo es preciso, sino que está bien balanceado para el problema de negocio:

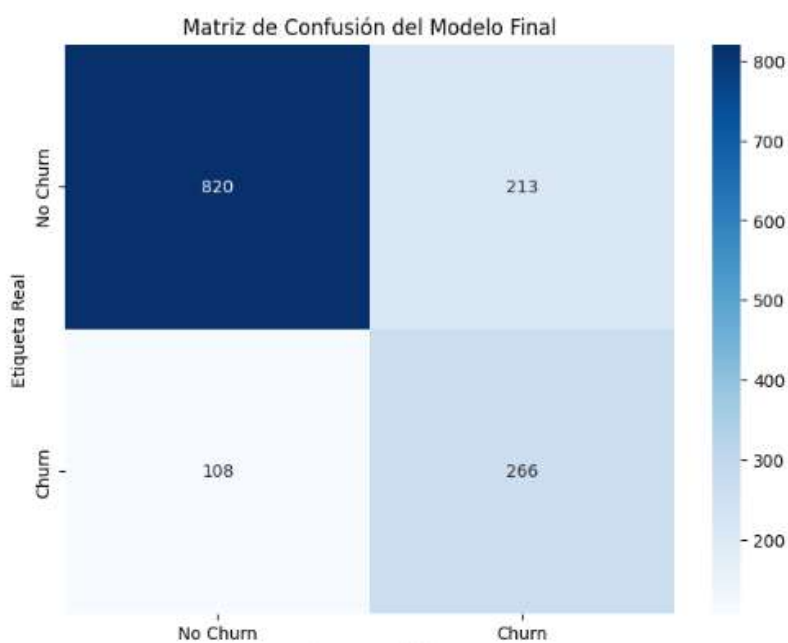
- **Accuracy (Exactitud) del 77%:** El modelo clasifica correctamente al 77% de los clientes en general.
- **Recall (Sensibilidad) para Churn del 71%:** Esta es la métrica más importante para el negocio. Significa que **de cada 100 clientes que realmente van a cancelar su servicio, el modelo identifica correctamente a 71**. Esto permite dirigir los esfuerzos de retención al grupo correcto.
- **Precision (Precisión) para Churn del 56%:** Cuando el modelo predice que un cliente se irá, acierta en el 56% de los casos. Aunque esto implica que algunos clientes no desertores serán contactados, el costo de un esfuerzo de retención innecesario es significativamente menor que el costo de perder un cliente.
- **ROC-AUC de 0.8322:** Este valor, cercano a 1, indica una excelente capacidad del modelo para distinguir entre clientes que se irán y los que se quedarán.

--- Evaluación del Modelo Final en el Conjunto de Prueba ---

Reporte de Clasificación:

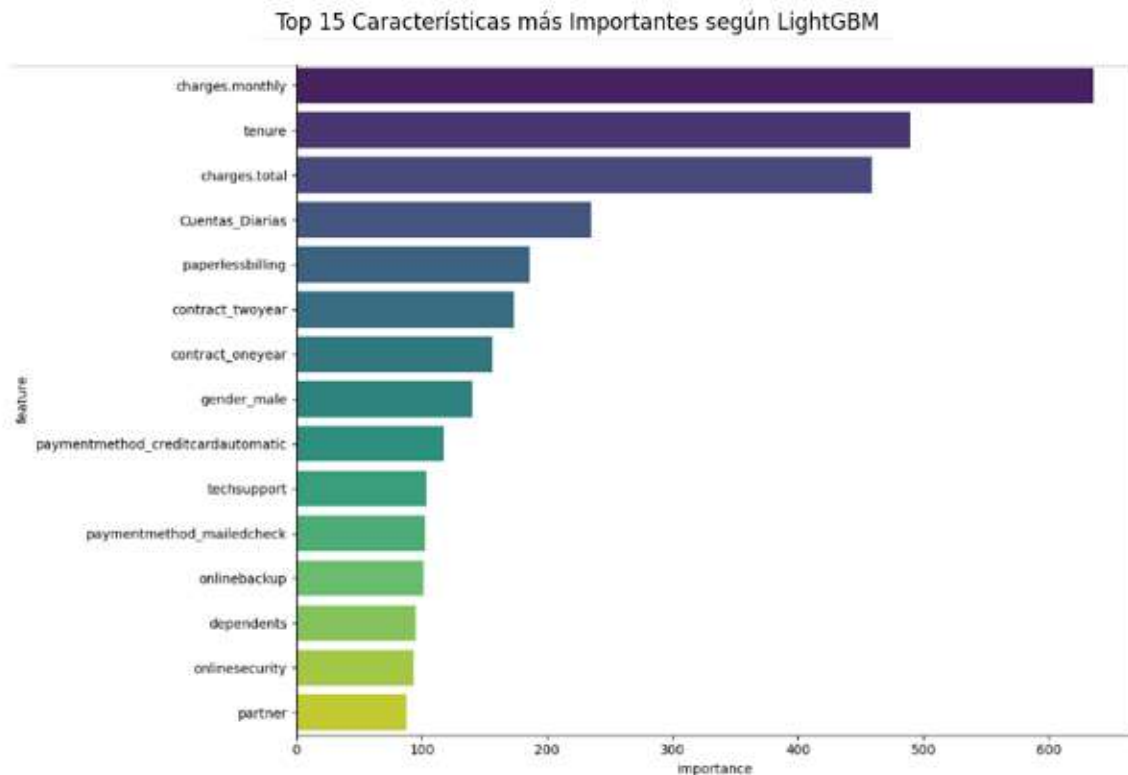
	precision	recall	f1-score	support
0	0.88	0.79	0.84	1033
1	0.56	0.71	0.62	374
accuracy			0.77	1407
macro avg	0.72	0.75	0.73	1407
weighted avg	0.80	0.77	0.78	1407

Puntuación ROC-AUC en el conjunto de prueba: 0.8322



3. Análisis de Variables Relevantes para la Predicción

3.1. Variables Más Importantes (Según el Modelo LightGBM)



El modelo final nos proporciona una jerarquía clara de los factores que más pesan en la decisión de un cliente de abandonar la compañía:

1. **Cargos Mensuales (charges.monthly):** Es el factor más decisivo. Clientes con cargos mensuales más altos son más propensos a cancelar.
2. **Permanencia (tenure):** El segundo factor más importante. Clientes con menor antigüedad (pocos meses) tienen una probabilidad de fuga mucho mayor que los clientes a largo plazo.
3. **Cargos Totales (charges.total):** Relacionado con la permanencia, pero también indica el valor histórico del cliente.
4. **Tipo de Contrato (contract...):** Los clientes con contratos mes a mes son drásticamente más propensos a irse que aquellos con contratos de uno o dos años. Esta es una de las palancas de retención más claras.
5. **Facturación Electrónica (paperlessbilling):** Indica un tipo de cliente posiblemente más digital y quizás más sensible a la experiencia de usuario online.

6. **Soporte Técnico y Servicios Adicionales (techsupport, onlinebackup, onlinesecurity):** La ausencia de estos servicios de valor agregado está correlacionada con una mayor tasa de churn. Los clientes que no los utilizan perciben menos valor en la oferta global.

3.2. Interpretación de Variables en Otros Modelos (Análisis Teórico)

Para un informe completo, es útil entender cómo se interpretaría la relevancia de las variables en los otros modelos que exploraste:

- **Regresión Logística:** La importancia se mide directamente a través de los **coeficientes**. Un coeficiente positivo grande significa que un aumento en esa variable incrementa fuertemente la probabilidad de churn (ej. charges.monthly). Un coeficiente negativo grande indica un factor de retención (ej. tenure).
- **Random Forest:** Al igual que LightGBM, utiliza la **importancia de características** (feature_importance_). Este valor se calcula midiendo cuánto contribuye cada variable a reducir la "impureza" de los nodos en los árboles de decisión. Una variable que crea divisiones más "puras" (que separan mejor a los clientes desertores de los no desertores) es más importante.
- **K-Nearest Neighbors (KNN):** Este modelo no ofrece una métrica de importancia directa. La relevancia es implícita: las variables que más influyen en el **cálculo de distancia** entre clientes son las más importantes. Por ello, el escalado de datos (como el StandardScaler que usé) es crucial. Variables con mayor varianza o que mejor separan los grupos en el espacio de características tendrán un mayor impacto.
- **Support Vector Machine (SVM):** Para un SVM lineal, los **coeficientes de la frontera de decisión** son análogos a los de la regresión logística e indican la importancia. Para SVM no lineales (con kernels), la interpretación es más compleja, pero las variables de los **vectores de soporte** (los puntos de datos más cercanos a la frontera) son las que definen la separación y, por lo tanto, son las más críticas.

CATBOOST:

```
→ === Matriz de Confusión ===
[[759 274]
 [ 98 276]]

=== Reporte de Clasificación ===
              precision    recall  f1-score   support

     0       0.89        0.73        0.80       1033
     1       0.50        0.74        0.60        374

   accuracy          0.74       1407
  macro avg          0.69        0.74        0.70       1407
 weighted avg          0.78        0.74        0.75       1407

=== ROC-AUC ===
0.8073805074260629
```

4. Estrategias de Retención Propuestas

Basado en los hallazgos del modelo, se proponen las siguientes estrategias accionables:

1. Estrategia: Optimización de Precios y Contratos.

- **Acción:** Identificar clientes con alto riesgo de churn (según el modelo) y altos cargos mensuales. Ofrecerles proactivamente un **plan personalizado** o un **descuento de lealtad** a cambio de pasar de un contrato mes a mes a uno anual.
- **Justificación:** El modelo muestra que contract y charges.monthly son los principales impulsores del churn. Trabajar ambos simultáneamente, es la estrategia más efectiva.

2. Estrategia: Programa de Fidelización para Nuevos Clientes.

- **Acción:** Crear un programa de "Onboarding y Acompañamiento" para clientes con tenure menor a 6 meses. Esto puede incluir llamadas de seguimiento, tutoriales sobre los servicios y una pequeña bonificación (ej. un mes de un servicio premium gratis) al cumplir 3 y 6 meses.
- **Justificación:** La variable tenure es el segundo factor de riesgo. Es crucial asegurar que los nuevos clientes perciban valor desde el principio para superar la fase inicial de alto riesgo.

3. Estrategia: Campaña de Venta Cruzada de Servicios de Valor Agregado.

- **Acción:** Lanzar una campaña dirigida a los clientes de alto riesgo que no tienen contratados servicios como techsupport, onlinesecurity u onlinebackup. Ofrecer un paquete con estos servicios a un precio muy atractivo durante los primeros meses.
- **Justificación:** Los clientes que utilizan más servicios de la compañía tienen una menor probabilidad de irse. Aumentar el "enganche" del cliente con el ecosistema de la empresa es una táctica de retención probada.

4. Estrategia: Digitalización y Simplificación de Pagos.

- **Acción:** Identificar a los clientes que aún usan métodos de pago no automáticos (como mailedcheck) y ofrecerles un incentivo (ej. un pequeño descuento único) por cambiarse a débito automático o tarjeta de crédito.
- **Justificación:** Los métodos de pago automáticos reducen la fricción y aumentan la probabilidad de que el cliente continúe con el servicio a largo plazo.

5. Conclusiones finales

Aunque se trabajó con varios modelos de entrenamiento y validación, a mi juicio, no se pudo alcanzar un resultado de alta asertividad, aunque también es cierto que las métricas del modelo son aceptables. En este sentido, se puede concluir que el modelo predictivo desarrollado es una herramienta de gran valor que puede permitir a la empresa pasar de una estrategia reactiva a una estrategia proactiva de retención, en cuanto a evitar el Churn, pero por otro lado, la Precisión para Churn del 56%, implica que el modelo no predice significativamente que un cliente se irá, como se mencionó con anterioridad. Esta característica, puede generar, por ejemplo, que erróneamente casi la mitad de los clientes no desertores sean contactados durante una campaña de retención, pero si se utiliza esta estrategia como un refuerzo positivo de fidelización de los clientes no desertores, se justifica plenamente el costo de la acción fallida del esfuerzo de retención innecesario, lo que además implica un costo significativamente menor que el costo de perder un cliente.

Al enfocarse en los factores clave identificados —principalmente el tipo de contrato, la permanencia y el costo mensual— y al implementar las estrategias propuestas, es posible reducir significativamente la tasa de fuga de clientes, maximizando así la rentabilidad y el valor de vida del cliente.

Los aspectos positivos del modelo, tales como: **ROC-AUC de 0.8322, Accuracy del 77%, y Recall para Churn del 71%, tienen mayor significancia que la baja precisión del 56%. Esto permite hacer predicciones útiles, para la retención de clientes, que era el objetivo principal de este trabajo de análisis de datos.**

Análisis de Resultados con todos los datos

Los resultados generales muestran las predicciones del modelo LightGBM aplicado a todos los datos originales, junto con las probabilidades estimadas de churn:

1. Consistencia en las Predicciones

- Distribución global vs. conjunto de prueba:
 - Todos los datos: 33.18% predichos como churn .
 - Conjunto de prueba: 34.04% predichos como churn.
 - Conclusión: Las proporciones son muy similares, lo que indica que el modelo no está sobreajustado y generaliza bien.

2. Ejemplos de Predicciones (Primeras 5 Filas)

Índice	churn_real	churn_pred	probabilidad_churn	¿Correcto?
0	0	0	0.13	(Acierto)
1	0	0	0.44	(Acierto)
2	1	1	0.79	(Acierto)
3	1	1	0.69	(Acierto)
4	1	1	0.65	(Acierto)

- Observación: En estas filas, el modelo acertó todas las predicciones.
- Probabilidades:
 - Los casos con churn_real= 0 tienen probabilidades bajas (<0.5, como era esperado).
 - Los casos con churn_real= 1 tienen probabilidades altas (>0.65), lo que sugiere que el modelo está seguro en estos ejemplos.