Análisis Predictivo de Churn de Clientes

Proyecto es desarrollar un modelo de machine learning que permita **predecir la pérdida de clientes (Churn)** en una empresa de telecomunicaciones llamada ***TelecomX***.

Este modelo tiene como objetivo anticipar qué clientes podrían abandonar la compañía, permitiendo así **diseñar estrategias preventivas de retención**.

Agenda

01

Análisis Exploratorio de Datos

Carga, limpieza y preparación de datos.

02

Preprocesamiento de Datos

Codificación y normalización de variables.

03

Modelado Predictivo

Creación y evaluación de modelos de clasificación.

04

Optimización y Conclusiones

Ajuste de hiperparámetros y recomendaciones estratégicas.

Análisis Exploratorio de Datos

Carga y Procesamiento

Se cargó el archivo de clientes y se realizó una inspección inicial para entender la estructura de los datos.

```
df=pd.read_csv('/content/TelecomX_clientes_churn.csv
')
```

Se eliminaron columnas no relevantes como 'customerID' y 'cuenta_diaria'.

```
df.drop(columns=['customerID','cuenta_diaria'],
inplace=True)
```

El dataset contiene información de clientes, incluyendo variables demográficas, de servicios contratados y comportamiento de pago. La variable objetivo es `Churn`, indicando si un cliente dejó la compañía (`Yes`) o no (`No`).

Información del DataFrame

El conjunto de datos contiene 7032 entradas y 22 columnas, con tipos de datos variados.

#	Column	Non-Null Count	Dtype
0	customerID	7032 non-null	object
1	Churn	7032 non-null	object
2	customer_gender	7032 non-null	object
3	customer_SeniorCitizen	7032 non-null	int64
4	customer_Partner	7032 non-null	object
5	customer_Dependents	7032 non-null	object
6	customer_tenure	7032 non-null	int64
7	phone_PhoneService	7032 non-null	object
8	phone_MultipleLines	7032 non-null	object
9	internet_InternetService	7032 non-null	object
10	internet_OnlineSecurity	7032 non-null	object
11	internet_OnlineBackup	7032 non-null	object
12	internet_DeviceProtection	7032 non-null	object
13	internet_TechSupport	7032 non-null	object
14	internet_StreamingTV	7032 non-null	object
15	internet_StreamingMovies	7032 non-null	object
16	account_Contract	7032 non-null	object
17	account_PaperlessBilling	7032 non-null	object
18	account_PaymentMethod	7032 non-null	object
19	account_Charges_Monthly	7032 non-null	float64
20	account_Charges_Total	7032 non-null	float64
21	cuenta_diaria	7032 non-null	float64

Preprocesamiento de Datos

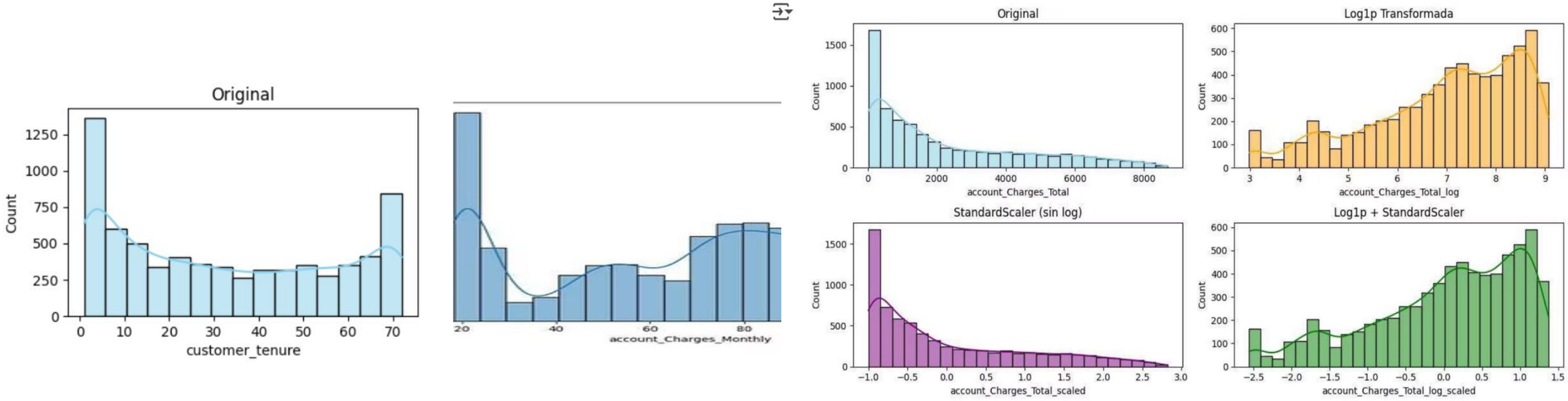
Codificación de Variables Categóricas

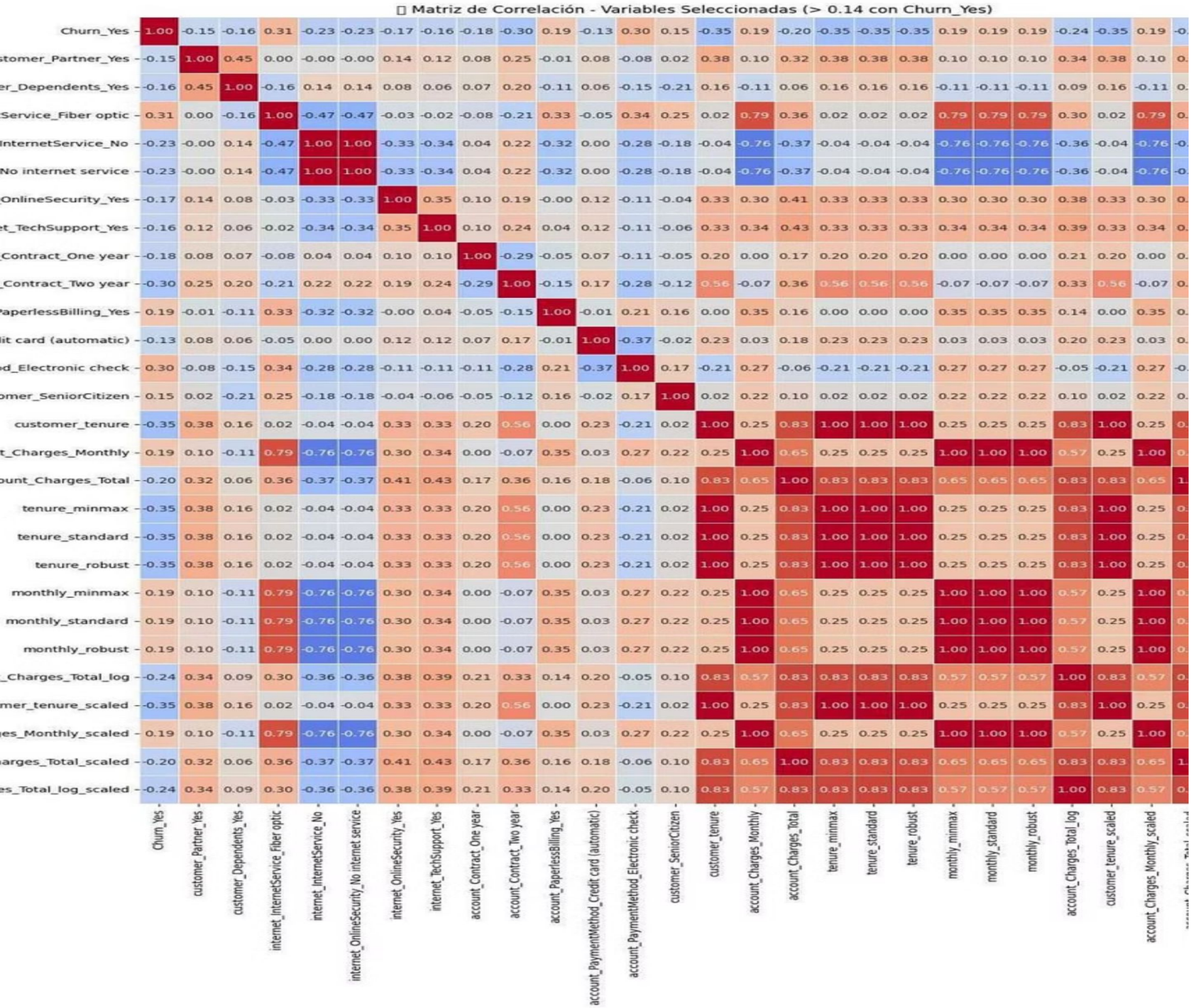
Las variables categóricas se transformaron a formato numérico usando **OneHotEncoder** para compatibilidad con algoritmos de Machine Learning.

Normalización de Variables Numéricas

Se identificaron y normalizaron las variables numéricas clave para asegurar una escala comparable.

Se aplicó **StandardScaler** a 'customer_tenure' y 'account_Charges_Monthly'. Para 'account_Charges_Total', se usó una transformación logarítmica seguida de StandardScaler debido a su asimetría.





Análisis de Correlación y VIF

Correlación con Churn

Las variables más positivamente correlacionadas con Churn son: **InternetService_Fiber optic** (+0.31), **PaymentMethod_Electronic check** (+0.30), **PaperlessBilling_Yes** (+0.19), **Charges_Monthly** (+0.19) y **SeniorCitizen** (+0.15).

Las variables más negativamente correlacionadas (mayor retención) son: **customer_tenure** (-0.35), **Contract_Two year** (-0.30) y **Charges_Total_log** (-0.24).

Análisis VIF (Variance Inflation Factor)

El VIF se usó para detectar multicolinealidad. Variables con VIF alto (>5) fueron revisadas para evitar distorsiones en el modelo.

Se eliminaron variables con VIF infinito (perfecta colinealidad) y duplicadas.

Variables Seleccionadas

Una vez Revisada la **Matriz de Correlacion** , la **Correlacion** de las Variables, **VIF** del Dataframe , y **analisis de Graficos** se seleccionaron las siguientes variable
algunos

```
[
    'internet_InternetService_Fiber optic',
    'account_PaymentMethod_Electronic check',
    'account_PaperlessBilling_Yes',
    'account_Charges_Monthly_scaled',
    'customer_SeniorCitizen',
    'account_Contract_Two year',
    'customer_tenure_scaled'
]
```

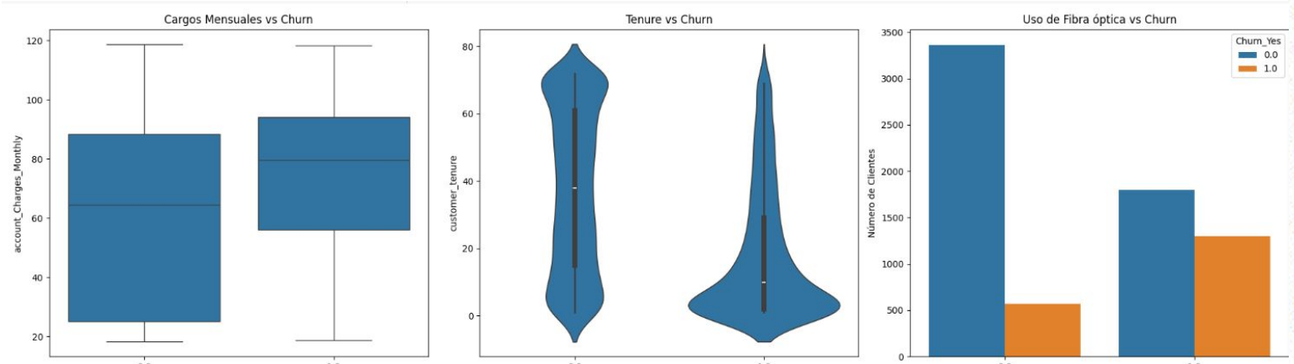
Variable	VIF
customer_gender_Male	1.002225
customer_SeniorCitizen	1.153332
account_PaperlessBilling_Yes	1.208495
customer_Dependents_Yes	1.383064
customer_Partner_Yes	1.465469
account_PaymentMethod_Credit card (automatic)	1.561089
account_Contract_One year	1.635221
account_PaymentMethod_Mailed check	1.857863
account_PaymentMethod_Electronic check	1.980015
account_Contract_Two year	2.697445

Análisis Exploratorio

Se realizaron análisis descriptivos y visuales para comprender el comportamiento de los clientes que hacen churn

- **Clientes con contrato mensual y pago electrónico** tienen mayor probabilidad de churn.
- **Clientes con fibra óptica** muestran mayor propensión al churn.
- El churn disminuye con mayor **antigüedad (`tenure`)**.
- Distribuciones con outliers y sesgos fueron observadas en cargos mensuales y Totales

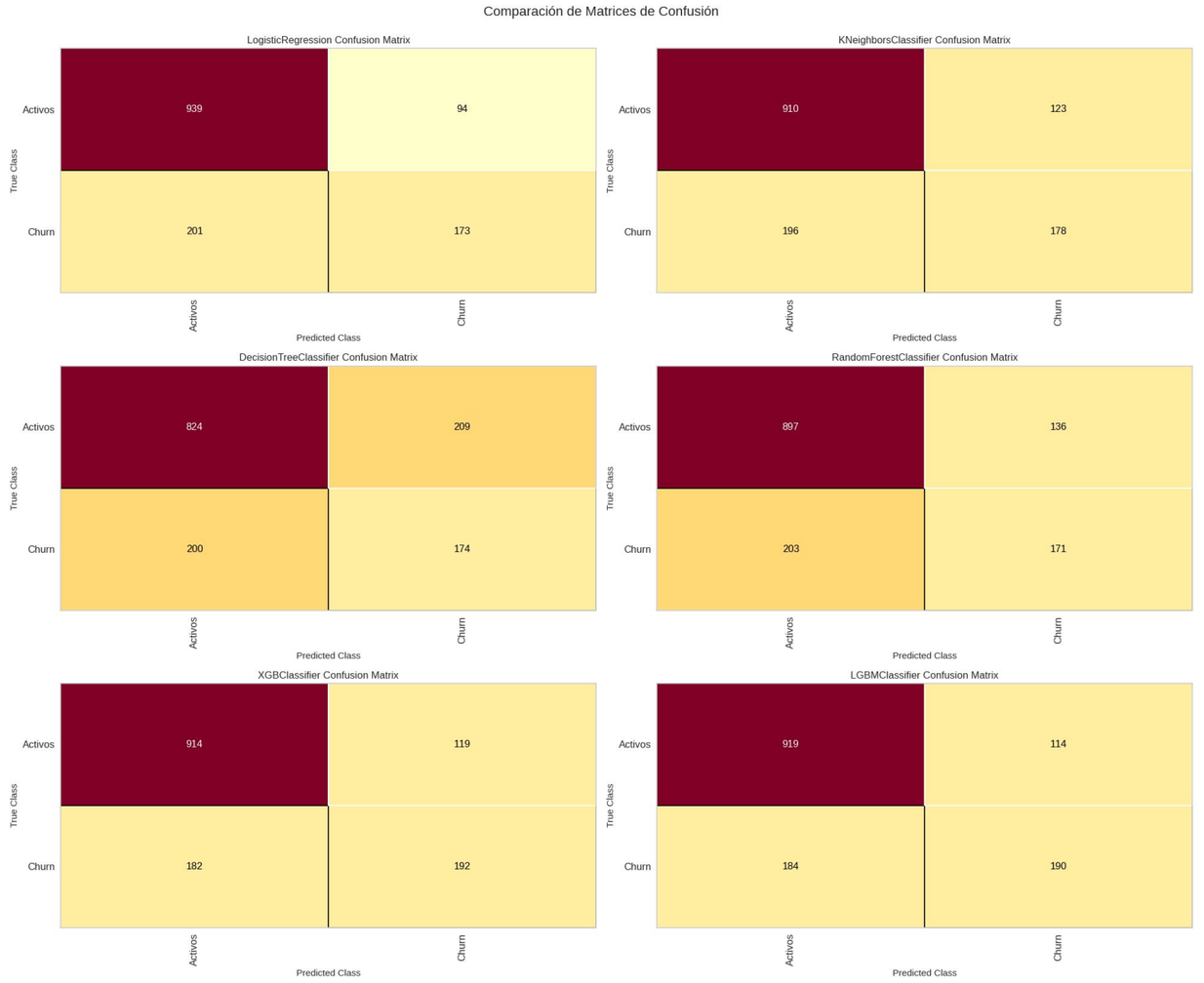
```
Correlación de variables con Churn_Yes (ordenadas):
internet_InternetService_Fiber optic              0.307463
account_PaymentMethod_Electronic check            0.301455
monthly_minmax                                    0.192858
account_Charges_Monthly_scaled                    0.192858
monthly_standard                                  0.192858
account_Charges_Monthly                           0.192858
monthly_robust                                    0.192858
account_PaperlessBilling_Yes                      0.191454
customer_SeniorCitizen                            0.150541
internet_StreamingTV_Yes                          0.063254
internet_StreamingMovies_Yes                      0.060860
phone_MultipleLines_Yes                           0.040033
phone_PhoneService_Yes                            0.011691
customer_gender_Male                             -0.008545
internet_DeviceProtection_Yes                    -0.066193
internet_OnlineBackup_Yes                        -0.082307
account_PaymentMethod_Mailed check               -0.090773
account_PaymentMethod_Credit card (automatic)    -0.134687
customer_Partner_Yes                             -0.149982
customer_Dependents_Yes                          -0.163128
internet_TechSupport_Yes                         -0.164716
internet_OnlineSecurity_Yes                      -0.171270
account_Contract_One year                        -0.178225
account_Charges_Total_scaled                     -0.199484
account_Charges_Total                            -0.199484
internet_InternetService_No                      -0.227578
internet_OnlineSecurity_No internet service      -0.227578
account_Charges_Total_log                        -0.241908
account_Charges_Total_log_scaled                 -0.241908
account_Contract_Two year                        -0.301552
tenure_standard                                  -0.354049
tenure_minmax                                    -0.354049
customer_tenure_scaled                           -0.354049
tenure_robust                                    -0.354049
customer_tenure                                  -0.354049
```



Modelado Predictivo: Modelos Base

Se evaluaron varios modelos de clasificación sin balanceo de clases para establecer una línea base.

Modelo	Accuracy	Precision (Churn)	Recall (Churn)	F1 (Churn)	RMSE	MAE	R^2
Baseline	0.73	0.00	0.00	0.00	0.5156	0.2658	-0.3621
Reg. Logística	**0.79**	**0.65**	0.46	**0.54**	0.4579	0.2097	-0.0743
KNN	0.77	0.59	0.48	0.53	0.4762	0.2267	-0.1618
Árbol Decisión	0.71	0.45	0.47	0.46	0.5392	0.2907	-0.4895
Random Forest	0.76	0.56	0.46	0.50	0.4909	0.2409	-0.2346
XGBoost	**0.79**	0.62	**0.51**	**0.56**	0.4625	0.2139	-0.0962
LightGBM	**0.79**	0.62	**0.51**	**0.56**	**0.4602**	**0.2118**	**-0.0853**



* Mejores modelos **LightGBM y XGBoost:**

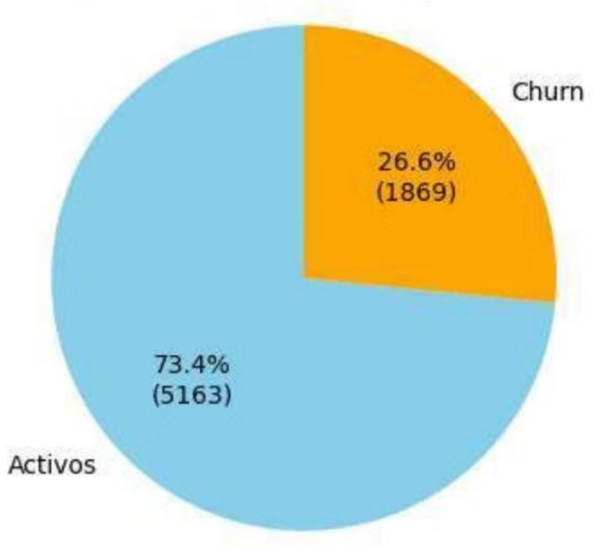
Tienen el mayor número de verdaderos positivos (TP).
El menor número de falsos negativos (FN), lo cual es clave para detectar Churn.
Buen balance entre precisión y recall.

Ajustando Modelos

Verificación de la Proporción de Churn

Se analizó la proporción de clientes que cancelaron (Churn) frente a los activos, revelando un desbalance.

Clase	Cantidad	Proporción (%)
Activo	5163	73.4
Churn	1869	26.6



Se realiza SMOTE , para mejorar la Proporcionalidad de los datos

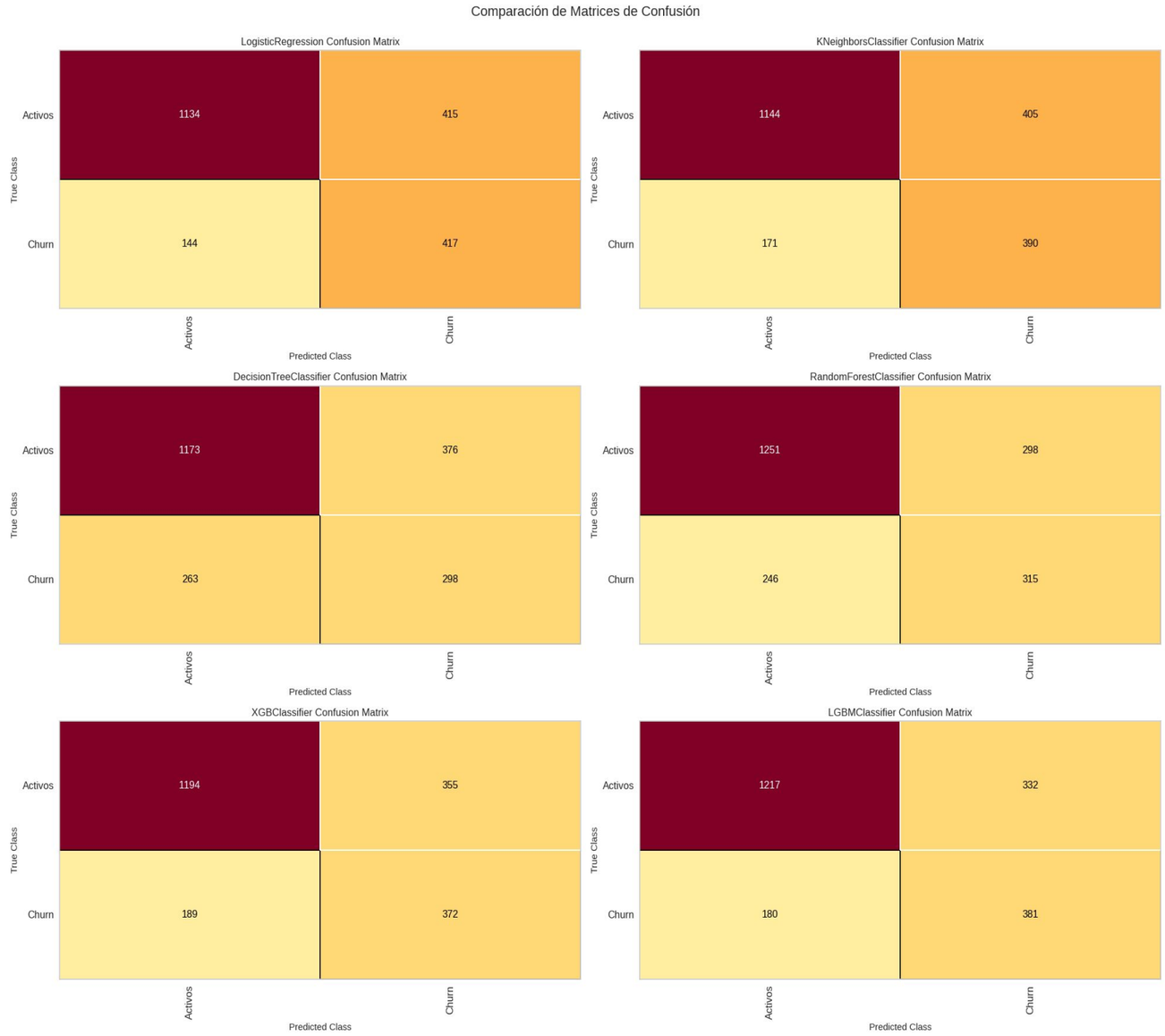
```
Antes de SMOTE:
 Churn_Yes
0.0    3614
1.0    1308
Name: count, dtype: int64

Después de SMOTE:
 Churn_Yes
1.0    3614
0.0    3614
Name: count, dtype: int64
```

Modelado Predictivo: Balanceo de Clases con SMOTE

Para abordar el desbalance de clases, se aplicó la técnica de **Oversampling (SMOTE)** a la clase minoritaria (Churn).

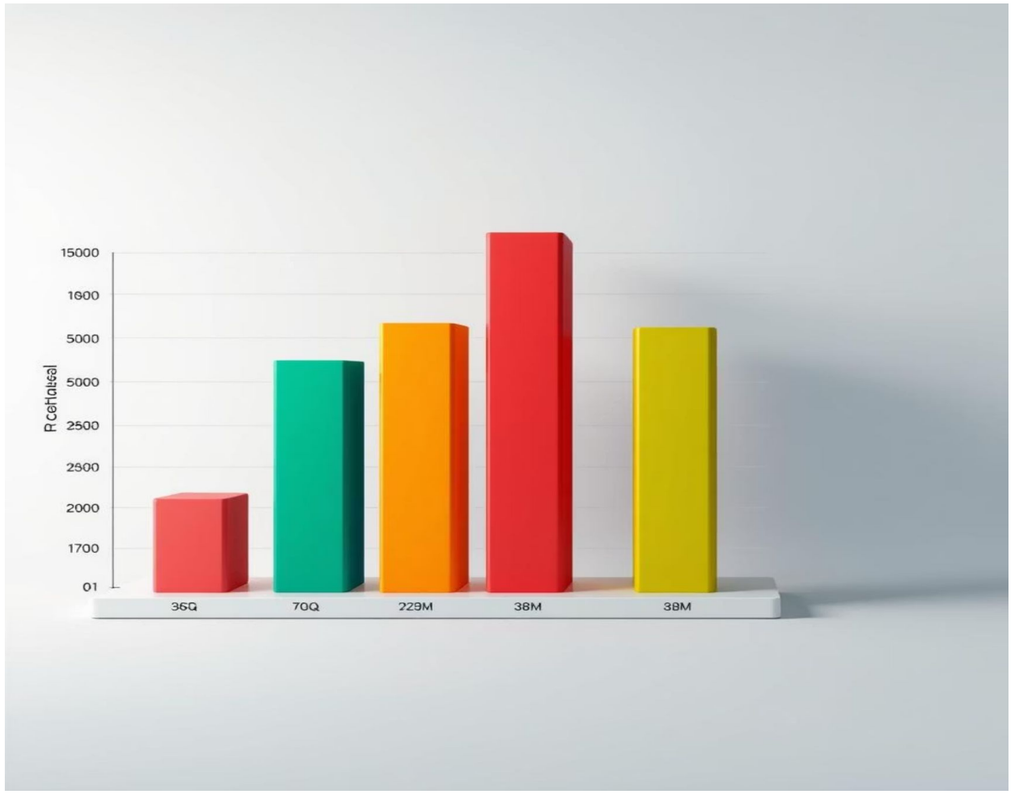
Modelo	Accuracy	Recall Churn	F1 Churn	RMSE	MAE	R2
Baseline	0.73	0.00	0.00	0.5156	0.2659	-0.3622
LogisticReg	0.74	0.74	0.60	0.5147	0.2649	-0.3573
KNN	0.80	0.53	0.58	0.4509	0.2033	-0.0417
Decision Tree	0.80	0.63	0.63	0.4446	0.1976	-0.0125
Random Forest	**0.84**	**0.63**	**0.67**	**0.4044**	**0.1635**	**0.1623**
XGBoost	0.74	0.66	0.58	0.5078	0.2578	-0.3209
LightGBM	0.80	0.55	0.60	0.4430	0.1962	-0.0052



Optimización y Conclusiones

Ajuste de Hiperparámetros

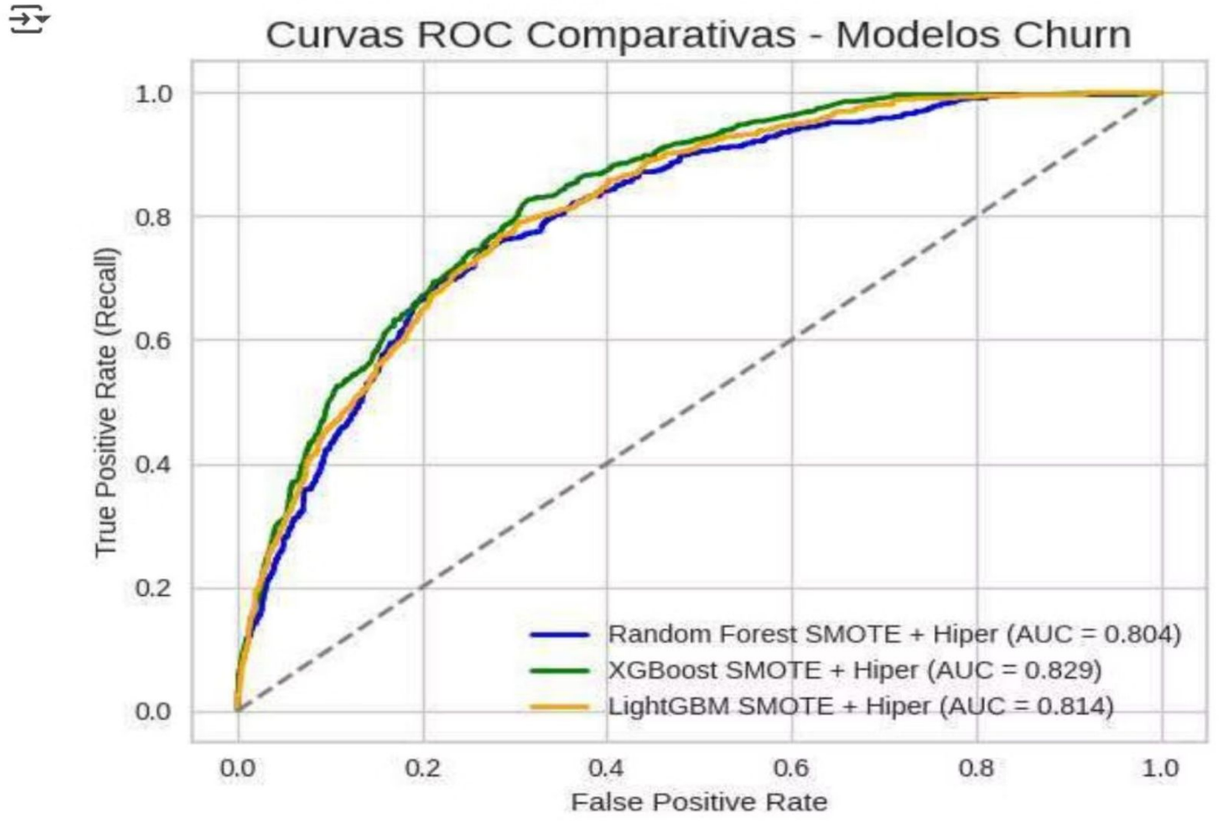
Se realizó un ajuste de hiperparámetros con GridSearchCV para **XGBoost**, **LightGBM** y **Random Forest**, buscando maximizar el recall.



XGBoost obtuvo el mejor recall (0.846), seguido de **Random Forest** (0.837) y **LightGBM** (0.833).

Curva ROC y AUC

La curva ROC evalúa el rendimiento de los modelos en la clasificación binaria.



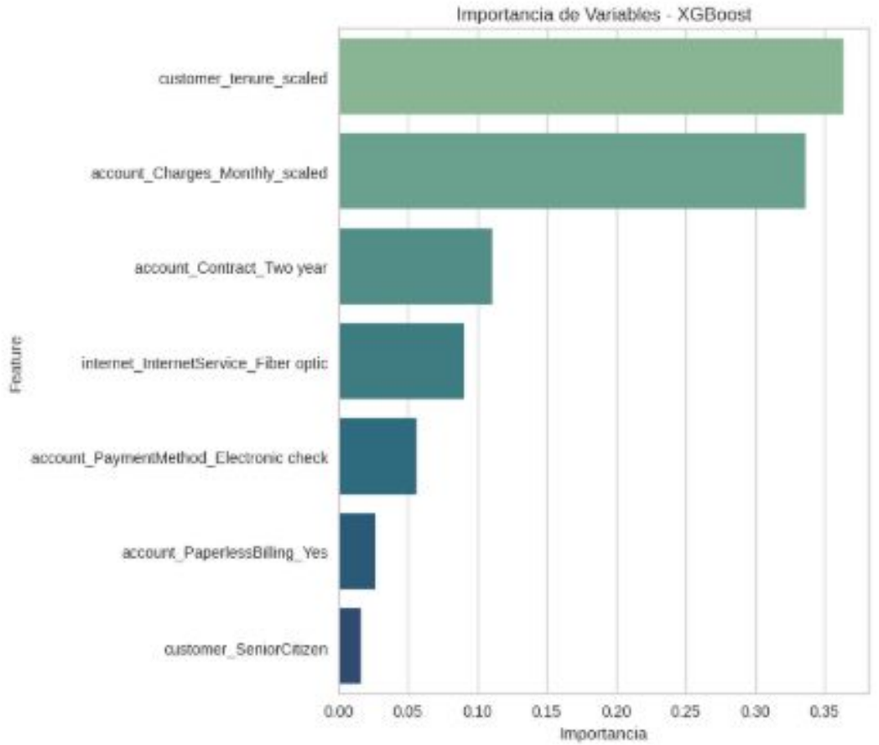
XGBoost también se destacó con el mejor AUC (0.834), indicando una excelente capacidad para distinguir entre clientes que hacen churn y los que no.

XGBoost es el modelo de mejor Comportamiento

Modelo	AUC-ROC	Recall	Precisión	F1-Score	Observaciones Clave
XGBoost	**0.834**	**0.846**	Moderada	Alta	• Mejor balance general entre recall y AUC. • Capta mejor a los clientes que hacen churn.
LightGBM	0.816	0.832	Similar	Alta	• Muy competitivo. • Ligeramente inferior a XGBoost en AUC y recall. • Más rápido de entrenar.
Random Forest	0.806	0.837	Similar	Alta	• Buen rendimiento general. • Menor capacidad discriminativa (AUC). • Puede sobreajustarse si no se regula.

Como influyen las variables en el modelo , analisis para conclusiones

```
      Listado de importancia de variables por modelo
------------------------------------------------------
Modelo XGBoost
                                  Feature  Importancia
5                 account_Contract_Two year     0.482214
0     internet_InternetService_Fiber optic     0.178167
6                   customer_tenure_scaled     0.152820
1   account_PaymentMethod_Electronic check     0.098010
3            account_Charges_Monthly_scaled     0.045237
2              account_PaperlessBilling_Yes     0.033684
4                   customer_SeniorCitizen     0.009868
```



Recomendaciones Estratégicas

Este proyecto proporciona un modelo efectivo para la predicción de churn en clientes **TelecomX**. El uso de **XGBoost** permite anticipar cancelaciones con alta precisión, ayudando a la compañía a **reducir la pérdida de clientes y mejorar la retención** mediante intervenciones estratégicas.

Contratos Largos

Incentivar contratos de dos años, ya que reducen significativamente el riesgo de churn.

Migración: A los clientes que hoy no tienen contrato y tienen unos de corta duración ofrecer planes atractivos para alargar su permanencia en la compañía.

Fidelizacion: puede ser a través de agregar servicios , descuentos por permanencia o beneficios complementarios como programas de descuentos en otra compañías.

Servicio de Fibra Óptica

Puede estar asociado a expectativas altas de servicio que no se cumplen se propone :

- Auditar la calidad del servicio de fibra óptica.
- Implementar un programa de monitoreo proactivo para detectar y resolver problemas técnicos antes de que afecten al cliente.
- Mejorar la atención postventa en clientes de fibra.
- Clientes nuevos tienen mayor riesgo de churn

Clientes Nuevos

- Establecer un programa de bienvenida + seguimiento intensivo en los primeros 3-6 meses.
- Comunicación personalizada, ofertas especiales, y soporte prioritario durante esta etapa crítica.

Cargos Mensuales Altos

Esto puede reflejar una percepción de mala relación costo-beneficio.

- Crear alertas para clientes con altos cargos mensuales + riesgo de churn.
- Ofrecer planes alternativos, descuentos por fidelidad o mejoras sin costo adicional.