

LABORATORIO X

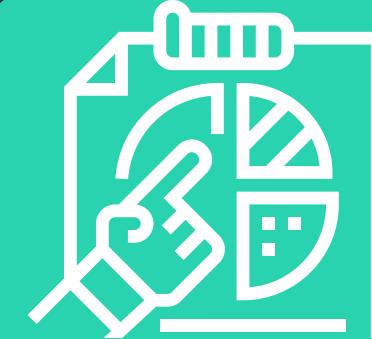
Grupo 1: Predicción de Abandono de Clientes



X-PLORACIÓN



X-PERIMENTACIÓN



X-PLICACIÓN



X-PERIENCIA

“EXPLORAMOS LO INCIERTO PARA PREDECIR LO IMPORTANTE”

Contexto

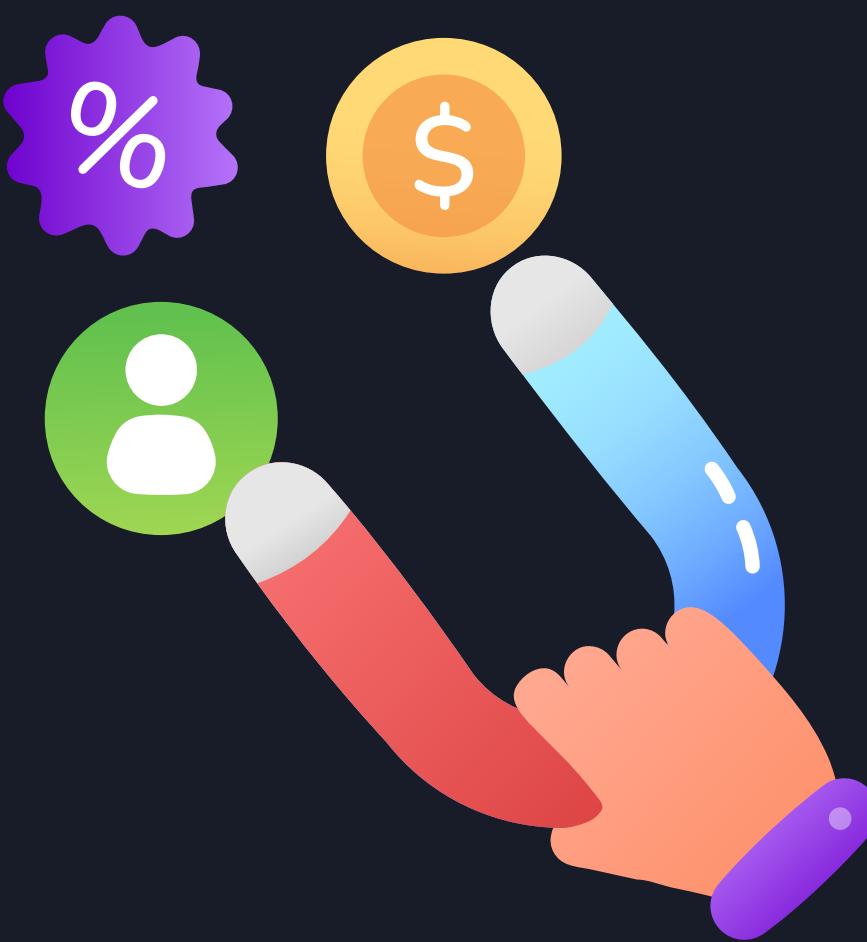
En un entorno altamente competitivo como el de las telecomunicaciones, retener a los clientes se ha convertido en una prioridad estratégica para las empresas.

Efectos del fenómeno de abandono de clientes (CHURN)

- Pérdida significativa de ingresos
- Inversiones constantes para captación de nuevos clientes

Mediante técnicas de Machine Learning podemos:

- Identificar patrones y comportamiento que predicen la probabilidad de abandono
 - Esto permitirá implementar acciones preventivas efectivas





Exploración Inicial

Análisis Exploratorio

- Comprender estructura y contenido del dataset.
- El dataset contiene datos:
 - demográficos
 - geográficos
 - servicios contratados
 - comportamiento de uso
 - Satisfacción y abandono de clientes

Calidad de los datos:

- Verificación de IDs repetidos.
- Análisis de valores nulos.
- Revisión de información faltante.

- 7.043 registros
- 50 columnas
- 31 Type Object
- 11 Type INT64
- 8 Type FLOAT 64



Exploración Inicial

Evaluación Preliminar

- Análisis de la categorías **Customer Status** y sus proporciones, así como su incidencia con otras columnas.
- Creación de nuevas columnas
 - **Churn Value** si Customer Status es
 - Churned → 1
 - Otro valor → 0
 - **Intervalo de Pemanencia** categorizando el tiempo de permanencia (Tenure in Months) en tres rangos:
 - Menor a un año
 - De uno a dos años
 - Mayor a 3 años.

Counts:		
	No	Yes
Customer Status		
Churned	1565	304
Joined	355	99
Stayed	3722	998

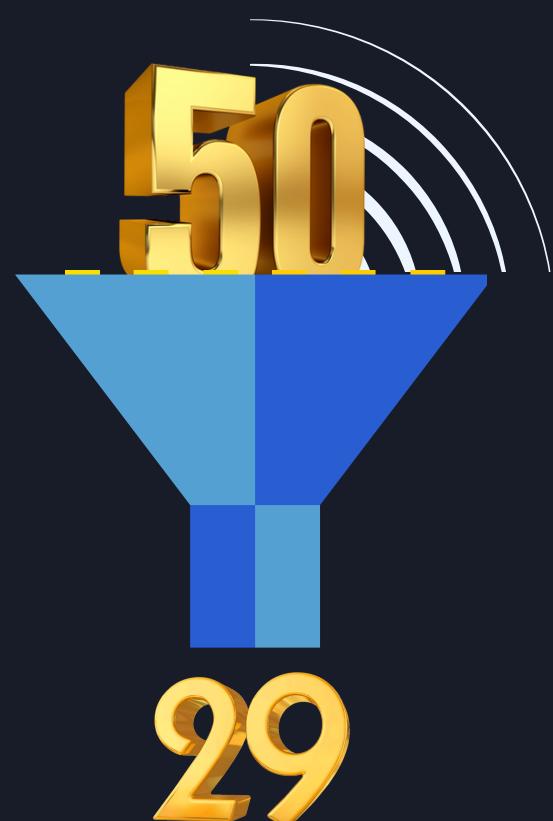
Percentages:		
	No	Yes
Customer Status		
Churned	83.734617	16.265383
Joined	78.193833	21.806167
Stayed	78.855932	21.144068



Exploración Inicial

Limpieza y Transformación

Se eliminaron las columnas que no representativas, pasando de 50 columnas a 29 que son con las que vamos a trabajar el modelo.



Data columns (total 29 columns):		
#	Column	Non-Null Count Dtype
0	Gender	7043 non-null object
1	Age	7043 non-null int64
2	Married	7043 non-null object
3	Number of Dependents	7043 non-null int64
4	Number of Referrals	7043 non-null int64
5	Tenure in Months	7043 non-null int64
6	Offer	7043 non-null object
7	Phone Service	7043 non-null object
8	Multiple Lines	7043 non-null object
9	Internet Type	7043 non-null object
10	Avg Monthly GB Download	7043 non-null int64
11	Online Security	7043 non-null object
12	Online Backup	7043 non-null object
13	Device Protection Plan	7043 non-null object
14	Premium Tech Support	7043 non-null object
15	Streaming TV	7043 non-null object
16	Streaming Movies	7043 non-null object
17	Streaming Music	7043 non-null object
18	Unlimited Data	7043 non-null object
19	Contract	7043 non-null object
20	Payment Method	7043 non-null object
21	Monthly Charge	7043 non-null float64
22	Total Charges	7043 non-null float64
23	Total Refunds	7043 non-null float64
24	Total Extra Data Charges	7043 non-null int64
25	Total Long Distance Charges	7043 non-null float64
26	Total Revenue	7043 non-null float64
27	CLTV	7043 non-null int64
28	Churn Value	7043 non-null int64

dtypes: float64(5), int64(8), object(16)
memory usage: 1.6+ MB



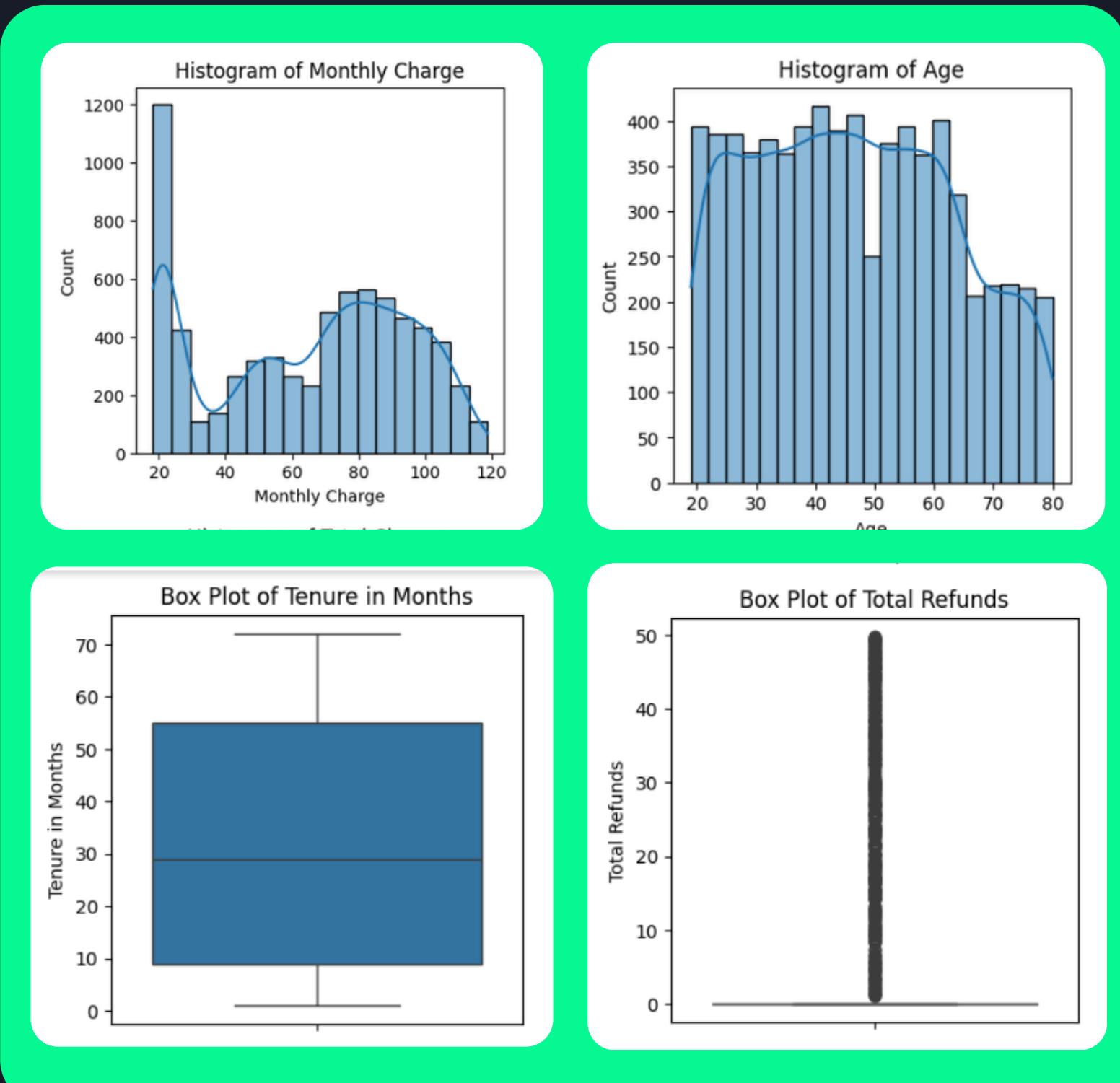
Exploración Inicial

Exploración Profunda

Se buscaron patrones de dispersión, extremos y comportamiento general para entender mejor cómo estas variables podían impactar en la permanencia o salida de los clientes.

Gráficos utilizados:

- Boxplots (gráficos de bigote):
 - Identificación de valores atípicos
 - Distribución general.
- Histogramas:
 - Distribución y concentración de valores.





Variable Objetivo:
Churn Value



Experimentación

Preparación del Modelo

- Codificación de variables categóricas para que interprete correctamente los datos no numéricos.
- Información relevante del cliente sea comprensible para el algoritmo.
- La variable objetivo se encontraba balanceada cerca de un 70/30 (26%) por lo tanto no se realizó el balanceo.
- Data Splitting fue de 70 / 15 / 15

Churn	Value
0	5174
1	1869

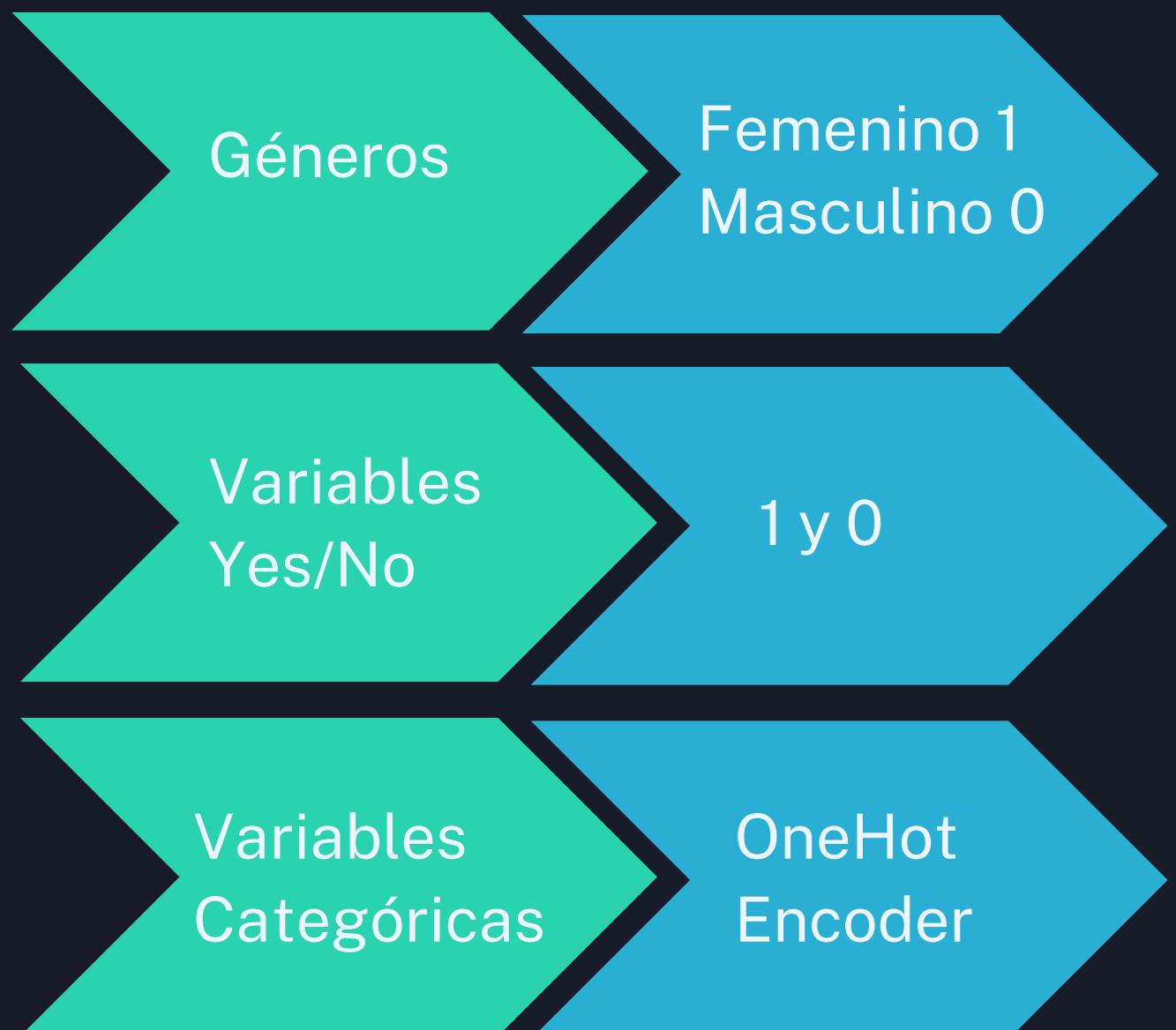
```
[ ] len(y_train),len(y_val),len(y_test)
```

→ (4930, 1056, 1057)



Experimentación

Encoding



```
ct = ColumnTransformer(transformers=[('genero', FunctionTransformer(gender_transform), Genero_col),
                                      ('yes_no', FunctionTransformer(yes_no_transform), YesNo_cols),
                                      ('OneHotEncoding', OneHotEncoder(), onehot_cols),
                                      ('dependent', FunctionTransformer(dependent_transform), dependent_col),
                                      ('robust', RobustScaler(), robust_cols),
                                      ('standard', StandardScaler(), standard_cols)],
                        remainder='passthrough')
```

- Escalamiento de variables
- Estandarización



Experimentación

```
▶ best_estimator_: RandomForestClassifier
  RandomForestClassifier
  RandomForestClassifier(class_weight='balanced', max_features=5,
    min_samples_leaf=2, n_estimators=50, random_state=4)
  |
```



```
▼ XGBClassifier
  XGBClassifier(base_score=None, booster=None, callbacks=None,
    colsample_bylevel=None, colsample_bynode=None,
    colsample_bytree=1.0, device=None, early_stopping_rounds=None,
    enable_categorical=False, eval_metric='logloss',
    feature_types=None, gamma=5, grow_policy=None,
    importance_type=None, interaction_constraints=None,
    learning_rate=0.05, max_bin=None, max_cat_threshold=None,
    max_cat_to_onehot=None, max_delta_step=None, max_depth=7,
    max_leaves=None, min_child_weight=None, missing=nan,
    monotone_constraints=None, multi_strategy=None, n_estimators=200,
    n_jobs=None, num_parallel_tree=None, random_state=1, ...)
```

Random Forest Classifier
RandomizedSearchCV

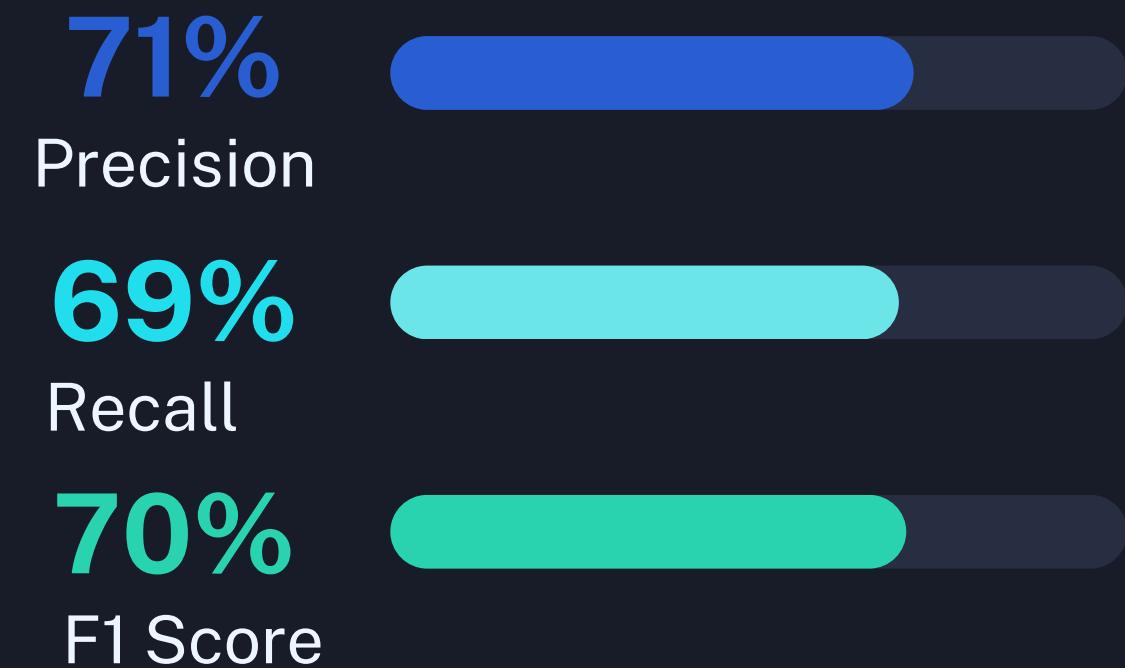
XGBClassifier
GridSearchCV

Balanceo de clases mediante hiperparámetros de modelo

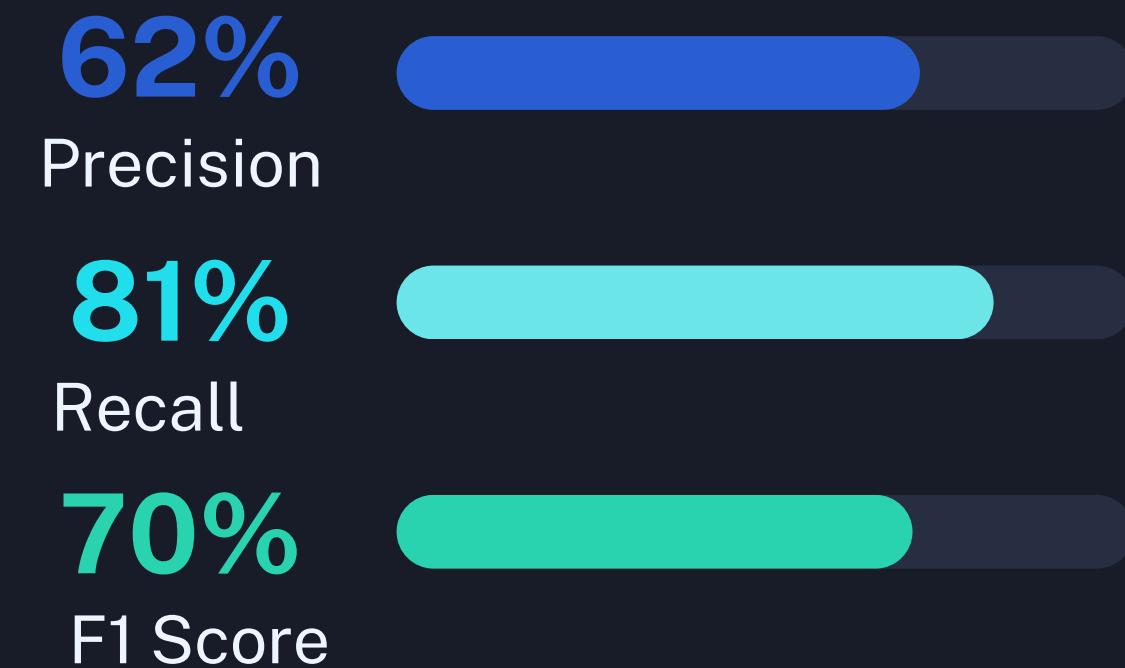


Experimentación

Random Forest Classifier

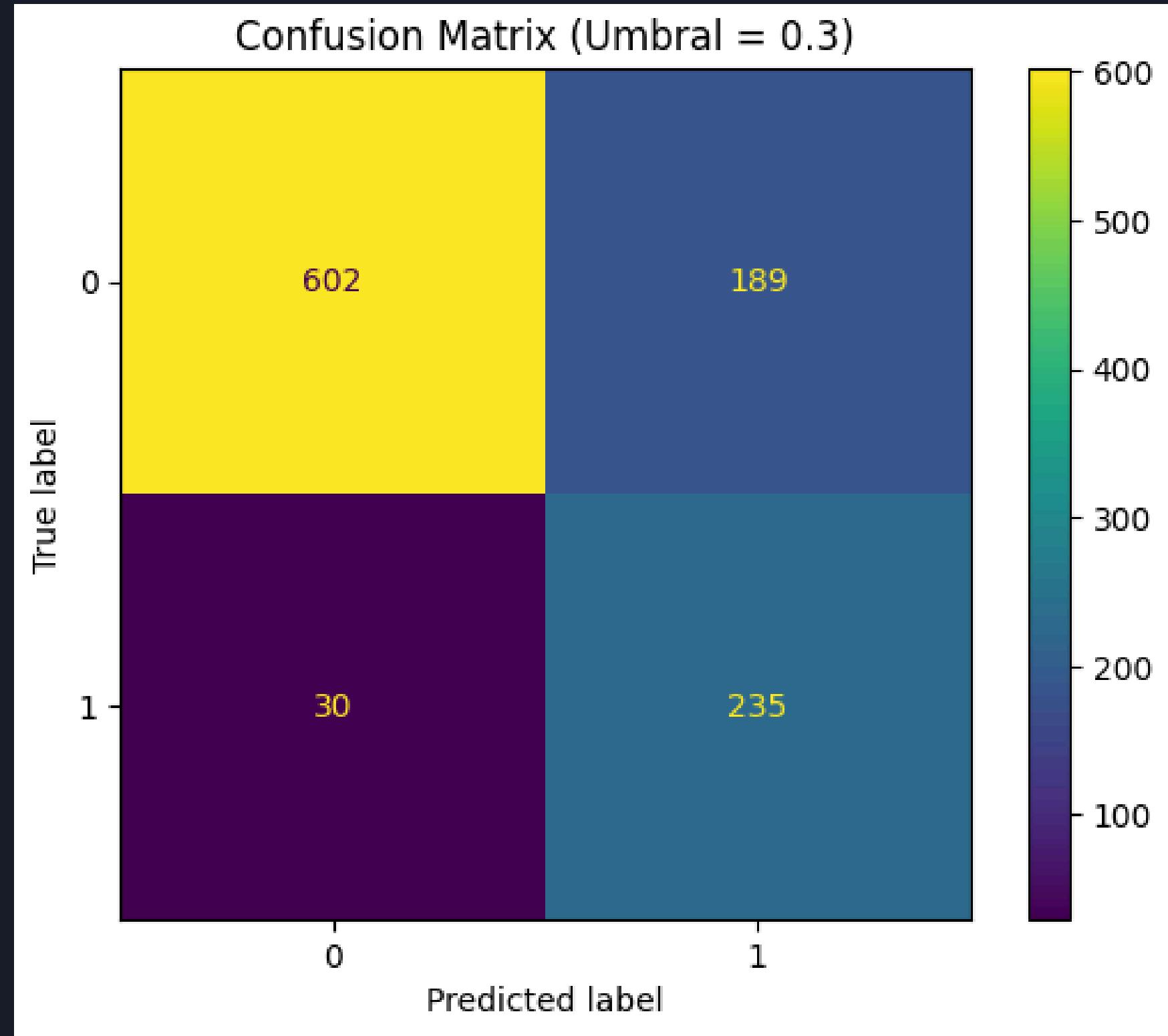


XGBClassifier





Experimentación



Se reduce umbral a 0.3 de decisión para poder detectar más casos 1

Métricas en el conjunto de validación:

55%

Precision

89%

Recall

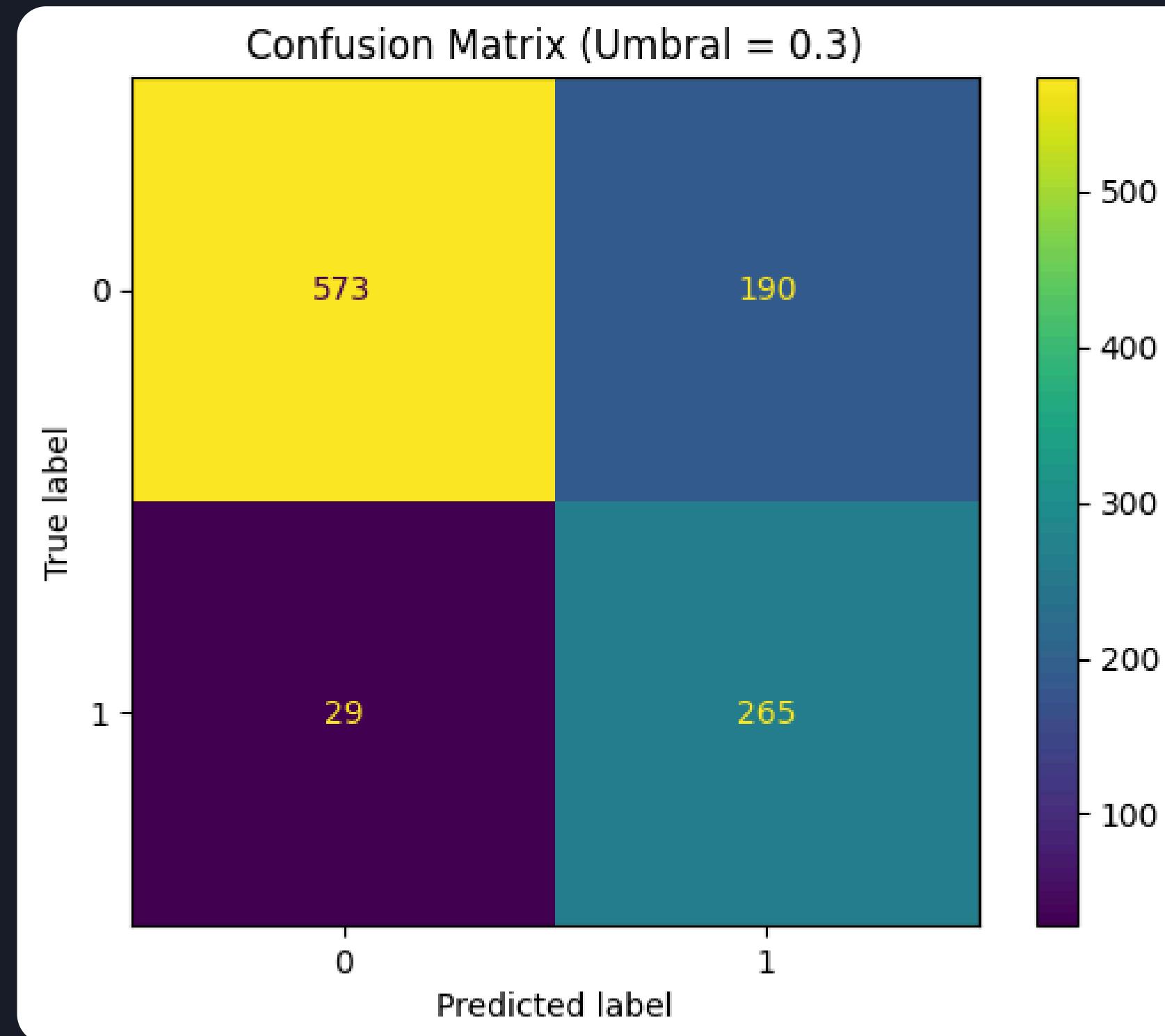
68%

F1 Score





Resultados Finales del Modelo (Test Data)



XGBClassifier

Métricas en el conjunto de test:

58%

Precision

90%

Recall

71%

F1 Score





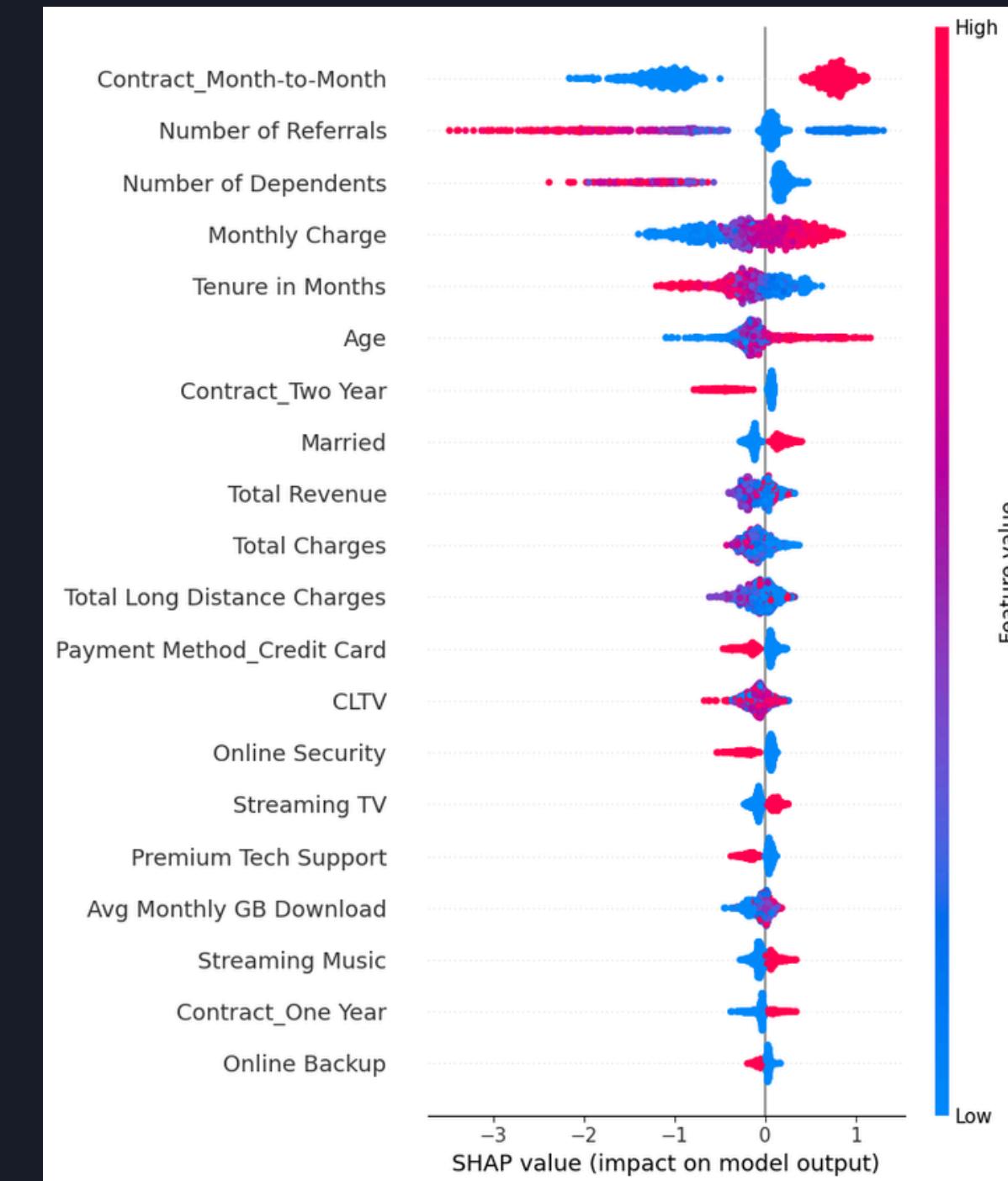
Explicación

Interpretabilidad y Explicabilidad

Variables que contribuyen mas al modelo



Comportamiento de la variable con respecto a X

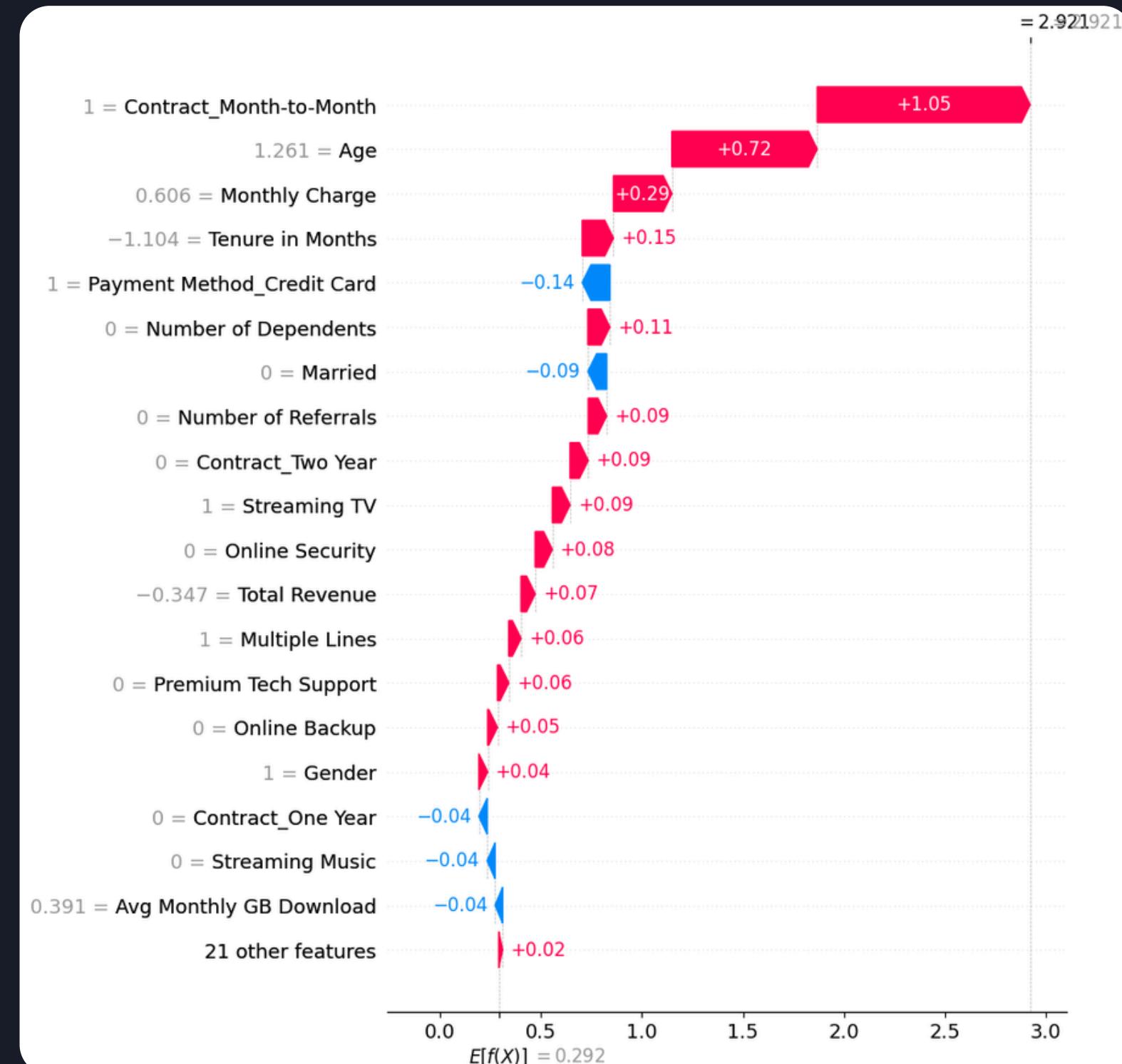
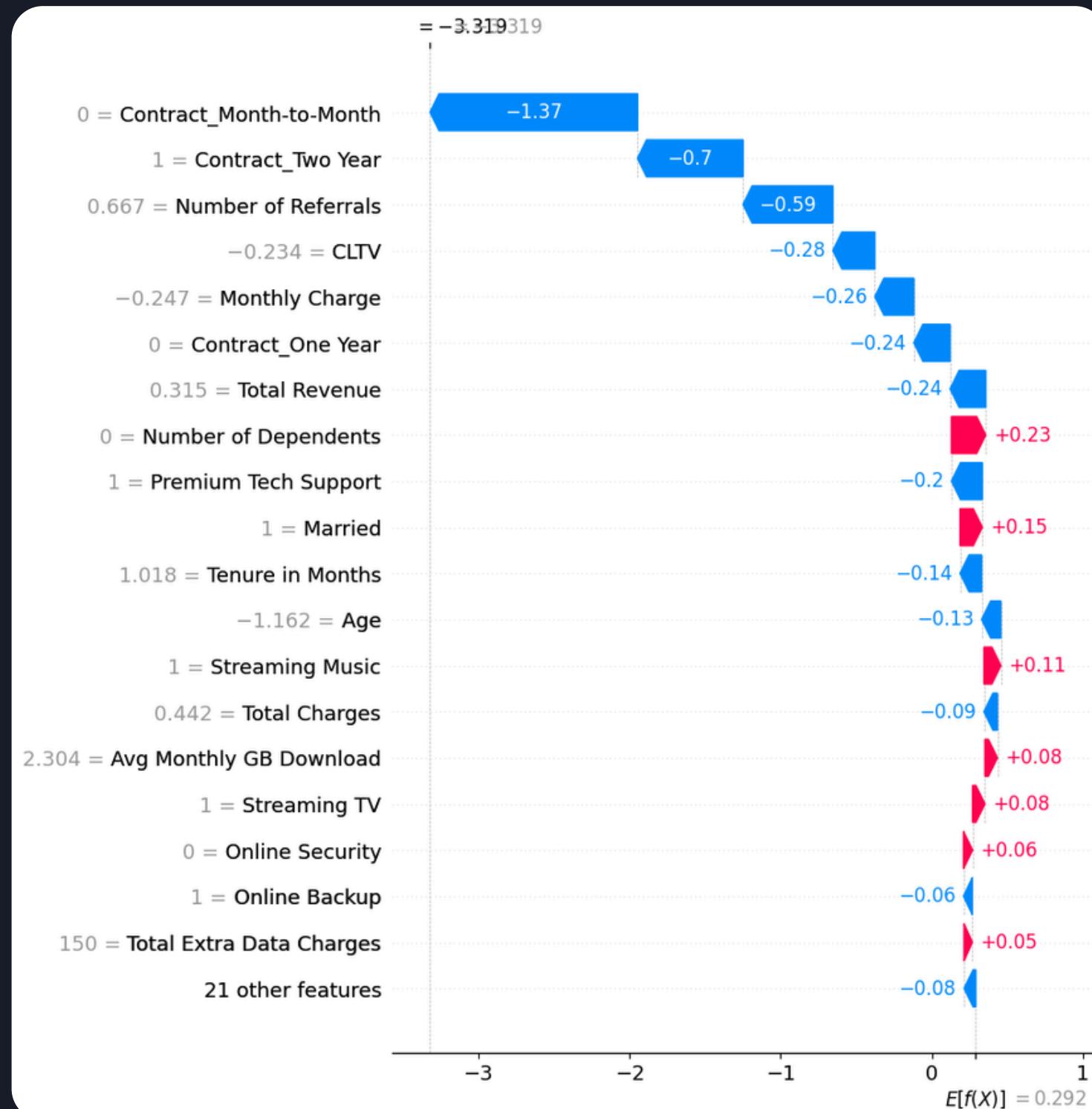




Explicación

Interpretabilidad y Explicabilidad

Explicabilidad a nivel de Cliente





Experiencia

Recomendaciones Estratégicas Basadas en el Modelo

SEGMENTACIÓN POR RIESGO

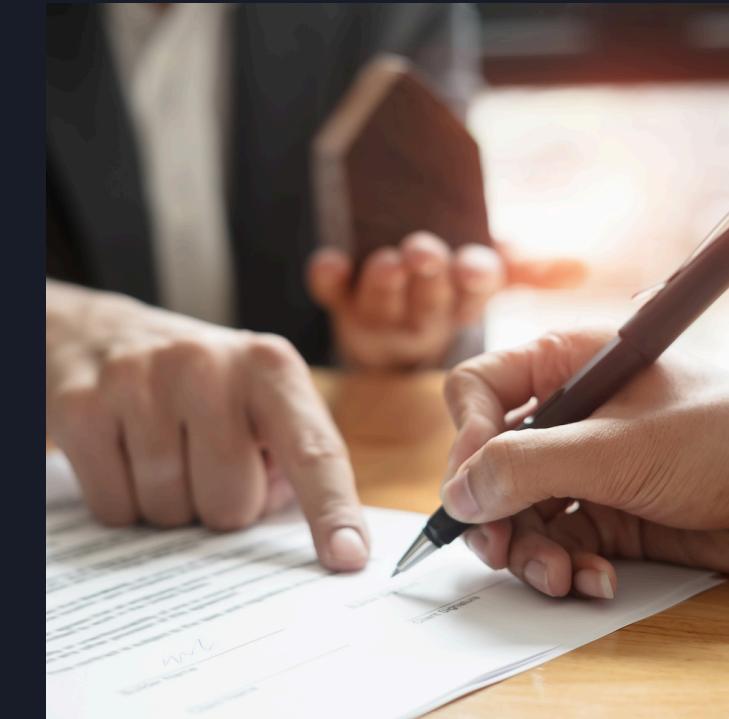
- Identificar grupos de clientes con alta probabilidad de abandono
- Aplicar estrategias de retención personalizadas

CONTRATOS A LARGO PLAZO

Factor clave: Tipo de contrato



Promover migración de contratos mensuales a anuales o bianuales, con beneficios adicionales



FIDELIZACIÓN

- Ofertas personalizadas por CLTV (Customer Lifetime Value)
- Clientes con alto CLTV deben recibir propuestas exclusivas premiando su fidelidad.



ALERTAS TEMPRANAS

Implementar sistemas automáticos que detecten señales de posible abandono que disparen acciones inmediatas, sobretodo en los primeros meses del servicio.





Consideraciones / Limitaciones

- Tiempo para realizar pruebas con más modelos



- Mayor poder computacional



- Mejorar histórico de datos





Conclusiones y Recomendaciones

- Se estableció un modelo XGBClassifier para predecir el comportamiento del usuario.
- El tipo de contrato, el número de referidos y dependientes, el tiempo de uso en meses y el cargo mensual son variables clave en la decisión de churn.
- Se ha trabajado en contar con un buen Recall para detectar la mayor cantidad de personas que abandona el servicio, sin embargo se recomienda mejorar el modelo para contar un Accuracy más robusto.
- La interpretabilidad es fundamental para convertir los resultados técnicos en decisiones estratégicas por lo que sería recomendable considerar otras variables como el nivel de ingreso y la competencia directa en la zona geográfica del cliente.