PROYECTO FINAL

Entrenamiento y optimización de modelos de machine learning para la predicción del abandono de clientes en una empresa de telecomunicaciones

Realizado por:
Carlos Parra
CODER HOUSE

INDICE

CONTEXTO Y AUDIENCIA

Contexto

Este análisis se centra en el comportamiento de los clientes de una empresa de telecomunicaciones para determinar que suscriptores tienen más probabilidades de abandonar la plataforma.

La idea es descubrir el comportamiento de los clientes a través del análisis exploratorio de datos (EDA) y luego utilizar algunas de las técnicas de análisis predictivo para determinar los clientes que tienen más probabilidades de abandonar.

Audiencia

Esta investigación esta dirigida a los responsables de la toma de decisiones del área comercial de la empresa de Telecomunicaciones, en virtud de la perdida de suscriptores activos en los últimos meses.

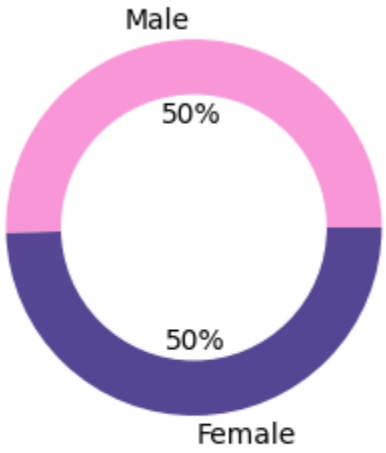
PREGUNTAS DE INTERÉS

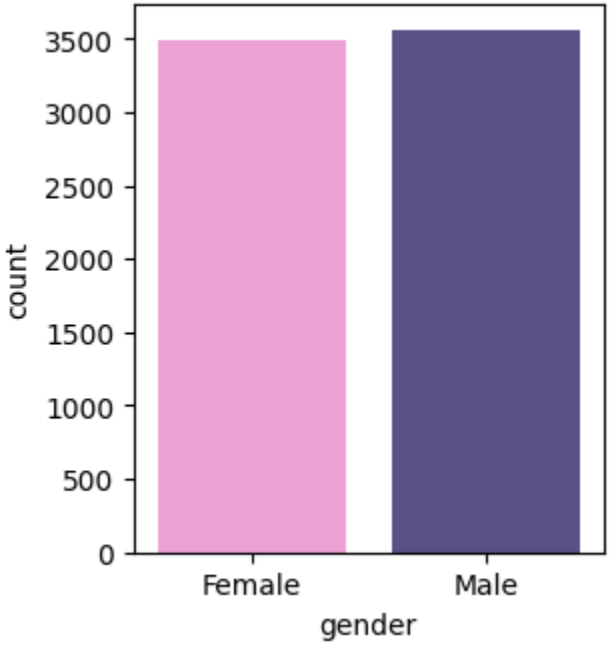
1.Determinar si la cantidad y el tipo de servicios contratados influye en la retención del cliente.
2.Determinar cómo se correlacionan las distintas variables del estudio con la duración de los contratos y la retención.
3.Determinar si el aumento en los cargos incide en la tasa de abandono
4.Predecir a través de un modelo si el cliente va a abandonar el servicio o no y cuáles son las variables que más influyen en la predicción.
5.¿Cuáles son los indicadores clave de la rotación de clientes?
6.Demostrar que el aumento de los servicios contratados incide en la permanencia de los clientes
7.¿Qué estrategias de retención se pueden implementar en función de los resultados para disminuir la pérdida de clientes potenciales?

ANÁLISIS EXPLORATORIO

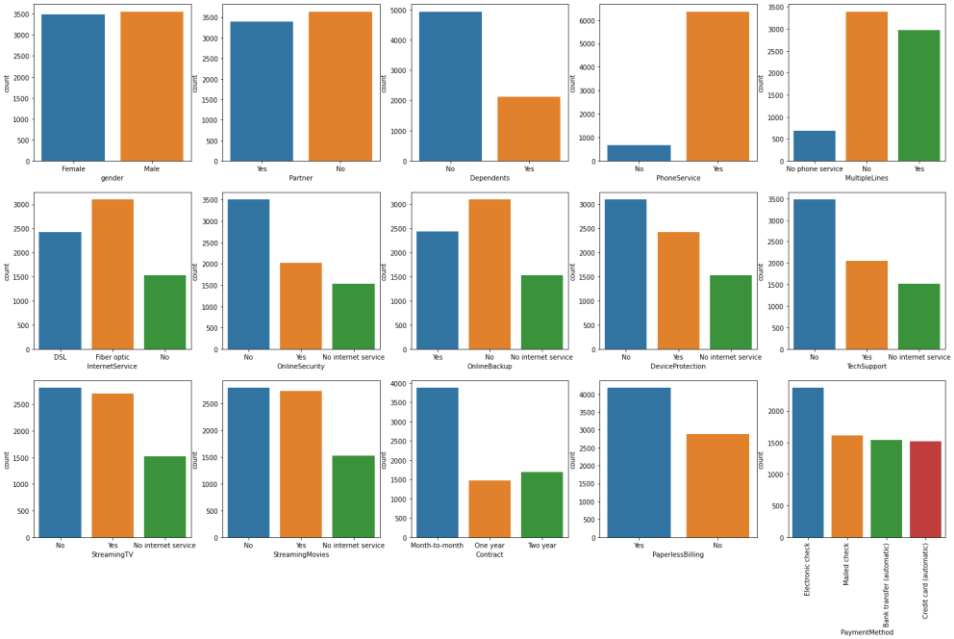
TASA DE ABANDONO DE CLIENTES:

- Se evidenció que casi 1/4 de los clientes ha abandonado de contrato en el periodo analizado
- El Genero no es un factor relevante y/o no influyen en la tasa de abandono

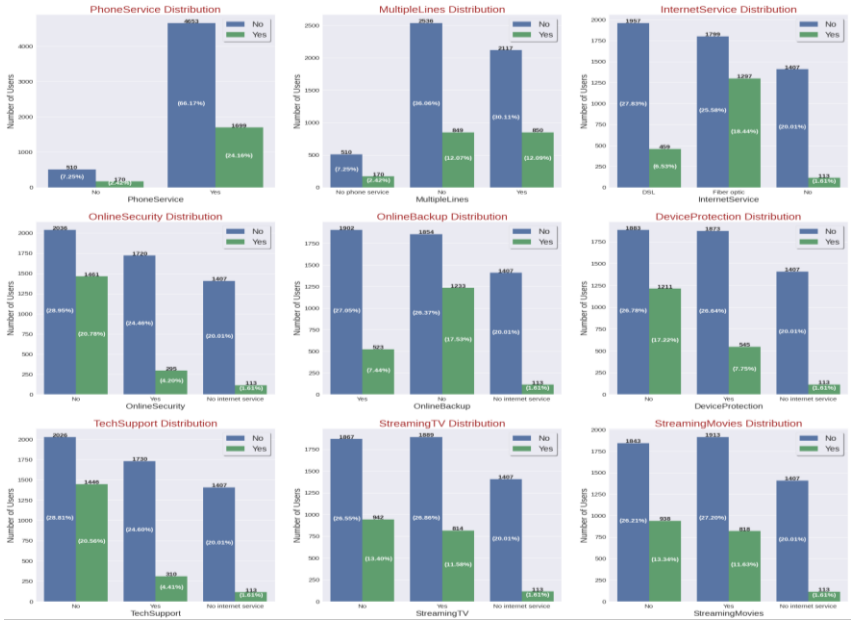




ANÁLISIS UNIVARIADO



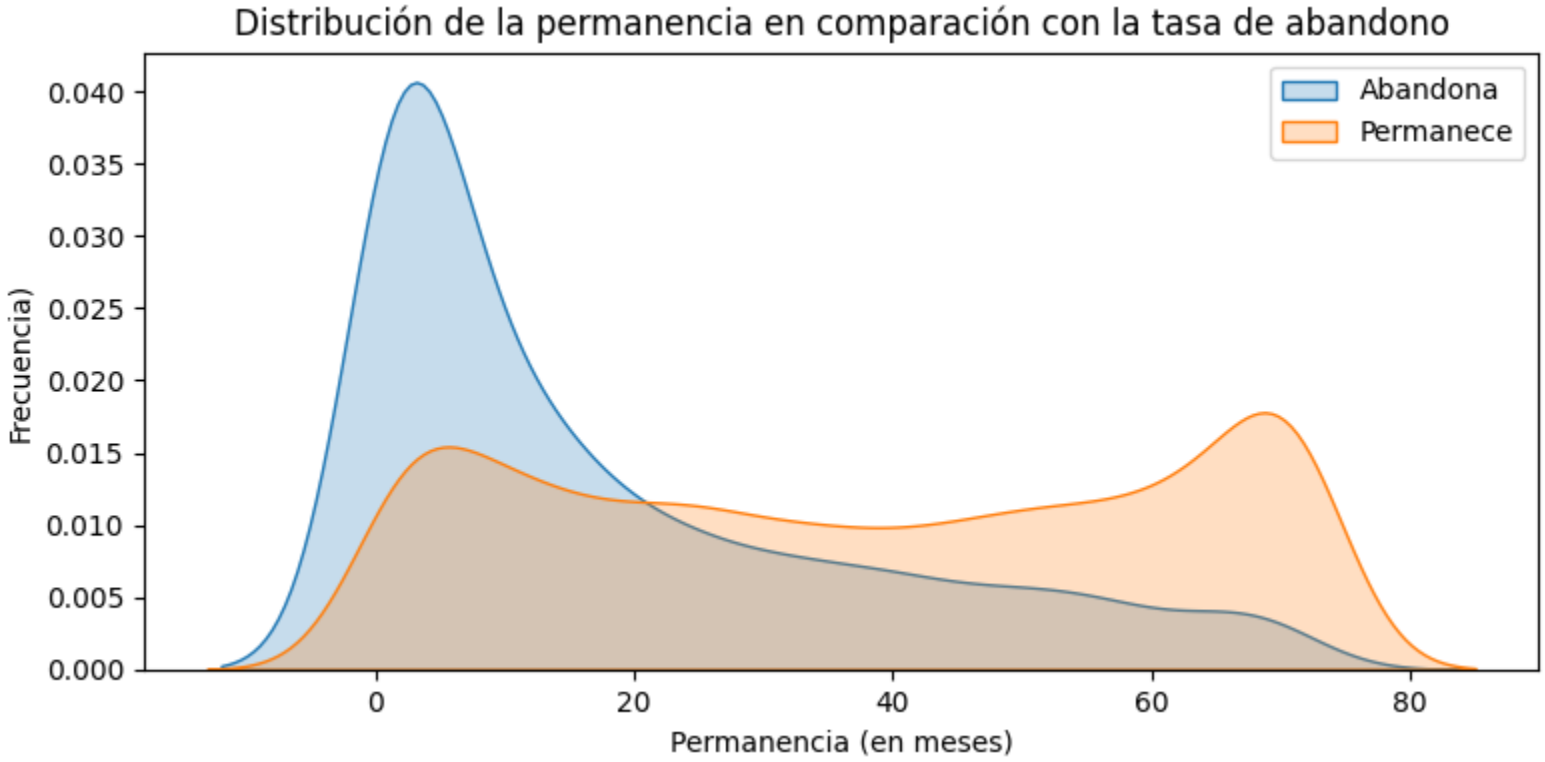
ANÁLISIS BIVARIADO



ANÁLISIS EXPLORATORIO

PERMANENCIA DE LOS CLIENTE (en el periodo analizado)

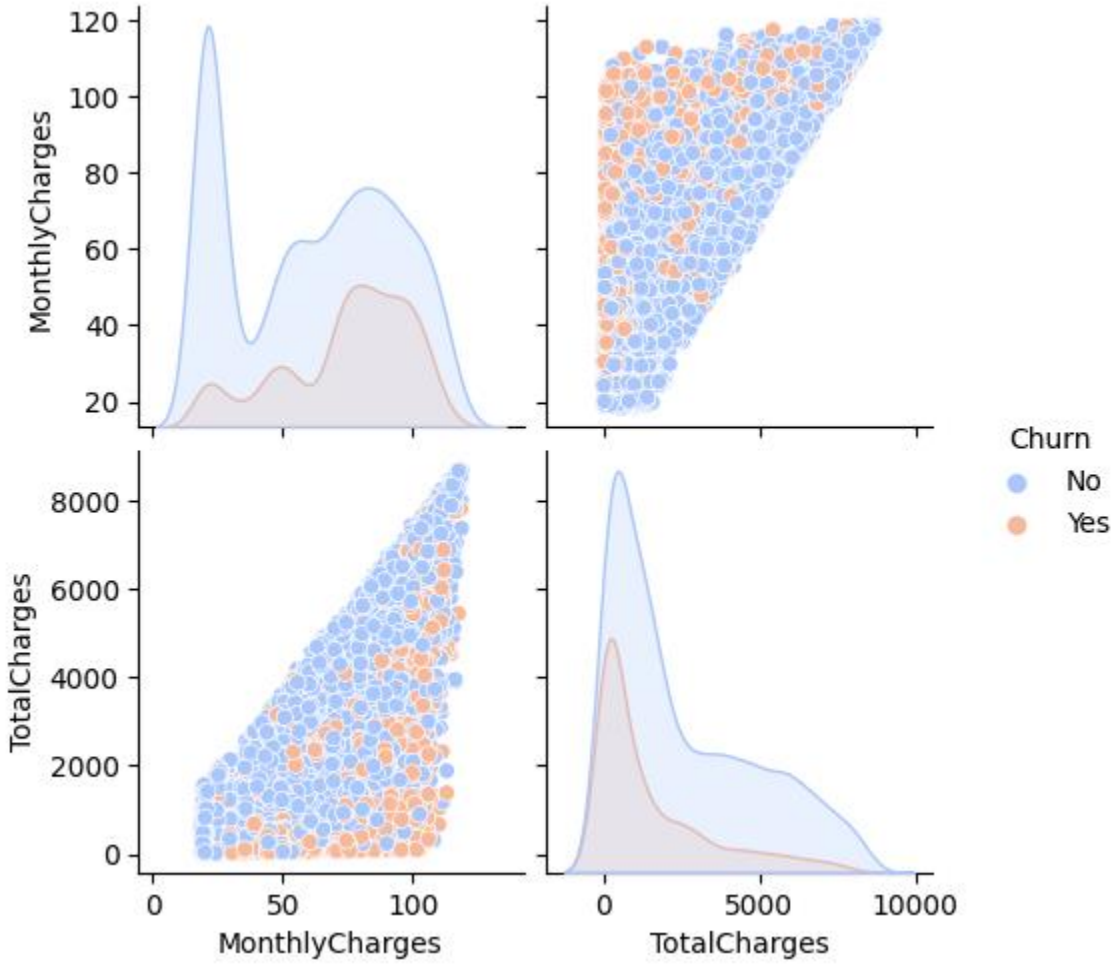
- Según la información suministrada, la mayoría de los clientes que abandonaron estuvieron en la empresa menos de 20 meses, a medida que aumenta la permanencia la probabilidad de abandonar disminuye.



ANÁLISIS EXPLORATORIO

INCIDENCIA DE LOS CARGOS (MENSUALES Y TOTALES) EN LA TASA DE ABANDONO:

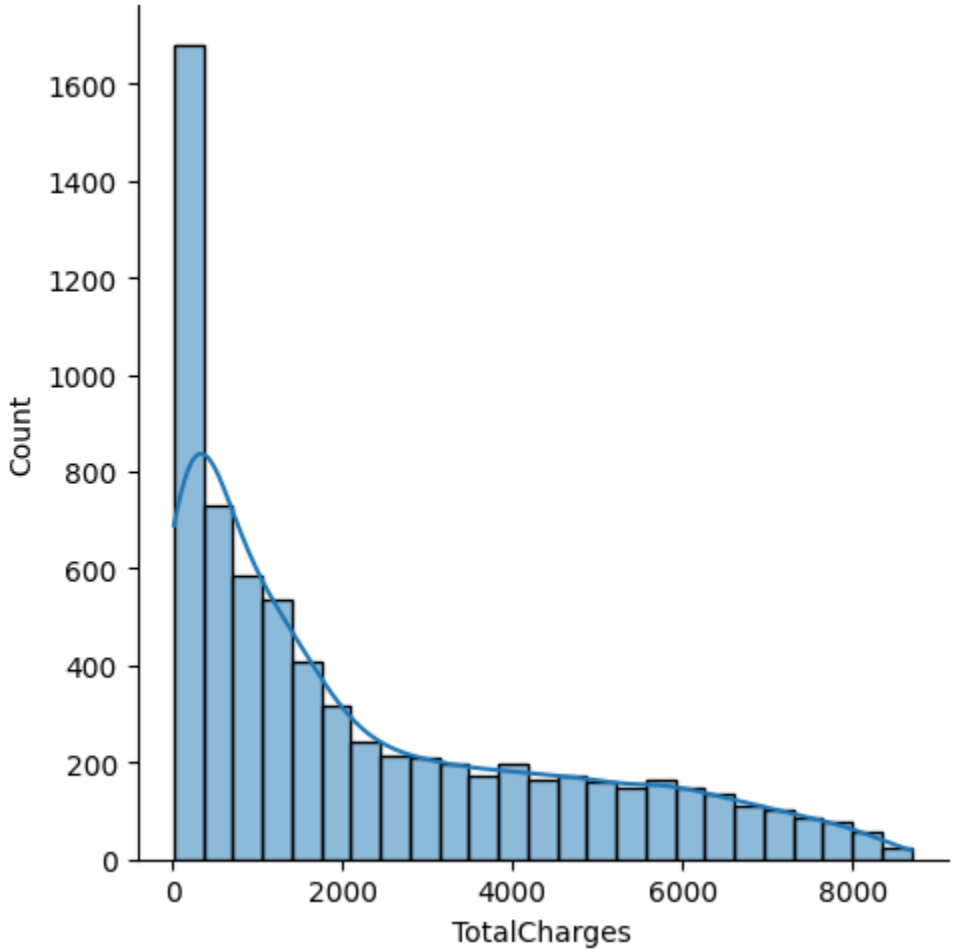
- Según la distribución de los cargos de los clientes, se observa un sesgo a la izquierda, lo que evidencia que a medida que aumentas los cargos, el abandono disminuye.



ANÁLISIS EXPLORATORIO

ANÁLISIS DE LOS CARGOS TOTALES:

- La media de los cargos totales es de aproximadamente $2283.30, mientras que la mediana es de 1397.47. La diferencia entre la media y la mediana sugiere que hay valores atípicos en el extremo superior de la distribución

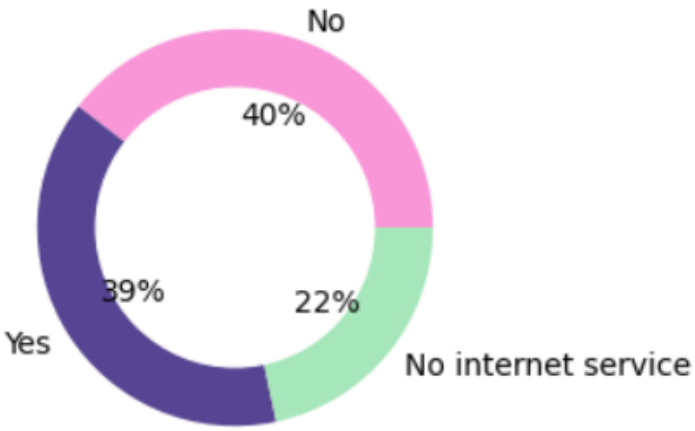


ANÁLISIS EXPLORATORIO

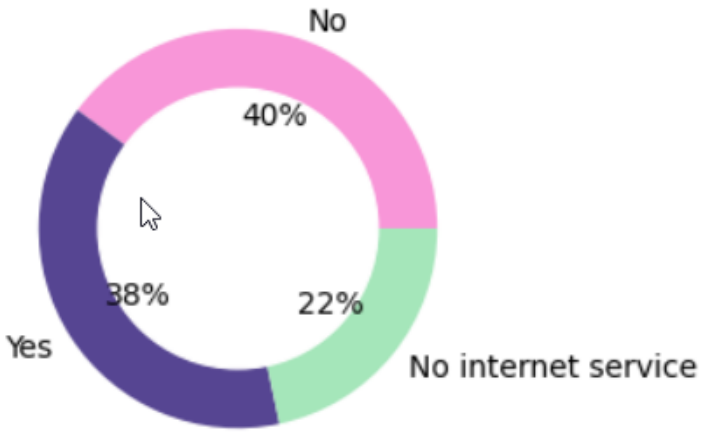
INFLUENCIA DE LOS SERVICIOS DE STREAMING:

- No se observa ninguna influencia de los servicios de streaming en el abandono.

```
------------------ STREAMINGMOVIES ---------
No                     2785
Yes                    2732
No internet service    1526
Name: StreamingMovies, dtype: int64
```




```
------------------ STREAMINGTV -------------
No                     2810
Yes                    2707
No internet service    1526
Name: StreamingTV, dtype: int64
```



ALGORITMOS DE CLASIFICACIÓN

Se utilizaron algoritmos de clasificación al dataset original obteniendo las siguientes métricas, que no son las esperadas en un modelo eficiente, por ello se procede a aplicar otras técnicas.

Classification report for Logistic Regression:

	precision	recall	f1-score	support
0	0.84	0.89	0.86	1033
1	0.63	0.53	0.57	374
accuracy			0.79	1407
macro avg	0.73	0.71	0.72	1407
weighted avg	0.78	0.79	0.79	1407

Classification report for Support Vector Machines:

	precision	recall	f1-score	support
0	0.81	0.93	0.86	1033
1	0.67	0.38	0.48	374
accuracy			0.78	1407
macro avg	0.74	0.66	0.67	1407
weighted avg	0.77	0.78	0.76	1407

Classification report for K-Nearest Neighbors:

	precision	recall	f1-score	support
0	0.82	0.85	0.84	1033
1	0.55	0.50	0.52	374
accuracy			0.76	1407
macro avg	0.69	0.68	0.68	1407
weighted avg	0.75	0.76	0.75	1407

Classification report for Linear Discriminant Analysis:

	precision	recall	f1-score	support
0	0.84	0.88	0.86	1033
1	0.62	0.54	0.58	374
accuracy			0.79	1407
macro avg	0.73	0.71	0.72	1407
weighted avg	0.78	0.79	0.78	1407

Classification report for Decision Tree:

	precision	recall	f1-score	support
0	0.73	1.00	0.85	1033
1	0.00	0.00	0.00	374
accuracy			0.73	1407
macro avg	0.37	0.50	0.42	1407
weighted avg	0.54	0.73	0.62	1407

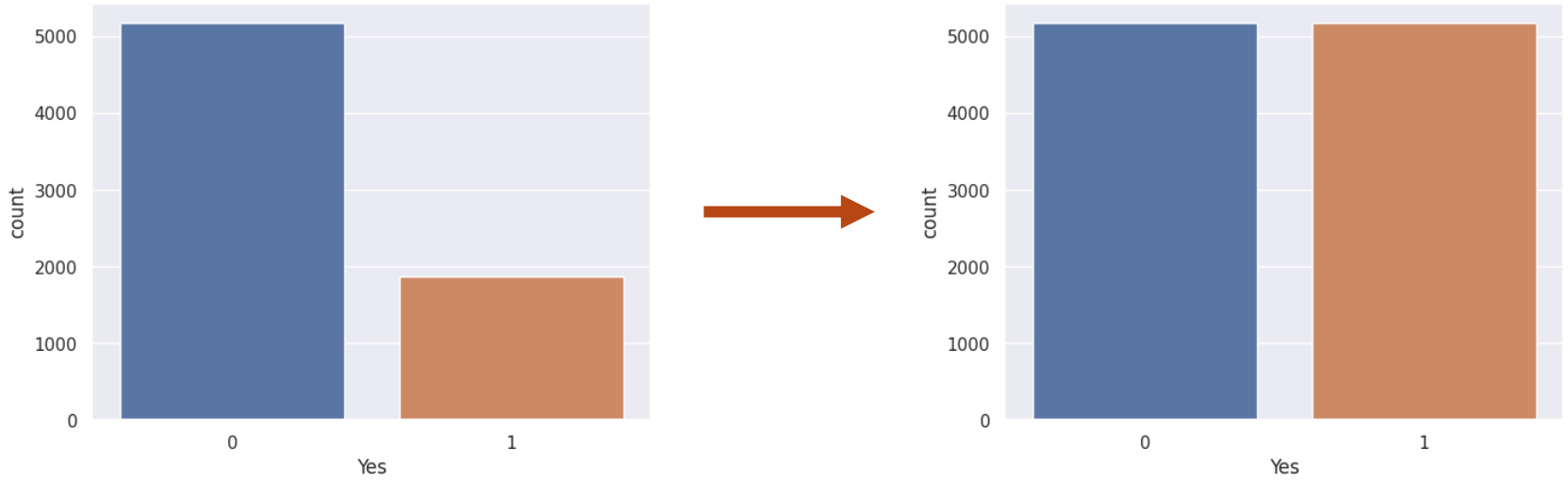
Classification report for Naive Bayes:

	precision	recall	f1-score	support
0	0.90	0.72	0.80	1033
1	0.50	0.77	0.61	374
accuracy			0.74	1407
macro avg	0.70	0.75	0.70	1407
weighted avg	0.79	0.74	0.75	1407

INGENIERÍA DE ATRIBUTOS

Luego de todos los algoritmos y técnicas aplicadas encontramos que los resultados obtenidos en función al accuracy no son aceptables, procederemos a cambiar el enfoque y centrarnos específicamente en mejorar las métricas en función del resultado deseado que es predecir más eficientemente la tasa de abandono, que basado en los datos seria predecir la variable Churn con resultado = 1, mejorando el recall, para ello haremos uso de otros algoritmos más ajustados en función de los hiperparametros y aplicando oversampling

Resultado del oversampling a la variable Churn para balancear la data.

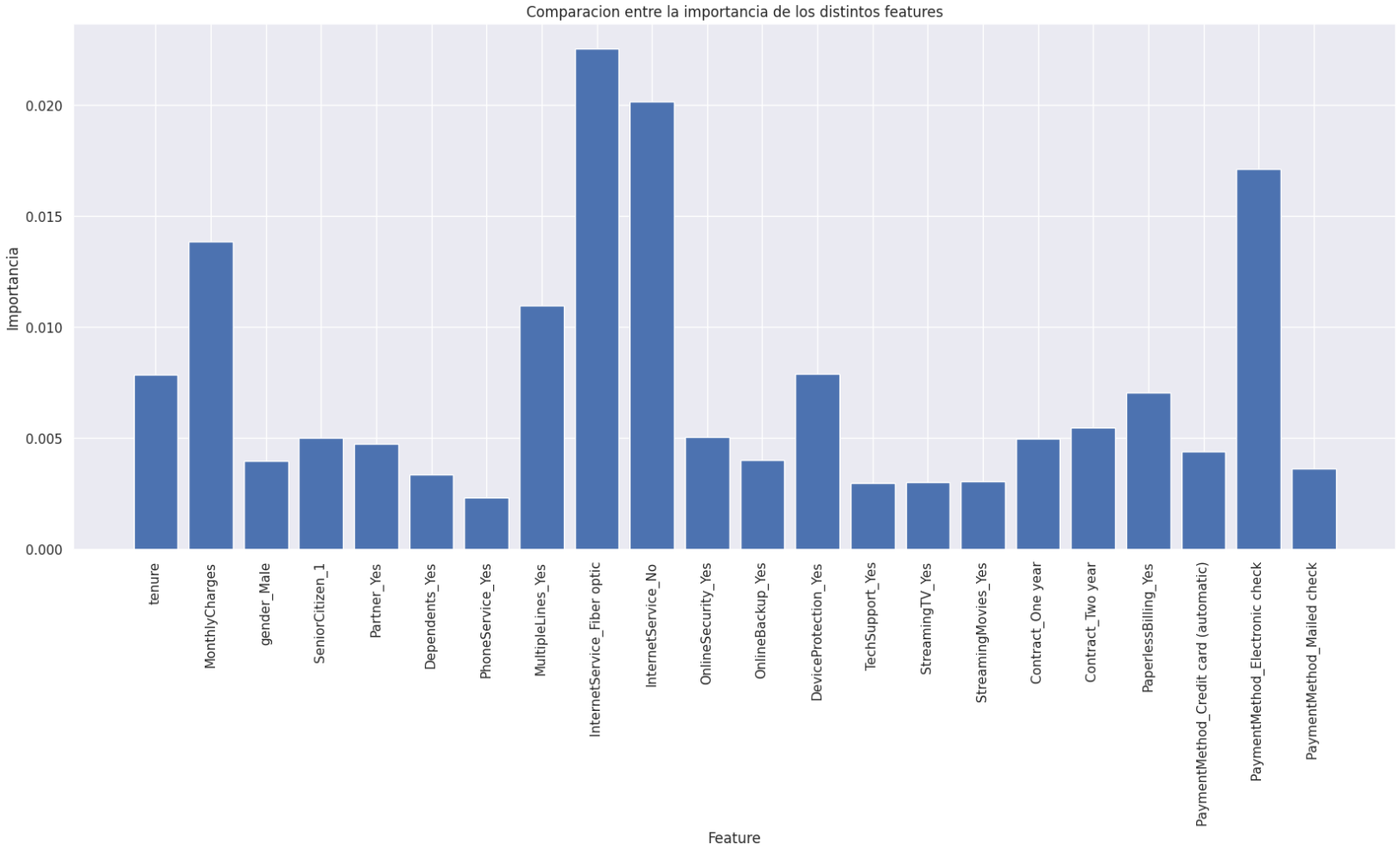


Antes

Después

ENTRENAMIENTO Y TESTEO

Se aplica adicionalmente un método árboles adicionales *(sklearn.ensemble.ExtraTreesClassifier)* para mejorar los resultados, resaltar los features más importantes y entrenar de nuevo los modelos



INSIGHTS

- El dataset contiene información sobre 7,043 clientes de una compañía de telecomunicaciones, con 21 variables diferentes.
- La tasa de abandono de clientes (churn rate) en el dataset es del 26.5%, lo que sugiere que la compañía de telecomunicaciones puede estar teniendo problemas para retener a sus clientes.
- La mayoría de los clientes (68.5%) son usuarios de servicios de telefonía, mientras que el resto son usuarios de servicios de Internet y servicios de televisión.
- Los clientes que tienen servicios de Internet de alta velocidad son más propensos a abandonar la compañía que los clientes que tienen servicios de Internet de baja velocidad.
- Los clientes que tienen contratos mensuales son mucho más propensos a abandonar la compañía que los clientes que tienen contratos a largo plazo.
- Los clientes que tienen cargos mensuales más altos y que han estado en la compañía por más tiempo tienden a tener cargos totales más altos.
- Las variables MonthlyCharges y TotalCharges están altamente correlacionadas, lo que sugiere que la compañía podría estar cobrando precios más altos a los clientes que han estado en la compañía por más tiempo.
- La mayoría de los clientes que abandonaron la compañía lo hicieron durante los primeros dos años, lo que sugiere que la compañía podría beneficiarse al retener a los clientes por más tiempo y mejorar la experiencia del cliente durante los primeros años de su contrato
- Se sugiere para mejorar la retención de sus clientes ofrecer contratos a largo plazo y mejorando la experiencia del cliente en los primeros años de su contrato.
- También podrían considerar la posibilidad de ajustar sus precios para retener a los clientes de mayor valor

CONCLUSIONES

Luego de todos los algoritmos y técnicas aplicadas encontramos que los resultados obtenidos en función al accuracy no son aceptables, procederemos a cambiar el enfoque y centrarnos específicamente en mejorar las métricas en función del resultado deseado que es predecir mas eficientemente la tasa de abandono, que basado en los datos seria predecir la variable Churn con resultado = 1, mejorando el recall, para ello haremos uso de otros algoritmos más ajustados en función de los hiperparametros.

Al observar que los datos están desabalceados se aplica oversampling para corregir esto y establecer las variables más importantes, para a partir de allí, reentrenar los modelos y elegir el mejor predictor.

Luego de este enfoque (Mejorar el Recall), podemos determinar que el mejor modelo es el Random Forest Classifier, con un Recall de 95% para la predicción acertada de los pueden abandonar y un resultado de 84% para los que no.

```
[ ]  from sklearn.ensemble import RandomForestClassifier
     param_grid = {
     'n_estimators': [50, 75,100, 150, 200,300],
     }
     rcv=RandomizedSearchCV(RandomForestClassifier(random_state=42),param_grid,cv=5)
     rcv.fit(X_train,y_train)
     y_pred_rcv=rcv.predict(X_test)
     print(classification_report(y_test,y_pred_rcv))
```

```
                  precision    recall  f1-score   support

               0       0.94      0.84      0.89      1298
               1       0.86      0.95      0.90      1289

        accuracy                           0.89      2587
       macro avg       0.90      0.90      0.89      2587
    weighted avg       0.90      0.89      0.89      2587
```