d:\DATA SCIENSE\HACKATHON\proyecto_final_mundose.ipynb

> Rodriguez Natacha - Estudiante de Ing. Química - Tester Jr
> Fabricio Sgro - Arquitecto de Datos Jr

[1]

<"churn prediction"/>

Predicción de bajas de clientes para empresas de telecomunicaciones indias.

Para que una empresa siga teniendo éxito, es fundamental que conserve su base de clientes existente. Para ello, las empresas deben conocer la tasa de abandono de sus clientes. Esto les permite reaccionar a las tendencias de forma temprana y, si es necesario, adaptar sus servicios a la demanda.

<"¿Qué significa churn?"/>

se traduce como pérdida de clientes o rotación de clientes. El término “churn” se compone de las palabras inglesas “change” (en castellano cambio) y “turn” (en castellano abandonar). Customer churn es la tasa de abandono, indicador clave de rendimiento (KPI). la relación entre el número de clientes que ya no utilizan un servicio y el número total de sus clientes. El resultado proporciona información sobre una posible disminución de la base de clientes en un período de tiempo determinado.

<"Importancia del customer churn"/>

Es inevitable que se produzcan pérdidas de clientes. No es realista pensar que el 100% de los clientes que le compraron el primer día de su negocio seguirán estando con usted varios años después. Pero cuando la tasa de abandono de sus clientes es muy alta o si muestra una tendencia a aumentar con el tiempo, querrá tomar medidas.

En términos generales, la pérdida de clientes es una mala noticia por las siguientes razones:

1. Los clientes insatisfechos afectan negativamente a su marca.
   La empresa podría recibir malas críticas y perjuicios que pueden afectar al valor de su marca.
2. La rotación de clientes le sale más cara.
   A menudo se dice que mantener un cliente existente cuesta menos y ofrece más valor que adquirir uno nuevo.
3. La rotación de clientes puede impactar a su crecimiento.
   Si la empresa está considerando traer nuevos productos y servicios al mercado, es probable que la mejor audiencia sean sus clientes existentes que ya conocen su marca. Sin embargo, si el valor de vida de sus clientes es bajo, su nuevo proyecto podría verse afectado.[1]

1. Fuente: https://www.qualtrics.com/es/gestion-de-la-experiencia/cliente/customer-churn/

[2] <"¿Por qué es importante este problema?"/>

Desarrollaremos un análisis de la base de datos de empresas una solución para entender y abordar factores que impactan la fidelización y la deserción de los clientes de las principales empresas de telecomunicaciones.

Analizaremos los datos que disponemos para revelar patrones, identificar áreas de mejora con las cual estas empresas puedan aplicar estrategias proactivas.

Nuestra solución basada en datos e hipótesis, proporcionarán claridad sobre las necesidades de los clientes, permitiendo decisiones estratégicas informadas. Nos comprometemos a documentar nuestros hallazgos para que las empresas del rubro puedan mejorar la calidad del servicio, fortalecer la retención y cultivar relaciones duraderas en el futuro.

Como científico de datos de este proyecto, su objetivo es explorar la intrincada dinámica del comportamiento de los clientes y la demografía en el sector de telecomunicaciones de la India para predecir la pérdida de clientes, utilizando dos conjuntos de datos completos de cuatro importantes socios de telecomunicaciones: Airtel, Reliance Jio, Vodafone y BSNL.

Se dispone de dos data frames:

> `telecom_demographics.csv` contiene información relacionada con la demografía de los clientes indios:

Variable	Descripción
`customer_id `	Identificador único de cada cliente.
`telecom_partner `	El socio de telecomunicaciones asociado con el cliente.
`gender `	El género del cliente.
`age `	La edad del cliente.
`state`	El estado indio en el que se encuentra el cliente.
`city`	La ciudad en la que se encuentra el cliente.
`pincode`	El código PIN de la ubicación del cliente.
`registration_event`	Fecha en la que se registró con el socio.
`num_dependents`	El número de dependientes (por ejemplo, niños) que tiene el cliente.
`estimated_salary`	El salario estimado del cliente.

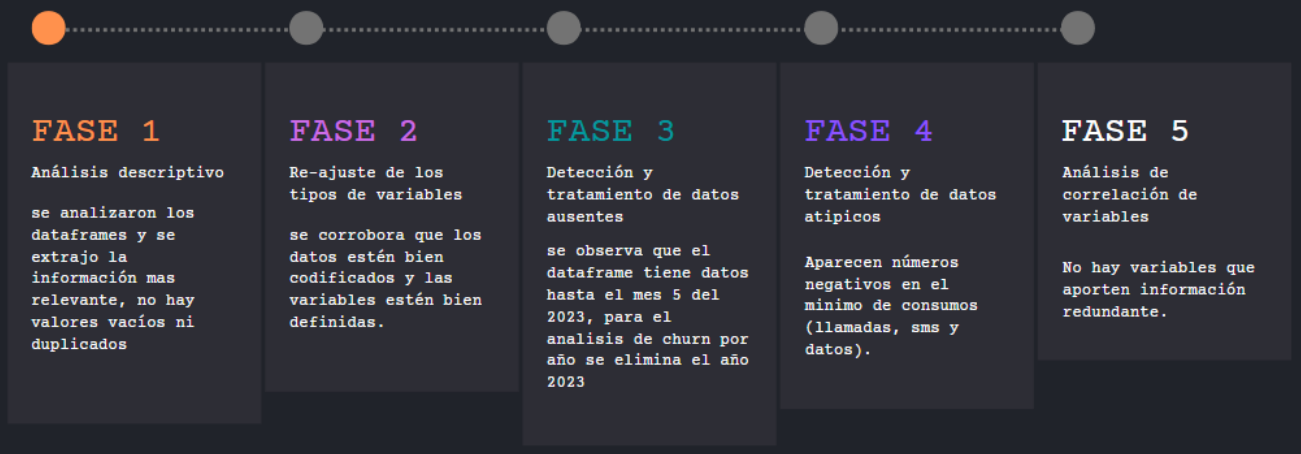
[3]

> `telecom_usage.csv` contiene información sobre los patrones de uso de los clientes indios:

Variable	Descripción
`customer_id `	Identificador único de cada cliente.
`calls_made`	El número de llamadas realizadas por el cliente.
`sms_sent`	El número de mensajes de SMS enviados por el cliente.
`data_used`	La cantidad de datos utilizada por el cliente.
`churn`	El estado indio en el que se encuentra el cliente.

<"ANALISIS EXPLORATORIO DE DATOS (EDA)"/>

Son el conjunto de técnicas estadísticas dirigidas a explorar, describir y resumir la información que contienen los datos maximizando su comprensión, presuponen las condiciones previas para garantizar la objetividad e interoperabilidad de los datos.



<"Herramientas y tecnologías utilizadas"/>

> Modelos de Machine Learning

Modelos de ensamble: Para capturar relaciones no lineales que mejoren la precisión.

> Procesamiento de datos

One Hot Encoding y Label Encoding para variables categóricas. Estandarización, normalización y scaling para variables numéricas. Análisis de 'churn' a lo largo del tiempo gracias variables de tiempo.

[4]

> Estadística aplicada

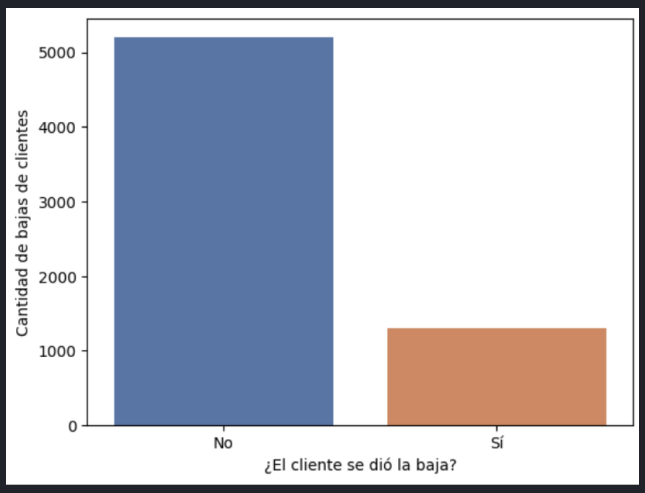
Pruebas estadísticas para la aceptación de hipótesis (p-valor).

<"¿Cómo fue el proceso?"/>

Nos encontramos con un desbalance de clases que tratamos con la metodología de undersampling:

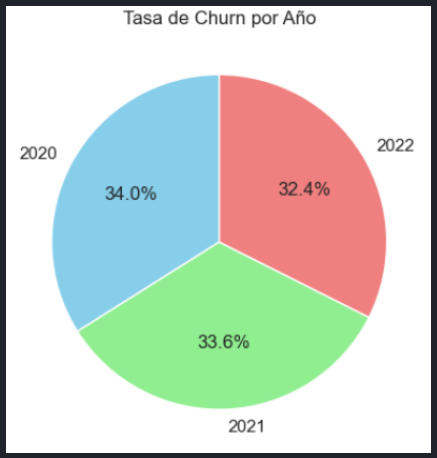
```
from imblearn.under_sampling import RandomUnderSampler
from collections import Counter

undersampler = RandomUnderSampler(sampling_strategy=0.28, random_state=42)
X_train, y_train = undersampler.fit_resample(X, y)

print ("Distribution before resampling {}".format(Counter(y)))

print ("Distribution labels after resampling {}".format(Counter(y_train)))
```

```
Distribution before resampling Counter({0: 5197, 1: 1303})
Distribution labels after resampling Counter({0: 4653, 1: 1303})
```

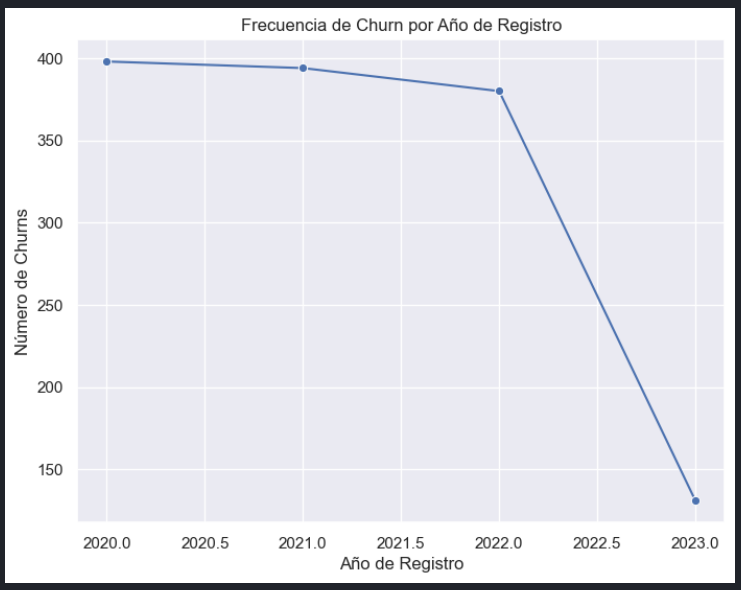


Continuamos con nuestro Análisis exploratorio de las variables:

Es muy interesante ver que el año en el que llega la pandemia (2020) se produce la mayor cantidad de bajas y esto puede ser debido a que mucha gente tuvo que dejar de trabajar debido al confinamiento y decidió darse de baja del servicio para reducir costos.



[5]



Investigamos cómo influye el estado al churn

Vemos que hay estados con mayores probabilidades de que sus habitantes renuncien, sin embargo, el porcentaje o probabilidad no difiere mucho de estado a estado

Preprocesamiento

Realizamos tareas de pre-procesamiento, para el entrenamiento de los modelos y revisamos con todas las variables del dataset que correlación, mediante spearman podía tener la variable churn, pero vemos que no existe una relación fuerte (negativa o positiva) con el resto de variables.

Función de entrenamiento + ploteo

Creamos un .py con funciones a importar en nuestro notebook que nos permite hacer predicciones (make_prediction) y luego verificar los resultados mediante una matriz de confusión (verify_results)

Modelos seleccionados

Elegimos modelos ensamble (Bagging y boosting) ya que vemos que la clasificación requería de una resolución no-lineal.

Randomforest los modelos sin hyper-param tunning y sin cross-validation presentaron altos niveles de resultados (overfitting)

<"Conclusión"/>

Pudimos comprobar que predecir el churn de los clientes de empresas de telecomunicaciones sí es posible.

A su vez, logramos detectar qué variables son imprescindible que las empresas de telecomunicación tengan siempre actualizada para poder tener en tiempo real, respuestas concretas sobre un cliente que pueda estar en riesgo de darse de baja, basándonos principalmente en su salario, edad y los consumos que ha realizado en su línea móvil (llamadas, sms y paquetes de datos).