DATA SCIENCE

CHURN EN TELECOMUNICACIONES

Análisis de la Empresa Telco

Elaborado por:
Florencia Garcia Palacio
Federico Hofmann
Gabriel Jourdan
Victoria Ramondelli

Supervisado por:
Damian Dapueto
Alejandro Pujol

Para:
CoderHouse - Data Science
Comision 29730

Tabla de contenido

Descripción del caso de negocio

Las empresas cuyo modelo de negocio refiere a la prestación de un servicio, tienen como principal objetivo lograr captar y retener clientes constantemente. Sin dudas que este es un desafío difícil teniendo en cuenta el acceso que tienen los consumidores a información de todo tipo; como sus derechos de usuario, funcionamiento del servicio, quejas o reclamos, innovación de la empresa o empresas competidoras, precios, etc.

De acuerdo al artículo publicado en el blog de Microtech (link), resulta de 6 a 7 veces más costoso captar nuevos clientes que retenerlos.. Por eso es importante contar con indicadores que permitan medir, controlar, y tomar acciones preventivas frente a posibles abandonos de servicio de parte de los consumidores.

La predicción de estos sucesos es, probablemente, uno de los casos de aplicación más importantes de la analítica de datos en el sector comercial. *Churn* hace referencia, en términos comerciales, al hecho de que un cliente cancele una suscripción a un servicio que ha estado usando. Un ejemplo común es la gente que cancela su suscripción a Spotify o Netflix. Por lo tanto, dicha predicción consiste esencialmente en qué clientes tienen más probabilidades de cancelar una suscripción.

En el presente trabajo abordaremos esta temática, tomando como base un dataset extraído de IBM, que cuenta con información de una cartera de clientes perteneciente a una empresa de telecomunicaciones.

Objetivos del trabajo

El objetivo del trabajo es predecir qué clientes tienen altas probabilidades de abandonar algún tipo de sus servicios mediante la aplicación de un algoritmo de clasificación.

Para poder llegar a este objetivo realizaremos, en primer lugar, un análisis preliminar y exhaustivo de las distintas variables presentes en nuestro dataset para descubrir y entender su comportamiento (EDA). Esto nos permitirá tomar decisiones relacionadas a la limpieza de datos y selección de variables, para luego poder aplicar los modelos.

Mediante el EDA se intentará encontrar tendencias, patrones y relaciones entre variables; desarrollándose tres análisis distintos denominados univariado, bivariado y multivariado. El primero explora el comportamiento de cada variable de forma individual, mientras que el segundo analiza la relación entre dos variables. Por último, el multivariado es necesario cuando se deben

analizar más de dos variables en forma simultánea. Cabe aclarar que se incluyen tanto variables numéricas como categóricas.

Una vez concluído el análisis exploratorio de datos, utilizaremos distintos algoritmos de clasificación para predecir la variable *Churn*. Realizaremos la selección del algoritmo en base a distintas métricas de performance que nos indicarán qué modelo predice mejor el abandono de clientes.

Datos

A continuación se presenta un listado con las variables que analizaremos a lo largo del proyecto. Es importante remarcar que trabajaremos utilizando el lenguaje de programación Python. La forma en que se almacena la información en un DataFrame u objeto Python afecta a lo que podemos hacer con él y también a los resultados de los cálculos. Hay dos tipos principales de datos que estaremos explorando en este trabajo: numéricos y de texto. Los tipos de datos numéricos incluyen enteros (integer) y números de punto flotante (float). Un número de punto flotante tiene puntos decimales incluso si el valor del punto decimal es 0. Un integer nunca tendrá un punto decimal. Así que, si quisiéramos almacenar 1.13 como un entero de tipo integer se almacenará como 1. En Python se ve el tipo de dato Int64 que representa un entero de 64 bits. Por último, el tipo de datos de texto se conoce como secuencia de caracteres (string). En Pandas se los conoce como objetos (object). Las secuencias de caracteres pueden contener números y / o caracteres.

Object (28)

1. Customer_id: ID del Cliente.
2. Interaction: Identificaciones únicas relacionadas con transacciones de clientes, soporte técnico e inscripciones.
3. City: Ciudad de residencia del cliente que figura en el estado de cuenta.
4. State: Estado de residencia del cliente como se indica en el estado de cuenta.
5. County: Condado de residencia del cliente como se indica en el estado de cuenta.
6. Area: Tipo de área (rural, urbana, suburbana), según datos del censo
7. Timezone: Zona horaria de residencia del cliente basada en la información de registro del cliente.
8. Job: Ocupación del cliente como se indica en la información de registro

9 Education: Grado de educación más alto obtenido por el cliente según lo declarado en la información de registro.
10 Employment: Estado de empleo del cliente según lo declarado en la información de registro.
11 Marital: Estado civil del cliente según lo indicado en la información de registro.
12 Gender: Autopercepción del cliente como hombre, mujer o no binario.
13 Churn: : El cliente interrumpió el servicio en el último mes (sí/no).
14 Techie: El cliente se considera tecnológico (según el cuestionario del cliente cuando se inscribió en los servicios) (sí/no).
15 Contract: Plazo del contrato del cliente (mes a mes, un año, dos años).
16 Port_modem: El cliente tiene un módem portátil (sí/no).
17 Tablet: El cliente posee una tableta como iPad, Surface, etc. (sí/no).
18 InternetService: Proveedor de servicios de Internet del cliente (DSL, fibra óptica, Ninguno).
19 Phone: El cliente tiene servicio telefónico (sí/no).
20 Multiple: El cliente tiene varias líneas (sí/no).
21 OnlineSecurity: El cliente tiene un complemento de seguridad online (sí/no).
22 OnlineBackup: El cliente tiene un complemento de copia de seguridad online (sí/no).
23 DeviceProtection: El cliente tiene un complemento de protección de dispositivos (sí/no).
24 TechSupport: El cliente tiene un complemento de soporte técnico (sí/no).
25 StreamingTV: El cliente tiene servicio de transmisión de TV (sí/no).
26 StreamingMovies: El cliente tiene películas en streaming (sí/no).
27 PaperlessBilling: El cliente tiene facturación electrónica (sí/no).
28 PaymentMethod: Método de pago del cliente (cheque electrónico, cheque enviado por correo, banco (transferencia bancaria automática), tarjeta de crédito (automática)).

Int64 (14)

1 CaseOrder: Variable de marcador de posición para conservar el orden original del archivo de datos sin procesar.
2 Zip: Código Postal de residencia del cliente que figura en el estado de cuenta.
3 Population: Población dentro de un radio de una milla del cliente, según datos del censo.
4 Email: Cantidad de correos electrónicos enviados al cliente en el último año (marketing o correspondencia).
5 Contacts: Número de veces que el cliente se comunicó con el soporte técnico.

6. Yearly_equip_failure: Cantidad de veces que el equipo del cliente falló y tuvo que reiniciarse/reemplazarse en el último año.

Las siguientes variables, representan las respuestas a una encuesta de ocho preguntas, en las que se pide a los clientes que califiquen la importancia de varios factores en una escala del 1 al 8 (1 = más importante, 8 = menos importante):

7. Item1: Respuesta oportuna
8. Item2: Reparaciones oportunas
9. Item3: Reemplazos oportunos
10. Item4: Fiabilidad
11. Item5: Opciones
12. Item6: Respuesta respetuosa
13. Item7: Intercambio cortés
14. Item8: Evidencia de escucha activa

Float64 (9)

1. Lat: Coordenada de latitud de la residencia del cliente (GPS), que figura en el estado de cuenta.
2. Lng: Coordenada de longitud de la residencia del cliente (GPS), que figura en el estado de cuenta.
3. Children: Cantidad de niños en el hogar del cliente según lo informado en el registro.
4. Age: Edad del cliente según lo informado en el registro
5. Income: Ingreso anual del cliente según lo informado en el momento del registro.
6. Outage_sec_perweek: Promedio de segundos por semana de interrupciones del sistema en el vecindario del cliente.
7. Tenure: Número de meses que el cliente se ha quedado con el proveedor.
8. MonthlyCharge: Importe cobrado al cliente mensualmente. Este valor refleja un promedio por cliente.
9. Bandwidth_GB_Year: Cantidad promedio de datos utilizados (GB), en un año por el cliente.

Manipulación de datos

En esta sección analizaremos la existencia de valores nulos y repetidos en todas las columnas del dataset, así como el tipo de dato que representa cada una, y comenzamos a visibilizar probables transformaciones en nuestro conjunto de datos.

- churn_bool: Creamos una variable booleana de churn con 0 y 1 para poder compararla contra variables numéricas
- total_encuesta: Variable que unifica los valores obtenidos en la encuesta (item 1 al 8)
- Cambiamos los nombres de las últimas columnas reemplazando itemx por el nombre del item:
  'item1': 'timely_response',
  'item2': 'timely_fixes',
  'item3': 'timely_replacements',
  'item4': 'reliability',
  'item5': 'options',
  'item6': 'respectful_response',
  'item7': 'courteous_exchange',
  'item8': 'active_listening'
- Creamos nuevas variables que cuentan la cantidad de cada tipo de servicios y el total de servicios que los clientes contratan:
  - ☐ q_online_serv: cuenta si el cliente tiene contratado internet service, online backup, onlinesecurity y/o tech support. Coloca un 1 por cada servicio contratado siendo el valor máximo de esta categoría = 4.
  - ☐ q_phone_serv: cuenta si el cliente tiene contratado phone y/o device protection. Coloca un 1 por cada servicio contratado siendo el valor máximo de esta categoría = 2.
  - ☐ q_streaming: cuenta si el cliente tiene contratado streaming tv y/o streaming movies. Coloca un 1 por cada servicio contratado siendo el valor máximo de esta categoría = 2.
  - ☐ q_total_serv: realiza una suma de las tres categorías nombradas anteriormente siendo el valor máximo de esta variable = 8.
- zip_zone: creación de variable geográfica agrupando los zip por su primer dígito. Este número indica una de las 9 zonas geográficas generales de EEUU.

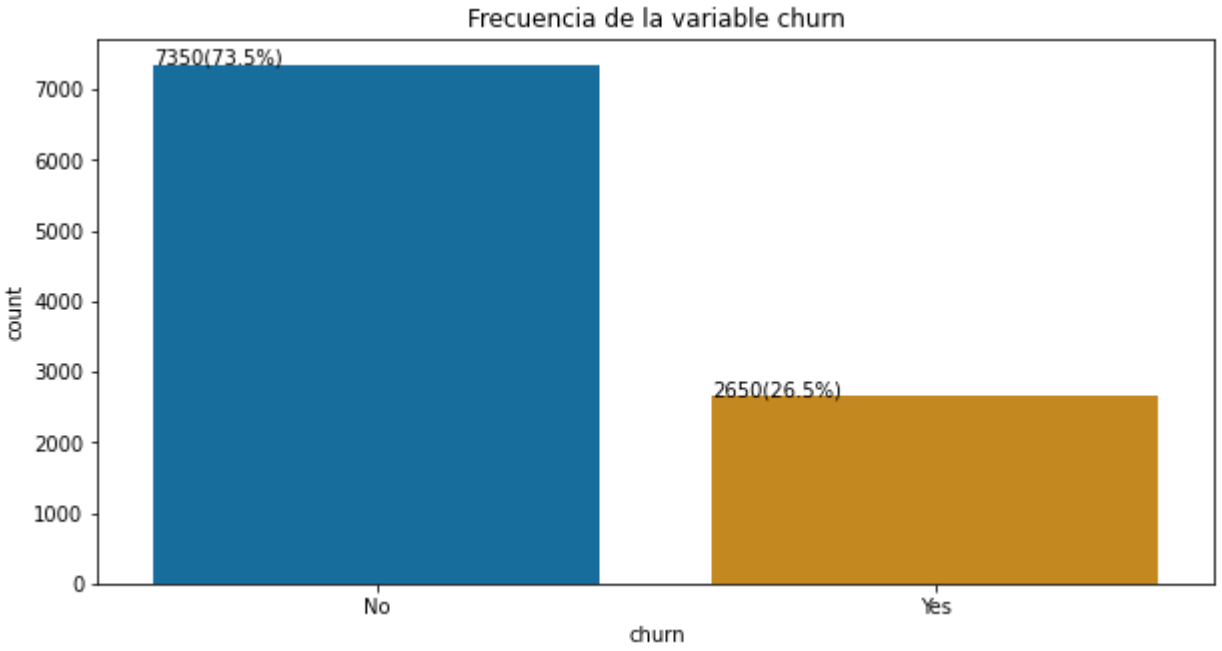
- Creamos variables string que se correspondan con las numéricas categóricas para poder realizar análisis de variables categóricas:
    - ☐ children_cat, age_cat, zip_catpop_cat, email_cat, contacts_cat, failure_cat, timely_response_cat, timely_fixes_cat, timely_replacements_cat, reliability_cat options_cat, respectful_response_cat, courteous_exchange_cat, active_listening_cat, total_encuesta_cat, q_online_serv_cat, q_phone_serv_cat, q_total_serv_cat, q_streaming_cat, zip_zone_cat.

Hallazgos encontrados por el EDA

Análisis Univariado

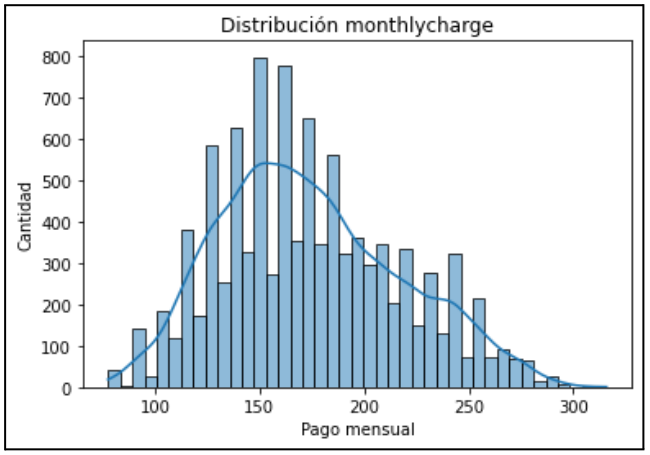
Como mencionamos anteriormente este análisis consiste en explorar cada variable individualmente. Un punto importante es conocer la distribución de las variables, y en caso no presentar una distribución normal, aplicarles una transformación para lograr su correcta manipulación. Como regla general, se trata de escoger una transformación que conduzca a una distribución simétrica, y más cercana a la distribución normal. Este punto es importante para mejorar el rendimiento de algunos algoritmos.

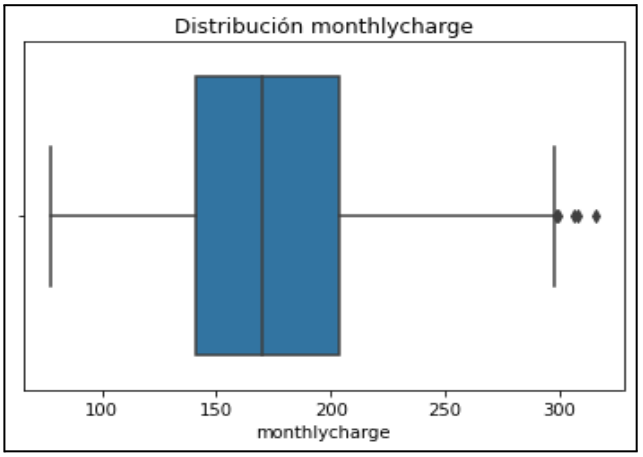
Variable Target: Churn



Hay un 26,5% de clientes que abandonan la empresa, mientras que un 73,5% continúan con el servicio, es decir la clase no-churn tiene 41 puntos porcentuales más que churn. Esta diferencia es significativa, por lo que nos encontramos ante un conjunto de datos desbalanceado. Es fundamental tener en cuenta esto a la hora de evaluar los algoritmos, ya que el modelo podría aprender a detectar la clase predominante, mientras que no es eficiente al predecir la otra clase.

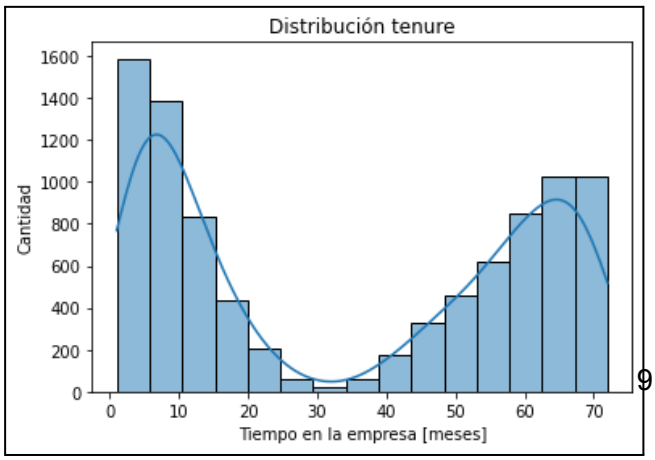
Monthly Charge

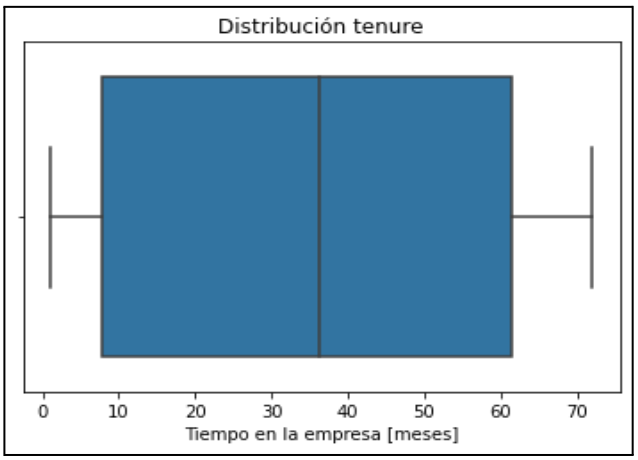




Se observa una variable relativamente normalizada, es decir, se aproxima a una normal, por lo que no hará falta aplicarle una transformación. Al ser la distribución normal una distribución simétrica, el valor de la moda, media y mediana, deben coincidir. En otras palabras, los valores más frecuentes deben estar alrededor de la media, y esto no sucede en su totalidad. Se observa en el histograma que la mayor cantidad de datos se encuentra concentrada en el orden de los 150 dólares mensuales, es decir, estamos en presencia de una distribución asimétrica hacia la izquierda y esto implica que sus sesgo es negativo. Además, los clientes que más aportan a los beneficios de la empresa pagan aproximadamente 300 dólares mensuales, mientras que los que menos aportan pagan aproximadamente 80 dólares. Por último, cabe recalcar que existen valores atípicos en el orden de los 300 dólares mensuales.

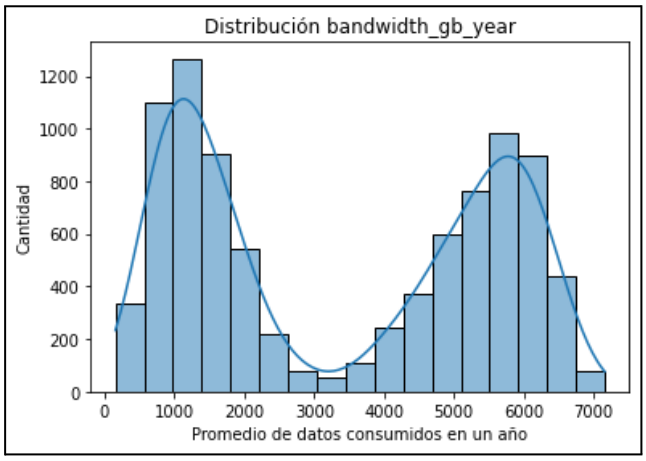
Tenure

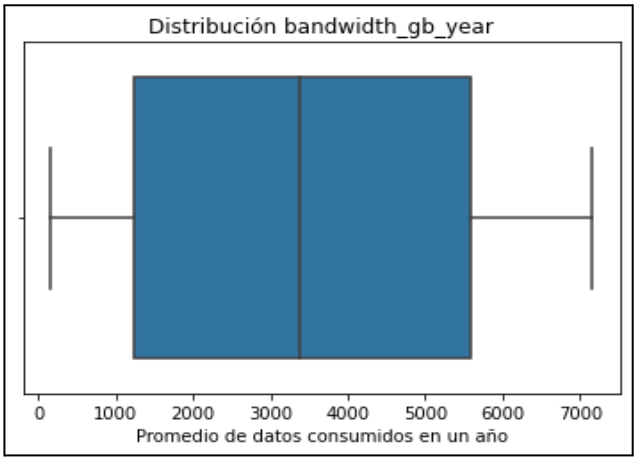




El histograma permite ver dos picos que representan distintas modas, es decir, los dos valores más comunes en nuestro conjunto de datos. Esto quiere decir que la variable sigue una distribución bimodal, lo que implica que se distinguen dos tipos de clientes principales: quienes están hace más de 40 meses y quienes no superan los 20 meses de antigüedad. Ahora bien, a la hora de analizar el porqué de esta distribución, podemos pensar en que la gente decide rápidamente si quiere conservar el servicio o rotar hacia la competencia. Es sabido que las empresas de telecomunicaciones compiten agresivamente por los clientes constantemente. Una mejor manera de analizar e interpretar las distribuciones bimodales es simplemente dividir los datos en dos grupos separados, para luego analizar el centro y la dispersión de cada grupo, pero no es algo que nos interese en este proyecto.

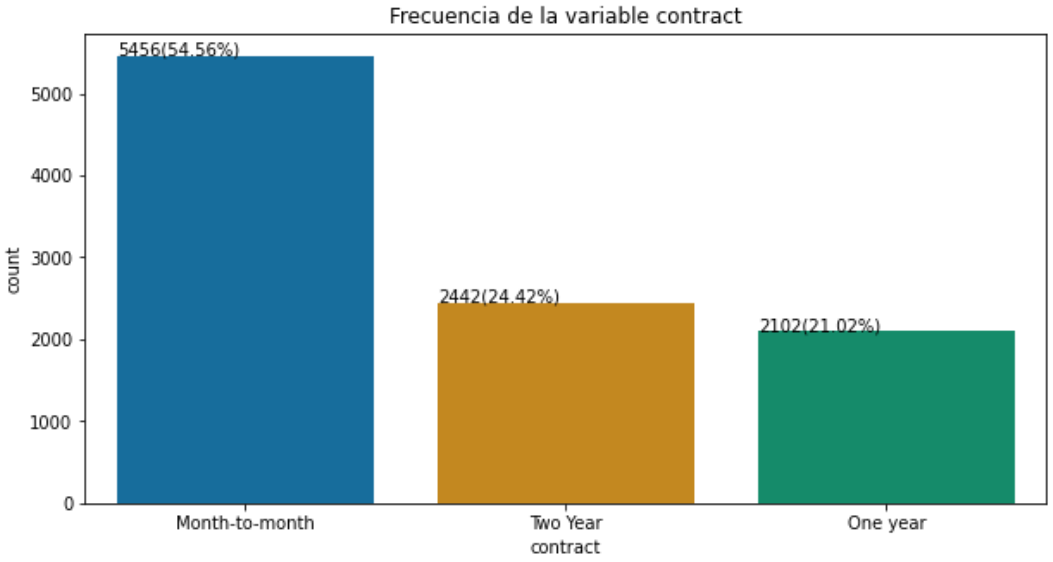
Bandwidth_Gb_Year





Esta variable también presenta una distribución bimodal, lo que implica que se distinguen dos grupos de clientes principales: quienes consumen anualmente alrededor de 5500 GB, y quienes consumen 1000 GB aproximadamente.

Contracts



Hay tres categorías de contratos: month to month, one year y two years. Hay un 54.56% de los clientes que contratan el servicio mensualmente, otro 21% que prefiere renovarlo anualmente, y un 24.42% que renueva cada dos años.

Análisis Bivariado

En términos generales, el análisis bivariado es la investigación de la relación entre dos conjuntos de datos, como pares de observaciones tomadas de una misma muestra o individuo. Para evaluar la relación entre variables numéricas se utilizó la correlación de Spearman considerando aquellas variables con relaciones +/- 0.5

A continuación presentaremos los principales hallazgos respecto de relaciones con la variable objetivo: **Churn**

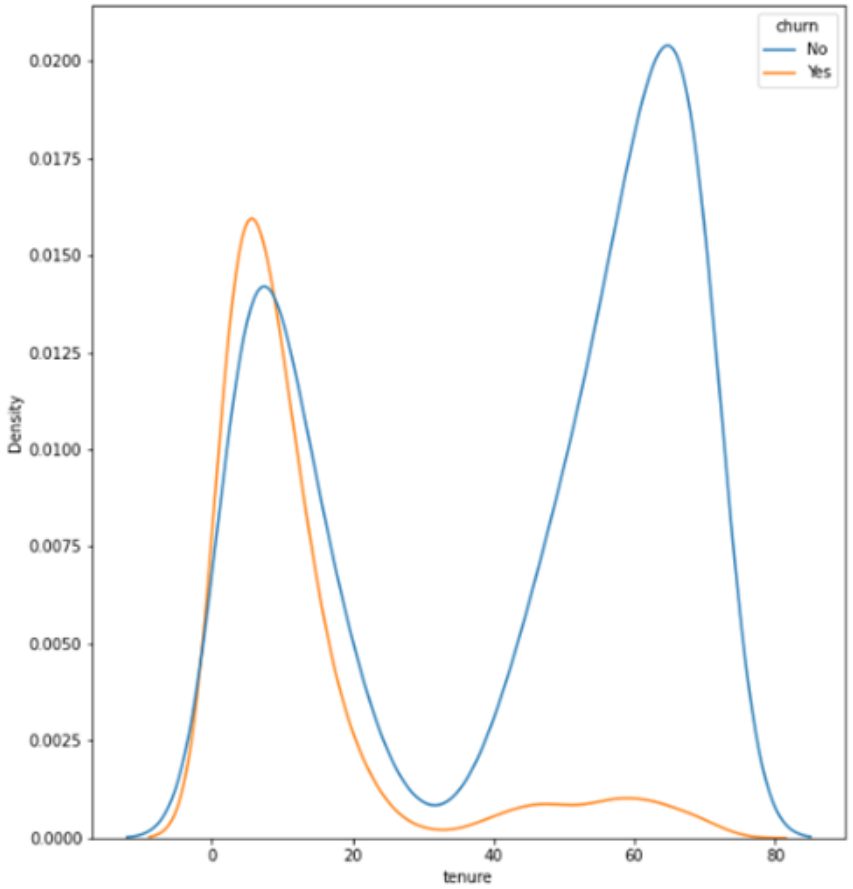
Principales relaciones entre variable Churn y otras variables

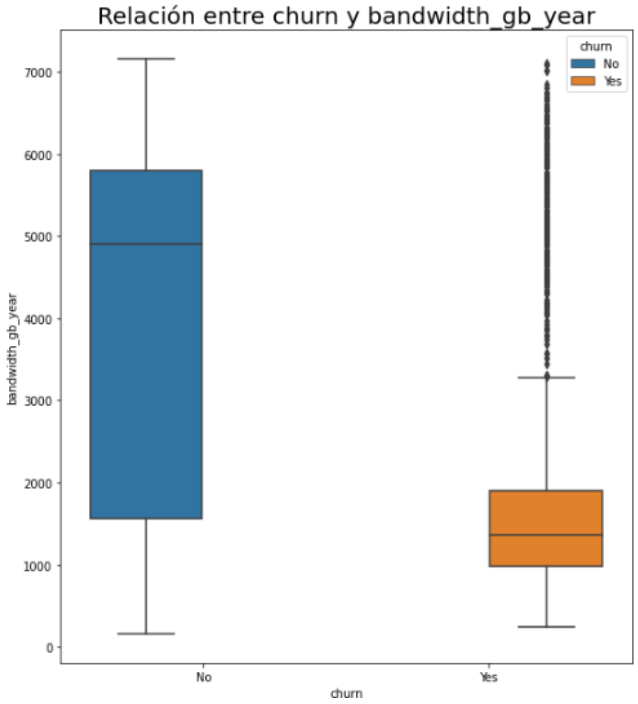
Para empezar con este análisis, creamos la variable churn_bool para poder tomar a churn cómo numérica, reemplazando los “Yes” por 1 y los “No” por 0. Luego, analizamos las relaciones entre churn_bool y todas las variables numéricas, quedándonos con aquellas que presentan comportamientos marcadamente diferentes contra no churn. También analizamos la relación entre churn y las variables categóricas y numéricas categóricas.

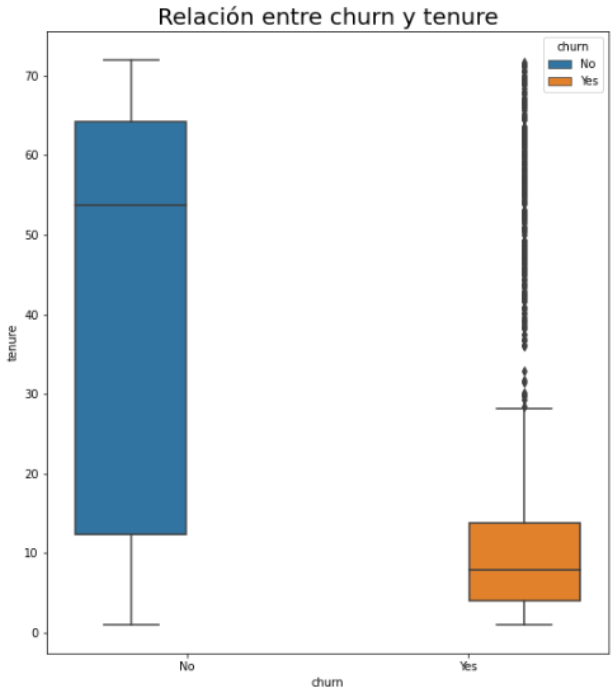
Por otro lado, analizamos la relación entre todas las variables del dataset, lo que nos permitió observar las diferentes relaciones entre las variables independientes.

Churn Vs. Tenure y Bandwidth GB Year

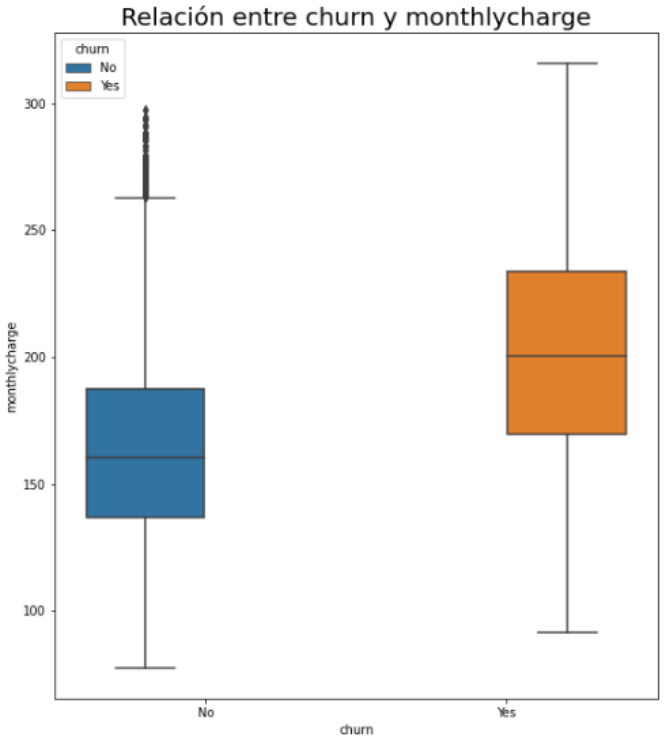
La mayoría de los churn se concentran en personas con menos de 30 meses de antigüedad. Hay outliers en los clientes más antiguos, pero en general estos se quedan. La mediana de los churn y no churn es muy distinta, al igual que su dispersión. Churn tiene un sesgo a la derecha concentrando valores en antigüedad baja pero con una cola importante en valores de tenure alta. La correlación entre churn y Tenure es de -0.46, siendo la variable mayormente correlacionada con el target.
Bandwidth tiene una correlación positiva casi perfecta con Tenure, por lo que su relación con Churn es muy similar. La correlación entre bandwith y churn es de -0.39.







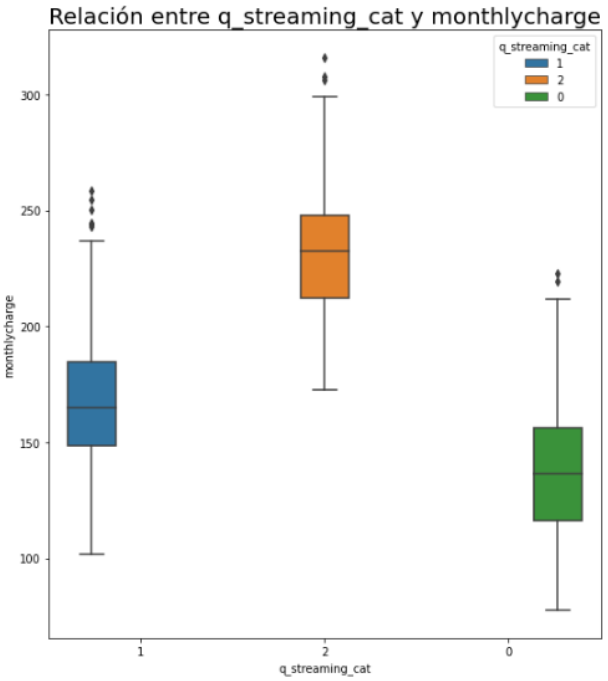
Churn Vs. Monthly Charge y servicios



Podemos ver que quienes abandonan tienen una mediana más elevada y su rango intercuartílico también se ubica en valores mayores que quienes no abandonan, pero su dispersión es mayor,

teniendo una distribución más uniforme. Su correlación es levemente positiva (0.36). Hay outliers con altos pagos mensuales en los que se quedan.

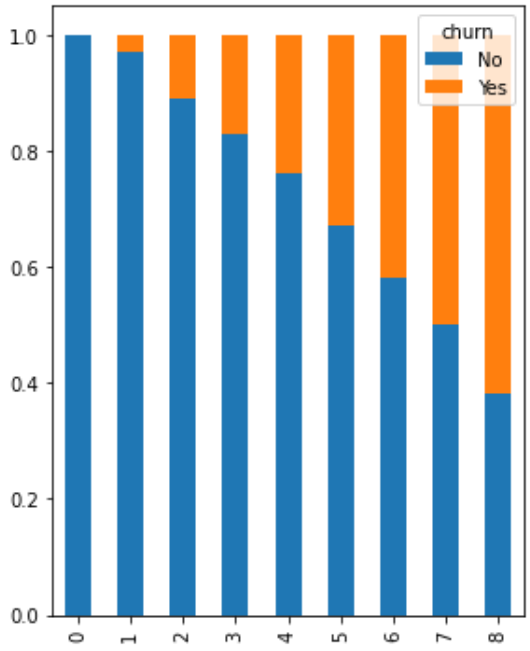
Además encontramos relaciones significativas entre churn y las variables correlacionadas con Monthly Charge: q_other_services (0.69), q_streming (0.75) y q_total_services(0.69). Todas ellas variables numéricas categóricas que indican cantidad de servicios y son calculadas en base a las variables de servicios originales del dataset. Es importante destacar que las correlaciones entre los pagos que realizan los clientes y los servicios se dan en dos categorías: otros servicios y servicios de streaming.



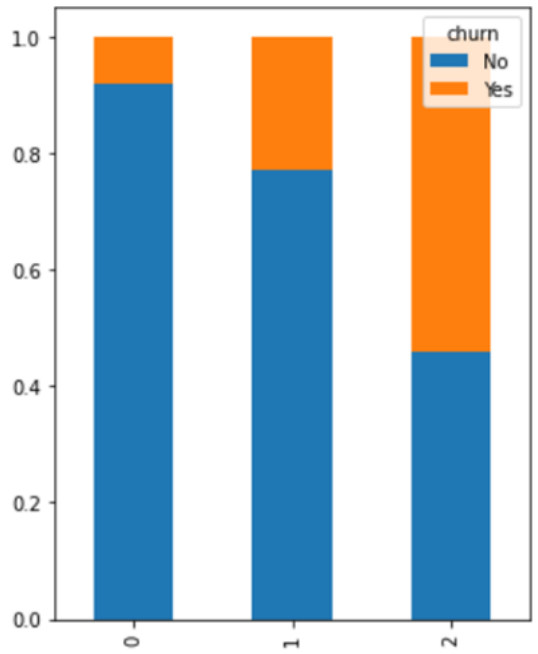
Quienes contratan los 2 servicios de streaming tienen una mediana mensual de USD 240. La mediana del abono es similar para quienes contratan streaming TV y Movies. La mediana en ambos casos es de USD 200, mientras que quienes no contratan alguno de estos servicios tienen una mediana de USD 150.

En cuanto a la relación de churn con estas variables podemos observar que a medida que se contratan más servicios aumenta la proporción de churn. Si ahondamos en los tipos de servicios este comportamiento se observa en los servicios de streaming, siendo constante la proporción de clientes que abandonan para las distintas categorías de los demás tipos de servicios.

Churn vs Cantidad de servicios contratados

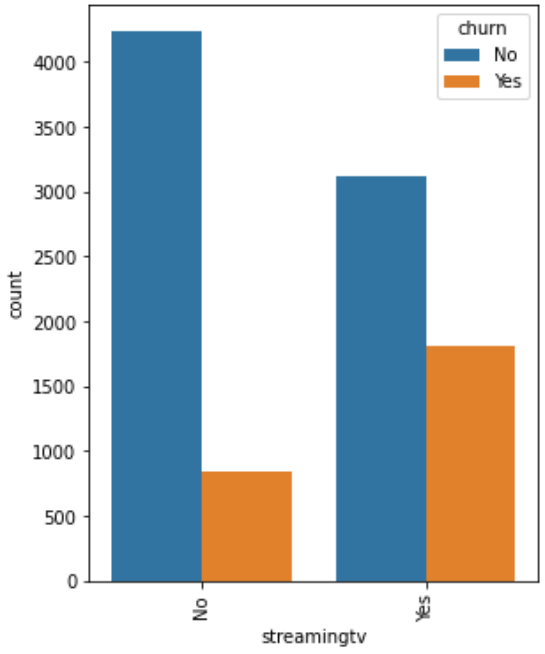


Churn vs Cantidad de servicios de streaming

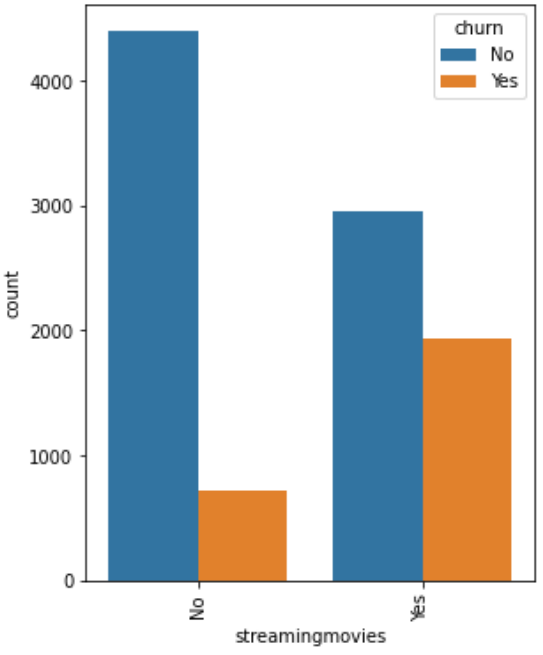


Las personas que contratan los dos servicios de streaming de la empresa, tienen una tasa de abandono del 54%.

Churn Vs. Streaming Tv



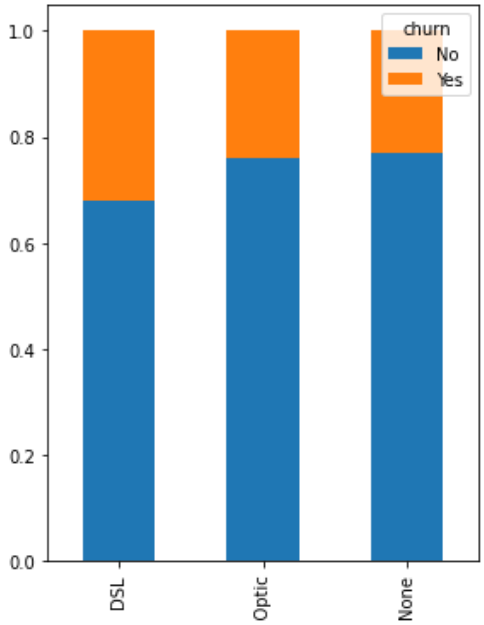
Churn Vs. Streaming Movies



Dada la alta tasa de abandono según servicio de streaming, decidimos profundizar en cada uno de ellos. En ambos casos observamos una alta tasa de abandono, siendo mayor la de streaming Movies. El 37% de quienes contratan el primer servicio abandonan, mientras que los churn de quienes contratan el segundo servicio es el 40%.

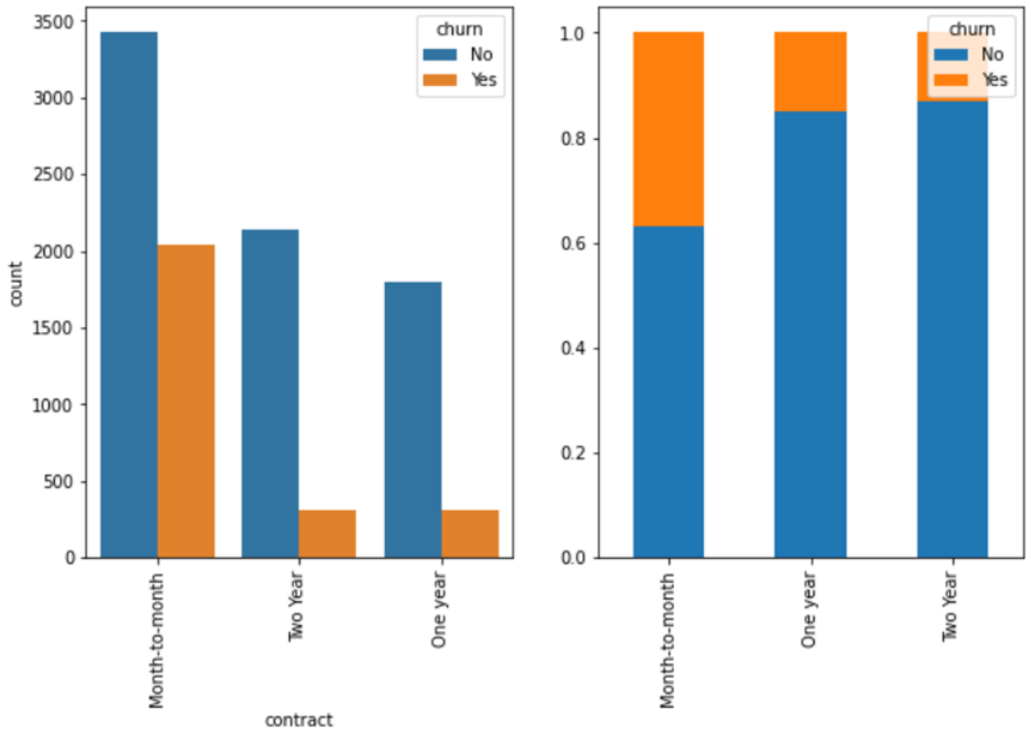
Por otro lado, cuando los clientes no contratan estos servicios el porcentaje de churn baja drásticamente siendo de 16% en el primer caso y 14% en el segundo.

Churn Vs. Internet Service



El tipo de conexión a internet afecta en la churn siendo la tasa de abandono de un 32% en las personas que contratan por DLS mientras que internet por Fibra Óptica tiene una tasa de abandono del 24%. Quienes no contratan internet, tienen una tasa de abandono del 25%.

Churn Vs. Contract

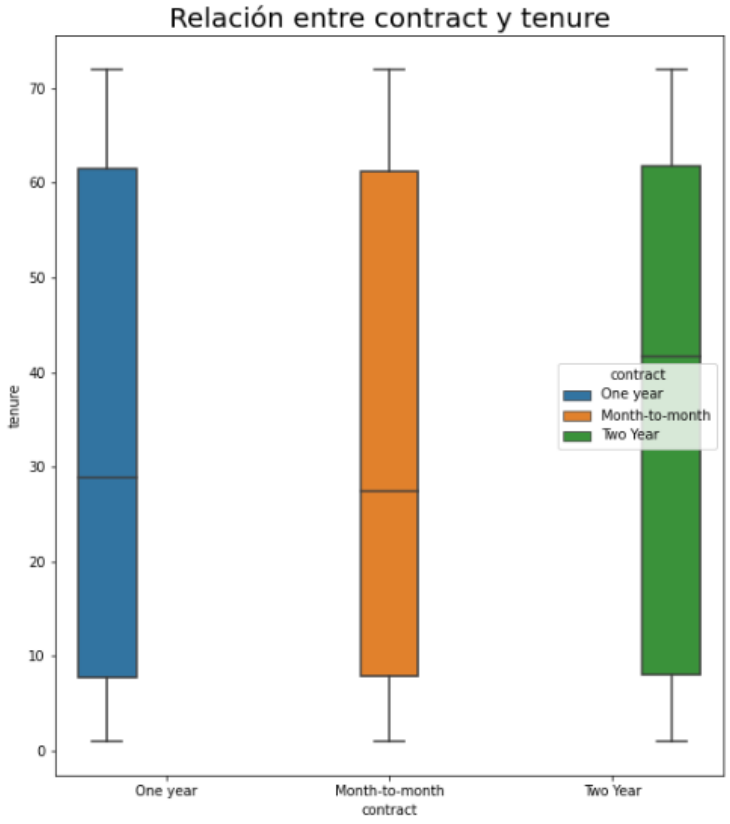


Month-to-month contract presenta un porcentaje más elevado de churn, superando ampliamente a quienes contrataron servicios de uno y dos años. Esto puede deberse a que los contratos de uno y

dos años son pagados por adelantado, por lo que la persona decidirá abandonar o no la empresa después de pasado el tiempo del contrato.

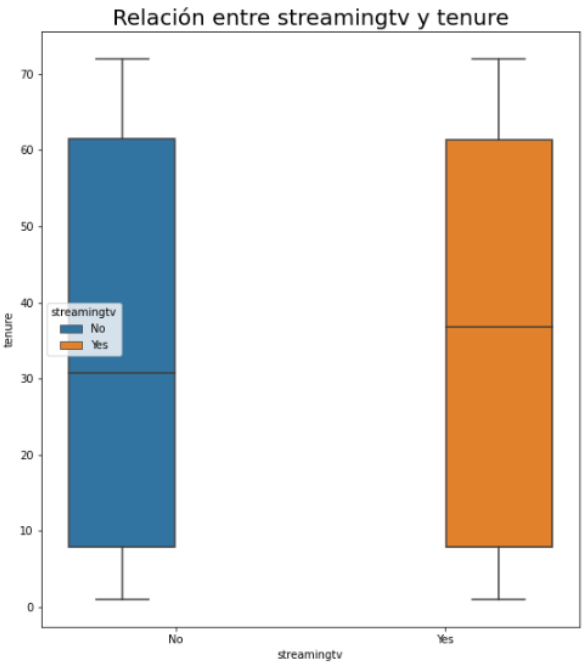
Otras relaciones significativas entre variables independientes

Tenure Vs. Contract



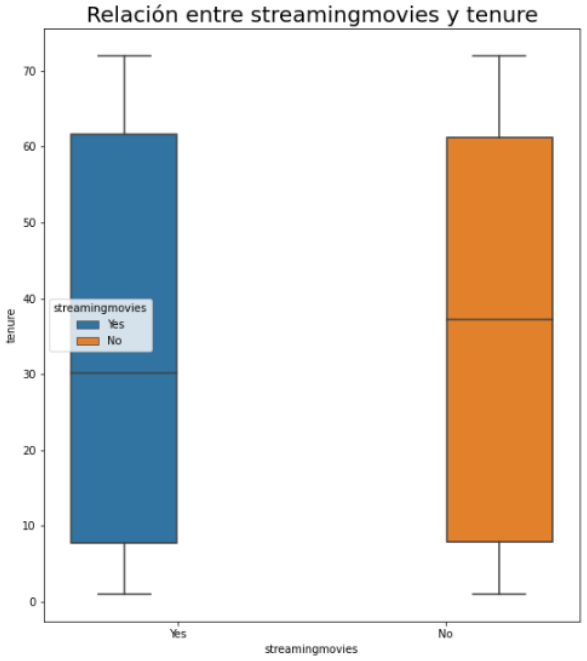
Month to month y one year tienen una mediana parecida, alrededor de 30. Quienes tiene contrato bianual, tienen una mediana de tenure mayor.

Tenure Vs. Streaming TV



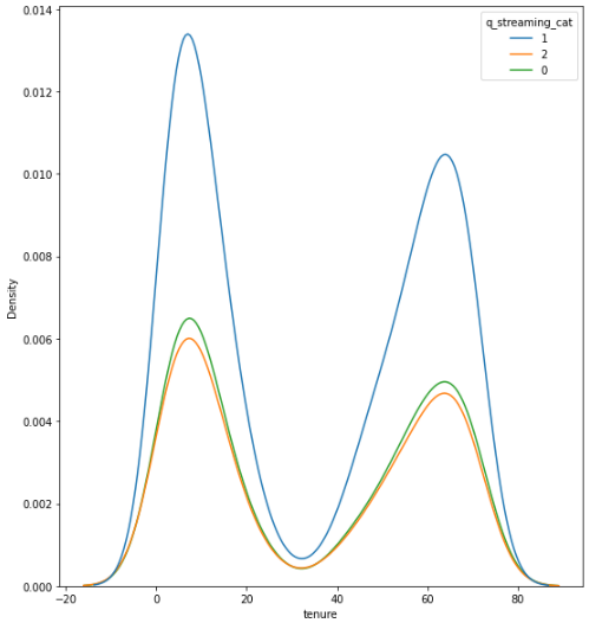
Quienes contratan streaming TV tienen un tenure mayor que quienes no contratan.

Tenure Vs. Streaming Movies



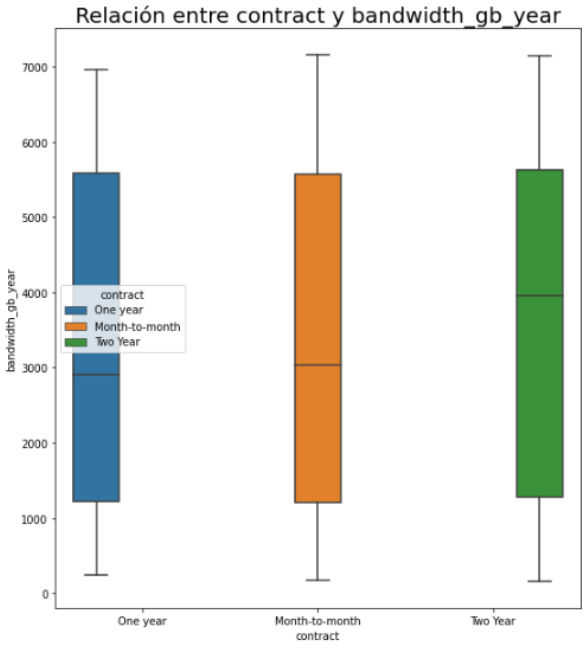
Quienes contratan streaming movies tienen un tenure menor que quienes no contratan este servicio.

Tenure Vs. Streaming



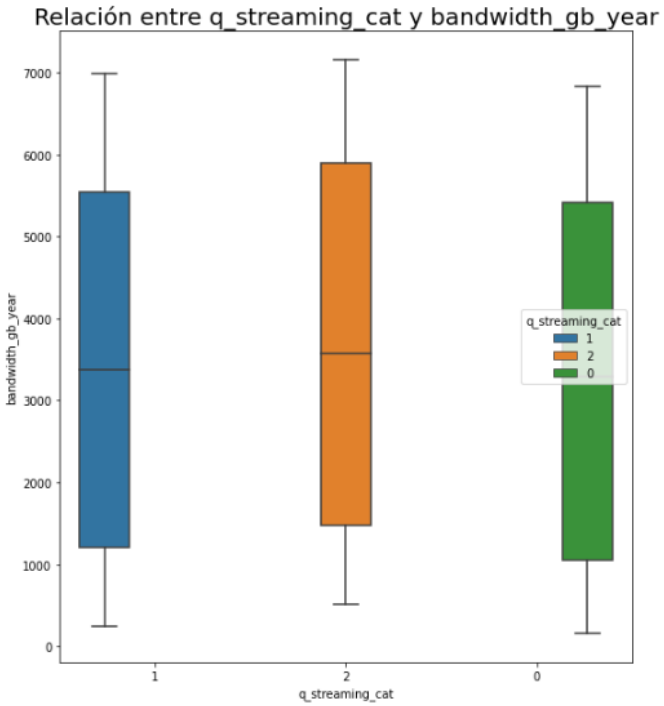
Quienes contratan un servicio de streaming tienen un tenure mayor que quienes contratan 2 servicios o ninguno. Es decir, tenemos dos tipos de clientes según consumo de streaming apenas contratan con la empresa: los que contratan y prueban todos los servicios y los que no contratan ninguno.

Bandwidth GB Year Vs. Contract



Quienes contratan por dos años tienen una mayor mediana.

Bandwidth GB Year Vs. Streaming



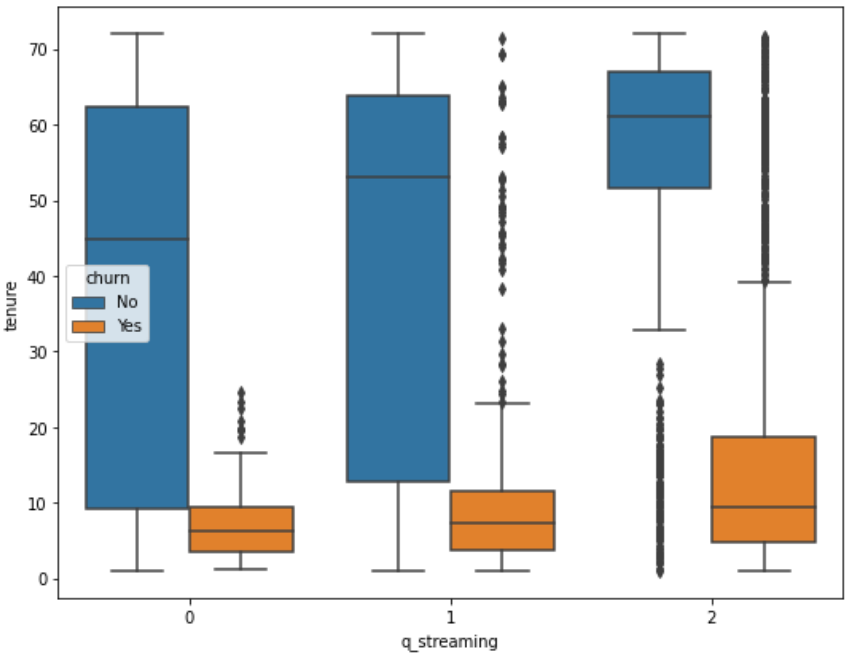
Quienes contratan streaming tienen mayor mediana, lo cual es esperado porque son servicios relacionados a videos con alto tráfico de datos.

Análisis Multivariado

En esta etapa del EDA se analizan 3 ó más variables al mismo tiempo para comprender en profundidad su comportamiento.

Buscaremos responder las preguntas que surgieron tanto en el análisis univariado como en el bivariado siguiendo de esta manera las pistas que nos va brindando la exploración del dataset.

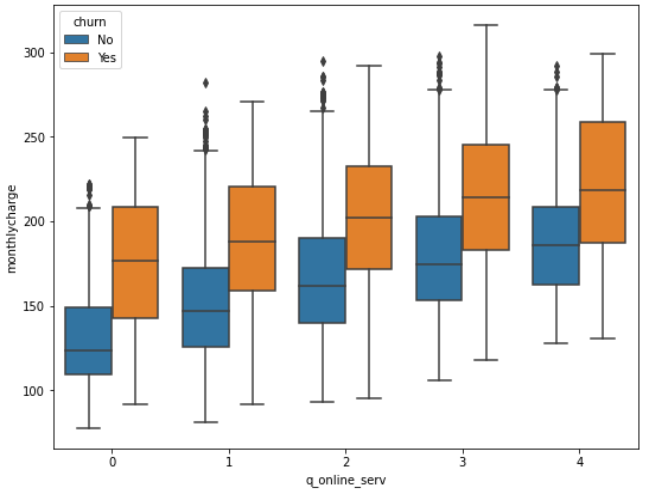
Streaming, tenure y churn:

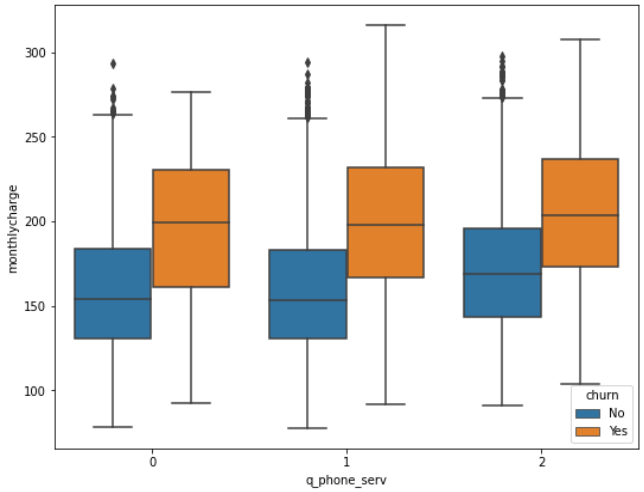


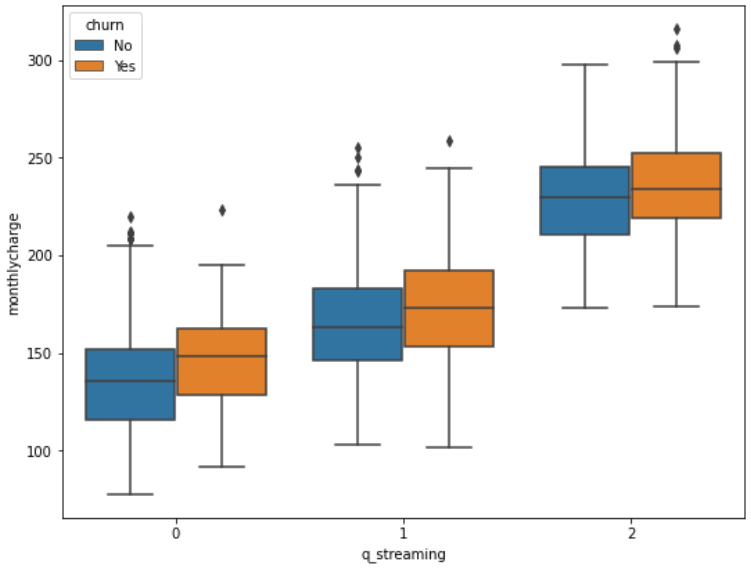
Es notable la concentración de datos en los casos que contratan dos servicios de streaming: se concentran en niveles bajos de tenure y en niveles altos de tenure. Aquellos que tienen muchos meses de antigüedad no abandonan al contratar streaming pero si lo hacen quienes tienen pocos meses de antigüedad. Una posibilidad es que se encarezca mucho el servicio y, al no estar fidelizados por antigüedad, deciden cambiar de empresa.

Se observan también muchos outliers: clientes con tenure por encima de los veinte meses y a pesar de la antigüedad, contratan uno o dos servicios de streaming y abandonan la empresa. En los casos de no churn, se observan muchos outliers en quienes contratan dos servicios de streaming y tienen pocos meses de antigüedad.

Services vs monthly charge y churn:





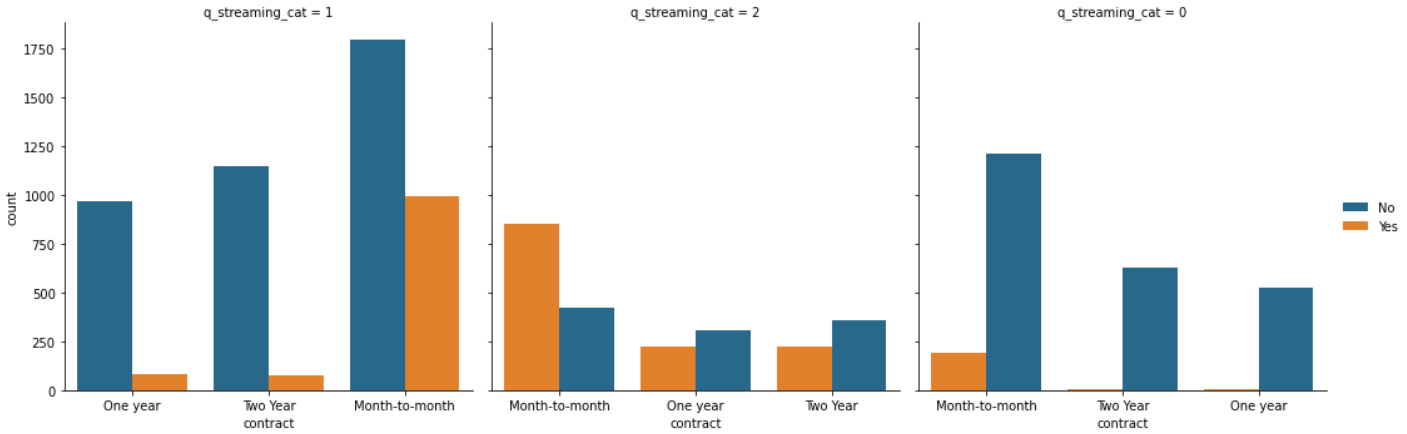


Del análisis bivariado surge la pregunta de qué tipo de servicio contratan los outliers que pagan mucho y no abandonan. Al observar estos gráficos podemos afirmar que tenemos outliers con pagos mensuales altos en todos los servicios. Solo no hay outliers en quienes consumen los dos servicios de streaming.

Al comparar estos tres gráficos que analizan como variable dependiente la categoría monthly_charge se observa una disparidad en la mediana: quienes abandonan tienen mediana considerablemente mayor en el abono.La diferencia en el abono en usd entre quienes abandonan y los que no es de 50usd aprox y se mantiene para los distintos tipos y cantidad de servicios contratados. Esta mediana superior en todos los casos confirma que para quienes contratan estos servicios si es determinante en aumento en el abono para abandonar.

Este gráfico nos permite observar que la diferencia en el costo del abono entre los que dejan y los que no dejan es de 10usd aprox en el caso de los servicios de streaming. Esto indicaría que en el caso de los servicios de streaming no es el costo del abono lo que hace que abandonen la empresa.

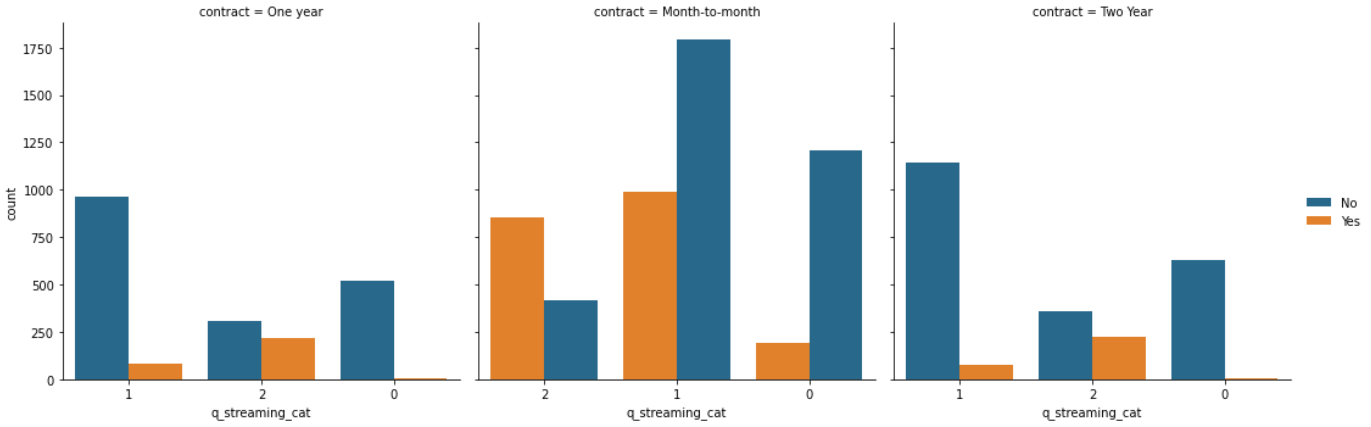
Servicio de streaming, contract y churn:



En todos los casos, la mayoría de los churn son los contratos mensuales. Quienes contratan 2 servicios de streaming, independientemente del tipo de contrato, son los que más abandonan.

En el caso de quienes contratan 1 servicio de streaming destacan los contratos month to month con una proporción mayor de churn, cuando se lo compara con los otros tipos de contratos en dicha categoría.

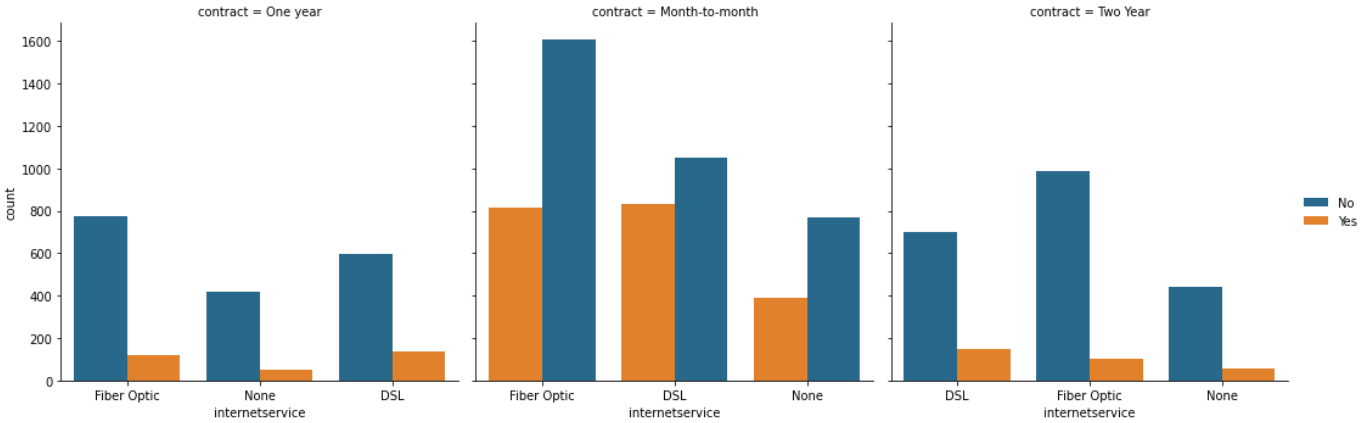
Quienes no contratan un servicio de streaming y tienen contrato anual o bianual, no abandonan la empresa en general.



En este gráfico hemos alternado la posición de contract y streaming lo cual nos permite observar con facilidad la concentración de abandono en quienes tienen contrato mensual.

También se observa que en todos los tipos de contrato cuando se contratan 2 servicios hay mayor abandono. Quienes contratan por uno o dos años y no tiene servicios de streaming tiene muy baja churn. Casi el total de la muestra continua en la empresa. Lo mismo para quienes contratan un solo servicio de streaming.Tal como vimos en el análisis bivariado, quienes tienen contrato anual y bianual abandonan menos la empresa.

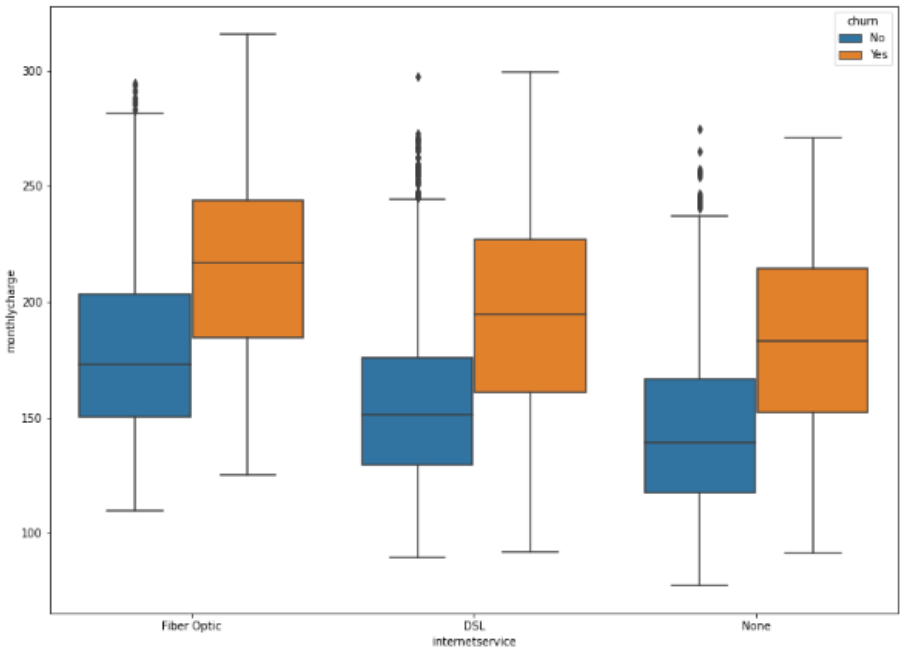
Servicio de internet, contract y churn:



Este gráfico analiza el comportamiento de los clientes respecto al tipo de servicio de internet contratado. Resaltamos la mayor cantidad de clientes con contrato mensual por sobre los otros contratos. Estas personas contratan en su mayoría DSL service, de ellos, la mitad abandona. Le siguen en cantidad de personas quienes no contratan internet. De estas personas, el 75%abandona la compañía. Son menos las personas con contrato mensual que contratan fibra óptica. De ellos, la mitad abandona.

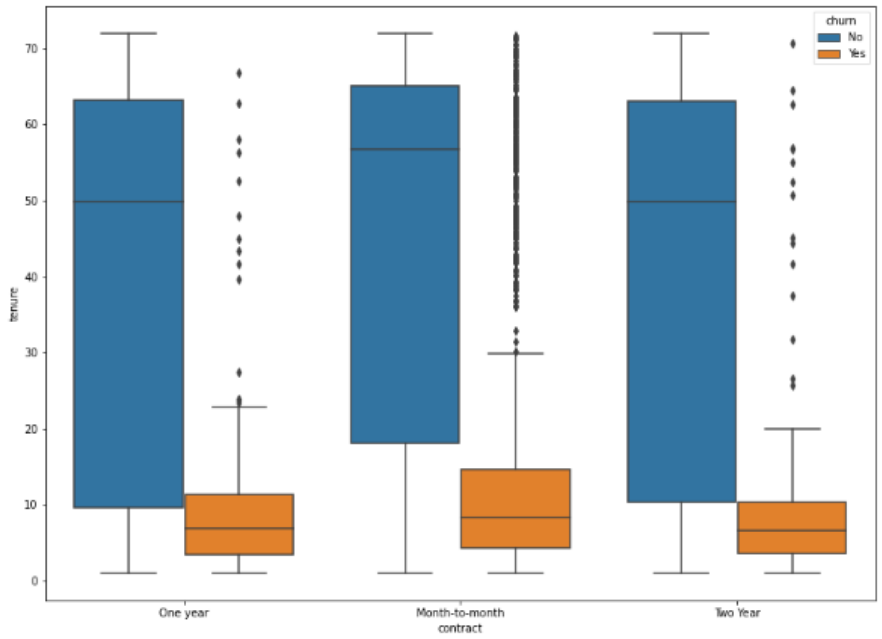
Los contratos anuales y bianules tiene muy baja tasa de abandono independientemente del tipo de internet.

Servicio de internet, monthly charge y churn:



Internet por fibra óptica tiene una mediana de abano superior a DSL. Quienes abandonan la empresa tienen en todos los casos una mediana de abono superior que quienes se quedan.

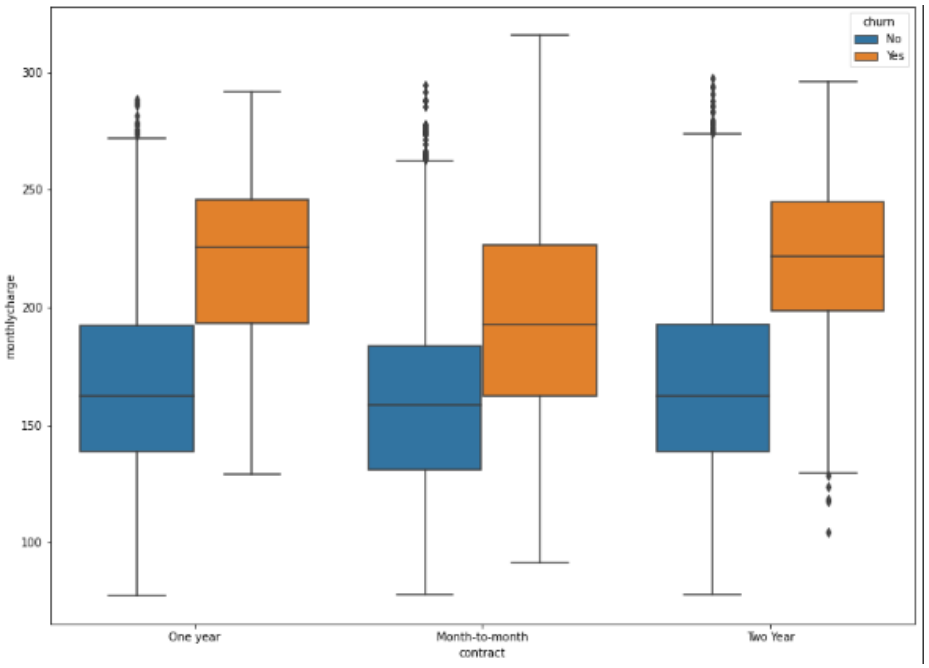
Contract, tenure y churn:



En este gráfico se observa una mediana de tenure mayor cuando el contrato es mensual tanto en quienes abandonan las empresas como los que se quedan. El hecho de tener un contrato por plazo más corto no indica que vayan a dejar antes la empresa.

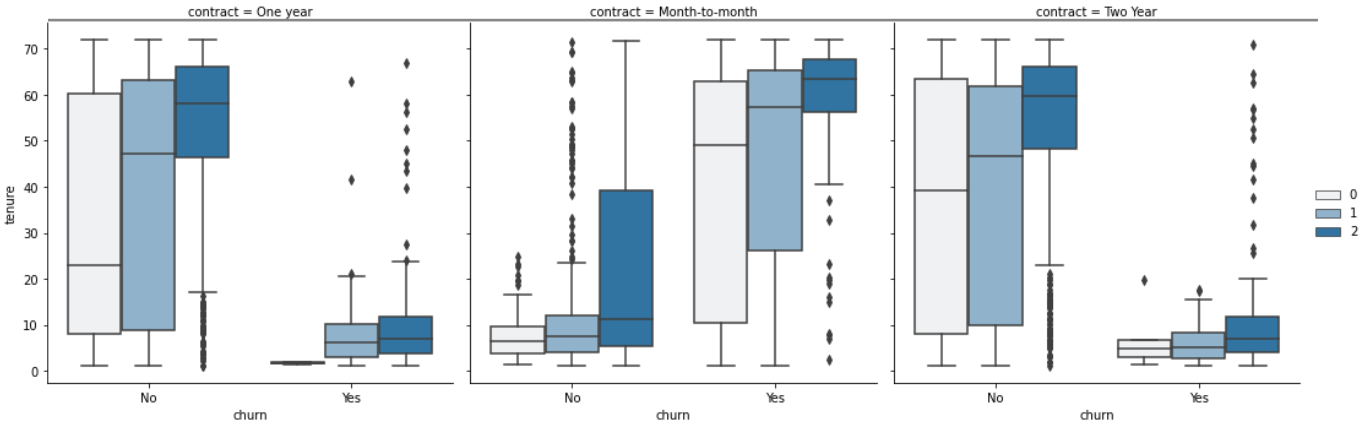
Llama la atención de este gráfico la concentración de los datos en tenures bajas en todos los casos de churn independientemente del tipo de contrato.

Contract, monthly charge y churn



Este gráfico demuestra que quienes abandonan la empresa siempre tienen medianas de abono mayor que quienes continúan. Observamos también que quienes tienen contrato mensual tienen una mediana menor que quienes tienen contrato anual y bianual. Por tanto, pagan menos.

Servicios de Streaming, tenure, contract y churn



Este gráfico nos permite observar el comportamiento de antigüedad en las personas que abandonan la empresa respecto al tipo de contrato que tienen y la contratación de streaming. Observamos que quienes tienen contrato mensual abandonan en tenures mucho más altas que quienes tienen contrato anual y bianual independientemente de si contratan streaming o no. Se observa que a medida que incrementa la cantidad de servicios, incrementa la mediana de la tenure.

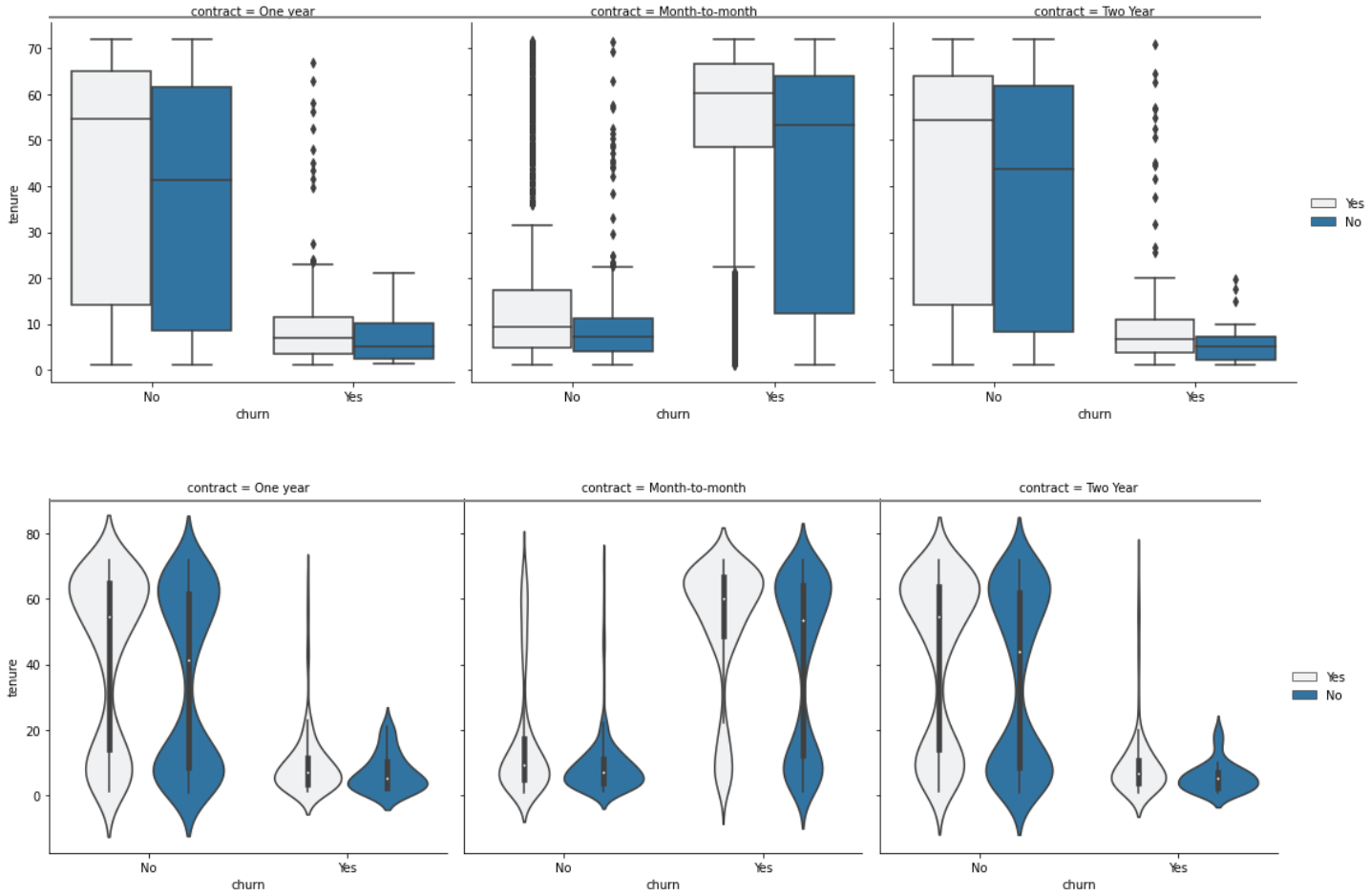
Los abandonos en contratos mensuales tienen una mediana de antigüedad de 60 meses mientras que en contratos anuales y bianuales los abandonos se producen con una mediana de 9 meses aprox. Abandona antes que termine el contrato. No parece verse muy afectado por la cantidad de servicios que contratan. Tanto en anual como bi anual, las personas que no abandonan tienen una tenure alta.

En cambio, en los contratos mensuales que no abandonan, la mediana de la tenure es considerablemente baja.

Cabe destacar el comportamiento de quienes tienen contrato por dos años y contratan 2 servicios de streaming, abandonan con tenures más altos que quienes nos abandonan con menos servicios. Quizá sí sea una variable de retención la cantidad de servicios contratados en este caso.

En los contratos anuales y bi anuales se observan datos muy dispersos respecto a la tenure en aquellos que no abandonan. Quienes si abandonan se concentran en tenures bajas, no hay dispersión de datos.

Streaming movies, tenure, contract y churn



Sabiendo que el 40% de las personas que contratan streaming movies abandonan la empresa mientras que quienes no contratan abandonan solo el 14%, hemos decidido profundizar en esta variable.

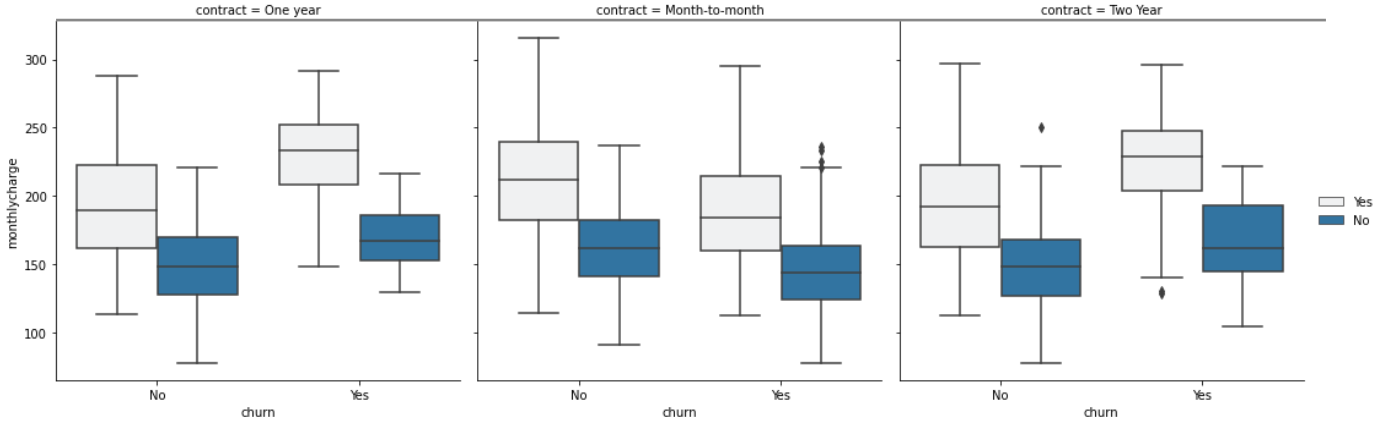
Quienes si contratan y nos abandonan tienen en su mayoría un contrato mensual. Estas personas nos abandonan con tenure alta. Existe un porcentaje de personas, consideradas outliers por el boxplot, que contratan mensual y abandona con tenures bajas.

En los contratos anuales y bi anuales, la mayor cantidad de personas que abandonan son las que contratan servicio de streaming movie. Ambos grupos abandonan generalmente antes de llegar al año de antigüedad.

Se observa gran dispersión de datos en las personas con contrato mensual que abandonan la empresa sin haber contratado servicio de streaming: tenemos personas que se van con tenures bajas y personas que se van con tenures altas.

En todos los tipos de contrato tanto como si son churn o no, quienes contratan servicio de streaming movies tienen una mediana de tenure mayor.

Streaming movies, monthly charge, contract y churn



Respecto al abono mensual,los que contratan streaming movie siempre tienen medianas mayores. En los contratos anuales y bianuales se cumple que abandonan con medianas de abono más altas que quienes no abandonan. El abono parecería ser influyente en estos tipos de contrato.
En los contratos mensuales, el abandono se produce en medianas más bajas que quienes permanecen. Por tanto, quienes no son churn, tienen un abono mensual mayor que los que abandonan, no parecía afectarles el pago en el caso de contratos mensuales.

Conclusion

Las variables que más se relacionan con el abandono o no la empresa son:

- Tenure: la mayor tasa de abandono se da en tenures bajas, con una mediana de 10 meses.
- Monthly charge: Quienes abandonan la empresa tienen una mediana de abono mensual mayor que quienes no abandonan. Solo para el caso de los contratos mensuales, el abandono se produce en medianas más bajas que quienes permanecen. Por tanto, quienes no abandonan, tienen un abono mensual mayor que los que abandonan, no parecía afectarles el pago en el caso de contratos mensuales.
- Contract: la mayoría de nuestros clientes tiene contrato mensual. Este tipo de contrato tiene una tasa de abandono cercana al 40%, mientras que los contratos anuales y bi anuales tienen una tasa de abandono del 15%. Los abandonos en contratos mensuales tienen una mediana de antigüedad de 60 meses mientras que en contratos anuales y bianuales los abandonos se producen con una mediana de 9 meses aprox. Abandona antes que termine el contrato.
- Streaming Movie: Las personas que contratan este servicio tienen una tasa de abandono del 40%. Estas personas tienen en su mayoría un contrato mensual y nos abandonan con

tenure alta. En los contratos anuales y bi anuales, la mayor cantidad de personas que abandonan son las que no contrataron servicio de streaming movie.

- Internet Service: quienes contratan DLS service tiene una tasa de abandono 10% mayor que quienes contratan fibra óptica o quienes no contratan internet. El servicio de DLS es más barato que la fibra óptica.

Desarrollo del Modelo

Tal como describimos en el objetivo del trabajo, el mayor desafío que enfrenta la empresa es lograr retener sus clientes actuales reduciendo al mínimo posible su tasa de abandono (variable 'churn'). Para lograr esto, debe identificar aquellos clientes con mayor probabilidad de churn para poder ofrecerles una propuesta de servicio superadora que los retenga. Es por ello que nuestro problema es del tipo supervisado y se resuelve con modelos de clasificación. Un problema de tipo supervisado es aquel en el que los algoritmos son entrenados utilizando ejemplos etiquetados, como una entrada donde se conoce el resultado deseado, en este caso nuestra etiqueta será churn.

En este proyecto realizaremos las predicciones utilizando los siguientes algoritmos:

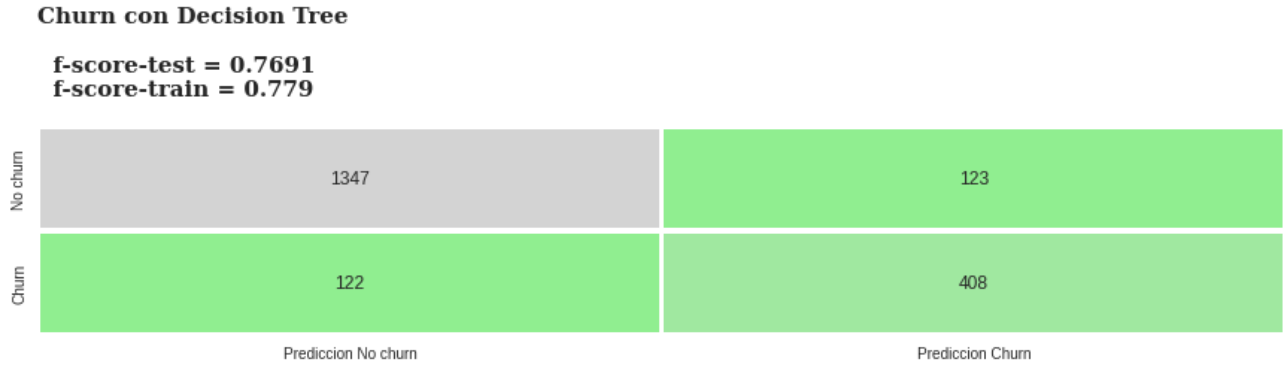
- Decision Tree
- Random Forest
- Regresion Logistica
- KNN

Una vez aplicados los algoritmos mencionados compararemos sus resultados mediante diferentes métricas de performance, en su mayoría resumidas en la matriz de confusión.
Ahora bien, definiremos ciertos conceptos que nos ayudarán a entender mejor esta parte del proyecto. En primer lugar, una matriz de confusión es una cuadrícula que desglosa las predicciones frente a los resultados reales, mostrando los verdaderos positivos, verdaderos negativos, falsos positivos (error tipo I) y falsos negativos (error tipo II). Generalmente, queremos que los valores de la diagonal (de arriba a la izquierda hacia abajo a la derecha) sean más altos porque estos reflejan clasificaciones correctas. Queremos evaluar cuántos clientes predichos de abandonar y continuar con el servicio realmente lo hicieron (verdaderos positivos y negativos). Las otras celdas reflejan predicciones erróneas.

La métrica que nos interesa sea lo mejor posible es el **f-score** pues se trata de un problema desbalanceado. Buscamos también un alto **recall** ya que queremos predecir con mayor exactitud la mayor cantidad de personas que nos van a abandonar, es decir nos importa reducir el número de falsos negativos, siendo menos importante en nuestro problema la precisión, aunque teniendo en cuenta esta métrica a la hora de evaluar el modelo. Es por esto que elegimos evaluar con f-score y no sólo el recall.

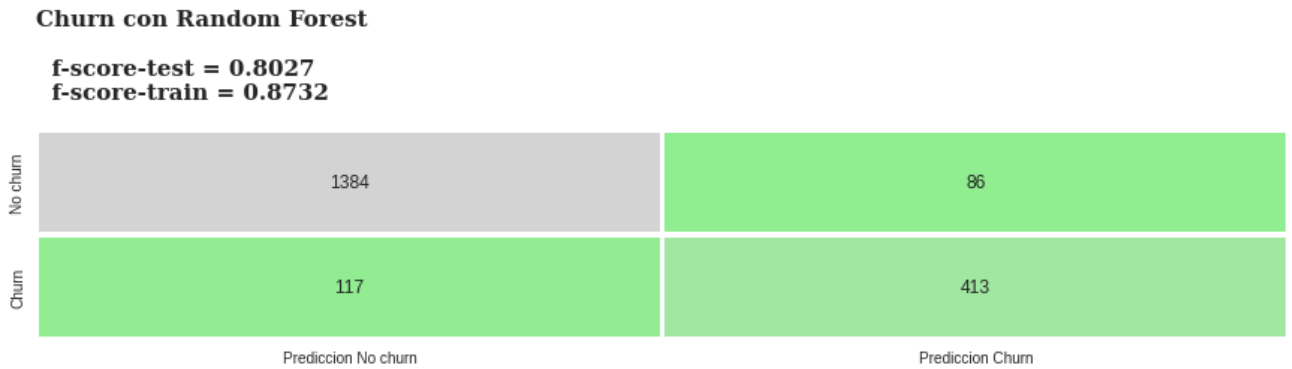
Decision Tree

Matriz de confusion:



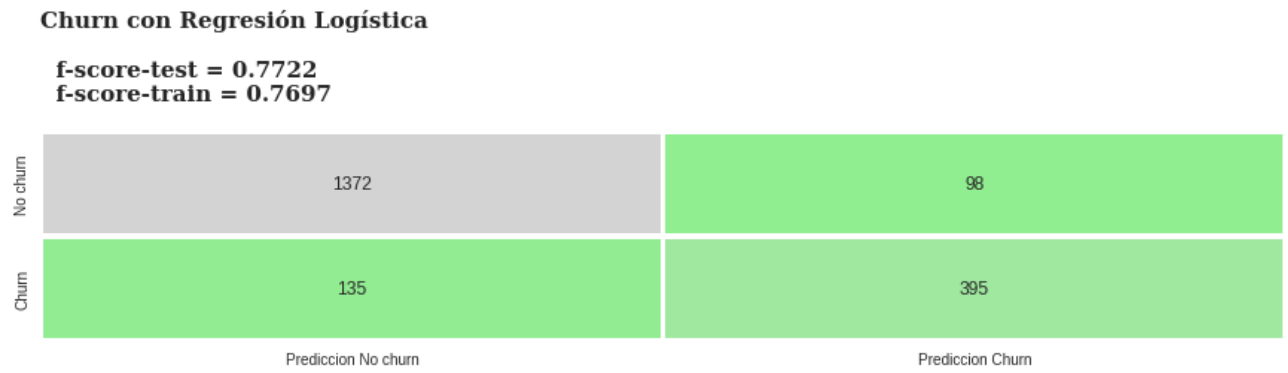
Random Forest

Matriz de confusion:



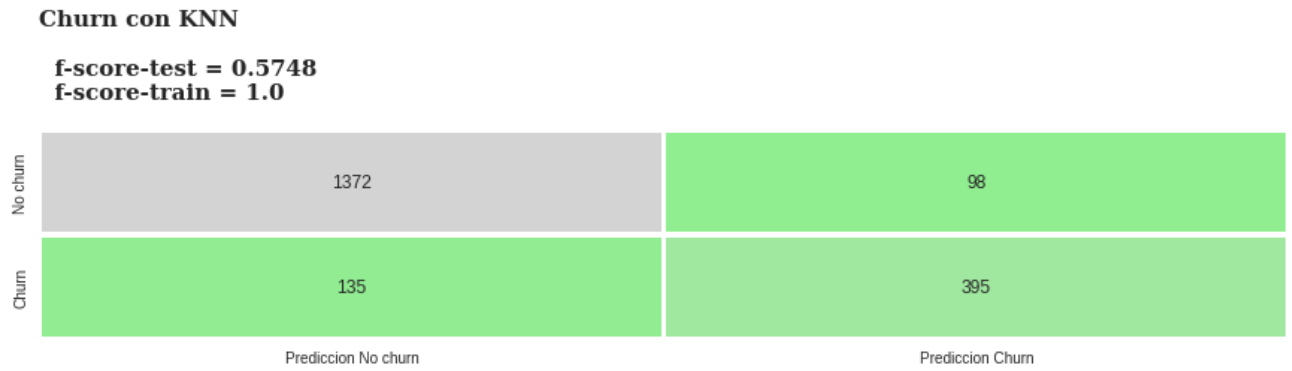
Regresion Logistica

Matriz de confusion:



KNN

Matriz de confusion:



Métricas

Algoritmo	F1-Score	F-Score-rain	Precision	Recall	AUC
Decision tree base	0.72	1	0.73	0.71	0.80
Decision Tree CV	0.78	0.77	0.77	0.77	0.84
Random Forest base	0.79	1	0.83	0.76	0.85
Random Forest CV	**0.87**	**0.80**	0.83	**0.78**	0.86
Logistic Regression	0.77	0.77	0.80	0.73	0.83
KNN	0.78	0.56	0.66	0.5	0.70

Debido a la importancia que tiene para este problema de negocio el recall y el f-score, seleccionamos el algoritmo de Random Forest con Cross Validation.

Los parámetros del modelo son:

RandomForestClassifier(criterion='entropy', max_features=0.50,
min_samples_leaf=8, random_state=42)

Futuros pasos

Consideramos que las métricas del modelo son aceptables para lograr una predicción de la variable objetivo que permita detectar una cantidad considerable de futuros casos de churn, es por esto que proponemos como próximos pasos:

- Productivizar el modelo
- Construir el back de una aplicación que subiremos a la nube (AWS), para que el equipo de Marketing conecte mediante una API al data lake de usuarios y actualice semanalmente la base y así poder dirigir sus campañas en pos de la reducción del churn.
- Como equipo de Data & Analytics, estaremos monitoreando y haciendo revisiones periódicas para ir reajustando.