UNIVERSIDAD
TORCUATO DI TELLA

Master in Management + Analytics

# Técnicas de aprendizaje automático aplicadas en empresas de pequeña escala


**Resumen**

Data Analytics es comúnmente utilizado por empresas de gran escala, pero también puede aportar información valiosa a empresas de pequeña escala. El objetivo de este trabajo será demostrar que se pueden aplicar técnicas de aprendizaje supervisado y no supervisado utilizando un volumen de datos pequeño y aun así obtener buenos resultados. Se realizó un análisis predictivo sobre la variable Churn, alcanzando un puntaje de 0,82 de área bajo la curva ROC. Asimismo, se realizó un análisis descriptivo sobre los consumidores, logrando la segmentación de estos en tres grupos con sentido de negocio. Los resultados obtenidos aportan información valiosa para las decisiones futuras que tomará la empresa para crecer.



Alumno: Agostina Toto Ruá

Director: Ramiro Gálvez

UNIVERSIDAD
TORCUATO DI TELLA

Master in Management + Analytics

# Machine learning techniques applied in small-scale companies

**Abstract**

Data Analytics is commonly associated to large companies; however, it can also provide valuable information to small scale companies. The objective of the present paper is to demonstrate that techniques of supervised and unsupervised learning can be also used in small data sets and obtain favorable results. A Churn prediction analysis was conducted, reaching a 0.82 AUC score. In addition, customers were segmented into three groups with business sense. The results obtained in this paper will contribute with valuable information for the future decisions that the company must take to grow.

Alumno: Agostina Toto Ruá

Director: Ramiro Gálvez

# Índice

# I. Introducción

Durante los últimos años “Data Analytics” se ha convertido en una tendencia adoptada por muchas organizaciones con el objetivo de construir información valiosa a partir de los datos recolectados *(Sivarajah, Kamal, Irani & Weerakkody, 2017)*. El volumen de los datos continúa creciendo exponencialmente y proviene de plataformas digitales, sensores inalámbricos, aplicaciones de realidad virtual y billones de teléfonos celulares *(Henke & Bughin, 2016)*. Se cree que, si las empresas logran recolectar, procesar y analizar grandes conjuntos de datos, entonces la información recolectada puede ser extremadamente valiosa para el crecimiento exponencial de las mismas (*Jelonek, 2017)*.

Big Data puede erróneamente asociarse a empresas de gran escala, capaces de generar grandes volúmenes de datos. Sin embargo, hoy en día empresas de pequeña escala pueden beneficiarse de la cantidad de datos generados tanto de manera online como offline, y tomar decisiones sabias basadas en datos para hacer crecer su negocio *(Ogechi Ogbuokiri, Agu & Udanor, 2015)*.

El objetivo de este trabajo es demostrar con un ejemplo concreto que no se necesita ser un gigante y contar con muchos datos para poder utilizar técnicas de aprendizaje automático. Adicionalmente, se buscará encontrar o generar información valiosa para la empresa, y que pueda mejorar su negocio al tomar medidas con objetivos claros. Para ello primero se creará un modelo supervisado de Churn para predecir qué clientes dejarán de comprar a partir de la información de su última transacción. Luego se realizará un segundo análisis en el cual se segmentará a los consumidores para de esta forma poder personalizar las estrategias de marketing. Este último objetivo contribuirá al primero, ya que, si la estrategia es efectiva, será mucho menos probable que el consumidor deje de comprar. Por último, se tratará de predecir el segmento al cual pertenecerá un cliente a partir de su primera compra.

**La empresa**

La empresa de la cual se obtuvieron los datos es una empresa pequeña que opera desde el año 2017, y no se la mencionará para preservar su privacidad. La misma produce

comidas caseras envasadas al vacío, y quien la compra debe hervir la bolsa en la que viene envasada por tan solo quince minutos en promedio. Cada comida tiene su tiempo específico y eso está especificado en su bolsa, junto a la información nutricional. Las ventas son únicamente online y hacen envío a domicilio dentro de los tres días luego de haber realizado el pedido. Se recibieron los datos para el período enero 2019 – marzo 2021, y en ese rango de tiempo, la empresa contó aproximadamente con 10.900 clientes.

La idea detrás del negocio es salvarles las comidas a los clientes. En un contexto normal, los clientes vuelven de trabajar cansados y lo último que quieren hacer es ponerse a cocinar. En esos casos, su "salvación" es tener comida ya preparada en el freezer, descongelarla y comer rápido sin perder tiempo en prepararla. En ese momento entra en juego esta empresa que no sólo ofrece rapidez, sino que tiene una carta amplia de productos para todos los gustos.

Para aquellos que se quieren cuidar, para aquellos que van mucho al gimnasio y necesitan energía, para aquellos que sólo consumen comidas veggies o para aquellos que simplemente quieren disfrutar un plato que nunca cocinarían, esta empresa viene a solucionarles las comidas.

La empresa comenzó vendiendo solo almuerzos o cenas, pero luego no sólo fueron expandiendo su carta de platos, sino que incorporaron nuevos segmentos. Entre ellos, producen jugos naturales y détox también envasados al vacío y para congelar. Por otro lado, ya ofrecen postres y snacks para completar el menú.

Los productos se pueden comprar individualmente o por packs armados personalmente por cada cliente según sus gustos. Los packs pueden ser de siete, catorce, veintiún o veintiocho comidas. En un caso ideal, los clientes compran suficientes platos para stockearse por uno o dos meses y luego vuelven a hacer pedidos.

**Motivación**

El comercio electrónico en Argentina durante el año 2020 creció un 124% con respecto al año anterior según el Estudio Anual de Comercio Electrónico en Argentina.[1] Sin dudas, este crecimiento fue impulsado por la pandemia, pero lo cierto es que las compras online van tomando cada vez más peso, y muchos prefieren realizar una compra de esta manera en vez de perder tiempo trasladándose al local físico y todo lo que implica una compra presencial.

Algunas ventajas adicionales de las compras online para el consumidor son las siguientes *(Maheshwari, 2020)*:

- Pueden comparar precios entre locales en cuestión de minutos
- No tienen que hacer fila para pagar
- Pueden comprar desde cualquier lugar si tienen un dispositivo electrónico
- El pedido se puede realizar cualquier día y en cualquier horario
- Pueden programar envío a domicilio dentro de los pocos días de haber realizado la compra
- Conocimiento de stock. Si buscan algo en particular que está agotado, no desperdician el viaje

Para la empresa también hay ventajas tales como que no necesitan local físico y pueden ofrecer muchos más productos, ya que no tienen problemas de espacio físico para exhibirlos. Por otro lado, pueden observar cuales son los productos más demandados y tienen mayor control sobre el inventario. No tienen los costos fijos del local físico tal como alquiler y personal de atención al público, ni tampoco los variables como luz, agua, etc. asociadas a un local a la calle.

Por otro lado, la venta online tiene un alcance distinto y mayor. Si bien tanto las tiendas online como las físicas pueden tener la misma estrategia de marketing (ya sea a través de *newsletters*, publicidad en televisión o radio, panfletos), lo cierto es que es mucho más fácil y rápido entrar a ver lo que se ofrece en una tienda online al ver la publicidad. Por el contrario, luego de ver una publicidad de un local físico, uno no puede corroborar

1 https://www.cace.org.ar/noticias-el-comercio-electronico-crecio-un-124-y-supero-los-novecientos-mil-millones-de-pesos-en-ventas#:~:text=El%20comercio%20electr%C3%B3nico%20en%20Argentina,cace.org.ar)

los productos que se venden automáticamente, sino que debería tener que trasladarse al local.

Teniendo en cuenta entonces el crecimiento que está experimentando el *e-commerce*, y teniendo de referencia a grandes empresas como Mercado Libre y Amazon, parece atractivo analizar una empresa de este rubro. Las empresas mencionadas anteriormente tienen otro tipo de exposición, y, por lo tanto, otra cantidad de datos disponibles.

# II. Datos

Como se mencionó anteriormente, los datos provienen de una pequeña empresa que vende sus productos únicamente de forma online. Se recibieron un total de tres conjuntos de datos: "Pedidos", "Órdenes" y "Quejas" para el período enero 2019 – marzo 2021. El conjunto "Pedidos" contiene la información agrupada por compra, detallando la cantidad de platos y el monto total pagado por cada cliente. El conjunto "Órdenes" contiene la información desagregada de cada pedido, es decir, una línea por cada producto adquirido dentro de un mismo pedido. Por último, el conjunto "Quejas" detalla las quejas realizadas por los clientes, especificando el pedido que recibió una queja y el motivo de esta. A continuación, se profundiza sobre cada uno de los mismos.

**Conjunto de datos "Pedidos"**

Este primer conjunto de datos contiene la información general de todos los pedidos realizados desde enero 2019 hasta marzo 2021. Tiene una extensión de 22.811 registros, y cada uno de ellos corresponde a una transacción distinta. Las columnas, y por lo tanto las variables con las que se cuenta son las siguientes:

- **# Pedidos**: Indica el número de pedido, junto al nombre y apellido del cliente.
- **Venta formato número**: Monto total pagado por la transacción.
- **Fecha de Ingreso**: Fecha en la que se realizó el pedido.
- **Método de pago**: Método de pago utilizado en cada transacción.
- **Transacción MPago**: Número de transacción de la plataforma Mercado Pago, si es que el pago se realizó de esta manera. Se puede observar que la variable posee errores de entrada, pero la misma no se utilizará para el análisis.
- **Comidas**: Cantidad de platos por pedido.
- **Recurrent? (si=1)**: Indica si el usuario realizó otras compras o si es la única. Sin embargo, no se tomó en cuenta a esta variable y en la siguiente sección se explicará la razón.
- **Compra única?**: Indica si es la primera compra realizada por ese cliente.
- **Fecha entrega (sin horario)**: Fecha en la que se realizó la entrega del pedido.
- **Fecha de Alta**: Fecha en la que cada usuario se registró en el sistema. Se pueden observar errores de formato que se corrigieron manualmente.
- **Nombre y Apellido**
- **Mail**
- **Número de cliente:** Es un código interno utilizado por la empresa para identificar a los clientes. A partir de este código, se pueden relacionar todos los conjuntos de datos.

A continuación, se muestran las primeras 10 filas del primer conjunto, ocultando las columnas de Nombre, Apellido y Mail de los usuarios para preservar su privacidad. En esta muestra ya se pueden observar algunas características de los datos a corregir que se ampliaran en la siguiente sección, como por ejemplo información que debería estar separada en dos variables o error de entrada.

**Tabla 1**: Conjunto de datos "Pedidos"

**1.A** Primeras 6 columnas

| # Pedido | Venta formato numero | Fecha de Ingreso | Método de pago | Transacción MPago | Comidas |
|---|---|---|---|---|---|
| #1114075 - EZEQU | $2.545 | 2019-01-01 | Mercado Pago | 4406702018 | 14 |
| #1114090 - MAGDA | $1.390 | 2019-01-02 | Mercado Pago | 4407417032 | 7 |
| #1114128 - MATIAS | $2.613 | 2019-01-02 | Mercado Pago | 4407479304 | 14 |
| #1114162 - BARBA | $2.634 | 2019-01-02 | Mercado Pago | 4407618974 | 0 |
| #1114177 - CAROL | $2.895 | 2019-01-02 | Mercado Pago | 44075528544407550000 | 14 |
| #1114192 - LISA C | $1.255 | 2019-01-02 | Mercado Pago | 4407646560 | 7 |
| #1114211 - MARGA | $3.040 | 2019-01-02 | Mercado Pago | 4408188241 | 14 |
| #1114233 - MARIA | $1.580 | 2019-01-02 | Mercado Pago | 4407736759 | 6 |
| #1114261 - IGNAC | $2.620 | 2019-01-02 | Mercado Pago | 4408056024 | 14 |
| #1114274 - TOMAS | $2.665 | 2019-01-02 | Mercado Pago | 4408259313 | 14 |

**1.B** Últimas 5 columnas

| Recurrent? (si=1) | Compra unica? | Fecha entrega (sin horario) | Fecha de Alta | Número de Cliente |
|---|---|---|---|---|
| 0 | NO | 2019-01-03 | 2019-01-02 | 2065 |
| 1 | NO | 2019-01-02 | 2018-12-04 | 1964 |
| 1 | NO | 2019-01-04 | 2018-08-17 | 1367 |
| 1 | NO | 2019-01-03 | 2018-07-15 | 1251 |
| 1 | NO | 2019-01-04 | 2018-09-17 | 1639 |
| 0 | SI | 2019-01-06 | 2019-01-02 | 2067 |
| 0 | NO | 2019-01-04 | 2019-01-02 | 2068 |
| 0 | NO | 2019-01-03 | 2019-01-02 | 2069 |
| 0 | NO | 2019-01-03 | 2019-01-02 | 2070 |
| 0 | SI | 2019-01-03 | 2019-01-02 | 2071 |

**Conjunto de datos "Órdenes"**

El segundo conjunto recibido contiene la información detallada acerca de cada uno de los pedidos. Se cuenta con una fila para cada ítem dentro de un pedido y, en caso de tratarse de un pack, primero se observa una fila resumen del pack y luego una fila por cada ítem. Es decir, si un cliente compra un pack de 7 comidas, entonces ese pedido contará con 8 filas dentro del conjunto Órdenes. Por el contrario, si un individuo realiza

un pedido de 7 comidas aisladas, entonces ese pedido contará con 7 filas. Cuenta con un total de 188.931 filas y las variables son las siguientes:

- **Orden:** Código interno que identifica cada ítem dentro de un mismo pedido.
- **nro cliente:** Código interno para identificar a los clientes. Coincide con la variable "Número de cliente" del conjunto Pedidos.
- **Pedido:** Código interno para identificar el pedido. Contiene el nombre y apellido del cliente. Coincide con la variable "# Pedidos" del conjunto Pedidos.
- **Producto:** Detalla el plato o pack comprado.
- **Cantidad**
- **Descuento por Producto**
- **Fecha Ingreso**
- **FECHA ENTREGA**
- **Cliente:** Nombre y apellido del cliente.
- **Nombre del producto:** Nombre del pack o plato
- **Nro pedido:** Código interno. Coincide con el número de la variable "# Pedidos" del conjunto "Pedidos" y la variable "Pedido" de este mismo conjunto. Esta variable no está acompañada del nombre del cliente como las mencionadas anteriormente.
- **Mail**
- **Código:** Código interno utilizado para identificar los platos.
- **Código 2:** Mismo código que la variable "Código" anterior. Se duplica información de forma innecesaria.
- **Suscripción:** Variable binaria que indica con el número 1 si el cliente está registrado en la página y 0 en caso contrario.
- **Día**
- **Mes**
- **Año**

A continuación, se muestran 10 filas del segundo conjunto de datos, nuevamente ocultando las columnas de Nombre, Apellido y Mail de los usuarios para preservar su privacidad. Por otro lado, se ocultan ciertas filas de los primeros pedidos para poder mostrar más de un pedido.

En este caso se puede observar la existencia de columnas que repiten información y podría resumirse en una columna.

**Tabla 2**: Conjunto de datos “Órdenes”

**2.A** Primeras 9 columnas

| Orden | nro cliente | Pedido | Producto | Cantidad | Descuento por Producto | Fecha Ingreso | FECHA ENTREGA | Cliente |
|---|---|---|---|---|---|---|---|---|
| 24515 | 2065 | #1114075 - EZE | Armalo Saludable & Fitness by Megatlon: 14 comidas | 1 | NINGUNO | 1/1/2019 | 3/1/2019 4:30pm | EZEQUIEL M |
| 24516 | 2065 | #1114075 - EZE | Pechuguitas rebozadas c/ batata rellena | 1 | PACK CUSTOM | 1/1/2019 | 3/1/2019 4:30pm | EZEQUIEL M |
| 24518 | 2065 | #1114075 - EZE | Pastel de vacío c/batata y zucchinis | 1 | PACK CUSTOM | 1/1/2019 | 3/1/2019 4:30pm | EZEQUIEL M |
| 24520 | 2065 | #1114075 - EZE | Pechuga al limón c/ soufflé espinaca | 1 | PACK CUSTOM | 1/1/2019 | 3/1/2019 4:30pm | EZEQUIEL M |
| 24521 | 2065 | #1114075 - EZE | Pechuga al limón c/ soufflé espinaca | 1 | PACK CUSTOM | 1/1/2019 | 3/1/2019 4:30pm | EZEQUIEL M |
| 24530 | 1964 | #1114090 - MA | Armalo Saludable by Megatlon: 7 comidas | 1 | NINGUNO | 2/1/2019 | 2/1/2019 7:00pm | MAGDALENA |
| 24531 | 1964 | #1114090 - MA | Hamburguesa de garbanzos c/ calabazas asadas | 1 | PACK CUSTOM | 2/1/2019 | 2/1/2019 7:00pm | MAGDALENA |
| 24532 | 1964 | #1114090 - MA | Soufflé de vegetales | 1 | PACK CUSTOM | 2/1/2019 | 2/1/2019 7:00pm | MAGDALENA |
| 24538 | 1367 | #1114128 - MAT | Suprema al horno con cremoso de papa | 1 | NINGUNO | 2/1/2019 | 4/1/2019 7:00pm | MATIAS MA |
| 24539 | 1367 | #1114128 - MAT | Pechuga portuguesa con cakes de papa | 1 | NINGUNO | 2/1/2019 | 4/1/2019 7:00pm | MATIAS MA |

**2.B** Últimas 8 columnas

| Nombre del producto | Nro pedido | Código | codigo2 | Suscripcion | DIA | MES | AÑO |
|---|---|---|---|---|---|---|---|
| PACK FITNESS | 1114075 | 80007 | 80007 | 0 | 1 | 1 | 2019 |
| PECHUGAS REBOZADAS CON BATATA RELLENA | 1114075 | 0 | 0 | 0 | 1 | 1 | 2019 |
| Pastel de vacío c/batata y zucchinis | 1114075 | 20104 | 20104 | 0 | 1 | 1 | 2019 |
| PECHUGA LIMÓN/HIERBAS CON SOUFFLÉ DE ESPINACA | 1114075 | 20105 | 20105 | 0 | 1 | 1 | 2019 |
| PECHUGA LIMÓN/HIERBAS CON SOUFFLÉ DE ESPINACA | 1114075 | 20105 | 20105 | 0 | 1 | 1 | 2019 |
| Armalo Saludable by Megatlon: 7 comidas | 1114090 | 0 | 0 | 0 | 2 | 1 | 2019 |
| HAMBURGUESA DE GARBANZO CON CALABAZAS ASADAS Y SEMILLAS DE GIRASOL | 1114090 | 20102 | 20102 | 0 | 2 | 1 | 2019 |
| SOUFFLE DE VEGETALES | 1114090 | 20043 | 20043 | 0 | 2 | 1 | 2019 |
| SUPREMA CON CREMOSO DE PAPA | 1114128 | 20101 | 20101 | 0 | 2 | 1 | 2019 |
| PECHUGA PORTUGUESA CON CAKES DE PAPA | 1114128 | 0 | 0 | 0 | 2 | 1 | 2019 |

## Conjunto de datos “Quejas”

El tercer y último conjunto recibido contiene información acerca de las quejas realizadas por los clientes. Cuenta con un total de 677 filas y las variables son las siguientes.

- **# Pedido:** Contiene el número de pedido junto al nombre y apellido del cliente. Coincide con las variables de los conjuntos anteriores
- **Mail**
- **Fecha Entrega**
- **Mes (Entrega)**
- **Tipo:** De qué tipo de queja se trata (precio, faltantes, bolsa, etc.)
- **Comentario:** Comentario realizado por el cliente al realizar la queja
- **Entrega 3:** Coincide con la variable “Fecha Entrega”
- **Año**
- **Mes interno**
- **Queja comida:** Variable binaria que indica si la queja en cuestión se debe a la comida
- **Queja bolsa:** Variable binaria que indica si la queja en cuestión se debe a la bolsa.
- **Queja tiempo:** Variable binaria que indica si la queja en cuestión se debe al tiempo que tardó la entrega.
- **Queja envío**: Variable binaria que indica si la queja en cuestión se debe a problemas con el envío.

- **Queja tamaño:** Variable binaria que indica si la queja en cuestión se debe a problemas con el tamaño de los platos.
- **Tipo de Queja:** Indica a cuál de las variables anteriores se debe la queja.

A continuación, se pueden observar las primeras 10 filas del conjunto "Quejas", ocultando la información personal.

**Tabla 3:** Conjunto de datos "Quejas"

**3.A** Primeras 5 columnas

| # Pedido | Fecha Entrega | Mes (entrega) | Tipo | Comentario |
|---|---|---|---|---|
| #1135554 - ARIE | 2020-01-02 | 1 | Precio | El servicio es muy bueno pero algo caro |
| #1135571 - MAU | 2020-01-07 | 1 | Queja faltantes | Faltaron dos fit |
| #1135643 - MIGU | 2020-01-02 | 1 | Queja faltantes | Se le envió un bowl de espinaca y no de choclo |
| #1135701 - SILV | 2020-01-03 | 1 | Queja comida | es muy practicoultimamente he notado una baja en la calidad de las comida |
| #1135890 - MAR | 2020-01-06 | 1 | Variedad | Me gustaría tener la opción de combinar platos ( la hamburguesa de carne sólo viene con puré de calabaza |
| #9384 - IVANA S | 2020-01-09 | 1 | Queja faltantes | faltó una suprema napo |
| #1136078 - FLOR | 2020-01-08 | 1 | Queja comida | Hay algunas particularmente como las papas, que me parece que se secan un poco en el proceso. Pero el |
| #1136157 - AGUS | 2020-01-10 | 1 | Precio | La degustación me pareció correcta relación calidad precio, ahora cuando vas a la página para comprar sale |
| #1136322 - FERN | 2020-01-10 | 1 | Tamaño,Precio | La comida es buena, no es muy abundante y es cara. |

**3.B** Últimas 10 columnas

| Entrega 3 | Año | mes interno | Queja comida | Queja Bolsa | Queja tiempo | Queja Envío | Queja faltantes | Tamaño | Tipo de queja |
|---|---|---|---|---|---|---|---|---|---|
| 2020-01-02 | 2020 | 25 | 0 | 0 | 0 | 0 | 0 | 0 | Precio |
| 2020-01-07 | 2020 | 25 | 0 | 0 | 0 | 0 | 1 | 0 | Queja faltantes |
| 2020-01-02 | 2020 | 25 | 0 | 0 | 0 | 0 | 1 | 0 | Queja faltantes |
| 2020-01-03 | 2020 | 25 | 1 | 0 | 0 | 0 | 0 | 0 | Queja comida |
| 2020-01-06 | 2020 | 25 | 0 | 0 | 0 | 0 | 0 | 0 | Variedad |
| 2020-01-09 | 2020 | 25 | 0 | 0 | 0 | 0 | 1 | 0 | Queja faltantes |
| 2020-01-08 | 2020 | 25 | 1 | 0 | 0 | 0 | 0 | 0 | Queja comida |
| 2020-01-10 | 2020 | 25 | 0 | 0 | 0 | 0 | 0 | 0 | Precio |
| 2020-01-10 | 2020 | 25 | 0 | 0 | 0 | 0 | 0 | 1 | Tamaño,Precio |

# III. Enriquecimiento de Datos

Como se mostró en la sección anterior, las variables de los conjuntos recibidos se encontraban muy desordenadas y contenían información repetida. Por ejemplo, se menciona el nombre del cliente junto al número de pedido en una misma columna, pero también se cuentan con variables específicas de nombre y apellido. Esta información se puede separar para relacionar los conjuntos entre sí a partir del número de pedido y para tener datos más limpios. Por lo tanto, se procedió con la limpieza de las variables aplicando las siguientes modificaciones:

1. Se eliminaron columnas que contenían información duplicada. Por ejemplo, el conjunto "Órdenes" contenía dos columnas con distinta nomenclatura que indicaban el ID del cliente.
2. Se separó información de una misma variable que podía ser mostrada en dos variables distintas como por ejemplo el número de pedido y el nombre y apellido de los clientes.
3. Se modificaron los nombres de algunas variables para poder manipularlas con facilidad en Python.
4. Se eliminaron filas que contenían "NA" en variables importantes a utilizar en los modelos. Lamentablemente se encontraron filas casi vacías que no contaban con información crucial para correr el modelo. En total se eliminaron 18 filas, quedando un total de 22.793 registros para utilizar.
5. Se agregaron columnas indicando información del cliente tales como "Sexo" y "Edad"
6. Se creó un nuevo conjunto de datos llamado "Clientes" con la información de cada individuo, para obtener un mayor entendimiento del tipo de cliente con el que cuenta la empresa y para poder realizar el análisis de Clustering.

A continuación, se detalla cómo quedaron los conjuntos una vez que se realizaron los cambios.

**Conjunto "Pedidos" modificado para el análisis de Churn**

Los cambios que se realizaron fueron los siguientes:

1. Se separó la variable "# Pedidos" en tres variables distintas: "Num_Pedido", "Nombre" y "Apellido" por dos razones. El número de pedido es una variable importante para relacionar los distintos conjuntos de datos, con lo cual debería estar por separado. Por otro lado, se necesitaba tener el nombre del cliente por separado para poder identificar el sexo de este, ya que no se contaba con esta información.
2. La variable "Nombre" se comparó contra la lista de nombres y sexos provista por el registro civil del Gobierno de la Ciudad de Buenos Aires.[2] A partir de esta información se creó la variable "Sexo". Lamentablemente no se pudieron obtener todos los sexos, ya que entre los clientes hay empresas y errores de escritura, o simplemente nombres de extranjeros que no se encuentran dentro de esta lista. Se definió "Empresa" como una de las nomenclaturas finales, junto a "M" y "F".
3. A la variable "Venta formato número" se le modificó el nombre a "Monto_total". Como se recibió información de las ventas de dos años y Argentina está caracterizada por ser un país con inflación alta, se indexó esta variable por la inflación mensual obtenida del INDEC, utilizando enero 2019 como base.[3] El resultado final se puede ver en la variable "Monto_total_indexado", indicando cuántos pesos hubiesen representado en enero 2019.
4. La variable "Fecha de ingreso" se modificó a "Fecha_Ingreso" y se eliminó la variable "Fecha entrega (sin horario)", ya que no era necesaria para el análisis. Por otro lado, a partir de la fecha de ingreso se crearon las variables "Mes", "Año" y "Dia_Semana" para identificar patrones de consumo.
5. No se recibió el DNI de los clientes para respetar información privada, pero se nos proveyó de un análisis interno realizado por la empresa para detectar la edad

2 https://buenosaires.gob.ar/areas/registrocivil/nombres/busqueda/imprimir.php

3 https://www.indec.gob.ar/indec/web/Institucional-Indec-InformacionDeArchivo-1

de los individuos a partir de su documento. Esta información fue agregada al conjunto, ya que podría ser una variable relevante en el análisis de Churn.

6. Se agregaron las variables “Quejas” y “Tipo” a partir de la unión de los conjuntos “Pedidos” y “Quejas” para poder identificar qué pedidos obtuvieron una queja y de qué tipo. Esta información puede ser relevante para el análisis de Churn.
7. Se agregó la variable “Cant_Compras” que indica cuántas veces el cliente en cuestión realizó una compra. Si bien se repetirá a lo largo del conjunto, será importante en el análisis de Churn teniendo en cuenta la definición que se utiliza. Al tener esta información, la variable “Compra única?” ya no es necesaria y se eliminó.
8. Se eliminó la variable “Recurrent? (si=1)” ya que al analizar los resultados de la misma, se encontraron incoherencias. Por ejemplo, algunos usuarios que realizaron muchos pedidos estaban mal clasificados como “0”, y algunos usuarios que realizaron una compra estaban clasificados como “1”. Por este motivo, se decidió eliminar la variable y observar esta información en “Cant_Compras”.
9. Se eliminaron las variables “Transacción MPago”, “Fecha de Alta” (luego de pasar la información al conjunto “Clientes”), “Nombre”, “Apellido” y “Mail”.
10. Se eliminaron las filas que no contaban con la identificación del usuario o información acerca de la compra, ya que no serán útiles para el análisis. Se pasó de tener 22.811 registros a 22.793 registros.
11. Por último, se modificaron otros nombres de variables y se reacomodó el orden de estas.

El resultado de los cambios que se realizaron es el siguiente, tomando febrero 2019 para observar la indexación.

**Tabla 4**: Conjunto Pedidos modificado

**4.A** Primeras 8 columnas

| ID_Cliente | Num_Pedido | Monto_total | Monto_total_indexado | Fecha_Ingreso | Mes | Año | Dia_semana |
|---|---|---|---|---|---|---|---|
| 2166 | 1116743 | 2330 | 2245 | 2019-02-01 00:00:00 | 2 | 2019 | 6 |
| 1940 | 1116856 | 1365 | 1315 | 2019-02-01 00:00:00 | 2 | 2019 | 6 |
| 2167 | 1116859 | 1795 | 1729 | 2019-02-01 00:00:00 | 2 | 2019 | 6 |
| 2169 | 1116899 | 2660 | 2563 | 2019-02-01 00:00:00 | 2 | 2019 | 6 |
| 2170 | 1116907 | 1310 | 1262 | 2019-02-02 00:00:00 | 2 | 2019 | 7 |
| 2171 | 1116910 | 1085 | 1045 | 2019-02-02 00:00:00 | 2 | 2019 | 7 |
| 662 | 1116913 | 990 | 954 | 2019-02-02 00:00:00 | 2 | 2019 | 7 |
| 2172 | 1116917 | 1575 | 1517 | 2019-02-02 00:00:00 | 2 | 2019 | 7 |
| 1755 | 1116922 | 1823 | 1756 | 2019-02-02 00:00:00 | 2 | 2019 | 7 |
| 2173 | 1116929 | 3480 | 3353 | 2019-02-02 00:00:00 | 2 | 2019 | 7 |

**4.B** Últimas 7 columnas

| Método_pago | Comidas | Sexo | Edad | Quejas | Tipo | Cant_Compras |
|---|---|---|---|---|---|---|
| Transferencia Bancaria | 14 | M | 41 | 0 | Sin Queja | 1 |
| Mercado Pago | 9 | M | 48 | 0 | Sin Queja | 3 |
| Transferencia Bancaria | 7 | F | | 0 | Sin Queja | 1 |
| Mercado Pago | 14 | F | 56 | 0 | Sin Queja | 2 |
| Mercado Pago | 7 | M | 58 | 0 | Sin Queja | 3 |
| Mercado Pago | 7 | M | 30 | 0 | Sin Queja | 2 |
| Mercado Pago | 7 | F | 89 | 0 | Sin Queja | 1 |
| Mercado Pago | 7 | M | 33 | 0 | Sin Queja | 1 |
| Mercado Pago | 8 | M | 58 | 0 | Sin Queja | 5 |
| Mercado Pago | 21 | F | 67 | 0 | Sin Queja | 1 |

### Conjunto “Órdenes” modificado

Si bien se realizaron modificaciones en el conjunto “Órdenes”, no se comentarán ya que no fue utilizado para ningún análisis. Se podría haber utilizado para un análisis que tuviera al producto ofrecido como foco, sin embargo, ese no es el foco de este trabajo.

### Conjunto “Quejas” modificado

Al conjunto de datos “Quejas” se lo utilizó para agregar información pertinente al conjunto “Pedidos”. Simplemente se relacionaron las quejas con los pedidos correspondientes, indicando únicamente si el pedido en cuestión recibió una queja y, en caso de haber recibido, de qué tipo de queja se trató.

**Conjunto "Clientes" creado para el análisis de Clustering**

A partir de la información obtenida, se creó un cuarto conjunto de datos con la información de los clientes únicos, para luego poder realizar el análisis de Clustering. El conjunto de clientes únicos contiene un total de 10.752 filas y las siguientes variables.

- **ID_Cliente**
- **Nombre**
- **Apellido**
- **Sexo**
- **Edad**
- **Alta**: Fecha en la que el cliente se registró en el sistema
- **Cant_compras**: Cantidad de compras totales realizadas por cada cliente
- **Monto_total**: Monto de dinero total gastado por cada cliente
- **Monto_Max**: Máximo monto gastado en una compra
- **Monto_Min**: Mínimo monto gastado en una compra
- **Gasto_promedio**
- **Última_compra**: Fecha en la que cada cliente realizó su última compra
- **Días_desde_última_compra**: Cantidad de días desde que el cliente realizó su última compra hasta el 24 de marzo de 2021, ya que ese es el último día para el cual se cuenta con información.

A continuación, se pueden observar las primeras 10 filas del conjunto, creado con el objetivo de analizar a los clientes en particular.

**Tabla 5**: Conjunto creado "Clientes"

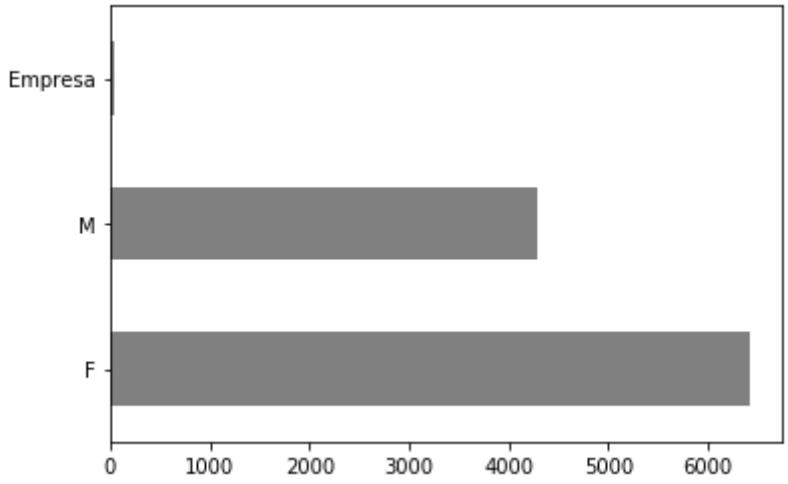
| ID_Cliente | Sexo | Edad | Alta | Cant_compras | Monto_total | Monto_total_indexado | Monto_max | Monto_min | Gasto_promedio |
|---|---|---|---|---|---|---|---|---|---|
| 61 | M | 30 | 2017-06-30 00:00:00 | 5 | 7584 | 5537 | 1440 | 693 | 1107 |
| 64 | F | 31 | 2017-06-30 00:00:00 | 23 | 32668 | 22966 | 2269 | 258 | 999 |
| 65 | M | | 2017-06-30 00:00:00 | 3 | 9636 | 6555 | 3995 | 1206 | 2185 |
| 70 | F | | 2017-06-30 00:00:00 | 2 | 2969 | 1594 | 1313 | 281 | 797 |
| 72 | F | 39 | 2017-06-30 00:00:00 | 1 | 1645 | 1002 | 1002 | 1002 | 1002 |
| 85 | M | | 2017-06-30 00:00:00 | 2 | 5955 | 5308 | 2670 | 2638 | 2654 |
| 86 | F | | 2017-06-30 00:00:00 | 2 | 5710 | 3529 | 2828 | 701 | 1764 |
| 88 | F | 65 | 2017-06-30 00:00:00 | 1 | 990 | 954 | 954 | 954 | 954 |
| 91 | F | 58 | 2017-06-30 00:00:00 | 5 | 18352 | 15016 | 4745 | 1587 | 3003 |
| 101 | F | | 2017-06-30 00:00:00 | 1 | 690 | 615 | 615 | 615 | 615 |

# IV. Análisis exploratorio

En esta sección se realizará un análisis exploratorio de los datos para obtener *insights* de patrones generales acerca del comportamiento de los usuarios y entender un poco más sobre el tipo de cliente que compone el conjunto. No se realizará el análisis de las variables enfocándose en la variable “Churn” en esta sección ya que se analizará en la sección V.

El primer conjunto analizado es el de “Clientes Únicos”, el cual se armó en base a los datos obtenidos para poder trabajar sobre los clientes, sin repetir su información. A continuación, se describen algunas de las variables.

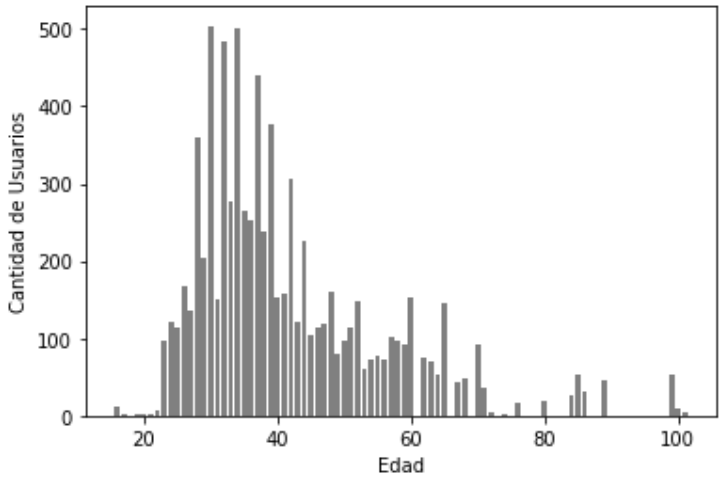
**Figura 1**: Distribución del sexo de los clientes




Se cuenta con un total de 10.752 clientes, de los cuales 6.418 son mujeres, 4.295 son hombres y 39 son empresas. El conjunto enviado por la empresa no contaba con esta variable, por lo tanto, como se mencionó anteriormente, se compararon los nombres de los clientes contra la lista de nombres y sexos provista por el registro civil del Gobierno de la Ciudad de Buenos Aires. Lamentablemente, el 4% de los usuarios no pudieron ser identificados automáticamente a partir de la comparación. Como se podía tratar de errores humanos al cargar la información personal, recorrimos aquellos que no fueron reconocidos y pudimos corregirlos. Por otro lado, se encontró que algunos de los clientes eran empresas y no personas físicas, y en ese caso se hizo la distinción.

Aproximadamente el 60% de los clientes son mujeres y cerca del 40% hombres. Este resultado sorprende a la empresa. Se esperaba encontrar más hombres que mujeres, ya que la idea de la empresa surgió de una necesidad propia y se observó que hombres recién mudados tenían la misma necesidad y falta de tiempo o ganas luego de un día largo de trabajo. Sin embargo, se apuntó a cualquier persona que tuviera la misma necesidad. Este resultado también deja en evidencia que nuestra sociedad actual lejos está de la sociedad anterior en la que las mujeres se dedicaban mucho a la cocina y tareas del hogar y no tanto al trabajo.

**Figura 2**: Distribución de edades de los clientes

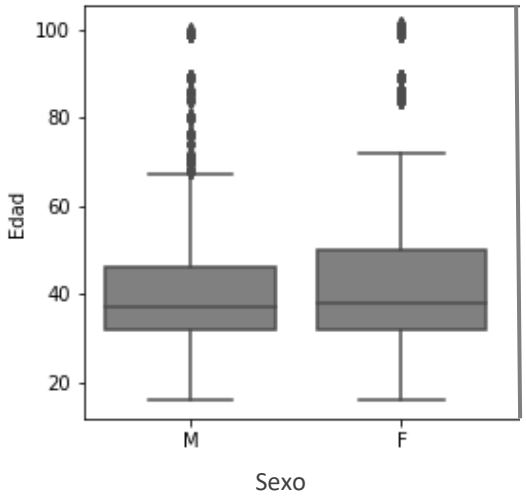



No es necesario completar la edad en la página al hacer el pedido, pero sí el DNI. La empresa se encargó de calcular la edad en base al DNI y envió directamente el dato. Por lo tanto, se trata de una edad aproximada y no se cuenta con el cálculo de esta.

Se encontraron *outliers* con edades que se aproximan a los 100 años, y se cree que estos fueron calculados erróneamente y en realidad se trataban de DNI de extranjeros, o simplemente fue un error humano al ingresar el DNI. Se los consideró como “NA”.

No se cuenta con información de la edad del 24% de los 10.752 clientes, pero a continuación se observa la distribución de edades por sexo.

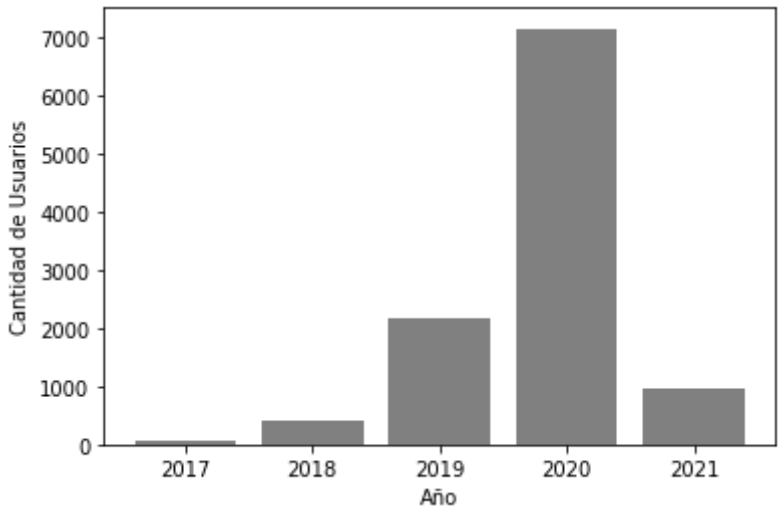
**Figura 3**: Boxplot de edades según sexo




Se puede observar que el rango de edad de las mujeres que compran estos productos es más amplio que el de los hombres. La mediana de ambos sexos es muy parecida y ronda los 37 años. En cuanto al primer cuartil (Q1), se encuentra alrededor de los 32 años en ambos casos, mientras que el tercer cuartil (Q3) se encuentra cerca de los 45 y 50 años para los hombres y mujeres respectivamente. El rango de edad comienza por debajo de los 20 años y continúa hasta alrededor de los 68 en el caso de los hombres y alrededor de los 72 en caso de las mujeres. Por último, se pueden observar varios *outliers* en el caso de los hombres, y menos cantidad en el caso de las mujeres.

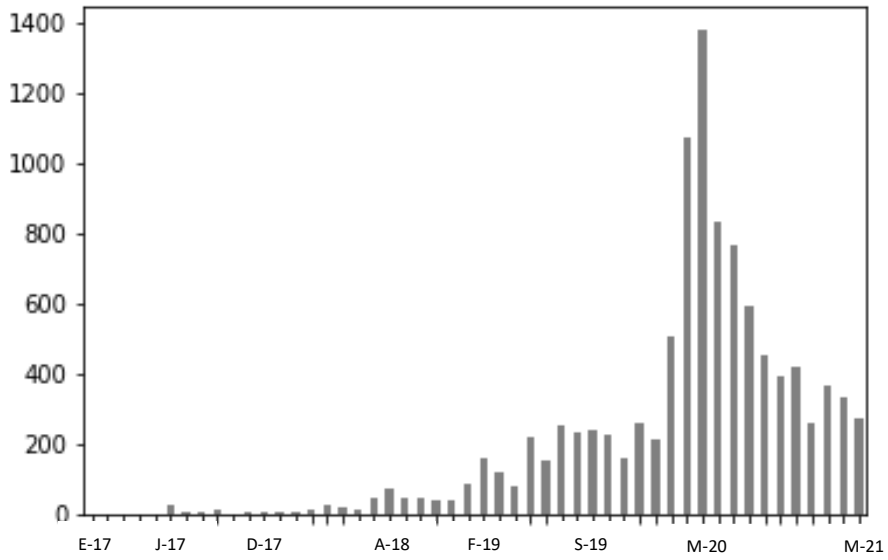
Nuevamente, no parece lógico que adultos mayores de 70 años estén utilizando esta herramienta, no solo por cuestiones de tecnología, sino cuestiones culturales acerca del *delivery* y la comida que no sea 100% casera. Por esta razón, se cree que la edad calculada por la empresa no debería tenerse en cuenta como una variable sumamente relevante, ya que no es certera.

**Figura 4**: Cantidad de altas de usuarios por año




Se puede observar un crecimiento de altas año a año de 492% entre 2017 y 2018, 455% entre 2018 y 2019 y 230% entre 2019 y 2020. Hasta el momento, el pico de nuevos clientes se alcanzó en el año 2020. Los datos del 2021 están incompletos, ya que contamos con información hasta marzo. Se puede corroborar que es una empresa chica, ya que el total de las altas no excede los 11.000 clientes.

**Figura 5**: Cantidad de altas de usuarios por mes




Por otro lado, se analizó la cantidad de altas por mes por cada año, para comprobar si existe un patrón, pero al ser un negocio relativamente nuevo, no se observa un patrón marcado. Lo que sí se puede observar es que el año 2020 fue un gran año para la

empresa, y el crecimiento en altas tuvo un auge con la extensión de la cuarentena en los meses de abril y mayo.

**Figura 6**: Cantidad de compras realizadas por año

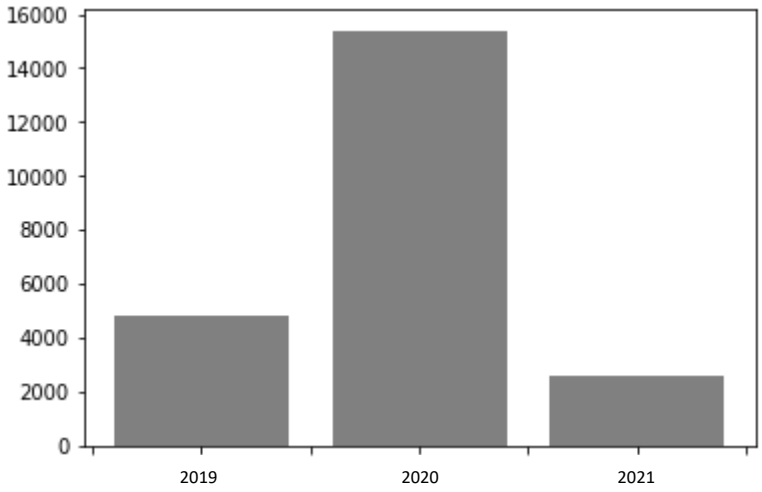
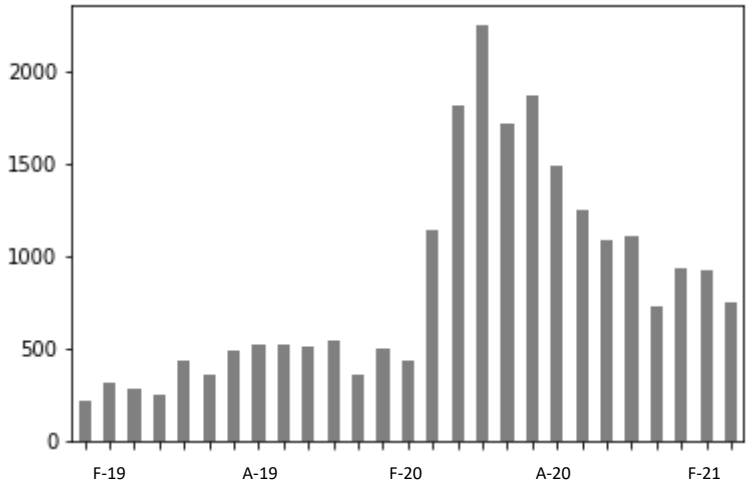



**Figura 7**: Cantidad de compras realizadas por mes

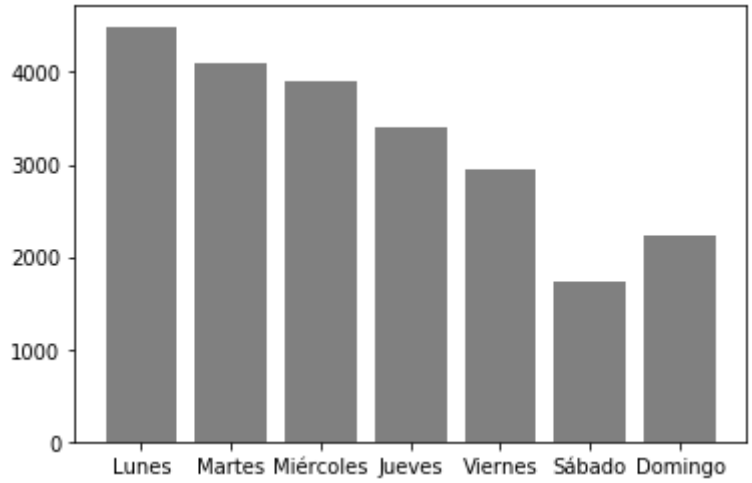



Para comprobar si existe cierta estacionalidad en el patrón de consumo, se calculó la cantidad de compras realizadas por mes por año para aquellos años para los cuales se obtuvo información.

No se observa ningún patrón para afirmar que existe estacionalidad. Se trata de una empresa relativamente nueva y continúa en etapa de crecimiento. Mucha gente todavía

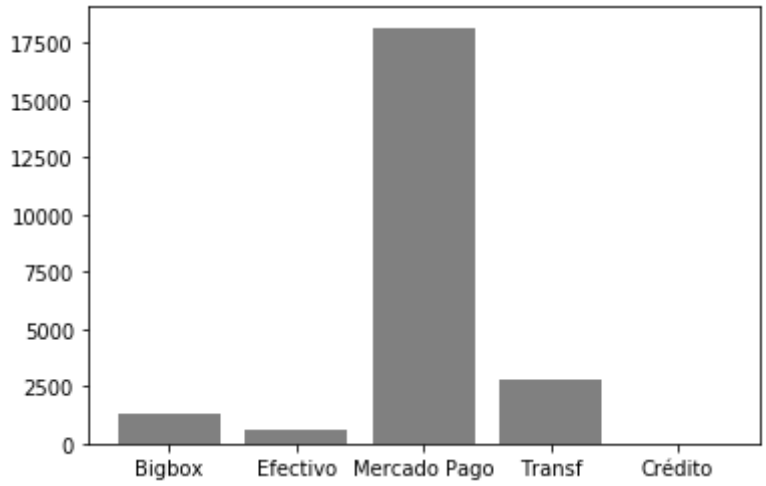
no conoce al emprendimiento, y es por esta razón que también se tratará de proporcionar información valiosa para que pueda ser usada por el equipo de Marketing.

**Figura 8**: Cantidad de compras realizadas según el día de la semana



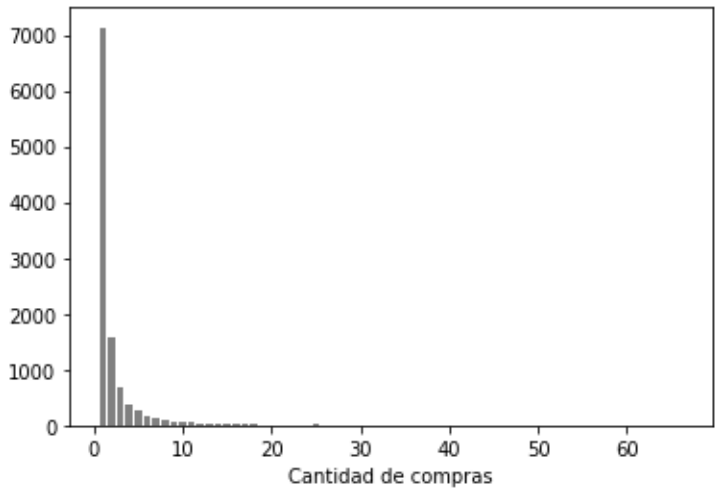

A partir de la información provista por la variable “Fecha de Ingreso”, se calculó el día de la semana en la cual se realizó cada compra para entender si hay una preferencia por parte de los clientes. Se puede observar que el caudal de pedidos llega a un pico los lunes y va disminuyendo a medida que avanza la semana, volviendo a crecer recién el domingo. Se puede atribuir a que los clientes querrán tener resuelta la compra antes de afrontar la semana, que es cuando menos tiempo de cocinar tienen.

**Figura 9**: Métodos de pago




Los métodos de pago ofrecidos por la empresa son diversos, pero el preferido por la mayoría de los clientes es “Mercado Pago”.

**Figura 10**: Frecuencia de compras

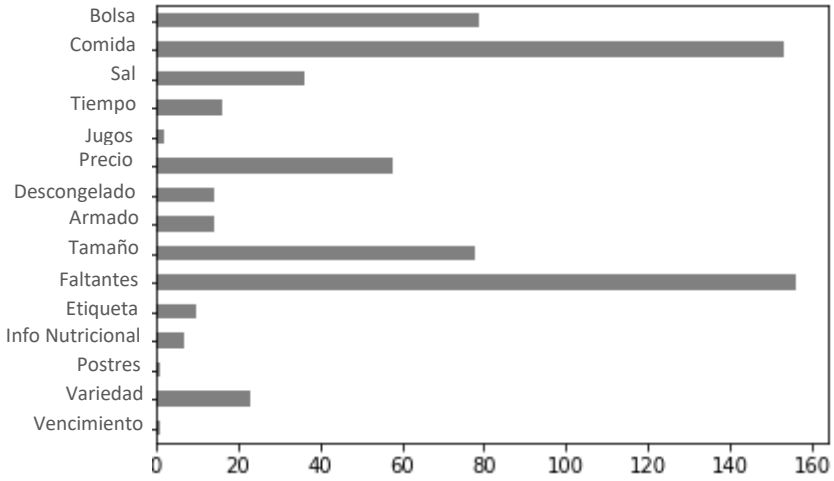



Se calculó la cantidad de compras realizadas por cada uno de los clientes para entender un poco más acerca de su patrón de consumo. Lo que se quiere ver es si la proporción de clientes que solo compran una vez y nunca más es significativa.

En la figura 10 se puede observar que 7.132 clientes, o el 66.3% de los clientes, compraron una única vez. Un total de 1.583 personas, o 14.7% de los clientes, realizaron únicamente dos compras. Sin embargo, cabe destacar que 961 clientes se dieron de alta en los tres meses correspondientes al año 2021 y solo realizaron una compra. Por lo tanto, algunos de los que se sumaron en enero podrían haber estado por hacer una segunda compra en marzo/abril. Lamentablemente, no se recibieron los datos y no se podrá comprobar. Por otro lado, el cliente más activo realizó un total de 66 compras hasta el 24 de marzo del 2021, última fecha para la cual se recibieron datos.

A partir de esta variable, se entiende que la empresa si tiene un problema a resolver y que su causa aún no está identificada. Con este trabajo no se apunta a identificar la causa del problema, pero sí a descubrir con tiempo qué clientes no volverán a comprar. De esta forma, se los podrá incentivar con alguna promoción para que continúen comprando y en el mejor de los casos, lograrán fidelizar consumidores que de otra forma hubiesen estado perdidos.

**Figura 11**: Tipo y cantidad de quejas recibidas




La cantidad de quejas recibidas luego de un total de 22.793 pedidos fue de 583, es decir que sólo el 2.5% de los pedidos tuvieron algún reclamo. Entre los tipos de reclamo se pueden encontrar razones como que la comida no es rica, que el precio es elevado, que los platos contienen mucha sal, que la bolsa llegó rota, que la etiqueta no es la correcta, entre otros.

Se buscó tener una dimensión de la cantidad de quejas de cada tipo, para entender cuáles son los puntos para mejorar. Dado que un cliente se puede quejar de varias cosas dentro de un mismo pedido, el total de la cantidad de quejas en los siguientes gráficos excederán al total de pedidos con quejas. Es decir, en caso de que un cliente se haya quejado de la comida y la sal, esa misma queja se contará dos veces, una vez en la columna “Comida” y otra vez en la columna “Sal”.

A partir del gráfico anterior, se puede observar que las quejas principales son por faltantes, es decir pedidos incompletos, por el sabor de la comida, por la bolsa rota, el tamaño y el precio.

La empresa debería tener en cuenta esta información. Que a los clientes no les guste la comida es un problema grave y a solucionar. Lo mismo sucede con el error al armar los envíos ya que a nadie le gusta pagar por productos que no reciben o recibir el producto incorrecto.

La proporción de quejas es muy baja con respecto a los pedidos totales. Sin embargo, no se puede confirmar que la razón por la cual los clientes que sólo compraron una única

vez, no haya sido alguna de estas razones. En ese caso, la empresa sí debería prestar más atención. Asimismo, podrían incorporar una encuesta de calidad, y seguramente recolectarían mucha más información de la que poseen actualmente.

# V. Modelos de aprendizaje supervisado y no supervisado

En un modelo de aprendizaje supervisado, para cada observación de los predictores $x_i$, $i$=1,…,$n$ existe una respuesta asociada $y_i$. Se quiere entrenar y ajustar un modelo que relacione la respuesta a los predictores, con el objetivo de predecir adecuadamente la respuesta de futuras observaciones (predicción) o entender la relación entre la variable respuesta y los predictores (inferencia).[4]

Por otro lado, en un modelo de aprendizaje no supervisado para cada observación $i$=1,…,n se observa un vector de mediciones $x_i$ pero no se cuenta con una variable respuesta $y_i$ que supervise el análisis.[5]

En este trabajo se utilizarán modelos de aprendizaje supervisado para predecir la variable "Churn" que se definió y luego se elegirá el modelo más preciso. Una vez concluido este análisis, se realizará otro análisis completamente distinto, en el cual se aplicarán modelos de aprendizaje no supervisado para entender si se pueden dividir en clusters a los clientes según su comportamiento al realizar compras. De esta forma, se podrá ver qué grupo de clientes es el más y menos activo y la empresa podrá elegir distintas formas de premiar, fidelizar o atraer a más clientes.

A partir de la información obtenida sobre los clusters, se aplicarán nuevamente modelos de aprendizaje supervisado para comprobar si es posible predecir a qué cluster pertenecerá un nuevo cliente con tan solo ver información sobre su primera compra, y en ese caso, tomar las medidas necesarias desde el inicio.

[4] An Introduction to Statistical Learning
[5] An Introduction to Statistical Learning

# VI. Churn

**Definición**

Una parte muy importante de Data Mining es el análisis de Churn. Se trata del cálculo de la tasa de abandono de clientes e identificación de los clientes que son más propensos a dejar de consumir *(Forhad & Rahman, 2014)*. Para predecir esta variable, se puede utilizar información histórica de compras que permitan identificar patrones. Para el caso de esta empresa, se definirá como Churn a aquel cliente que no haya vuelto a hacer un pedido por un período mayor a dos meses. Es decir, un mismo cliente puede ser considerado como Churn en un momento, pero como un cliente activo en otro momento.

Por ejemplo, si un cliente realizó una compra el 21 de marzo 2020 y su siguiente compra la realizó el 4 de abril 2020, entonces en la primera compra mencionada será considerado como Churn = 0, ya que el tiempo transcurrido entre compra y compra fue menor a dos meses. Si su siguiente compra luego del 4 de abril la realiza el 10 de octubre 2020, entonces en la compra del 4 de abril será considerado como Churn = 1 ya que pasaron seis meses entre compra y compra.

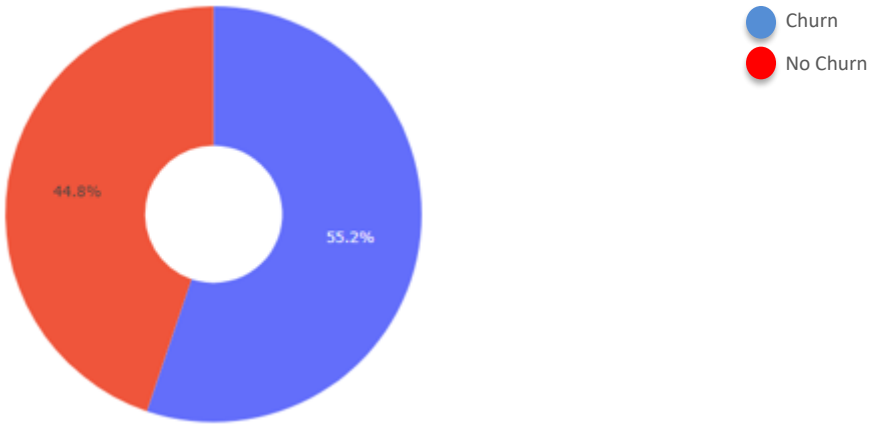
Como se recibieron datos hasta marzo 2021, se podrán utilizar los datos de las compras realizadas hasta enero 2021, ya que para estos últimos sabremos si efectivamente volvieron a hacer un pedido o no antes de marzo 2021. Contrariamente, no se podrán utilizar los datos de febrero y marzo 2021, ya que deberíamos tener información de abril y mayo 2021 respectivamente para confirmar si realizaron o no otra compra.

Siguiendo la lógica planteada anteriormente, se recorrió el conjunto de datos “Pedidos”, en el que se detalla la información de todos los pedidos realizados por cada cliente y la fecha en la que se realizó. En base a las compras de cada cliente, se definió la variable objetivo “Churn”, para luego entrenar modelos supervisados.

**Análisis exploratorio sobre la variable "Churn"**

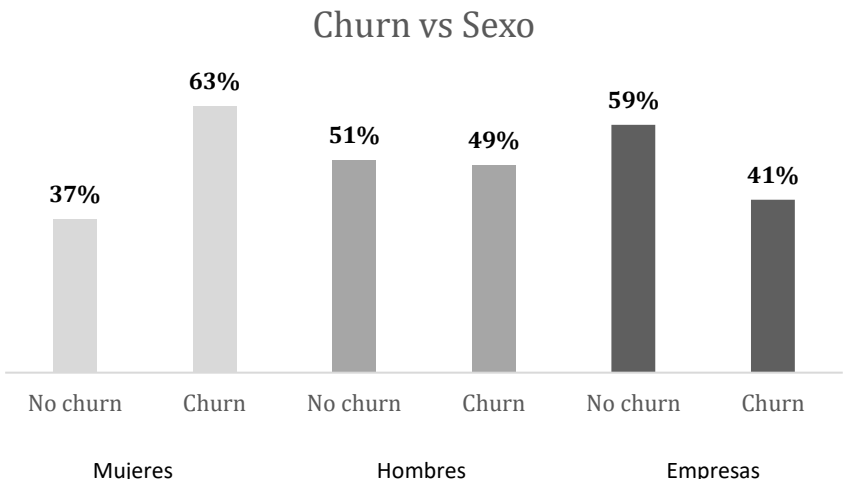
Como se mencionó anteriormente, si un cliente no realizó una compra dentro de los dos meses siguientes a la compra que se está analizando, entonces se lo considerará "Churn = 1". Por el contrario, los clientes que realizan compras dentro de los dos meses siguientes a la compra que se está analizando, serán considerados "Churn = 0". Es por esta razón que el cliente se analizará a nivel pedido, y no a nivel cliente. Por lo tanto, en vez de tener información para 10.752 clientes en los siguientes gráficos, se tendrá información para 22.793 pedidos.

**Figura 12**: Proporción de Churn



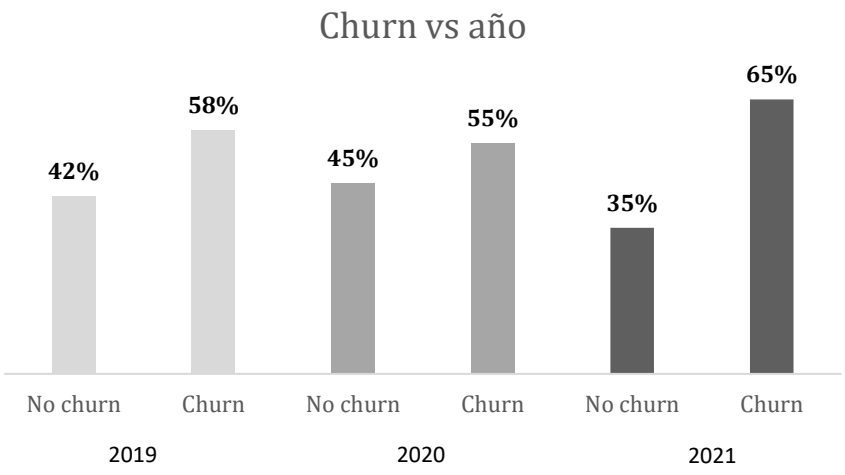

Una vez definida la variable, se puede observar que el conjunto de datos no está muy desbalanceado. En el 44.8% de los pedidos, es decir en 9.464 pedidos, los clientes se consideraron No Churn, mientras que el 55.2%, 11.654 pedidos, se consideraron Churn.

**Figura 13:** Churn vs Sexo



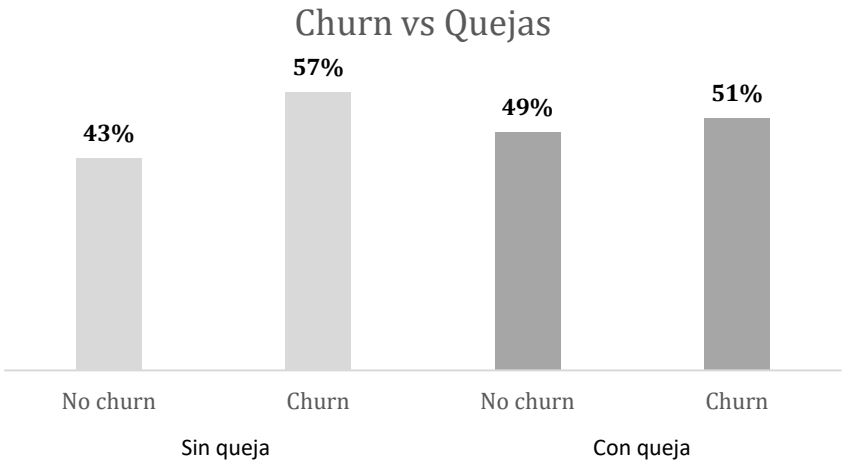
A partir del gráfico anterior, se puede observar que la proporción de Churn en los hombres es mucho más pareja que la proporción de Churn en las mujeres. En los hombres predominan los usuarios activos, mientras que en las mujeres predominan los abandonos. Asimismo, la cantidad de usuarios activos de sexo masculino supera a los usuarios activos de sexo femenino.

**Figura 14:** Churn vs Año




El comportamiento de la variable Churn parece ser estable en los dos años para los cuales se tiene información completa. Predominan los usuarios que dejan de comprar, pero no se trata de una diferencia abismal.

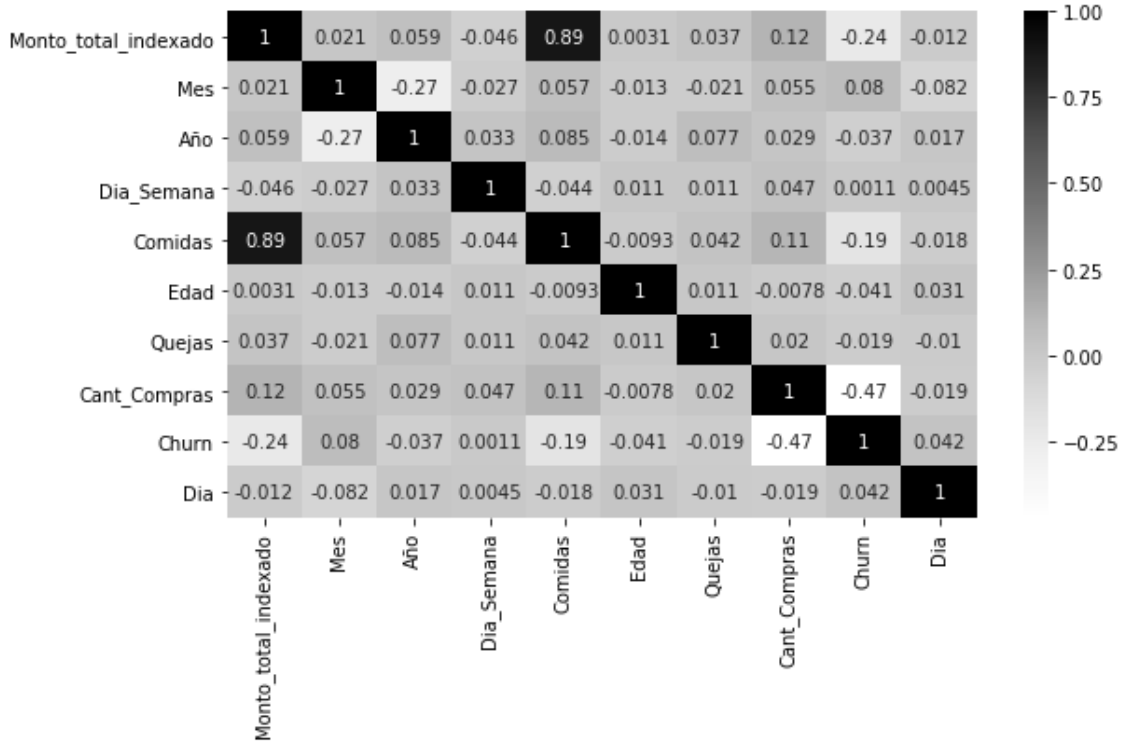
**Figura 15**: Churn vs Quejas



Si bien la cantidad de quejas es mínima, puede ser una clara razón para explicar a la variable Churn. Es de interés la variable “Quejas”, pero está sumamente desbalanceada, ya que solo 583 pedidos cuentan con una.

Como se mencionó en la sección anterior, no se puede asegurar que la cantidad de quejas sea la real, ya que muchos usuarios no compraron una segunda vez y se desconocen las razones al respecto.

**Figura 16**: Correlación de variables



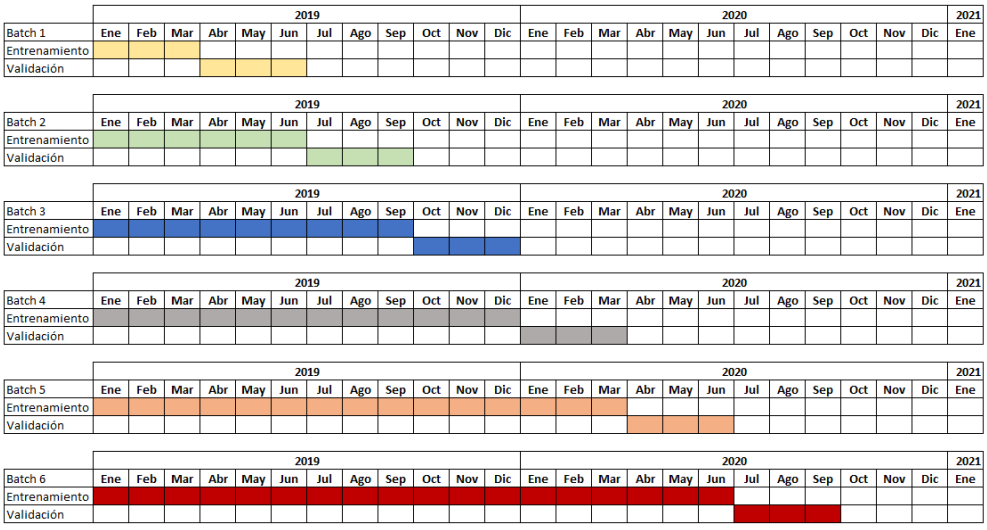

Se analizó la correlación entre las variables y se hizo foco en las más relevantes. Se puede observar que la variable “Churn” tiene una correlación negativa con las variables “Cant_Compras”, “Monto_total_indexado” y “Comidas”. Este resultado tiene sentido, ya que es de esperar que, si un cliente realiza muchas compras, significa que los productos le gustan, por lo tanto, es menos probable que deje de realizar compras. Lo mismo ocurre con el monto gastado y las comidas pedidas. Si el cliente pide muchas comidas, y por lo tanto aumenta el monto gastado, también quiere decir que los productos le gustan y disminuye la probabilidad de que se convierta en un cliente inactivo.

El resto de las variables no son tan significativas. Sin embargo, se utilizarán todas las variables para comenzar el análisis, y luego se seleccionarán algunas para mejorar los resultados. La variable “Edad” sólo se considerará cuando el modelo que se utilice pueda manejar correctamente los NA. En caso contrario, se quitará del conjunto de variables predictoras.

**Esquema de validación propuesto**

La validación cruzada es una metodología que se utiliza para escoger el mejor modelo, seleccionando los hiper parámetros correctos o eligiendo las variables más relevantes.[6] Como la definición planteada depende de una variable temporal, no se podrán separar los datos de manera totalmente aleatoria a los fines de conformar los grupos de entrenamiento, validación y testeo. Por lo tanto, se utilizarán los datos entre enero 2019 hasta septiembre 2020 para entrenar el modelo y los datos de octubre 2020 hasta enero 2021 para testear el modelo. Dentro del conjunto de entrenamiento se propone el esquema de validación cruzada *TimeSeriesSplit*, que en cada iteración divide el set de entrenamiento en dos grupos, respetando que el conjunto de validación siempre esté por delante del grupo de entrenamiento para respetar la variable temporal.[7]

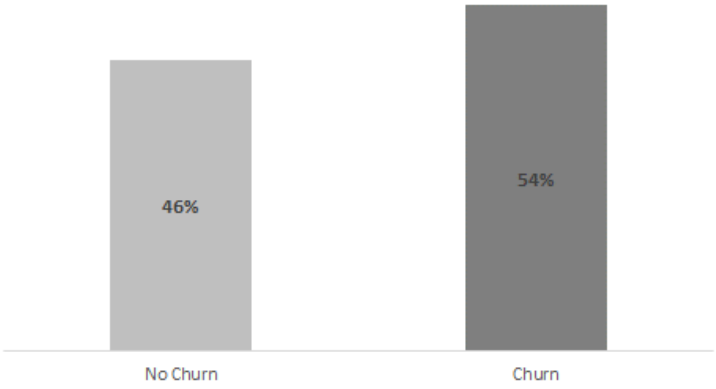
**Figura 17:** Esquema de validación cruzada




[6] https://hub.packtpub.com/cross-validation-strategies-for-time-series-forecasting-tutorial/
[7] https://hub.packtpub.com/cross-validation-strategies-for-time-series-forecasting-tutorial/

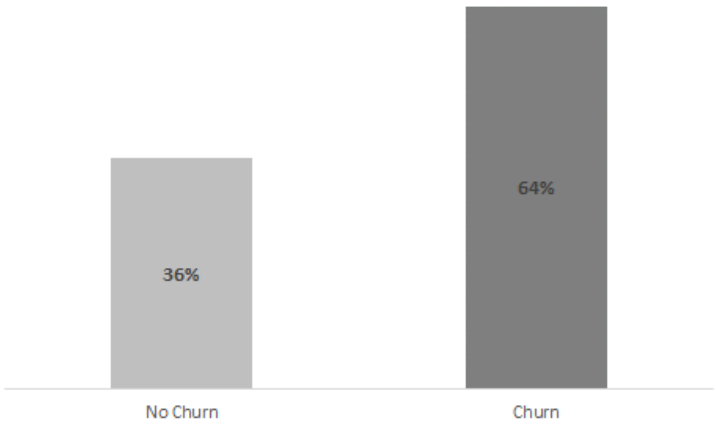
Teniendo en cuenta la definición de Churn, no se podrán utilizar los datos correspondientes a los meses de febrero y marzo del año 2021, ya que no se cuenta con la información acerca de los pedidos realizados en abril y mayo respectivamente. Por lo tanto, se utilizarán los datos desde enero 2019 hasta enero 2021.

**Figura 18**: Conjunto "*Train*": enero 2019 – septiembre 2020.




En total, este conjunto cuenta con 17.252 pedidos de los cuales el 54% son considerados "Churn" y el 46% son considerados "No Churn".

**Figura 19**: Conjunto "*Test*": octubre 2020 – enero 2021.




Cuenta con un total de 5.541 pedidos de los cuales el 64% fueron considerados "Churn" y el 36% fueron considerados "No Churn". Este conjunto cuenta con un desbalanceo notorio en los datos, mientras que el conjunto *Train* no tanto. Este desbalance puede generar un peor resultado en el modelo, pero lamentablemente no es un punto que se

pueda solucionar ya que en este caso se debe respetar la temporalidad de los datos. Este conjunto contiene información de los últimos meses del 2020 y el primer mes del 2021, período en el cual las restricciones por pandemia no eran tan fuertes como durante todo el 2020. Esta puede ser una razón para explicar el aumento de Churn.

## Modelos

En esta sección se comentarán los modelos que se utilizaron y los resultados que se obtuvieron. Se probaron distintos modelos y diferentes variantes de los mismos para luego poder comparar y elegir aquél que obtenga mejores resultados.

Para evaluar la performance de los modelos, se utilizará el área bajo la curva ROC (AUC por sus siglas en inglés). La curva ROC es una representación bidimensional del rendimiento del clasificador *(Fawcett, 2006)*. Una forma de reducir la performance a un único valor es calcular el área bajo la curva ROC. El AUC de un clasificador es equivalente a la probabilidad de que el clasificador posicione a una clase positiva elegida aleatoriamente por encima de una clase negativa elegida aleatoriamente *(Fawcett, 2006)*. Una clasificación aleatoria produce una línea diagonal entre (0,0) y (1,1), es decir un área de 0,5. Por lo tanto, un buen clasificador tendrá un valor AUC más cercano a 1 en el conjunto de *test*.

### I. Regresión Logística

La regresión logística es un modelo de aprendizaje automático y es de clasificación. Se lo utilizará para predecir la probabilidad de una variable binaria dependiente. Como se quiere encontrar a los clientes que no volverán a comprar, el caso de "éxito" definido es Churn = 1. Por lo tanto, el algoritmo nos ayudará a predecir $P(y_i = 1|x_i)$ como función de nuestras variables.

Para poder utilizar el algoritmo y entrenar el modelo, fue necesario transformar a todas las variables categóricas en numéricas, ya que la regresión logística necesita valores numéricos. Para lograrlo, se utilizó el método *One hot encoding*, uno de los esquemas de codificación más utilizados. El mismo consiste en transformar una variable categórica

con n observaciones y m valores distintos en m variables binarias con n observaciones cada una, indicando la presencia (1) o ausencia (0) de la variable binaria *(Potdar & Pai, 2017)*. Esta transformación se realizó en las variables "Método_pago", "Sexo", "Tipo" y "Mes_Año". En este caso, la transformación de variables categóricas es útil porque el modelo necesita variables numéricas. Asimismo, ayuda a que el modelo no asuma relaciones, en este caso inexistentes, entre las distintas categorías de una misma variable.

Las variables utilizadas para el modelo fueron las siguientes: *"Monto_total_indexado", "Dia_semana", "Método_pago", "Comidas", "Sexo", "Tipo", "Cant_compras", "Mes_Año".*

Se eliminaron las siguientes variables:

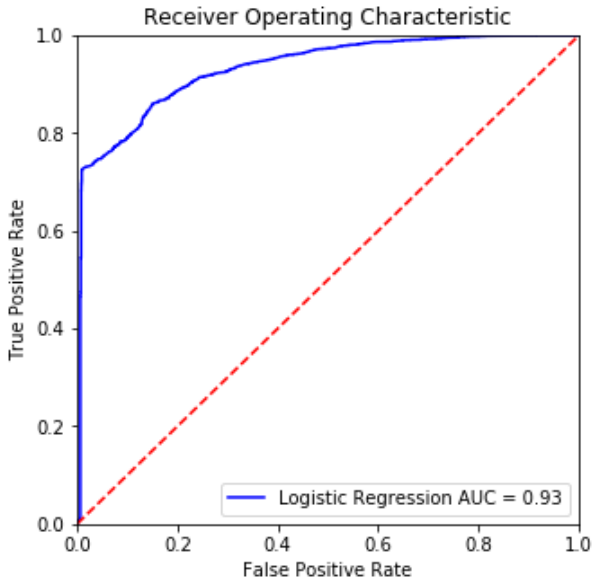
- "ID_Cliente" y "Num_pedido". No se considera que tengan valor para el modelo.
- "Fecha_Ingreso", "Mes" y "Año" ya que se creó la variable "Mes_Año" agrupando la información que se considera relevante de la fecha
- "Monto_total". Se utilizará el monto total indexado.
- "Edad". Contenía muchos "NA" y al no conocer el método para calcularlo, se decidió no tenerla en cuenta.

Al convertir las variables categóricas en variables *dummy*, se pasó de tener 8 variables a un total de 89. Las variables numéricas se escalaron para obtener mejores resultados. Se utilizó el método "Recursive Feature Elimination", el cual elimina las variables más débiles que no van a afectar la performance del modelo utilizado y mantiene aquellas variables que lo afectan positivamente. Es un proceso iterativo y comienza creando un modelo con la totalidad de las variables y le asigna un peso e importancia a cada una de ellas dependiendo el efecto que tenga esa variable sobre la variable a predecir. Luego elimina de a una las variable con menos importancia y recalcula los pesos e importancias de las variables restantes hasta alcanzar el rendimiento máximo del modelo *(Akkaya, 2021)*. Se utilizó este método y se seleccionaron las 20 variables más importantes para entrenar el modelo sobre los 6 grupos de datos planteados para realizar validación cruzada.

| | AUC |
|---|---|
| Batch 1 | 0,90 |
| Batch 2 | 0,88 |
| Batch 3 | 0,80 |
| Batch 4 | 0,89 |
| Batch 5 | 0,93 |
| Batch 6 | 0,91 |

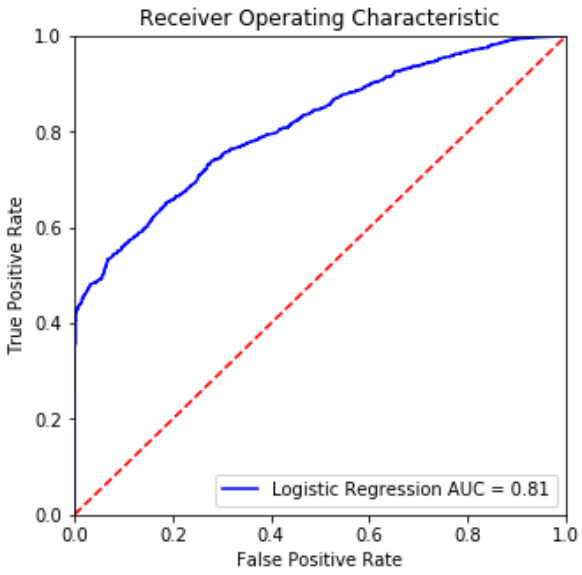
El conjunto que obtuvo un valor más alto de AUC fue el 5, es decir al utilizar los datos desde enero 2019 – Marzo 2020 como entrenamiento y abril 2020 – junio 2020 como validación.

**Figura 20:** Validación del *batch* 5 – AUC 0,93




Se utilizó el modelo entrenado con el *batch* 5 sobre los datos de testeo y se obtuvo un score de 0,81. Si bien el valor AUC baja considerablemente al predecir sobre datos del conjunto test, los resultados mejoraron respecto al AUC obtenido en un escenario donde el conjunto de validación se dejó fijo en el período junio 2020 – septiembre 2020.

**Figura 21**: Test – AUC: 0,81




## II. XGBoost

XGBoost (*eXtreme Gradient Boosting*) es un algoritmo de aprendizaje automático que produce una serie de árboles simples que son utilizados para asignarle un valor a los nodos "hojas" cuando genera la división *(Chen, Liu & Zhang, 2018)*. En XGBoost, los árboles individuales se crean utilizando múltiples núcleos y los datos se organizan para minimizar los tiempos de búsqueda, disminuyendo de esta forma el tiempo de entrenamiento y aumentando el rendimiento de los modelos *(Santhanam, Raman, Uzir & Banerjee, 2017)*.

Algunas de las ventajas del algoritmo son las siguientes *(Espinosa-Zúñiga, 2020)*:

- Puede manejar grandes conjuntos de datos con múltiples variables
- Puede manejar valores perdidos
- Sus resultados son muy precisos
- Tiene una gran velocidad de ejecución

Por otro lado, se encuentran las siguientes desventajas *(Espinosa-Zúñiga, 2020)*:

- Puede consumir muchos recursos computacionales
- Se deben ajustar correctamente los parámetros del algoritmo para minimizar el error de precisión
- Solo trabaja con vectores numéricos

Se entrenaron seis modelos con los datos de los distintos conjuntos *train* propuestos para realizar validación cruzada y se evaluó la precisión de cada modelo sobre los datos de validación, pidiéndole al modelo que calcule la probabilidad de pertenecer a cada una de las clases.

Se realizó *random search CV* sobre cada conjunto, un enfoque que toma distintos valores de hiper parámetros de forma aleatoria dentro de una grilla establecida para encontrar una buena combinación de hiper parámetros y de esta forma, aumentar el *score*.[8] Los hiper parámetros buscados fueron los siguientes:

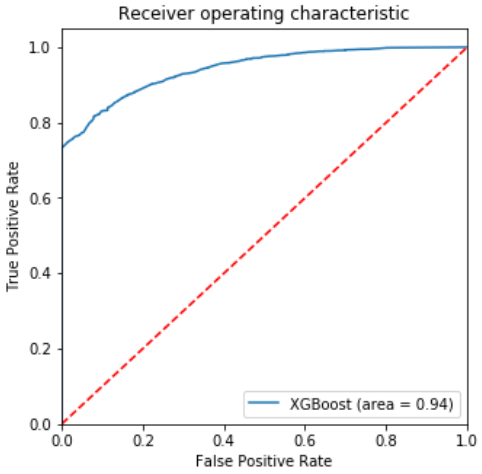
- Min_child_weight: Es el mínimo peso requerido para poder crear un nuevo nodo en el árbol. Cuanto más chico este parámetro, menos muestras se permitirán en el nodo, por lo tanto, se crearán árboles más complejos, y aumentará la probabilidad de hacer *overfitting*. El sobreajuste u *overfitting* se produce cuando el modelo tiene buena *performance* sobre los datos de entrenamiento, pero no puede predecir correctamente sobre datos que no conoce, es decir, sobre el conjunto de *test*. El modelo memoriza la data de entrenamiento en vez de aprender la lógica detrás de los datos *(Ying, 2019)*.
- Gamma: Corresponde a la mínima mejora permitida para generar una nueva partición en la hoja de un árbol. Cuanto más grande es, más conservador será el árbol.
- Learning_rate: Indica la velocidad a la que aprende el modelo.
- Subsample: Corresponde a la cantidad de observaciones que se considerarán en cada uno de los pasos. Si se define este parámetro como 1, entonces se utilizarán todas las observaciones del conjunto de datos.
- Colsample_bytree: Corresponde a la cantidad de características o variables que se utilizarán. Nuevamente, si se define como 1, entonces se utilizarán todas las columnas.
- Max_depth: Es la cantidad máxima de nodos permitidos desde la raíz hasta el final del árbol. Cuanto más profundo es el árbol, el modelo podrá generar relaciones más complejas al añadir más nodos. Sin embargo, a partir de cierto punto, la profundidad solo generará ruido y las relaciones serán menos relevantes, y aumentará la probabilidad de hacer *overfitting*.

Se obtuvieron los siguientes resultados en los distintos conjuntos de validación, utilizando los mejores hiper parámetros encontrados en cada caso:

[8] https://towardsdatascience.com/machine-learning-gridsearchcv-randomizedsearchcv-d36b89231b10

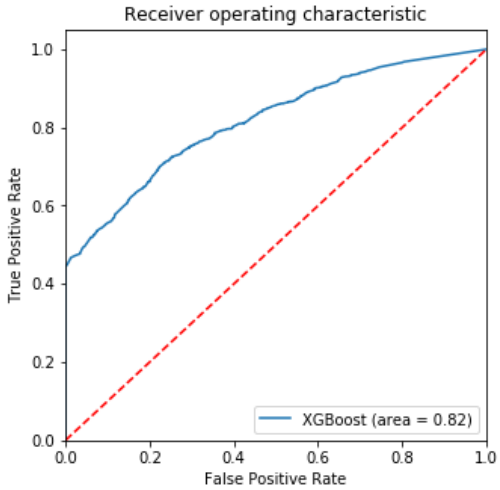
| | AUC |
|---|---|
| Batch 1 | 0,88 |
| Batch 2 | 0,88 |
| Batch 3 | 0,81 |
| Batch 4 | 0,89 |
| Batch 5 | 0,94 |
| Batch 6 | 0,92 |

**Figura 22:** Predicción de probabilidades de pertenecer a cada clase en el *batch* 5 de validación – AUC: 0,94




Se utilizó el modelo entrenado con el *batch* 5 sobre los datos de testeo y se obtuvo un score de 0,82. Nuevamente se puede observar una caída significativa entre las predicciones sobre el conjunto de validación y sobre el conjunto test.

**Figura 23**: Predicción de probabilidades de pertenecer a cada clase en test – AUC: 0,82



### III. Random Forest

El algoritmo de Random Forest utiliza el enfoque “Divide y vencerás”. Genera múltiples árboles de decisión sobre un conjunto de datos de entrenamiento, eligiendo de forma aleatoria las variables a utilizar. Los resultados de todos los árboles se combinan para obtener un único modelo más robusto a utilizar sobre todos los datos de entrenamiento. Los resultados se combinan con el fin de obtener un único modelo más robusto

Alguna de las ventajas del algoritmo son las siguientes *(Ali, Khan & Ahman, 2012)*:

- Reduce el problema de sobreajuste
- Puede manejar miles de variables e identificar a las más importantes
- Puede manejar bien los valores atípicos

Por otro lado, cuenta con las siguientes desventajas:

- Pérdida de interpretación
- Poco control sobre lo que hace el modelo (caja negra)
- Requiere mucha más potencia y recursos computacionales

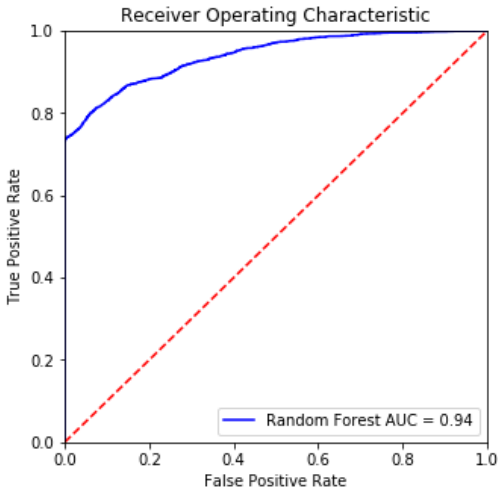
Nuevamente, se entrenó el modelo con los distintos grupos de datos de entrenamiento y se evaluó la precisión del modelo sobre los distintos grupos de datos de *validación*. Se realizó *random search CV* para encontrar una buena combinación de hiper parámetros sobre cada conjunto y de esta forma, aumentar el *score*. En este caso, los hiper parámetros buscados fueron los siguientes:

- N_estimators: Es el número de árboles que se van a construir antes de tomar decisiones. Cuanto más alto es este número, mejor performance va a tener el modelo, pero va a ser más lento.
- Max_features: Corresponde al máximo número de variables que el algoritmo puede probar en cada árbol.
- Max_depth: Indica la profundidad de los árboles. Cuanto más alto es, más divisiones tienen los árboles y pueden capturar más información.
- Min_samples_split: Representa el mínimo de observaciones requeridas para poder crear una nueva división de nodos.
- Min_samples_leaf: Corresponde a la cantidad mínima de observaciones que tiene que haber en cada una de las hojas de los árboles.
- Bootstrap: Al entrenar el modelo, cada árbol aprende sobre un conjunto de datos random, y ese conjunto de datos se crea aleatoriamente. Es decir que muchos árboles pueden trabajar con una misma observación en caso de que este hiper parámetro este activado.

Se obtuvieron los siguientes resultados sobre los distintos grupos de validación:

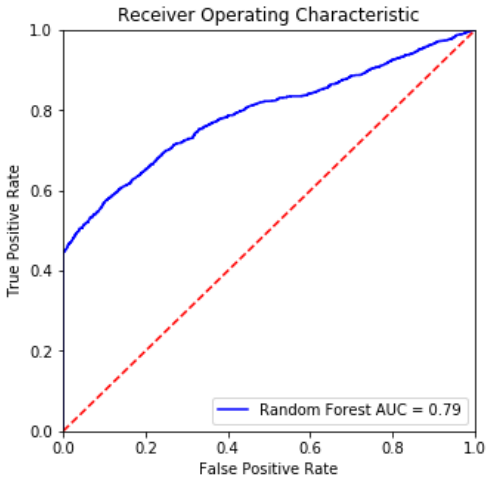
| | AUC |
|---|---|
| Batch 1 | 0,89 |
| Batch 2 | 0,86 |
| Batch 3 | 0,80 |
| Batch 4 | 0,89 |
| Batch 5 | 0,94 |
| Batch 6 | 0,92 |

**Figura 24:** Predicción de probabilidades de pertenecer a cada clase en el *batch* 5 de validación – AUC: 0,94



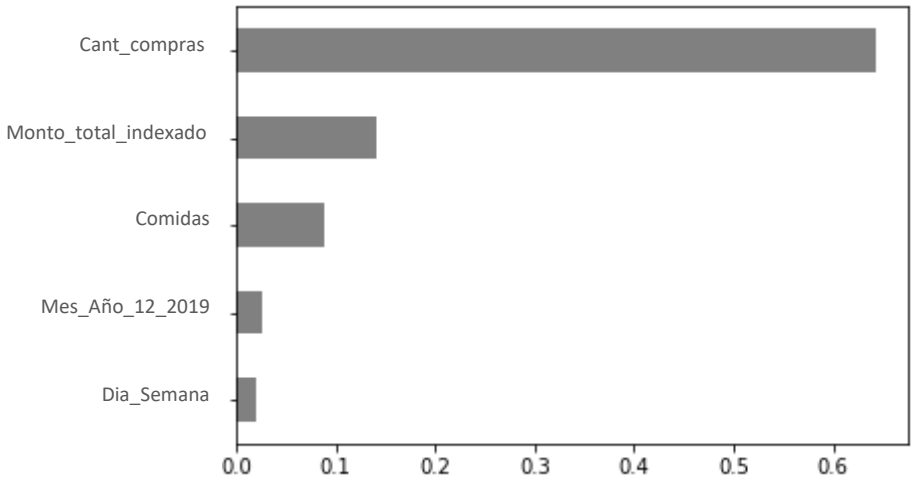

Nuevamente se utilizó el modelo entrenado con el *batch* 5 ya que fue el que obtuvo mejor resultado y se obtuvo un AUC de 0,79 en test.

**Figura 25:** Predicción de probabilidades de pertenecer a cada clase en test – AUC: 0,79



**Interpretaciones, resultados y conclusiones del modelo de Churn**

**Figura 26:** Las variables más importantes utilizadas por el modelo




Las variables más relevantes para predecir la variable "Churn" utilizadas por el algoritmo de "Random Forest" fueron las mencionadas en la figura 26. Si se suma la importancia de las tres primeras, se obtiene el 86% de la importancia total (0,64, 0,14 y 0,08 respectivamente). Este resultado era de esperar, ya que eran las tres variables que más correlación tenían contra la variable a predecir (Figura 16). Asimismo, tiene sentido ya que las tres variables son un claro indicador de qué tanto les gusta el producto a los clientes. Cuantas más compras realice, más platos pida y por ende más dinero gaste, la probabilidad de dejar de comprar disminuirá. Por el contrario, cuanto menos realice alguna de estas acciones, mayor será la probabilidad de dejar de comprar.

Afortunadamente, todos los modelos que se corrieron arrojaron muy buenos resultados al realizar predicciones sobre los datos que fueron separados para testear. Sin embargo, en los tres modelos presentados se puede observar una caída significativa en el *score* AUC entre las predicciones realizadas sobre datos de validación y sobre datos de testeo.

| AUC | Validación | Test |
|---|---|---|
| Regresión logística | 0,93 | 0,81 |
| XGBoost | 0,94 | 0,82 |
| Random Forest | 0,94 | 0,79 |

La caída observada se genera cuando el modelo se sobre ajusta a los datos de entrenamiento y luego al predecir sobre datos desconocidos, no tiene tan buena *performance* como en datos de validación. El claro desbalance que se observa en la variable a predecir del conjunto *test* puede estar afectando negativamente a los resultados. Esto puede representar un problema a la hora de generalizar el modelo propuesto. En este trabajo se propuso un esquema de validación para disminuir este sobre ajuste, pero en trabajos futuros podrían probarse otros esquemas compatibles con datos que dependan de una variable temporal y mejorar de esta forma el score.

El modelo seleccionado será el de XGBoost, ya que es el modelo que obtuvo un mejor *score* sobre los datos de testeo. Utilizando un algoritmo de este estilo, la empresa tendrá un 82% de probabilidad de asignarle un valor más alto a una observación aleatoria cuando la misma sea positiva, y un valor más bajo cuando la misma sea negativa. De esta forma la empresa podrá tener un mayor entendimiento acerca del comportamiento de sus clientes y tomar medidas de forma proactiva y no reactiva para lograr que el cliente vuelva a comprar.

El análisis de Churn sumado al análisis de segmentación de clientes que se detallará en la próxima sección debería ofrecer información muy valiosa acerca del momento particular en el que se encuentra la relación entre el cliente y la empresa. Para cada momento de la relación se deberían aplicar técnicas distintas para retener al cliente. A su vez, es muy importante definir a qué clientes se quiere retener y para los cuales se invertirá esfuerzo y dinero en hacerlo.

Para aquellos que sólo hayan realizado una compra, una sugerencia podría ser ofrecer algún tipo de descuento o beneficio que lo atraiga nuevamente. Los descuentos suelen ser atractivos, pero también se pueden ofrecer muestras gratis para aumentar la exposición de la variedad de productos ofrecidos y lograr que los mismos sean probados por más gente. Por otro lado, a clientes más maduros se le pueden ofrecer beneficios para premiar su fidelidad. Algunas ideas pueden ser crear un sistema de puntos y premios tales como envío gratis, incluir alguna comida bonificada, etc. Esto no sólo beneficiará a aquellos que compran seguido ya que lograrán los objetivos de forma

rápida, sino que incentivará a que aquellos que compran esporádicamente, lo hagan más seguido.

Más allá de los beneficios a ofrecer y las estrategias de marketing a seguir, es muy importante que la empresa invierta en entender por qué tantos usuarios no vuelven a comprar luego de realizar su primera compra. Hacer un seguimiento acerca de la experiencia del usuario y pedir feedback luego de la compra puede ser un paso importante a seguir. Si bien las quejas recibidas por escrito son pocas respecto a la cantidad de ventas realizadas, no se deben tomar a la ligera. Se podría definir un plan de acción a realizar cuando se reciba una queja y de esta forma tratar de no perder al cliente en cuestión. A su vez, la empresa debería analizar si la queja se trata de un caso aislado o si es una tendencia entre otros consumidores y tomarlo como un punto a mejorar.

# VII. Clusters

Actualmente las preferencias de los consumidores se modifican continuamente, y las empresas deberían estar preparadas, o incluso adelantarse a estos cambios para no quedarse atrás en un mercado sumamente competitivo. Dado que los clientes leales son el activo más importante de una empresa, las mismas han puesto más atención en el desarrollo de programas de fidelización y retención de clientes *(Nasir, 2017)*. Un primer paso hacia este camino podría ser el entendimiento de qué tipo de consumidores compone el conjunto de clientes de una empresa.

La segmentación de clientes es el proceso de división del conjunto de datos de clientes en varios grupos (o clusters), de manera que cada cluster consista en clientes que cuenten con características similares *(Kushawaha & Prajapati, 2008)*. Estas características pueden ser diversas, entre ellas sexo, edad, intereses o comportamiento de consumo. La segmentación de los consumidores es muy importante para las empresas ya que al hacerlo, tendrían un mayor entendimiento acerca de las necesidades y patrón de consumo de los individuos que componen cada cluster. De esta forma, podrían satisfacer las necesidades de forma precisa y entenderán de qué manera podrían atraerlos y generar fidelidad. Se utilizará en conjunto el modelo RFM para analizar el patrón de consumo de los clientes y el algoritmo K-means para crear los potenciales clusters.

**Algoritmo K-means**

El algoritmo K-means es un método que trata de agrupar los datos en K grupos que cumplan con cierto criterio, minimizando la suma de distancias entre cada observación y el centroide de su cluster *(Li & Wu, 2012)*. La idea es minimizar la varianza dentro de cada cluster y maximizar la varianza entre cada cluster.

**Modelo RFM**

El modelo RFM (cuyas siglas corresponden a Recency, Frequency, Monetary), es una técnica empleada para evaluar a los consumidores basándose en su patrón de consumo *(Christy, Umamakeswari, Priyatharsini & Neyaa, 2018)*. Se ha demostrado que el modelo RFM fue exitoso en varias áreas y que ayuda a identificar a los consumidores valiosos y desarrollar estrategias de marketing efectivas *(Wei, Lin & Wu, 2010)*. De esta forma, se pueden reconocer a los clientes más y menos activos y tomar medidas al respecto.

- Recency: Mide el tiempo transcurrido desde la última compra. En este caso, se calculó el tiempo transcurrido desde la última compra realizada por cada uno de los clientes hasta el 24 de marzo 2021, ya que no contamos con los datos posteriores a la fecha mencionada.
- Frequency: Frecuencia de compras de cada consumidor
- Monetary: Total gastado por cada consumidor

La metodología utilizada fue la de calcular el valor de las tres variables RFM para cada consumidor, teniendo en cuenta la información agregada de sus compras. A partir de los valores obtenidos, se corrió un modelo de K-means sobre cada variable, especificando un número total de k=3 clusters y de esta forma, se le asignó un numero de cluster a cada variable y a cada individuo. Los clusters van de 0 a 2, donde 0 representa un cluster inactivo, mientras que 2 representa un cluster activo. Esta graduación se respeta para las tres variables. Es decir, si un individuo obtuvo R=1, F=2 y M= 0, significa que la última compra del cliente fue hace un tiempo intermedio, que el cliente suele comprar con frecuencia alta y que no suele gastar mucho en sus compras. El mejor escenario es un cliente que obtuvo 2 en cada uno de los clusters.

Una vez obtenidos los clusters para cada variable y para cada cliente, se creó la variable "Overall Score", que suma los valores de los clusters individuales. Es decir, en el ejemplo anterior, el individuo tendría un Overall Score de 3 (R=1 + F=2 + M=0). De esta forma, se pasa de tener 3 clusters, a 6. Se puede realizar de esta forma dado que la numeración del cluster indica lo mismo para las tres variables, ya que el cluster 2 es el más valorado y el cluster 0 es menos valorado. Para simplificar el análisis, se agruparon los 6 clusters

finales en 3 a partir de los datos promedios de R, F y M, y ese cluster final es el que define en qué grupo se encuentra cada cliente.

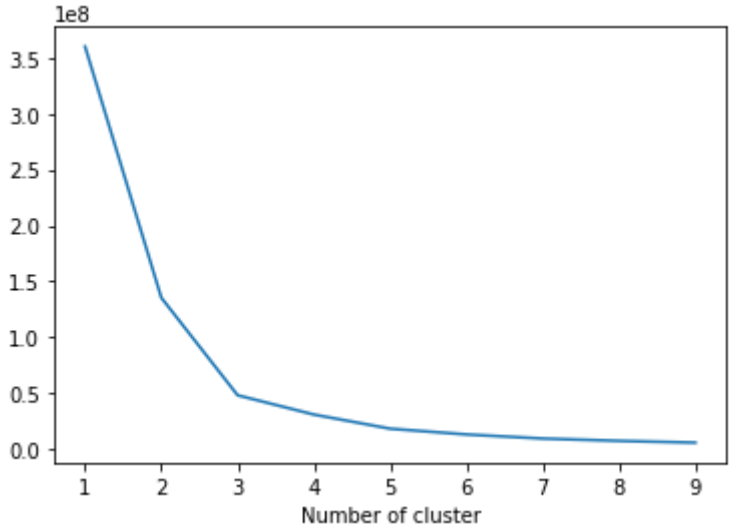
**Recency**

```
count    10752.000000
mean       284.454520
std        183.011124
min          0.000000
25%        153.000000
50%        276.000000
75%        355.000000
max        812.000000
Name: Recency, dtype: float64
```

Se cuenta con 10.752 clientes y en promedio pasaron 284 días desde la última compra. Como se mencionó anteriormente, lo ideal es que los clientes vuelvan a comprar dentro de dos meses, por lo tanto, es evidente que la empresa tiene un problema a solucionar, ya que ese promedio debería ser mucho menor.

Se corrió el algoritmo K-means probando distintos números de grupos k para observar cuál es la cantidad correcta de grupos a utilizar. Se utilizó el método del codo para seleccionar k. A partir del siguiente gráfico, se puede observar que k=3 podría ser una buena elección.

**Figura 27:** Elbow




Por lo tanto, se corrió el algoritmo sobre la variable “Recency” utilizando k=3 grupos y se obtuvieron los siguientes resultados:

| RecencyCluster | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| 0 | 1787.0 | 603.388920 | 95.341987 | 453.0 | 521.0 | 590.0 | 672.00 | 812.0 |
| 1 | 5556.0 | 301.070914 | 58.465225 | 196.0 | 258.0 | 306.0 | 337.25 | 452.0 |
| 2 | 3409.0 | 90.187445 | 61.524818 | 0.0 | 32.0 | 82.0 | 144.00 | 195.0 |

El cluster más activo es el N°2, compuesto por 3.409 clientes que en promedio realizaron su última compra hace 90 días. Luego se pueden identificar dos clusters bastante inactivos, pero uno más que el otro. El cluster N°1 se compone de 5.556 clientes y en promedio hicieron su última compra hace 301 días. Si bien es mucho tiempo, es menos de un año y la empresa podría reactivarlos. En cambio, los 1.787 clientes que componen el cluster N°0 en promedio realizaron su última compra hace 603 días, es decir más de un año y medio, y seguramente va a ser más difícil lograr que realicen una nueva compra.

**Frequency**

```
count    10752.000000
mean         2.119885
std          2.999682
min          1.000000
25%          1.000000
50%          1.000000
75%          2.000000
max         66.000000
Name: Frequency, dtype: float64
```

Observando la frecuencia de las compras realizadas, queda claro que el gran problema de la empresa es que la gran parte de los clientes que compran una vez, no lo vuelven a hacer. Este es el caso para el 50% - 75% de los clientes.

No es el objetivo de este trabajo identificar la razón de este comportamiento, pero sí la de llamar la atención al respecto y proporcionar información acerca de los grupos existentes para que se puedan tomar acciones enfocadas a cada grupo de clientes y, en el mejor de los casos, revertir esta situación.

Nuevamente se corrió el algoritmo K-means utilizando k=3 grupos y se obtuvieron los siguientes resultados:

| FrequencyCluster | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| 0 | 9769.0 | 1.416010 | 0.781998 | 1.0 | 1.0 | 1.0 | 2.0 | 4.0 |
| 1 | 868.0 | 7.345622 | 2.437290 | 5.0 | 5.0 | 7.0 | 9.0 | 14.0 |
| 2 | 115.0 | 22.469565 | 9.104001 | 15.0 | 16.5 | 19.0 | 25.0 | 66.0 |

El cluster N°2 es el más activo, ya que los clientes que lo componen realizaron en promedio 22 compras. Lamentablemente, solo 115 personas pertenecen a este grupo. Por otro lado, observamos al cluster N°1 que si bien no es tan activo como el N°2, los clientes realizaron en promedio 7 compras. Seguramente se puedan tomar acciones para que eventualmente se muevan hacia el cluster activo, ya que evidentemente los productos les gustan y por eso compraron más de una vez. Por último, el cluster N°0 es el más problemático, y dentro de él seguramente se encuentren algunos clientes que pertenecen al cluster N°1 en Recency, y no es muy tarde para reactivarlos.

**Monetary**

Como se mencionó en secciones anteriores, se indexó la variable “Monto_total” para poder comparar de forma adecuada los montos gastados por cada cliente. Por lo tanto, asumiendo que la empresa ajustó sus precios en base a la inflación mensual publicada por el INDEC, se obtuvo el valor total que el cliente hubiese gastado en valores de enero 2019.

```
count     10752.000000
mean       4144.630110
std        8069.288168
min          14.000000
25%         514.117365
50%        1490.046705
75%        3921.955032
max      145364.222999
Name: Monetary, dtype: float64
```

En promedio, el monto total gastado por los clientes considerando todas sus compras realizadas es de $4.144.

Luego de correr el algoritmo nuevamente con k=3 grupos, se obtuvieron los siguientes resultados:

| MonetaryCluster | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| 0 | 9545.0 | 1973.656166 | 2017.301263 | 14.000000 | 477.976215 | 1353.687444 | 2766.587496 | 9275.433053 |
| 1 | 1042.0 | 16634.671974 | 6212.196323 | 9328.370580 | 11581.404204 | 14854.814183 | 20149.819791 | 33648.552099 |
| 2 | 165.0 | 50855.676595 | 18055.504015 | 33843.447496 | 38325.867626 | 45280.652759 | 57186.622039 | 145364.222999 |

Se puede observar la gran diferencia que hay en cuanto al monto total gastado entre los tres clusters obtenidos. Por un lado, se encuentran los clientes del cluster N°2 que gastaron una suma alta, en promedio $50.855, pero lamentablemente solo 165 clientes pertenecen a este cluster. Se puede percibir cierta relación con la variable “Frecuencia”. Por otro lado, se encuentran los clientes que componen el Cluster N°1, que en promedio gastaron $16.634. Si bien no es tanto como los del cluster N°2, es mucho más que lo que gastaron en promedio los clientes del cluster N°0.

**Overall Score**

A partir de los clusters obtenidos en las tres variables que se detallaron, se sumó el valor de los clusters obtenidos para crear un resultado final. Por lo tanto, se pasó de tener 3 clusters por variable a tener 6 en total, ya que, en el mejor de los casos, el cliente pertenece al cluster 2 en las 3 variables y, por lo tanto, obtuvo un valor de cluster final de 6. Se tomó el promedio de los valores de Recency, Frequency y Monetary para cada uno de los 6 clusters y se obtuvieron los siguientes resultados:

| OverallScore | Recency | Frequency | Monetary |
|---|---|---|---|
| 0 | 604.754376 | 1.308635 | 1894.076989 |
| 1 | 304.249329 | 1.318565 | 1653.607268 |
| 2 | 112.068864 | 1.670330 | 3231.207633 |
| 3 | 170.865639 | 5.319383 | 14099.153079 |
| 4 | 83.432609 | 7.469565 | 19739.493427 |
| 5 | 89.973214 | 14.473214 | 37930.886593 |
| 6 | 52.544118 | 24.161765 | 60942.177553 |

Se decidió agrupar estos resultados en tres clusters, ya que se inició el análisis con esos mismos grupos, y se le puede encontrar una interpretación para el negocio.

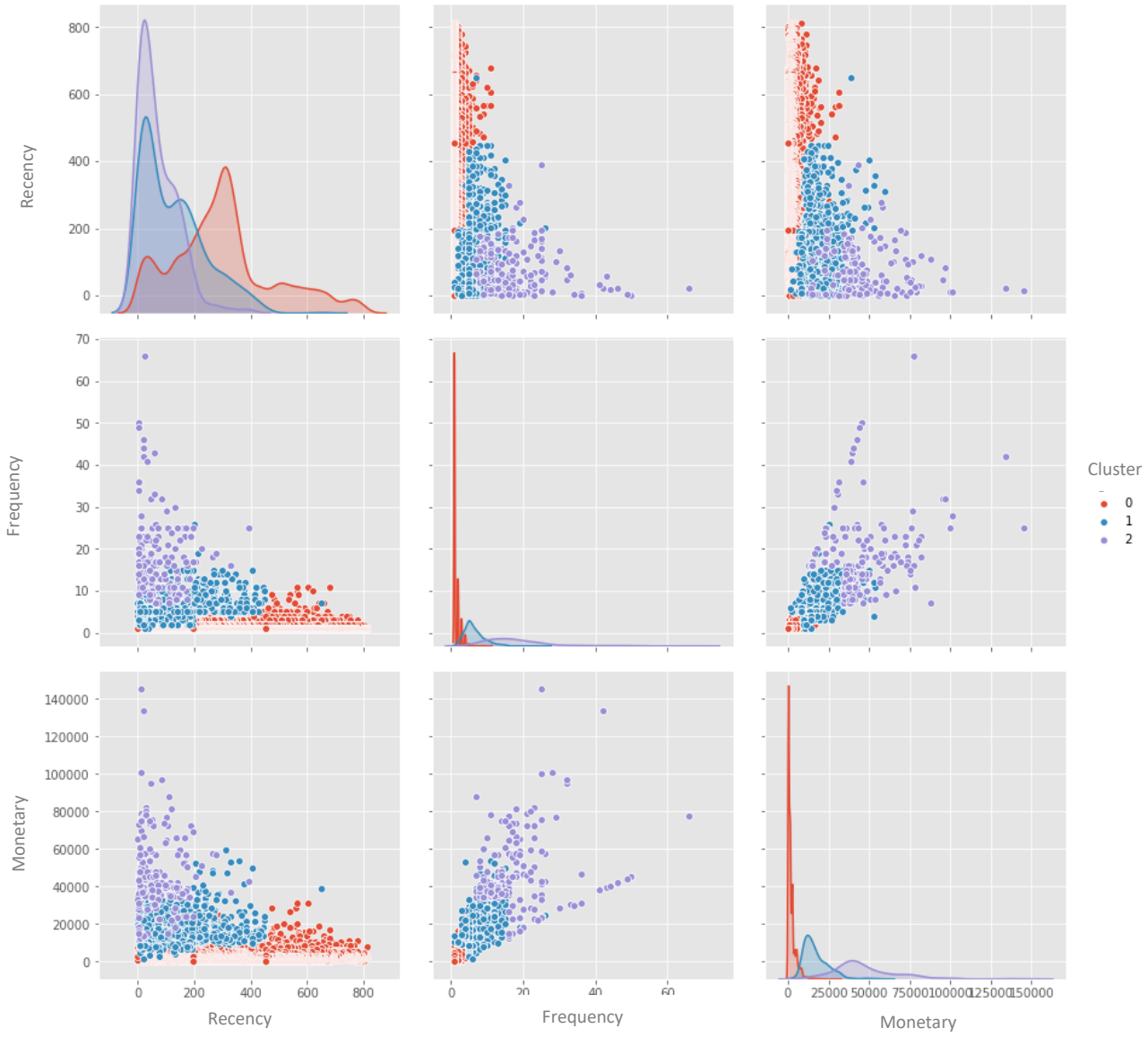
Los clientes que pertenecen a los clusters N°0, N°1 y N°2 van a representar a los clientes inactivos. Se puede ver que estos individuos realizaron su última compra hace un tiempo ya, pero lo que más determina la división elegida es que en promedio realizaron solo una compra. Por lo tanto, el monto gastado también es bajo.

Por otro lado, los clientes que pertenecen al cluster N°3 y N°4 serán considerados como los clientes intermedios, que no están completamente perdidos. Ellos compraron más de una vez y por lo tanto su monto gastado no es tan bajo. Esto quiere decir que los productos son de su agrado y va a ser más fácil recuperarlos que a los mencionados anteriormente.

Por último, los clientes que componen el cluster N°5 y N°6 serán considerados como los clientes activos. En promedio, su última compra fue realizada entre uno y dos meses atrás y la cantidad de compras realizadas es alta, al igual que el monto gastado.

A partir de este análisis, la empresa contará con más información acerca de sus clientes. Podrán tomar decisiones personalizadas para cada cluster, ya que no todos necesitan el mismo incentivo. Seguramente podrán premiar al cluster más activo para que continúen comprando e intentarán reactivar al cluster que no está completamente perdido.

**Figura 28:** Distribuciones bivariadas entre las variables Recency, Frequency y Monetary




**Panel 1**: Las tres figuras superiores

En la figura superior izquierda se puede observar la distribución de los clientes para la variable “Recency”. El cluster N°2 es el más activo, y por lo tanto la curva de distribución es más angosta, es decir que no pasó mucho tiempo desde su última compra. Por el contrario, el cluster N°0 se extiende hacia la derecha, indicando que pasó mucho tiempo desde la última compra.

En la figura superior central, se muestra la división de clusters cuando se tiene en cuenta a las variables “Recency” y “Frequency” en conjunto. El cluster N°0 alcanza valores altos

en “Recency” y valores bajos en “Frequency”. Por otro lado, el cluster N°2 alcanza valores que no superan los 200 días desde la última compra, y la frecuencia de compras es mayor a 5 compras.

Por último, en el cuadro superior derecho, se realiza la comparación entre la variable “Recency” y “Monetary”. Nuevamente, el cluster N°0 alcanza valores altos en la variable “Recency” y valores bajos en la variable “Monetary”, indicando que los individuos compraron por última vez hace varios días y el monto gastado es pequeño.

Los paneles 2 y 3 muestran las figuras centrales e inferiores respectivamente, y se sigue la misma lógica. Todas las figuras indican que el cluster N°2 es claramente el más activo, los clientes más valiosos para la empresa. Contrariamente, los clientes del cluster N°0 son clientes que se podrían considerar como perdidos. Recuperarlos podría significar un esfuerzo y una inversión alta. Los clientes del cluster N°1 no son tan activos, pero no se los considera como perdidos.

Como se mencionó en la sección V, la información sobre la predicción de Churn sumado a la segmentación en clusters debería proporcionar información muy relevante a la empresa. La misma podrá tomar acciones personalizadas para cada uno de estos clusters, entendiendo qué es necesario para cada grupo y así recuperar a aquellos clientes que puedan ser recuperados e incrementar la fidelidad de aquellos más valiosos. El objetivo de esta sección fue identificar los segmentos, pero el plan de acción a partir de esta información será lograr que usuarios del cluster N°0 pasen a ser usuarios del cluster N°1 y que los usuarios del cluster N°1 pasen a ser parte del cluster N°2. Se propone concentrar energías en retener a los usuarios a partir de la implementación de estrategias de atracción como descuentos, envíos gratis, muestras, etc. en vez de salir a buscar nuevos clientes.

**Predicción de cluster**

En el análisis anterior, se logró segmentar a los clientes en distintos clusters, asignándole un sentido de negocio útil para que la empresa pueda entender un poco más el comportamiento de sus clientes y tomar las medidas necesarias. Sin embargo, un

análisis adicional podría aportar información fundamental para que la empresa sea reactiva a partir de la primera compra realizada por un nuevo cliente, y no espere a que un usuario deje de comprar para tomar acciones.

Por lo tanto, proponemos entrenar un modelo predictivo que, a partir de la información recolectada en una primera compra, pueda identificar a qué cluster pertenecerá el usuario en el futuro. De esta forma, la empresa podrá tomar medidas preventivas y podrá lograr que un usuario que según el modelo terminará en el cluster inactivo, no lo haga y sea un cliente activo.

**Metodología**

Como se mencionó anteriormente, el conjunto de datos recibido contiene información acerca de las compras realizadas desde enero del 2019 hasta marzo del 2021. Se analizó la cantidad de altas de cada mes, y se identificó que el mes en el que más clientes nuevos hubo fue mayo del 2020. Por lo tanto, se tomó ese mes como punto de corte para tener más datos para testear.

**División de datos**

**i. Conjunto de Entrenamiento**

Se realizaron un total de 8.685 pedidos desde enero del 2019 hasta abril del 2020. Se repitió el análisis RFM y se le pidió al algoritmo que divida los datos en 3 clusters.

Al haber realizado nuevamente el análisis RFM sobre estos datos, se generó nuestra variable respuesta, es decir, el cluster al que corresponde cada usuario. Sin embargo, se seleccionó la información correspondiente a la primera compra realizada por los usuarios y quedaron un total de 3.546 filas. Se procedió de esta manera para que el modelo se entrene sobre datos de primeras compras.

Las variables que se utilizarán para la predicción serán las mismas que se utilizaron para el análisis de Churn. Es decir, información de los usuarios (como sexo) e información particular acerca de la compra que se necesita (en este caso, la primera).

### ii. Conjunto de Validación y Testeo

En mayo del 2020, la empresa contó con 1.394 clientes nuevos. Por lo tanto, el conjunto de validación y testeo contará con 697 primeras compras cada uno. Para identificar la variable respuesta, se realizó nuevamente un análisis RFM con la información particular de estos clientes y para sus compras desde mayo 2020 hasta marzo 2021. Se cuenta con un total de 2.207 pedidos realizados por estos clientes. El análisis RFM arrojó un valor para cada variable, pero para poder identificar a qué cluster pertenece cada cliente, se utilizó el modelo que se entrenó sobre el conjunto de entrenamiento sobre los datos de validación y testeo. De esta forma, se construyó la variable "Y" para los clientes nuevos de mayo 2020, la cual se intentará predecir.

**Modelo RFM – Predicción de clusters**

- **Recency – Entrenamiento**

```
count    4616.000000
mean      199.990251
std       135.746267
min        61.000000
25%        82.000000
50%       149.000000
75%       300.000000
max       545.000000
Name: Recency, dtype: float64
```

Se cuenta con 4.616 clientes y en promedio pasaron 199 días desde la última compra realizada y hasta el 30 de junio del 2020. Se eligió esta fecha para que aquellos que hayan realizado una compra el 30 de abril 2020 no tengan Recency = 0 días.

| RecencyCluster | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| 0 | 844.0 | 429.029621 | 58.360501 | 346.0 | 377.0 | 412.0 | 484.0 | 545.0 |
| 1 | 1225.0 | 261.368163 | 45.623951 | 178.0 | 225.0 | 258.0 | 300.0 | 345.0 |
| 2 | 2547.0 | 94.573223 | 30.877002 | 61.0 | 70.0 | 84.0 | 107.0 | 177.0 |

Para los datos de entrenamiento, podemos observar que al contrario de lo identificado hasta el momento, el cluster más activo es el que más clientes tiene. Esto puede

significar que el negocio se encontraba bastante activo en ese momento, y los clientes compraban regularmente.

- **Frequency – Entrenamiento**

```
count    4616.000000
mean        1.881499
std         2.041508
min         1.000000
25%         1.000000
50%         1.000000
75%         2.000000
max        25.000000
Name: Frequency, dtype: float64
```

En promedio los clientes realizaron una compra en el período analizado, y el máximo de compras realizadas fueron 25.

| FrequencyCluster | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| 0 | 3816.0 | 1.187893 | 0.390678 | 1.0 | 1.0 | 1.0 | 1.0 | 2.0 |
| 1 | 672.0 | 4.028274 | 1.210032 | 3.0 | 3.0 | 4.0 | 5.0 | 7.0 |
| 2 | 128.0 | 11.289062 | 3.330017 | 8.0 | 9.0 | 10.5 | 13.0 | 25.0 |

En cuanto a los clusters obtenidos, la mayor cantidad de clientes pertenece al cluster inactivo, lo cual es consistente con lo observado anteriormente.

- **Monetary – Entrenamiento**

```
count     4616.000000
mean      3425.101168
std       5781.460310
min         14.000000
25%        464.286812
50%       1447.921303
75%       3638.469947
max      75291.619752
Name: Monetary, dtype: float64
```

En promedio, el monto total indexado gastado fue de $3.425, y el máximo gastado fue $75.291.

| MonetaryCluster | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| 0 | 4043.0 | 1776.136216 | 1630.351492 | 14.000000 | 370.678742 | 1360.855563 | 2672.095039 | 6744.631135 |
| 1 | 487.0 | 11704.625019 | 4001.014045 | 6754.796663 | 8235.619501 | 10690.765590 | 13925.460437 | 22710.823258 |
| 2 | 86.0 | 34060.417262 | 11190.127665 | 23199.919712 | 26295.495156 | 29532.581598 | 38058.411847 | 75291.619752 |

Nuevamente, se puede observar que el cluster más activo, es decir el N°2, es el que menos clientes tiene. La gran mayoría pertenecen al cluster N°0, que en promedio gastó $1.776.

- **Overall – Entrenamiento**

| OverallScore | Recency | Frequency | Monetary |
|---|---|---|---|
| 0 | 431.578811 | 1.153747 | 1866.524304 |
| 1 | 266.240266 | 1.240266 | 1535.979736 |
| 2 | 107.820123 | 1.320361 | 1824.695660 |
| 3 | 150.382671 | 3.718412 | 7383.161969 |
| 4 | 97.367857 | 4.607143 | 11736.368193 |
| 5 | 115.338710 | 9.709677 | 22025.123607 |
| 6 | 87.539683 | 12.412698 | 35026.668904 |

A partir del número de cluster al cual pertenecía cada cliente para cada variable, se creó nuevamente el “Overall cluster”, y se obtuvieron las medias de estos. Los clusters 0-2 se caracterizan por baja frecuencia y bajo monto gastado. Los clusters 3 y 4 se caracterizan por una frecuencia y monto gastado intermedio. Por último, los clusters 5 y 6 se caracterizan por alta frecuencia y monto gastado si se lo compara con el resto. Por lo tanto, se decidió agrupar de esta forma y obtener 3 clusters finales.

| Segment | Recency | Frequency | Monetary |
|---|---|---|---|
| 0 | 213.922217 | 1.266141 | 1755.645711 |
| 1 | 123.732496 | 4.165171 | 9571.488257 |
| 2 | 101.328000 | 11.072000 | 28577.902436 |

**Modelo**

Para realizar la predicción del cluster a partir de una primera compra, se utilizó nuevamente un modelo de aprendizaje supervisado, particularmente XGBoost ya que los resultados de este algoritmo suelen ser buenos. Se realizó la búsqueda de los mejores hiper parámetros posibles, y se utilizaron los siguientes:

- Min_child_weight: 7
- Gamma: 0
- Learning_rate: 0,001
- Subsample: 0,8
- Colsample_bytree: 0,1
- Max_depth: 40

Los resultados obtenidos no fueron alentadores y se pueden encontrar dos explicaciones para los mismos. Por un lado, el conjunto de entrenamiento contenía tan solo 3.546 filas, es decir que el modelo contaba con pocos datos para encontrar patrones y poder predecir correctamente. Por otro lado, el conjunto de entrenamiento estaba muy desbalanceado ya que el 82% de las observaciones pertenecían al cluster N°0, el 15% al cluster N°1 y el 3% restante al cluster N°2.

Considerando la distribución de clusters, se podría concluir que en este caso puntual las primeras compras no son un buen indicador predictivo del cluster al que podría llegar a pertenecer un usuario en el futuro. Se cuenta con pocos datos para realizar el análisis, y a su vez, los datos se encuentran desbalanceados. Por un tema de temporalidad, no se aplicaron técnicas para balancear los conjuntos. Este análisis se podría considerar para un trabajo futuro cuando la empresa cuente con más datos.

# VIII. Conclusiones

**Resumen de los resultados y aplicaciones**

A lo largo de este trabajo se llevaron a cabo tres análisis, obteniéndose resultados favorables en dos de ellos. A partir de estos resultados, la empresa tendrá un mayor entendimiento del estado actual de su negocio y podrá empezar a tomar decisiones para el futuro de este.

Los modelos implementados para el análisis de Churn obtuvieron un nivel de precisión alto, pero se elegirá el modelo XGBoost ya que el mismo alcanzó un AUC de 82% al implementarlo sobre los datos de test. Teniendo en cuenta la situación actual en la que la mayoría de los clientes realizó una única compra en su historial, este análisis tendrá un valor muy importante para la empresa, ya que podrán adelantarse a la pérdida del cliente y tomar acciones necesarias para retenerlo.

Por otro lado, la segmentación realizada deja en claro que existen tres grupos de clientes: los que sólo compraron una vez y un monto chico, los que compraron más de una vez, pero no más de cinco veces y gastaron un monto intermedio pero chico y por último aquellos que suelen comprar con mayor frecuencia y montos grandes. Este análisis se complementa con el anterior, y si se utilizan juntos, se podrán tomar acciones concretas de marketing para retener a los consumidores y logra que los inactivos pasen al grupo intermedio y los intermedios al grupo activo.

A partir de los resultados obtenidos, queda claro que los modelos utilizados si pueden ser aplicados por startups o empresas de pequeña escala. Tener más datos va a contribuir favorablemente a los resultados, pero se alcanzaron resultados muy alentadores con la cantidad de datos provista.

**Limitaciones**

Los datos que se obtuvieron contaban con información valiosa y fueron muy útiles. Sin embargo, el hecho de que el período de análisis en cuestión coincidiera no solo con la etapa de crecimiento inicial de una empresa nueva, pero también con el período de

pandemia, seguramente haya influido en los resultados obtenidos. El hecho de haber salido de las restricciones durante el conjunto de test pudo haber ocasionado el desbalance de clases y afectar los resultados. Lamentablemente no se pudo acceder a los datos más recientes para incrementar el conjunto de datos y observar el comportamiento en post pandemia.

Por otro lado, se considera que la empresa podría recolectar información adicional como la edad, el sexo y lugar de residencia de los clientes para poder segmentar aún más. Si bien la edad y el sexo se utilizaron como variables en este trabajo, las mismas fueron estimadas. El lugar de residencia podría tratarse de una variable importante en la segmentación de los clientes. Podría encontrarse que cierta zona de la provincia o capital tiene clientes más activos y en ese caso tomar medidas con respecto al *delivery* de los pedidos y de esta forma disminuir costos y aumentar las ganancias.

**Trabajo futuro**

No se obtuvieron resultados favorables en la predicción del cluster a partir de la información de la primera compra, y se cree que puede llegar a ser un resultado muy poderoso para la empresa en cuestión. Contar con esta información desde la primera compra, les permitirá ser reactivos y en el mejor de los casos evitar la pérdida de un cliente.

Conocer las razones por las cuales los clientes no vuelven a comprar podría ser un punto de partida. La empresa podría incluir una encuesta de satisfacción tanto de producto como servicio e ir descifrando cuales son los problemas para resolver. Seguramente esa variable contribuiría con el resultado de la predicción.

Por último, en un trabajo futuro se podría plantear una definición distinta de “Churn”, ampliando la ventana de tiempo para su definición en caso de observar un cambio en el tiempo promedio entre compra y compra. Asimismo, podría probarse otro esquema de validación cruzada para obtener mejores resultados.

**Conclusión**

Como se mencionó anteriormente, Big Data puede estar erróneamente relacionado con empresas de grandes ganancias y volúmenes de datos. Las aplicaciones de estos modelos en empresas pequeñas no son tan comentadas como la aplicación en empresas grandes, y por eso no queda claro hasta qué punto pueden ser útiles. En este trabajo se demostró como puede ser de utilidad para una startup que no cuenta con un gran volumen de datos y que podría afinar aún más las variables a considerar.

# Referencias

[1] Critical analysis of Big Data challenges and analytical methods - *(Sivarajah, Kamal, Irani & Weerakkody, 2017)*
[2] The age of analytics: Competing in a data-driven world - *(Henke & Bughin, 2016)*
[3] Big Data Analytics in the Management of Business - (*Jelonek, 2017)*
[4] Implementing Big Data analytics for small and medium enterprise (SME) regional growth - *(Ogechi Ogbuokiri, Agu & Udanor, 2015)*
[5] Online vs offline shopping - *(Maheshwari, 2020)*
[6] Churn analysis: Predicting churners - *(Forhad & Rahman, 2014)*
[7] An introduction to ROC analysis - *(Fawcett, 2006)*
[8] An introduction to ROC analysis - *(Fawcett, 2006)*
[9] A comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers - *(Potdar & Pai, 2017)*
[10] The Effect of Recursive Feature Elimination with Cross-Validarion Method on Classification Performance with Different Sizes of Datasets - *(Akkaya, 2021)*
[11] XGBoost-Based Algorithm Interpretation and Application on Post-Fault Transcient Stability Status Prediction of Power System - *(Chen, Liu & Zhang, 2018)*
[12] Experimenting XGBoost Algorithm for Prediction and Classification of Different Datasets - *(Santhanam, Raman, Uzir & Banerjee, 2017)*
[13] Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito - *(Espinosa-Zúñiga, 2020)*
[14] Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito - *(Espinosa-Zúñiga, 2020)*
[15] An Overview of Overfitting and its Solutions - *(Ying, 2019)*
[16] Random forest and decision trees - *(Ali, Khan & Ahman, 2012)*
[17] Customer Retention Strategies and Customer Loyalty - *(Nasir, 2017)*
[18] Customer Segmentation using K-Means Algorithm - *(Kushawaha & Prajapati, 2008)*
[19] A Clustering Method Based on K-Means Algorithm - *(Li & Wu, 2012)*.
[20] RFM Ranking – An Effective approach to Customer segmentation - *(Christy, Umamakeswari, Priyatharsini & Neyaa, 2018)*
[21] A review of the Application of RFM model - *(Wei, Lin & Wu, 2010)*