BODY TECH®

↘ Reto

BODYTECH: Más fuerte cada dato

+ Julio 2024

ÍNDICE

INTRODUCCIÓN

Con toda la información claramente estudiada y organizada, en este informe presentaremos la continuación al informe de Exploración. Este informe irá directamente a la preparación y entrenamiento del modelo de aprendizaje autómatico no supervisado.

Escogimos este tipo de algoritmo debido a que no teníamos una característica objetivo y los algoritmos no supervisados son los apropiados en este caso. Quisimos dividir los grupo de forma que termináramos con 3 grupos principales sobre los que pudiéramos hacer las estrategias comerciales que recomendamos al final.

Se escogió específicamente este algoritmo, porque a partir de muchas pruebas, fue el que mejor resultado dio en términos de exactitud y nos acercó más a los resultados que esperábamos. Es un algoritmo de clusterización jerárquico aglomerativo que crea los clusters a partir de la eliminación de uno a uno de la anterior iteración, comenzando con centroides en todos los datos y eliminando en cada iteración los más cercanos entre ellos a partir de la distancia euclidiana.

SELECCIÓN DE CARACTERÍSTICAS

Vamos a analizar una a una las columnas que tenemos:

- **ID:** Es una columna de registro que no nos ayuda para el proyecto, pues el número de factura no nos dice nada respecto al cliente.
- **Product_ID:** Es una columna ambigua, como lo vimos en el cuaderno de eda.ipynb, hay valores que no tienen ninguna similitud que lo comparten en la columna, lo que podría alterar el algoritmo en contra.
- **Products_Prices_ID:** Igual que la columna anterior, es muy ambigua, por lo que será necesario eliminarla.
- **Label:** Tiene muchos valores y muchas variables, es necesario crear una columna a partir de esta llamada Label_Type con los filtros que aplicamos en 'eda.ipynb'.
- **Description:** Es una columna de caracteres que no dan información respecto a los clientes.
- **Total_Sub:** Subtotal del pago, que para lo que necesitamos, nos convienen otras más importantes y no saturamos al modelo con procesos innecesarios.
- **Discount_Percent_(COP):** El tipo de descuento nos podría dar información sobre ciertos pacientes, así que la mantendremos.
- **Discount_Absolute_(COP):** Será el valor subtotal menos el descuento, así que podría ser innecesaria.
- **Discount_(COP):** Será el subtotal menos el descuento absoluto, que no siempre está en porcentaje, entonces habrá que eliminarse y crear una nueva de todos los descuentos.
- **Total_Ht:** Es el total que usaremos, porque es el valor real que se le cobra al usuario.

- **Tva:** Lo modificaremos por un porcentaje, pues es más efectivo a la hora de definir el tipo de cliente y por qué se le cobran esos impuestos.
- **Total_Ttc:** La eliminaremos porque será redundante con las dos columnas anteriores.
- **Venue_ID:** Nos da información importante sobre los usuarios, sólamente hay que tener en cuenta los que eliminamos en 'eda.ipynb'.
- **Create_At_Db:** Al tener tan pocos días en nuestro dataframe, no considero que sea una columna relevante en los análisis, la persona pudo asistir el 8 o el 9 de junio sin diferencia.
- **Users_ID:** Dado que no sabemos si es un valor consecutivo, o tiene que ver con la sede, o con la edad, o con alguna característica, la eliminaremos también.
- **Assigned_Executive_Secundary:** Tiene muchos valores nulos, no nos habla del cliente, y el principal nisiquiera fue entregado, entonces esta columna no aporta información.
- **Invoice_ID:** El ID de la factura, aparentemente no nos brinda información adicional del cliente.
- **Promo_ID:** Nos da una referencia de promoción a la que podría estar vinculado el cliente, aunque no sabemos especificamente de qué se trata, es información importante.
- **Agreement_Line_Deferred_Payment_ID:** Aunque hay muchos valores nulos, si hay clientes que comparten un contrato que nos podrian decir información valiosa sobre ellos y sus hábitos de consumo con la empresa.
- **Promotion_Price_ID:** Podría resultar redundante con la columna de Promo_ID.
- **Ruler_Member_Status_ID:** Ya demostramos en 'eda.ipynb' que esta era una columna importante en relación a unas sedes y a un promedio de valor de la compra, entonces la dejaremos, pero la codificaremos.
- **Promotion_Price_Mes_ID:** Es redundante al igual que la otra que ya eliminamos que también hablaba de promoción.
- **Ruler_Product_Status_ID:** No tenemos mayor contexto de esta columna, podría ser irrevelante junto a la otra de estado.

- **Brand_ID, Company_ID and Organization_ID:** Aunque no nos da mayor información, junto al valor de los impuestos, podría aportar información importante referente al tipo de cliente.
- **Status:** Es importante, porque podríamos no limitar nuestra búsqueda a cliente activos únicamente.
- **Venue_Use:** Podría ser redundante junto con la otra columna de Sede.
- **Is_Prepaid:** También lo analizamos en 'eda.ipynb' y podría darnos información sobre los clientes.

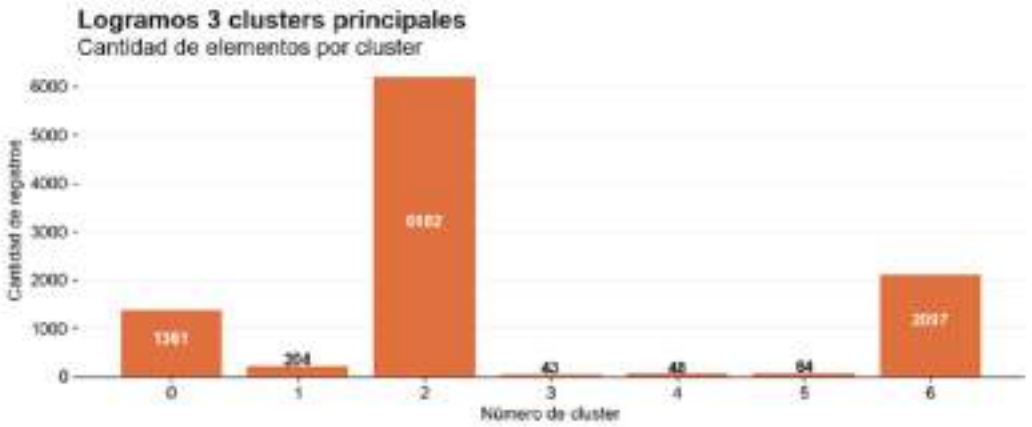
ENTRENAMIENTO DEL MODELO

Tomaremos las columnas actuales y comenzaremos por formatearlas en un formato entendible, luego le aplicaremos algún tipo de codificación de acuerdo con el tipo de datos que tenga la columna.

Antes de la codificación, nuestro dataframe tenía 9.999 filas y 13 columnas. Las columnas numéricas van a pasar por una estandarización debido a que tenemos rangos y distribuciones muy variadas y no queremos que eso distorsione la forma en la que el modelo entrena a partir de la información, por el contrario, que todas las características tengan el mismo peso en él. Las columnas categóricas pasarán por codificación One-Hot en la que se crearán columnas nuevas booleanas por cada registro diferente que encuentre en la columna original, esto nos ayudará a eliminar los números originales como por ejemplo el número de la sede, pues no es cierto que la sede 30 sea más importante que la 29 sólo porque el número es más alto, esto nos ayudará a eliminar esa distorsión.

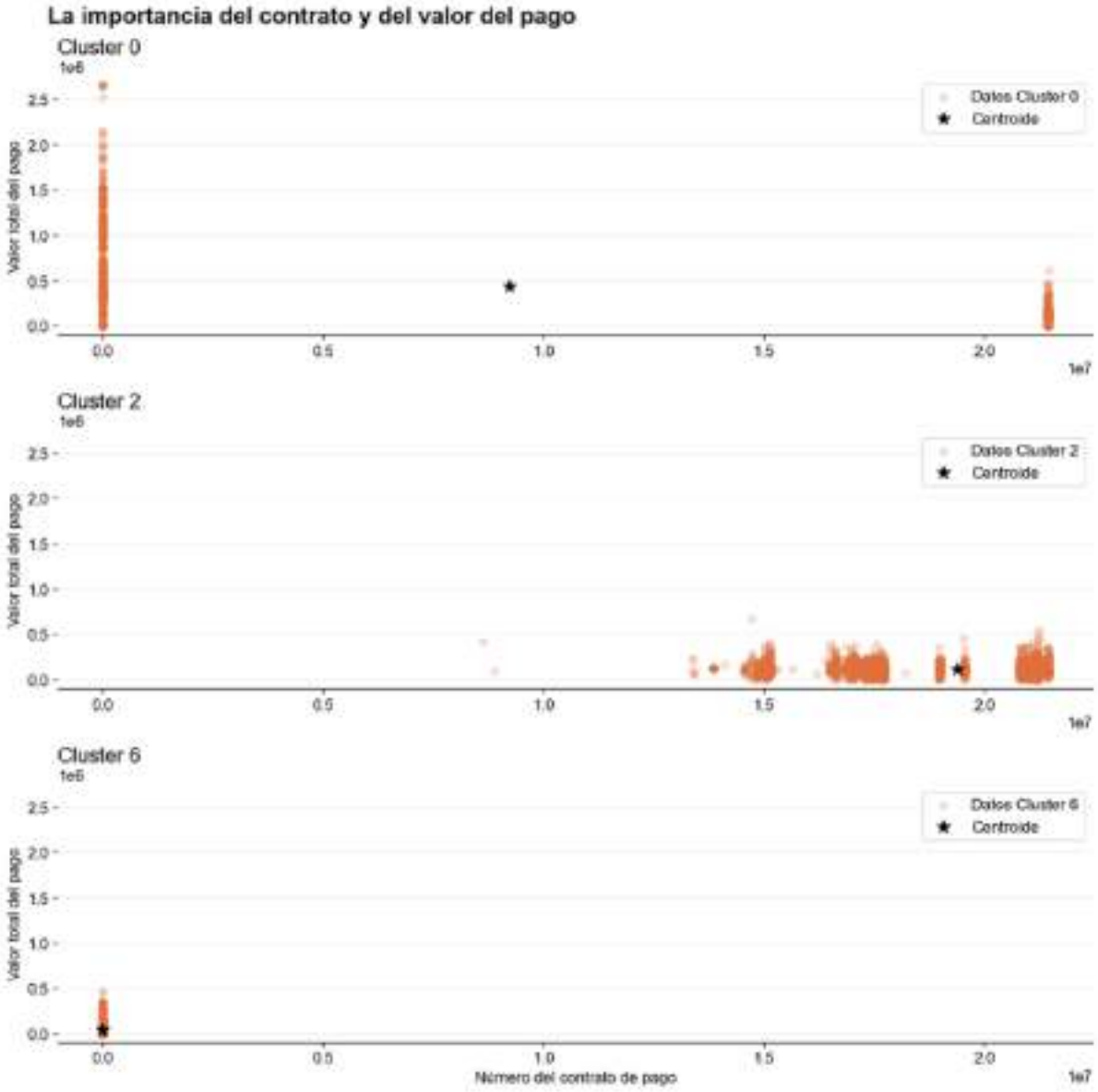
Debido a que tenemos muchos datos, incluso atípicos, y con diferentes escalas, nos pareció la mejor decisión estandarizarlos, es decir que todos quedaran con un promedio en 0 y una desviación estándar de 1 para, de esta forma, evitar que nuestro modelo le de más peso a unas características sobre otras y pueda ser más imparcial. Así mismo, se aplicó a dos columnas la codificación `One-Hot` con la eliminación de la primera columna sobre las columnas categóricas como el ID de la Sede, pues no conocemos si existe algún tipo de jerarquía en esos números, con la idea de evitar sesgos.

El algoritmo que utilizaremos proviene de la librería SciKit Learn y tiene el nombre de AgglomerativeClustering, se puede encontrar en la carpeta Cluster de SciKit Learn.



Cuando se probaron algoritmos de cluster más altos que el definitivo, se observaron clusters demasiado específicos donde teníamos, por ejemplo, con 10 clusters, los dos clusters con menor cantidad de elementos contenían 48 y 43 elementos respectivamente. Aunque no tenemos información sobre los presupuestos, tener 10 estrategias de marketing nuevas a partir de esta base de datos podría ser más desafiante de lo necesario. Nos enfocamos en definir 3 clusters que nos definan muy bien nuestros registros, esto se logró con 7 clusters iniciales creados a partir del modelo.

DEFINICIÓN DE CLUSTERS



Estas 4 columnas nos aportan las características más importantes para nuestros cluster de pagos que podríamos definir de la siguiente forma:

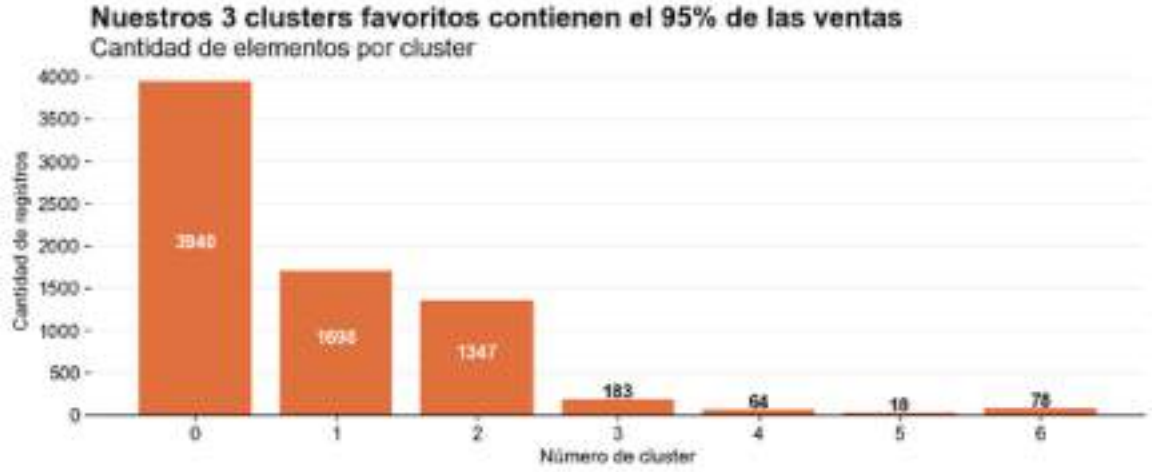
- **Cluster 0:** Los miembros de este grupo son los únicos que tienen los pagos superiores a los $1'000.000 COP, aunque si cumplen las otras condiciones y tienen un pago inferior también podrían pertenecer. Estos registros, o no tienen un número de contrato o tienen un número de contrato muy alto (por encima de 20'000.000). Además, son en su mayoría usuarios activos, el 99.92% de ellos. Y finalmente estos miembros podrían no tener un código de promoción como el 29.09% de ellos o sus códigos son superiores a 833.
- **Cluster 2:** Los miembros de este grupo no tienen un código de promoción, además tienen números de contrato superiores a 8'642.641 con un pago inferior a los $671.000 COP. En este cluster tenemos usuarios activos (el 71.36%) y encontramos la mayoría de inactivos.
- **Cluster 6:** Los miembros de este grupo no hacen pagos superiores a $464.000 COP y no tienen número de contrato. El 98% de ellos son miembros activos y sus códigos de promoción son inferiores a 463.

REENTRENAMIENTO DEL MODELO

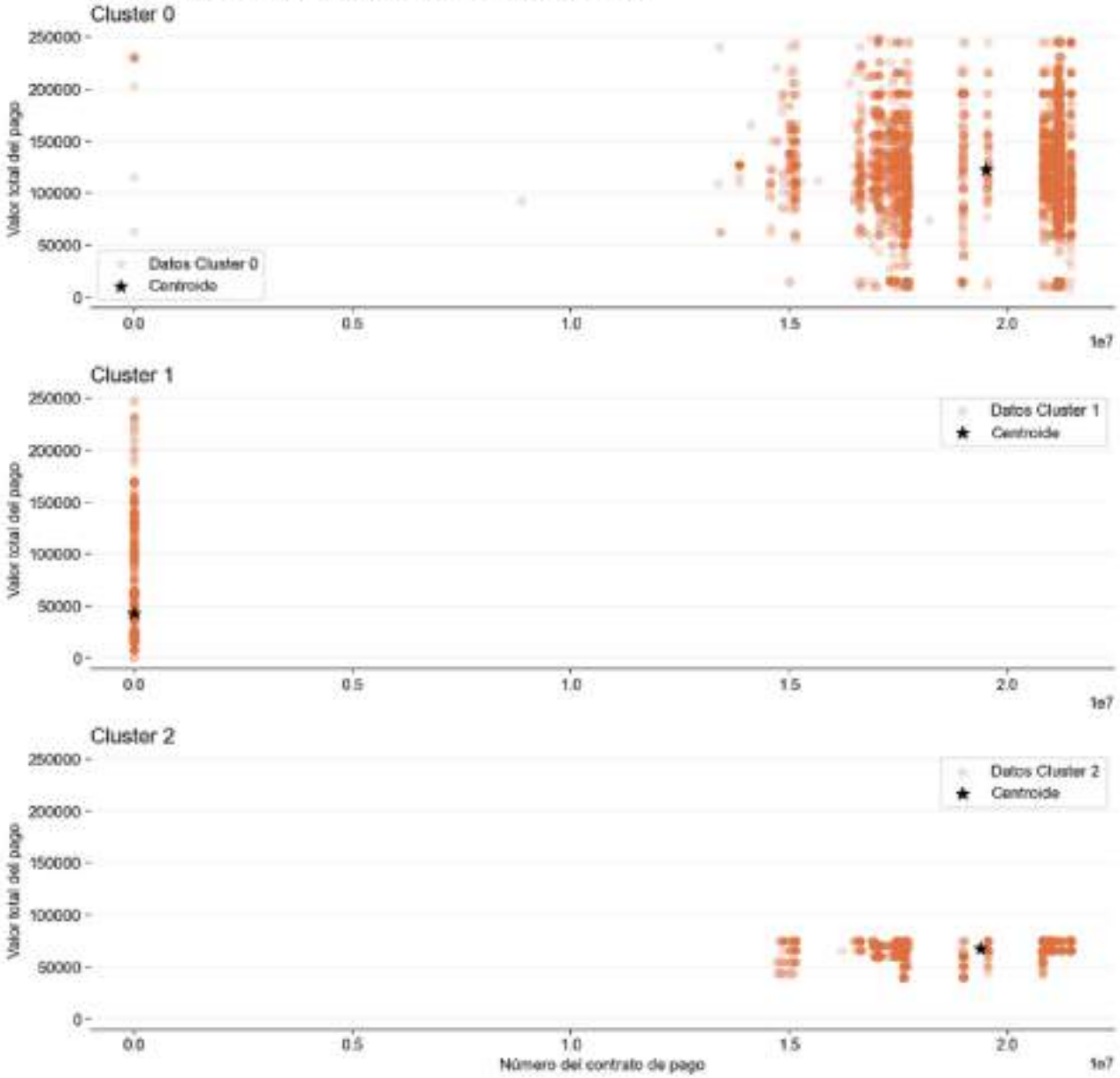
Aunque sentimos que estos grupos ya están lo suficientemente definidos, construiremos una estrategia mucho más detallada dejando de lado los valores atípicos de cada columna.

Las dimensiones del dataframe después de eliminar los valores atípicos es (7328, 135).

Y de nuevo, volvemos a codificar nuestras características y a construir un modelo nuevo sin los valores atípicos tanto en los porcentajes, como en los valores del pago.

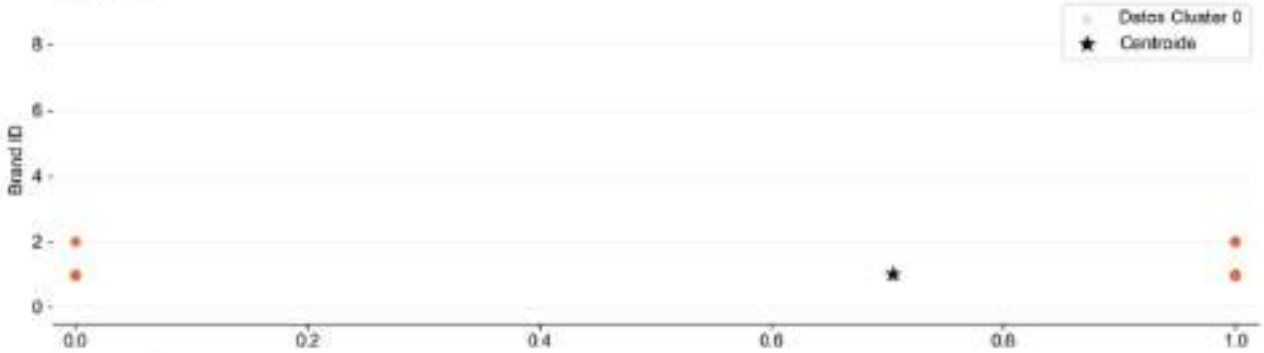


Logramos dividir los usuarios por costo de pago

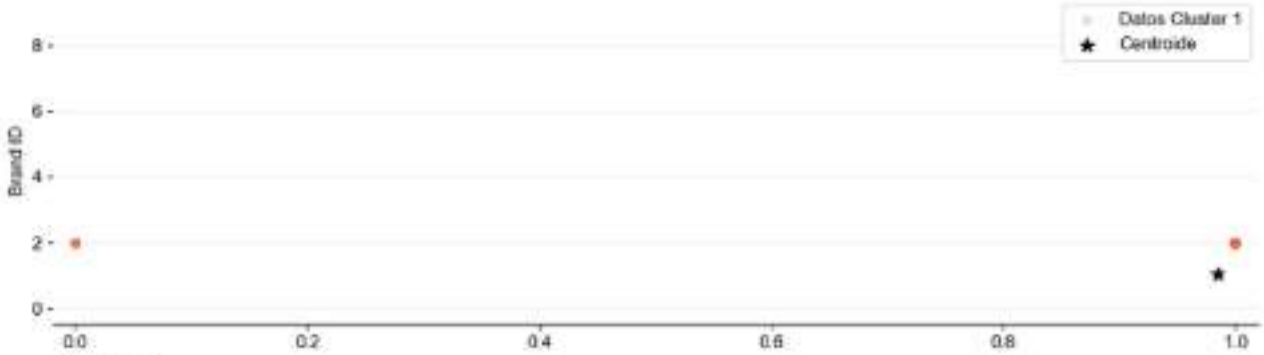


¿Qué significará Brand 2? Parece importante

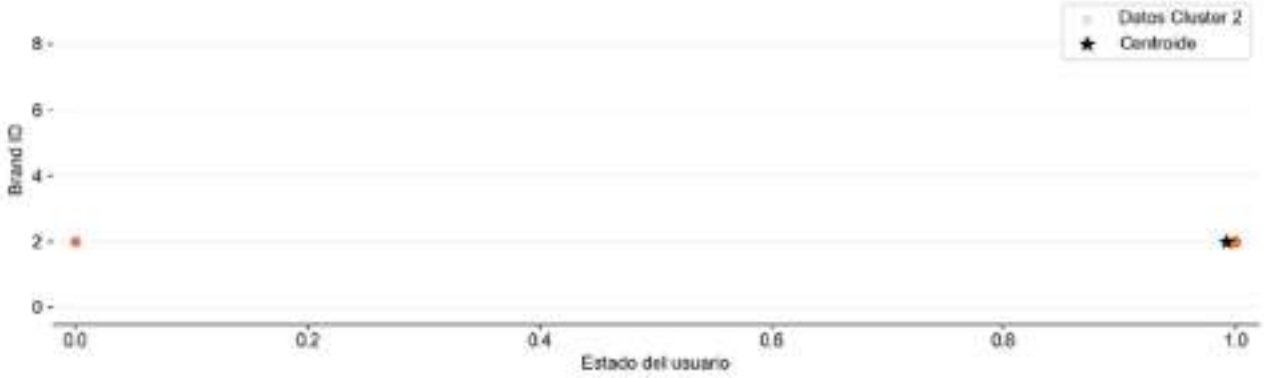
Cluster 0



Cluster 1

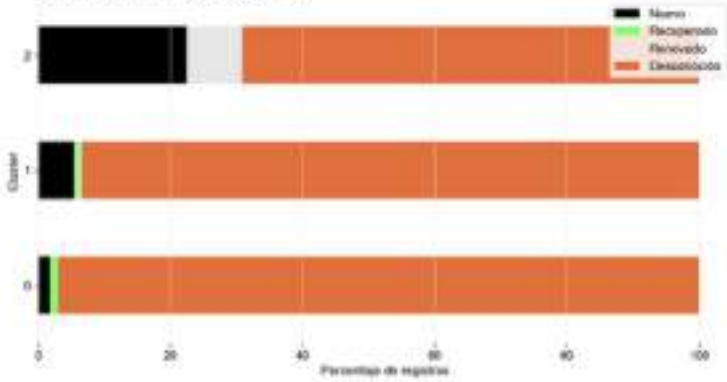


Cluster 2



Renovados, los que más ingresos traen, un 94.64% estarán en el cluster 2

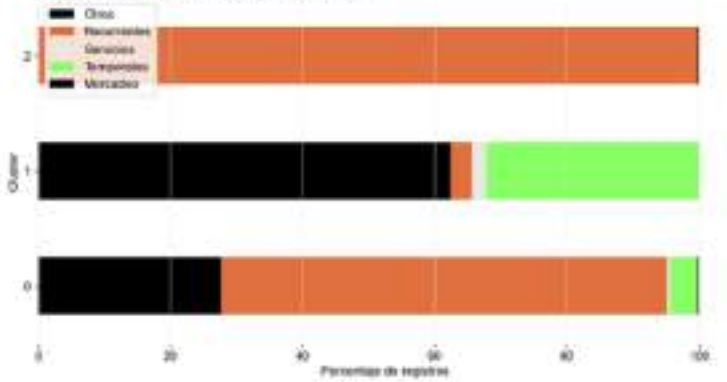
Porcentaje de estado de los usuarios



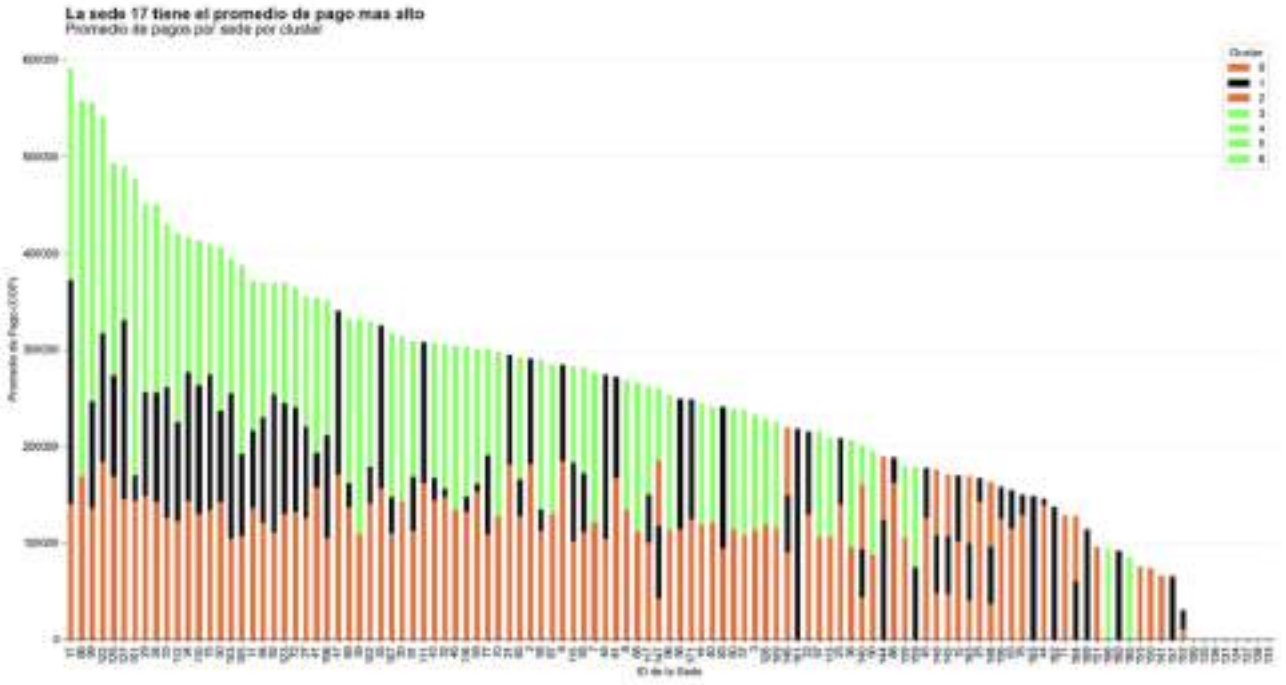
Cluster	Nuevo	Recuperado	Renovado	Desconocido
0	1.903553	1.065990	0.000000	97.030457
1	5.418139	0.706714	0.471143	93.404005
2	22.568671	0.000000	8.314774	69.116555

99.77% de seguridad que si es recurrente es del cluster 2

Porcentaje de tipos de pago de los usuarios



Cluster	Otros	Recurrentes	Servicios	Temporales	Mercadeo
0	27.715736	67.411168	0.431472	4.238579	0.203046
1	62.426384	3.239105	2.179894	32.153477	0.000000
2	0.000000	99.777283	0.000000	0.000000	0.222717



La intención de esta gráfica está en poder entender las sedes específicas donde mayor cantidad de cada cluster hacen sus pagos para que, de esta forma, las recomendaciones comerciales que determinamos al final, puedan ser aún más específicas. Es decir, si la recomendación es para el cluster 2, sería valioso desplegarlo principalmente para las sedes donde ya sabemos que la mayor cantidad de pagos los realiza ese cluster justamente, y así con las otras recomendaciones y las otras sedes.

Cuando armamos 7 clusters a partir de los datos típicos de venta, encontramos 3 principales, el 0, 1 y 2, éstos contienen el 95.31% de los registros y serán los que vamos a analizar y sobre los que basaremos nuestras sugerencias.
Teniendo en cuenta que, una vez eliminamos los valores atípicos, el valor de pago máximo fue de $248.000 COP, podemos definir estos clusters importantes así:

- **Cluster 0:** Usuarios cuya mediana de compra es de $120.800 COP, es decir que aquí encontramos los pagos de costo alto, además que el 0.22% de ellos no tienen número de contrato, es decir que la mayoría si lo tiene. El 29.49% de los miembros de este cluster son usuarios inactivos y el 99.01% de ellos también tiene valor Brand_ID de 1. También es importante definirlos porque el 97.03% de miembros de este cluster son usuarios desconocidos.
- **Cluster 1:** La mediana de compra de los usuarios de este cluster es de $21.100 COP, es decir aquí encontramos los pagos con costo bajo, ninguno de los miembros de este grupo tiene un número de contrato y el 98.52% de los miembros de este cluster son usuarios activos, la gran mayoría. El 97.29% de los miembros tiene un valor Brand_ID de 1 y el 83.46% de los pagos son de servicios, que a su vez, tienden a ser los más altos en valor promedio como lo vimos en eda.ipynb
- **Cluster 2:** El promedio de compra de los usuarios de este cluster es de $69.900 COP y todos tienen un número de contrato. En este cluster, todos los miembros tienen un valor Brand_ID de 2 y el 99.18% de usuarios son activos. El 94.64% de todos los usuarios renovados, los que más ingresos nos generan en total, están en este cluster y el 99.77% de estos, realizan pagos recurrentes.

CONCLUSIONES

Cuando armamos 7 clusters a partir de los datos típicos de venta, encontramos 3 principales, el 0, 1 y 2, éstos contienen el 95.31% de los registros y serán los que vamos a analizar y sobre los que basaremos nuestras sugerencias.
Teniendo en cuenta que, una vez eliminamos los valores atípicos, el valor de pago máximo fue de $248.000 COP, podemos definir estos clusters importantes así:

- **Cluster 0:** Usuarios cuya mediana de compra es de $120.800 COP, es decir que aquí encontramos los pagos de costo alto, además que el 0.22% de ellos no tienen número de contrato, es decir que la mayoría si lo tiene. El 29.49% de los miembros de este cluster son usuarios inactivos y el 99.01% de ellos también tiene valor Brand_ID de 1. También es importante definirlos porque el 97.03% de miembros de este cluster son usuarios desconocidos.
- **Cluster 1:** La mediana de compra de los usuarios de este cluster es de $21.100 COP, es decir aquí encontramos los pagos con costo bajo, ninguno de los miembros de este grupo tiene un número de contrato y el 98.52% de los miembros de este cluster son usuarios activos, la gran mayoría. El 97.29% de los miembros tiene un valor Brand_ID de 1 y el 83.46% de los pagos son de servicios, que a su vez, tienden a ser los más altos en valor promedio como lo vimos en eda.ipynb

- **Cluster 2:** El promedio de compra de los usuarios de este cluster es de $69.900 COP y todos tienen un número de contrato. En este cluster, todos los miembros tienen un valor Brand_ID de 2 y el 99.18% de usuarios son activos. El 94.64% de todos los usuarios renovados, los que más ingresos nos generan en total, están en este cluster y el 99.77% de estos, realizan pagos recurrentes.

RECOMENDACIONES COMERCIALES

Dado que no sabemos exactamente qué significan los números de contrato, ni los códigos de promoción, ni los Brand_ID, ni Company_ID, ni Organization_ID,no podemos dar recomendaciones en esa vía, más que la explicación matemática que divide los grupos.

Cluster 1

Perfil: Usuarios con compras de costo alto, la mayoría tiene un número de contrato, muchos son inactivos y desconocidos para el sistema, pero casi todos tienen Brand_ID de 1.

1. **Ofertas VIP:** Crear una membresia VIP con beneficios exclusivos como acceso a áreas premium, entrenadores personales y eventos exclusivos.
2. **Descuentos en servicios de lujo:** Ofrecer descuentos en servicios premium como sesiones de entrenamiento personal, kickboxing y fisioterapia.
3. **Bonos de reactivación:** Enviar bonos de descuento para motivar a los usuarios inactivos a volver al gimnasio.
4. **Comunicación personalizada:** Enviar comunicaciones personalizadas destacando los beneficios de los servicios de alto costo que suelen adquirir.

Cluster 2

Perfil: Usuarios con compras de costo bajo, la mayoría son activos y realizan pagos para servicios.

1. **Membresías combinadas:** Ofrecer membresías que combinen el acceso al gimnasio con servicios a un precio especial.
2. **Sesiones de seguimiento gratuitas:** Incluir sesiones de seguimiento gratuitas para los usuarios de servicios de salud.
3. **Evaluaciones de progreso:** Ofrecer evaluaciones de progreso gratuitas para usuarios regulares de servicios.
4. **Ofertas flash:** Implementar ofertas flash con descuentos significativos en servicios durante un tiempo limitado.
5. **Descuentos en horas no pico:** Ofrecer descuentos para servicios en horas no pico para maximizar el uso de las instalaciones.

Cluster 3

Perfil: Usuarios con compras de costo medio-alto, todos tienen contrato, son mayoritariamente activos y realizan pagos recurrentes.

1. **Bonos por lealtad:** Enviar bonos de descuento o servicios gratuitos como recompensa por su lealtad.
2. **Descuentos en pagos anuales:** Ofrecer descuentos significativos para aquellos que opten por pagos anuales.
3. **Descuentos por volumen:** Ofrecer descuentos para aquellos que compren múltiples paquetes de servicios.
4. **Ofertas especiales de temporada:** Crear ofertas especiales de temporada para incentivar la renovación y la lealtad.

BODY TECH®

GRACIAS

ELABORADO POR:
JUAN PABLO RAMOS BEDOYA

El repositorio con la documentación y el código completo lo encuentra en:
https://github.com/juanramosdataexpert/Bodytech-MasFuerteCadaDato