Universidad ORT Uruguay

Facultad de Ingeniería

Taller de aprendizaje automático

Alfredo Rodríguez - 5515

Adrián Arredondo - 129641

Federico Ramis - 170081

2021

Contenido

I - Presentación del caso de negocio

PedidosYa es una plataforma de pedidos online de lo que el usuario necesite, cuando el usuario lo necesite. En la plataforma se conectan a los usuarios con los socios de negocio conocidos como “partners”, que pueden vender desde comida, hasta productos de supermercado, farmacia, veterinaria y muchos más.

Durante el 2020 PeYa alcanzó más de 30 millones de órdenes en el segundo trimestre y superaron los 60 mil comercios adheridos. Además, adquirieron las operaciones de Glovo en América Latina, por lo cual en 2021 van a operar en 14 mercados. La pandemia, además, la impulsó aún más como un App multi-vertical, potenciando las compras en supermercados y llegando a inaugurar PedidosYa Market.

PeYa contacta a los partners desde operaciones, marketing y ventas con diferentes motivos: ya sea para asesorarlos con su comportamiento operativo, como para generar acciones que permitan aumentar la concurrencia a sus perfiles (visibilidad en primeras posiciones, participación en campañas de marketing, etc).

Frente a un crecimiento exponencial de la cantidad de partners por la expansión regional que tuvo la empresa, se plantea el objetivo de potenciar la generación de estrategias segmentadas sobre los diferentes tipos de partners, considerando si son buenos comercial y operativamente. Esto les permitirá enfocar mejor las acciones de diferente tipo, en base a las necesidades que tengan. Para esto se plantea el desafío de lograr una herramienta que permita a PeYa entender qué tipos de partners se tienen en la vertical de Groceries, para poder definir las mejores acciones.

II - Caso de Negocio y Planificación del Trabajo

II.I - Desafío planteado.

Hoy en día las personas prefieren cada vez servicios y ofertas más personalizadas, para esto es fundamental conocer el comportamiento no solo de los clientes sino también de los Partners.

Tomar definiciones en base a la automatización en el proceso de análisis de la información, es un camino fundamental en esta era donde se recolecta a gran velocidad, grandes volúmenes de datos.

De forma genérica, algunos de los beneficios esperables son:

- Aumentar las ventas mediante la identificación de nuevos canales.
- Aumentar la fidelidad de los clientes.
- Identificar posibles estrategias o acciones para realizar desde áreas como marketing, operaciones y ventasl que permitan agregar valor al negocio.
- Aumentar las posibilidades de llegar a más usuarios.
- Generar mayor satisfacción de los clientes actuales y nuevos clientes a través de mejoras operativas.
- Optimizar y potenciar los canales de pago Online.

Para PedidosYa la necesidad de profundizar en el conocimiento del Partner, tal como se vio, fue impulsada por el crecimiento de la empresa en los últimos meses. En este contexto, los procedimientos manuales de análisis de los datos generan que muchas áreas trabajen sobre los mismos partners. Es decir, no pueden realizar esfuerzos focalizados en aquellos que realmente requieren un apoyo diferenciado.

Optimizar los procesos relacionados con el análisis de los partners, implica obtener mejores resultados con menos recursos.

La identificación de acciones específicas en función de una clasificación de los Partners permitirá potenciar sus puntos fuertes y ayudarlos a superar sus dificultades. Es una estrategia ganar-ganar, ya que todo lo que permita que el Partner genere una mejor experiencia al usuario, o cliente final, se traducirá en beneficios para PedidosYa como canal de venta y distribución asociado.

Segmentar la cartera de partners, permitirá definir estrategias sobre cada cluster, ayudando a optimizar las acciones a realizar, ahorrando costos y mejorando las probabilidades de éxito.

II.II - Modelo a implementar

El objetivo de este trabajo es segmentar los partners en grupos afines, para luego identificar estrategias y acciones específicas para cada grupo, creando así un seguimiento y apoyo personalizado para cada socio de negocio.

Para resolver el problema de segmentación de Partners, se plantea utilizar un algoritmo de Clustering. Los algoritmos de clustering, son un modelo de Machine Learning de aprendizaje automático, no supervisado. El algoritmo aplica un conjunto de técnicas descriptivas (no explicativas) que tiene por objetivo formar grupos a partir de un conjunto de elementos u observaciones. Por tanto, se buscará establecer grupos donde los partners que pertenezcan a un conjunto, sean muy semejantes entre sí, y a la vez lo suficientemente diferente del resto de los grupos.

II.III - Hipótesis.

Hipótesis 1: Si el partner gestiona muchos productos, puede ser que no pueda tener suficiente stock o que no esté bien dimensionado su límite de recepción de órdenes, y entonces rechace pedidos más frecuentemente o que sea necesaria la asistencia por parte de PeYa.

Hipótesis 2: Un bajo tiempo en la confirmación del pedido, sumado a que no tenga entregas tardías, producirá un alto volumen de pedidos, y debería de ir acompañado por una mejor QoE (calidad de experiencia).

Hipótesis 3: Si un partner genera muchas consultas y acciones de asistencia, acompañado de un alto rechazo de órdenes, bajará la QoE y el volumen de transacciones. Esto se puede analizar además desde un punto de vista de la antigüedad del partner, dado que esta situación se puede dar en mayor medida en los partners más nuevos debido a que se encuentran en fase de aprendizaje.

Hipótesis 4: Aquellos partners que tienen muchos productos y muchas fotos, generan un alto volumen de órdenes a causa de que el usuario en parte, podrá realizar todas las compras en una sola operación.

Hipótesis 5: Los partners que tienen costo adicional de envío, que además tienen un mínimo de compra y que no cuentan con opciones de pago online, tendrán un volumen de ventas bajo.

Hipótesis 6: Partners que son relativamente nuevos en la plataforma, se entiende que aún no han logrado el máximo potencial a nivel de transacciones debido a que se encuentran en proceso de aprendizaje y no han adoptado herramientas de estímulo (por ejemplo el uso de vouchers y descuentos) que tienen disponibles para optimizar su propuesta y mejorar su performance.

II.IV - Variables relevantes y tipos de datos necesarios

Como se mencionó anteriormente, el problema se resolverá utilizando un algoritmo de Clustering. Para este tipo de algoritmos, que se basan en la medición de distancias, es mejor que los datos sean cuantitativos, que no sean cadenas de texto ni categóricos. A su vez, deberán estar normalizados y es conveniente que no tengan correlaciones.

De todas formas, si existen variables categóricas relevantes que se quieran considerar durante el modelado se puede definir una matriz de disimilitud usando la métrica de Gower.

Variables	Tipo de Variable
Tipo de Partner	Cualitativa
Tipo de delivery	Cualitativa
Cantidad de productos que gestiona	Cuantitativa
Antigüedad como partner	Cuantitativa
Alerta de saturación	Cuantitativa
Órdenes rechazadas	Cuantitativa
Cantidad de Disparadores	Cuantitativa
Tiempo de respuesta	Cuantitativa
Cantidad de Órdenes	Cuantitativa
Cantidad de encuestas positivas	Cuantitativa
Cantidad de encuestas negativas	Cuantitativa
Cantidad de encuestas	Cuantitativa
Cantidad de imágenes	Cuantitativa
Costo de envío	Cuantitativa
Mínimo de compra	Cuantitativa

II.V - Plan de trabajo y Metodología aplicada.

La metodología utilizada en la presente investigación, es la metodología CRISP-DM (Cross Industry Standar Process for Data Mining). Ésta proporciona una descripción del ciclo de vida de un proyecto estándar de análisis de datos. El ciclo de vida del proyecto consiste en seis fases mostradas en la figura siguiente.

Entendimiento del negocio: En la primera fase del proyecto, se tiene que realizar la comprensión de los objetivos. Ver el desafío planteado, y de qué manera se puede resolver. Realizar las primeras hipótesis y ver cómo será utilizado el resultado del proyecto.

Entendimiento de los datos: La fase de entendimiento de datos comienza con la colección de datos inicial y continúa con las actividades que los permiten entender. El objetivo es tener un conocimiento preliminar sobre los datos.

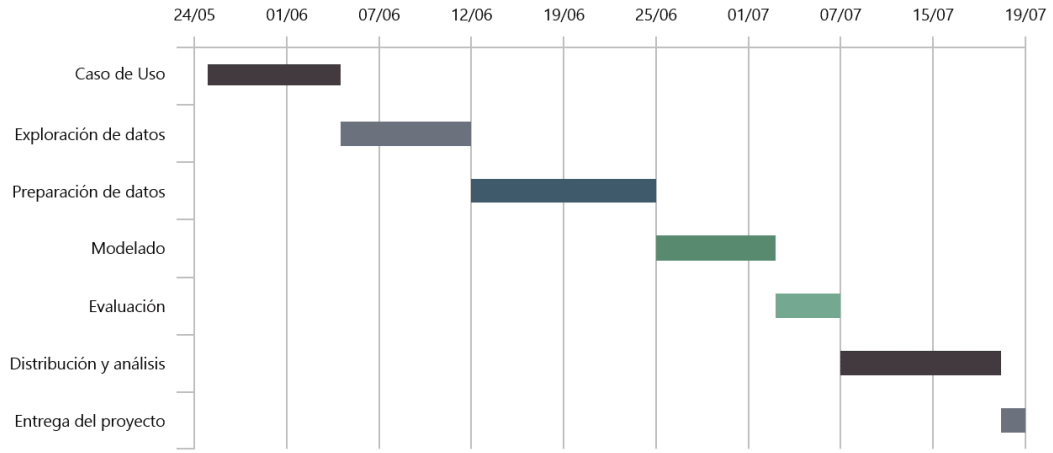
Preparación de los datos: La fase de preparación de los datos, abarca todo el recorrido realizado hasta construir la base de datos final. Desde la construcción de la base de datos analítica, la transformación, hasta la limpieza de datos.

Modelado: En esta fase se aplican las técnicas de modelado elegidas para resolver el problema. En este caso se definió utilizar 2 modelos.

Evaluación: En esta fase del proyecto, se evalúan los modelos que se han construido de acuerdo a los objetivos del negocio. Al final de esta fase, se debería obtener una decisión sobre la aplicación de los resultados del proceso de análisis de datos.

Distribución y Análisis: En la fase de implementación se establecen accionables en base a los resultados y análisis obtenidos. En esta etapa además se describe el mecanismo mediante el cual se evaluará periódicamente el modelo.

Se detalla la hoja de ruta a seguir en el presente proyecto, el plan de trabajo que se presenta no es estricto en su procedimiento, dado que toda fase puede volver a un paso anterior a medida que surge nueva información de cada fase.



II.VI -Aplicación de este trabajo por Pedidos Ya.

Como resultado de este trabajo, se entregará una base de datos con los partners clasificados o segmentados mediante los algoritmos de clusterización. Adicionalmente se entregarán un conjunto de accionables tendientes a solucionar problemas detectados o potenciar su performance.

Otro entregable de este trabajo son un conjunto de piezas de software que permitirán clasificar a los partners en tiempo real o mediante procesos batch según corresponda.

II.VII - Áreas involucradas

Es importante que participen o estén representadas las áreas de dirección de la empresa en el rol de sponsors del proyecto y como integradores entre las demás áreas involucradas, definiéndose un responsable (o comité responsable) que pueda arbitrar las situaciones en las que no haya acuerdo, que fije las restricciones de tiempo y recursos, así como los objetivos de alto nivel.

Deben participar también personas designadas del área de Operaciones y Comercial. Estos son quienes finalmente pueden accionar sobre los resultados de este trabajo y que además, poseen el conocimiento de sus respectivas áreas de negocio, de forma de validar con ellos tanto los supuestos iniciales como los resultados a los que se arribe en las diferentes etapas del proyecto.

Se debe contar también con la participación del personal de TI, tanto para el acceso a los datos así como el desarrollo de la infraestructura necesaria. Seguramente con ellos en un futuro se deba acordar cómo se aplica la lógica del modelo una vez entregado, permitiendo incorporar nuevos datos en los siguientes meses.

III - Estrategia y Preparación de Datos

En esta etapa se preparan los datos de forma de dejar pronta la tabla analítica para realizar el tratamiento y luego poder modelar. Esta base va a tener una fila por cada uno de nuestros Partners.

III.I - Estructura de la tabla de datos analítica.

Para este proyecto se cuenta con dos tablas de datos analíticas con 41.852 observaciones, "datos_caso_PEYA" y "datos_capacity_PEYA". La primera base tiene información del partner comprendida en 26 variables cuantitativas y cualitativas, mientras que la segunda base contiene información acerca de si el local cuenta con un límite de órdenes que puede recibir en determinado tiempo.

A continuación se observan las variables de ambas bases de datos.

datos_caso_PEYA:

- **partner_id**: Identificador de cada partner, es una variable numérica. Los partners van del 1 al 41.852.
- **business_type_name**: es una variable categórica de 7 niveles, indica el tipo de negocio: Restaurante, Market, Coffee, Shop, etc.
- **qty_orders**: Es una variable numérica, indica la cantidad de órdenes totales en histórico.
- **delivery_time**: Es una variable categórica de 9 niveles, indica el rango de tiempo prometido de entrega (en minutos).
- **has_shipping_amount**: Es una variable categórica de 2 niveles, indica si cobra costo de envío.
- **has_online_payment**: Es una variable categórica de 2 niveles, indica si acepta pago online.
- **is_logistic**: Es una variable categórica de 2 niveles, indica si el partner tiene delivery a cargo de PeYa.
- **is_important_account**: Es una variable categórica de 2 niveles, indica si es una cuenta importante.
- **first_date_online**: Es una variable categórica de 2.309 niveles, indica la fecha desde que el local está en la App.
- **qty_products**: Es una variable numérica, indica la cantidad de productos cargados en histórico.
- **qty_picts**: Es una variable numérica, indica la cantidad de fotos cargadas en histórico.
- **accepts_pre_order**: Es una variable categórica de 2 niveles, indica si el local acepta pre órdenes.
- **accepts_vouchers**: Es una variable categórica de 2 niveles, indica si el local acepta voucher.
- **has_discount**: Es una variable categórica de 2 niveles, indica si el local acepta descuentos.
- **has_mov**: Es una variable categórica de 2 niveles, indica si el local tiene valor mínimo de compra.
- **has_custom_photo_menu**: Es una variable categórica de 2 niveles, indica si el local tiene fotos reales en menú.
- **qty_triggers**: Es una variable numérica, indica la cantidad de triggers.
- **qty_order_late_10**: Es una variable numérica, indica la cantidad de órdenes entregadas tarde por parte del partner (más de 10 minutos) al rider.

- **response_time_minute**: Es una variable categórica con 37.688 niveles, indica el tiempo promedio de respuesta (tiempo que ocurre entre la realización del pedido y la confirmación o rechazo de la orden).
- **voucher_order**: Es una variable numérica, indica la cantidad de órdenes con voucher asociado.
- **rejected_orders**: Es una variable numérica, indica la cantidad de órdenes canceladas o rechazadas.
- **qty_passive**: Es una variable numérica, indica la cantidad de encuestas Pasivos en histórico. Calificaciones 7 y 8 (que tanto recomendarías del 1 al 10 PEYA)
- **qty_promoter**: Es una variable numérica, indica la cantidad de encuestas Promotores en histórico. Calificaciones 9 y 10 (que tanto recomendarías del 1 al 10 PEYA)
- **qty_detractor**: Es una variable numérica, indica la cantidad de encuestas de detractores en histórico. Calificaciones entre 1 y 6 (que tanto recomendarías del 1 al 10 PEYA)
- **answers**: Es una variable numérica, indica la cantidad de encuestas en histórico.
- **chats:** Es una variable numérica, indica la cantidad de chats asociados a alguna orden.
- **sessions**: Es una variable numérica, indica la cantidad de sesiones asociadas al partner, generadas en el periodo por los usuarios al acceder al centro de ayuda.

datos_capacity_PEYA

- **partner_id**: Identificador de cada partner, es una variable numérica, Los partners van del 1 al 41.852.
- **business_type_name**: es una variable categórica de 7 niveles, indica el tipo de negocio: Restaurante, Market, Coffee, Shop, etc.
- **capacity_check**: Es una variable categórica, con variables "NULL", "TRUE" y "FALSE". Indica si el local tiene un límite de órdenes a recibir en un determinado periodo de tiempo.

Para hacer la unión de ambas bases de datos se utilizó la variable “partner_id”.

III.II - Construcción de variables y transformaciones de datos.

Creación de variables

Días

Para poder normalizar las variables, en una primera etapa se creó la variable días, siendo el máximo de días de cada partner 180, ya que éste es el período de tiempo del cual se tiene información.

Antigüedad

Se calcula como la cantidad de meses que el partner está dentro de la plataforma.

Normalización

Dado que se están analizando partners que tienen distintas antigüedades dentro de la ventana de tiempo, se entiende importante realizar una normalización de las variables para poder hacerlas comparables entre ellos.

Para esto se toman algunas de las variables cuantitativas y se relativizan en función de los días que tiene el Partner en la aplicación. Otras variables cuantitativas se relativizan por la cantidad de órdenes del partner.

En base a lo mencionado anteriormente se construyeron las siguientes variables:

Promedio de órdenes diarias:

Nombre de la variable: prom_order_dias

Cálculo: qty_orders /dias

Triggers por órdenes:

Nombre de la variable: porcentaje_triggers

Cálculo: qty_triggers / $qty_orders

Porcentaje de órdenes con vouchers:

Nombre de la variable: porcentaje_voucher

Cálculo: voucher_order/qty_orders

Porcentaje de órdenes canceladas.

Nombre de la variable: porcentaje_ordenes_canceladas

Cálculo: rejected_orders / qty_orders

Porcentaje de órdenes confirmadas.

Nombre de la variable: porcentaje_ordenes_confirmadas

Cálculo: (qty_orders - rejected_orders) / qty_orders

Promedio de chats por día.

Nombre de la variable: chats_dias

Cálculo: chats / dias

Cantidad de sesiones por órdenes.

Nombre de la variable: porcentaje_sessions

Cálculo: sessions / qty_orders

Porcentaje Detractores

Nombre de la variable: porcentaje_detractores

Cálculo: qty_detractor / answers

Porcentaje Promotores

Nombre de la variable: porcentaje_promotores

Cálculo: qty_promoter / answers

Porcentaje Pasivos

Nombre de la variable: porcentaje_pasivos

Cálculo: qty_passive / answers

Indicador NPS

Nombre de la variable: NPS

Cálculo: porcentaje_promotores - porcentaje_detractores

Cantidad de fotos por producto.

Nombre de la variable: porcentaje_fotos

Cálculo: qty_picts / qty_products

Porcentaje de órdenes con demora en la entrega al rider mayor a 10 minutos

Nombre de la variable: porcentaje_orders_late

Cálculo: qty_order_late_10 / qty_orders

III.III - Tabla de datos analítica.

Durante el proceso de normalización o relativización de las variables se descartan las variables utilizadas para los cálculos y se seleccionan únicamente las normalizadas.

Las variables que se eliminan son: qty_triggers, voucher_order, rejected_orders, chats, sessions, qty_detractor, qty_passive, qty_promoters, answers, qty_picts, y qty_order_late_10.

Finalmente quedan las siguientes variables:

partner_id,

qty_products,

business_type_name

response_time_minute,

order_dias,

porcentaje_triggers

porcentaje_voucher

porcentaje_ordenes_canceladas

porcentaje_ordenes_confirmadas

chats_dias

porcentaje_sessions

NPS

porcentaje_fotos

porcentaje_orders_late

has_online_payment,

capacity_check,

has_shipping_amount

is_logistic

is_important_account

antigüedad

accepts_pre_order

has_discount

has_mov

has_custom_photo_menu

accepts_vouchers

IV - Exploración y Análisis Descriptivo

A continuación realizamos el Análisis Exploratorio de los Datos, en busca de un universo primario de análisis y las variables más relevantes.

Universo primario de análisis

En los siguientes puntos se analiza el universo de datos y se enfoca el universo de análisis en función de algunas de las variables.

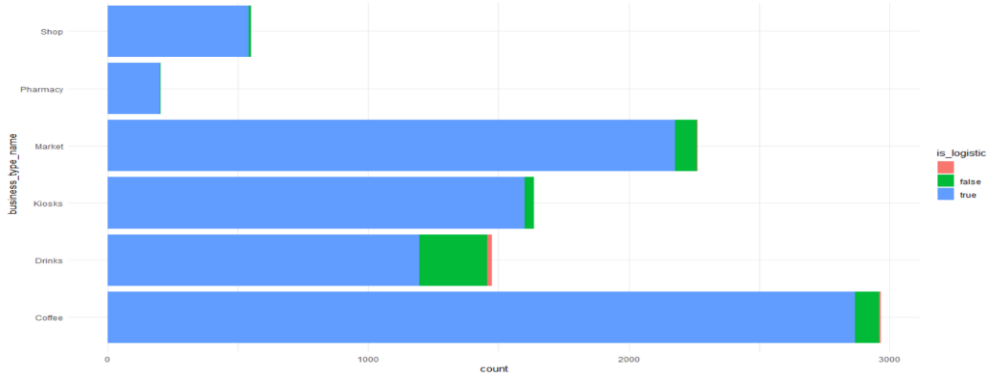
Análisis por tipo de partner

Se analizó la distribución de los partners en función del tipo de negocio para tomar la definición de cuáles se tomarán dentro de la vertical Groceries.

En la tabla siguiente se muestra la distribución del promedio de órdenes por tipo de partner.

Tipo de Negocio	Prom. Órdenes
Coffee	**3.000**
Drinks	**1.484**
Kiosks	**2.945**
Market	**2.714**
Pharmacy	**1.557**
Restaurant	**3.610**
Shop	**228**

En el siguiente gráfico se muestra la cantidad de partners por tipo de negocio (se quitó Restaurant para lograr una mejor visualización sabiendo que queda afuera de la vertical Groceries por definición), indicándose además la proporción de locales con logística de envío propia y a cargo de PeYa.



Los partners dentro Market y Kiosks tienen un promedio de órdenes similares; por otro lado, el porcentaje de los partners que cuentan con logística de envío de PeYa son similares para ambos tipos de negocio.

Se entiende que en el caso de Drinks, Pharmacy, Restaurant y Coffee son negocios que desde el punto de vista de los productos que ofrecen, la frecuencia de compra, y la premura en la entrega de los mismos no entran dentro de la categoría de Groceries.

Por tanto para el análisis de la vertical Groceries, se tomarán solamente dos categorías de partners: Market y Kiosks.

Análisis por antigüedad

La siguiente decisión que se tomó en base a la preparación de datos, fue dejar por fuera de este análisis a los Partners que tienen menos de 3 meses en la aplicación. Esta decisión se realizó considerando que en los primeros 3 meses los partners siguen dentro de la curva de aprendizaje y adaptación a la aplicación, y muchos de los datos observados no serán comparables con el resto con mayor antigüedad.

Tratamiento de NAs

Luego del filtrado por tipo de partner visto anteriormente, las variables con NAs son qty_triggers y qty_order_late_10.

Los NAs en qty_triggers son reemplazables por cero, debido a que cuando no hay datos es que no se disparó el trigger.

Para la variable qty_order_late_10 solo se cuenta el dato cuando PeYa se encarga del delivery. Consultado con el área funcional, si la entrega es de PeYa entonces, los NAs en qty_order_late_10 se consideran como cero. Quedan los NAs cuando el delivery está a cargo del partner.

En el caso de **capacity_check** cuando no tiene valor se asume que es que no no se estableció un límite de capacidad y se reemplaza por FALSE.

Análisis por logística de entrega.

Se observa que luego de seleccionar las categorías Kiosks y Market, solamente un 3,42% de los partners son los que tienen delivery propio. Si bien en un principio se pensó en descartar estas observaciones para poder utilizarlas en el modelo, finalmente se vio que no tienen una fuerte incidencia, por lo que se decidió no utilizar la variable ***qty_order_late_10*** para la clusterización y considerar el 100% de la muestra.

IV.I -Variables descartadas

Se descarta también la variable cualitativa **accepts_vouchers** porque se tiene esta información en forma cuantitativa del uso de vouchers en la columna **promedio_voucher.**

En este punto identificamos una cantidad de variables cualitativas, que degradan los algoritmos de clusterización. Solo podemos mantener unas pocas, utilizando la distancia de Gower para considerarlas, pero no demasiadas.

Se descartan las siguientes variables cualitativas para el modelado:

- has_shipping_amount
- is_logistic
- is_important_account
- accepts_pre_order
- has_discount
- has_mov
- has_custom_photo_menu

Se mantienen dos variables cualitativas (**has_online_payment** y **capacity_check),** que no serán utilizadas en ciertos algoritmos de clusterización porque no se pueden calcular las distancias sino solamente las disimilitudes, pero que a priori se entiende pueden aportar al modelo. Se evaluarán luego 2 modelos de manera de comprender si resulta conveniente tenerlas en cuenta o si se deben descartar.

El resto de las variables del set de datos que no están en esta lista se utilizarán en la caracterización de los modelos.

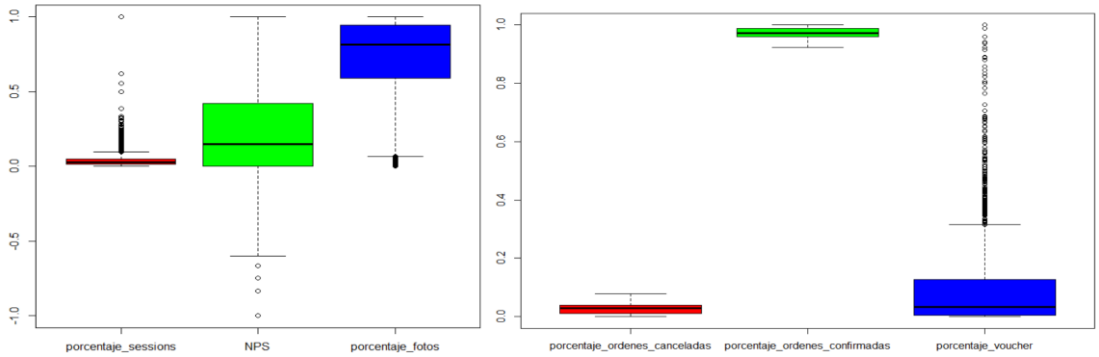
En este punto no se descartan más variables cuantitativas más allá de las eliminadas en la confección de las tablas de datos analíticas. Luego de eliminar las columnas indicadas más arriba que no serán consideradas en el modelo de segmentación, finalmente nos quedamos con las siguientes variables:

- partner_id,
- qty_products,
- response_time_minute,
- antiguedad,
- order_dias,
- chats_dias,
- porcentaje_triggers,
- porcentaje_voucher,
- porcentaje_ordenes_canceladas,
- porcentaje_sessions,
- porcentaje_fotos,
- NPS,
- has_online_payment,
- capacity_check)

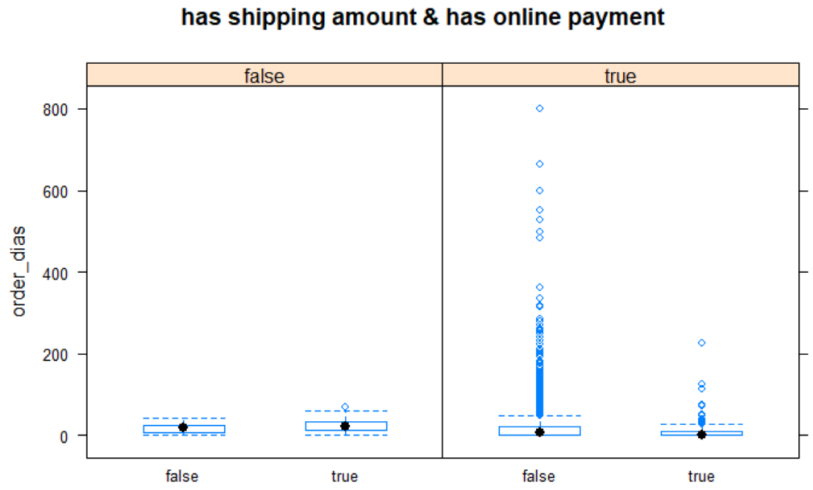
IV.III -Valores atípicos (outliers)

Los modelos de clusterización son sensibles a los valores atípicos u outliers, sin embargo, se debe ser muy cuidadoso en realizar este análisis en escenarios multivariados.

A continuación se muestran box plots de variables cuantitativas que fueron preseleccionadas para el análisis. Si bien este tipo de gráficos es utilizado para un análisis univariado, aporta cierta información acerca de la posible relación entre outliers.

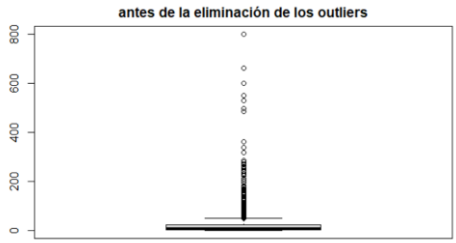


Se entiende que la variable cuantitativa promedio de órdenes diario es una variable muy relevante para entender a un partner, por lo que se muestran algunos box plot de esta variable en función de algunas variables cualitativas.



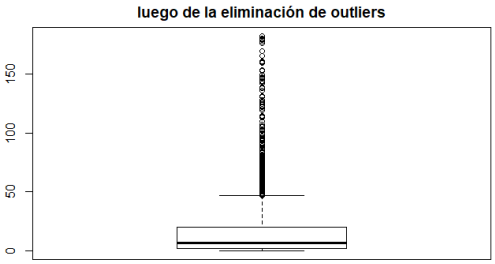
Estos gráficos muestran que en distintas variables del modelo, hay candidatos a outliers. Considerarlos como candidatos a outliers observando estas dimensiones de forma univariada puede privar al modelo de información en las otras variables, es por eso que se decidió tomar un criterio más conservador y realizar este estudio únicamente para una de las variables críticas del negocio como es la cantidad de órdenes por día.

Cantidad de Órdenes por Día

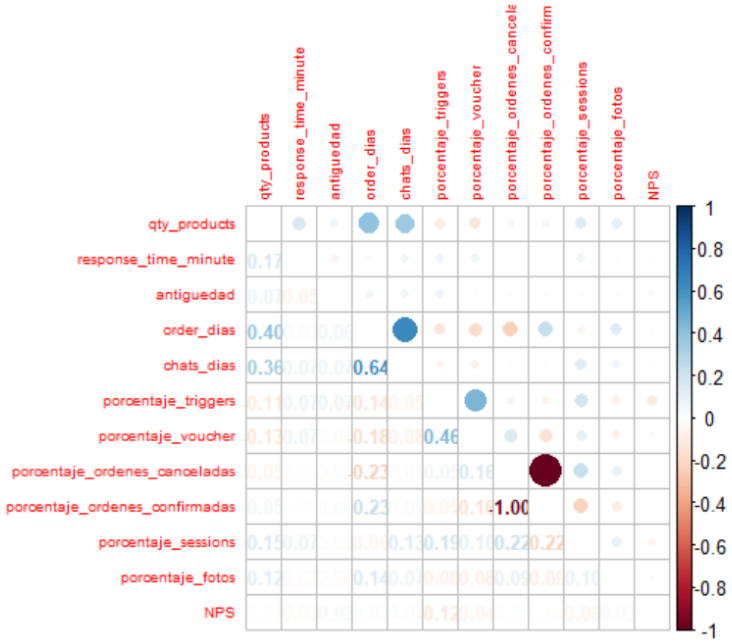


Se consideran candidatos a outliers aquellas observaciones que están por encima de las 190 órdenes diarias como los valores más lejanos.

El box plot para esta variable luego de eliminar estos outliers, se muestra a continuación.



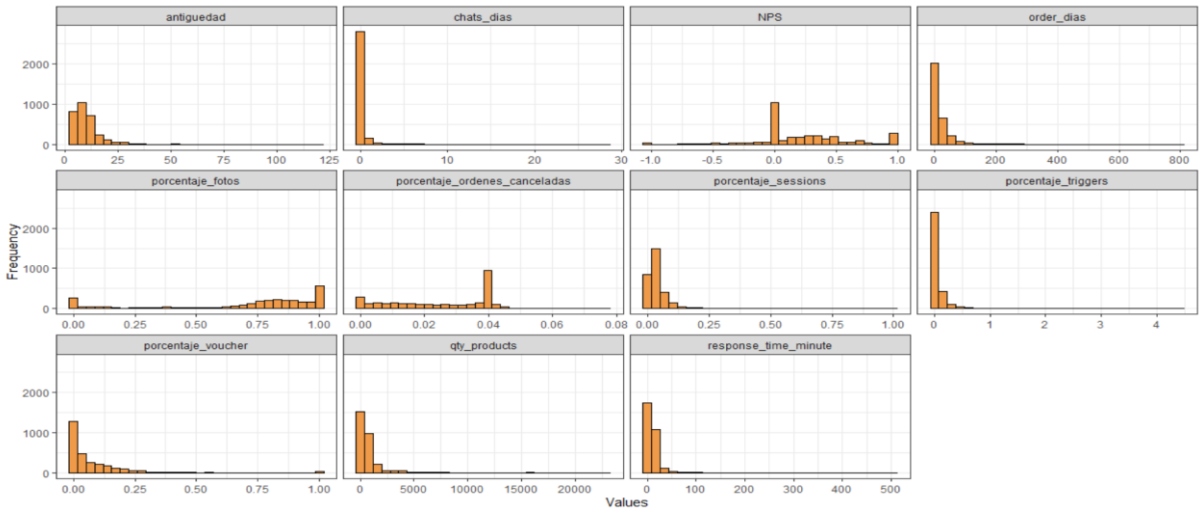
A continuación se analiza la matriz de correlación de las variables cuantitativas.



Se observa que el porcentaje de órdenes canceladas está correlacionada con el porcentaje de órdenes confirmadas, tal como era de esperarse. Se elimina la variables porcentaje_ordenes_confirmadas.

El gráfico siguiente muestra la distribución de las variables cuantitativas restantes en sendos histogramas. Luego de eliminados los outliers encontramos la siguiente distribución de variable

Se realiza un análisis de correlación, para ver qué tan correlacionadas están las variables que se utilizaran.



Se observa que la variable porcentaje_fotos presenta una distribución aplanada y por tanto será candidata a ser descartada durante el modelado.

V - Modelado y Evaluación

V.I - Variables pre-seleccionadas en base al análisis funcional y estadístico

Luego de realizados todos los pasos en función de las consideraciones mencionadas anteriormente, pre-seleccionamos el siguiente conjunto de variables para realizar el análisis:

- response_time_minute
- antigüedad
- qty_products
- order_dias
- chats_dias
- porcentaje_triggers
- porcentaje_voucher
- porcentaje_sessions
- porcentaje_fotos
- NPS
- porcentaje_ordenes_canceladas
- has_online_payment
- capacity_check

V.II - Estimación y comparación de modelos

Para el desarrollo de los modelos se definió utilizar 2 modelos con características diferentes, permitiendo abordar el problema permitiendo utilizar tanto variables mixtas (Cuantitativas y Cualitativas) así como variables únicamente Cualitativas.

En el caso de considerar variables mixtas se seleccionó un algoritmo que permitiese trabajar con la similitud de Gower; es por esta razón que se definió utilizar PAM.

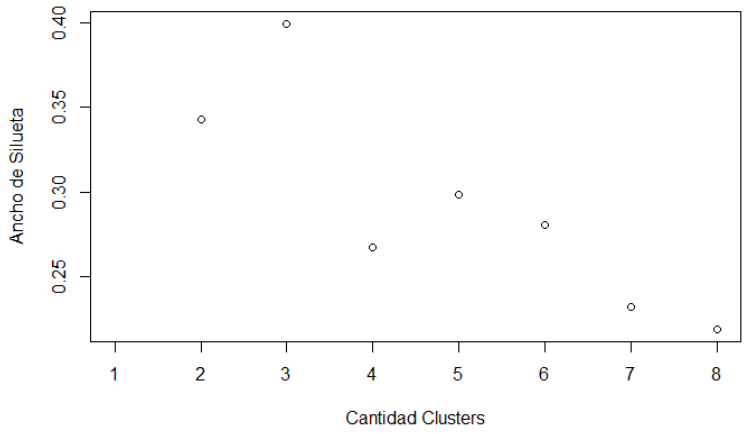
Para el análisis de las variables únicamente cuantitativas decidimos utilizar k-medias.

Comparando ambos modelos, además de que trabajan sobre diferentes tipos de datos, k-medias encuentra los centroides de forma más libre, mientras que PAM restringe los medioides a coincidir con una de las observaciones, PAM ademas es mas robusto frente a ruidos y outliers. PAM plantea como desventaja el tiempo de ejecución en relación a k-medias, sin embargo, para el conjunto de datos a analizar se entiende que no va a ser un obstáculo.

Ejemplo con todas las variables preseleccionadas.

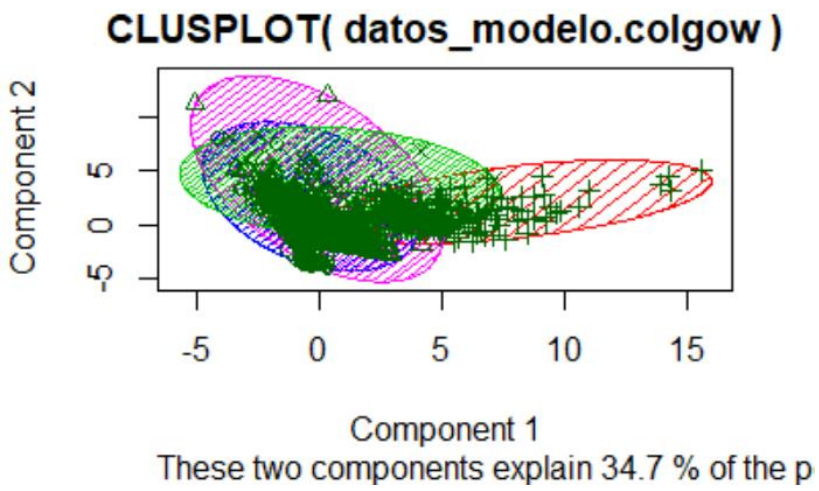
PAM

Mediante el método de silueta se busca determinar el número óptimo de clusters.



Observando el gráfico se presenta en 3 el mayor valor para segmentar, sin embargo, con 3 clusters, uno presenta muchas observaciones respecto de los otros dos. Por tanto, se modela con 4 clusters.

Gráfico de clusters y composición

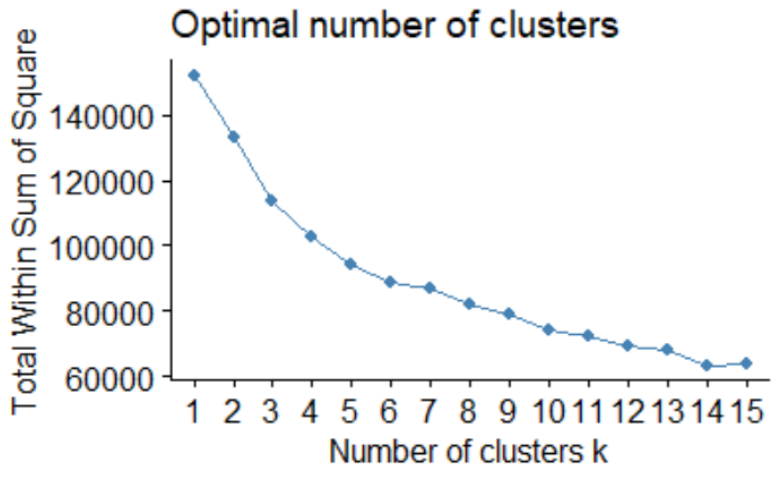


Con este modelo se obtienen la siguiente cantidad de individuos por cluster

Cluster 1	Cluster 2	Cluster 3	Cluster 4
593	983	226	1208

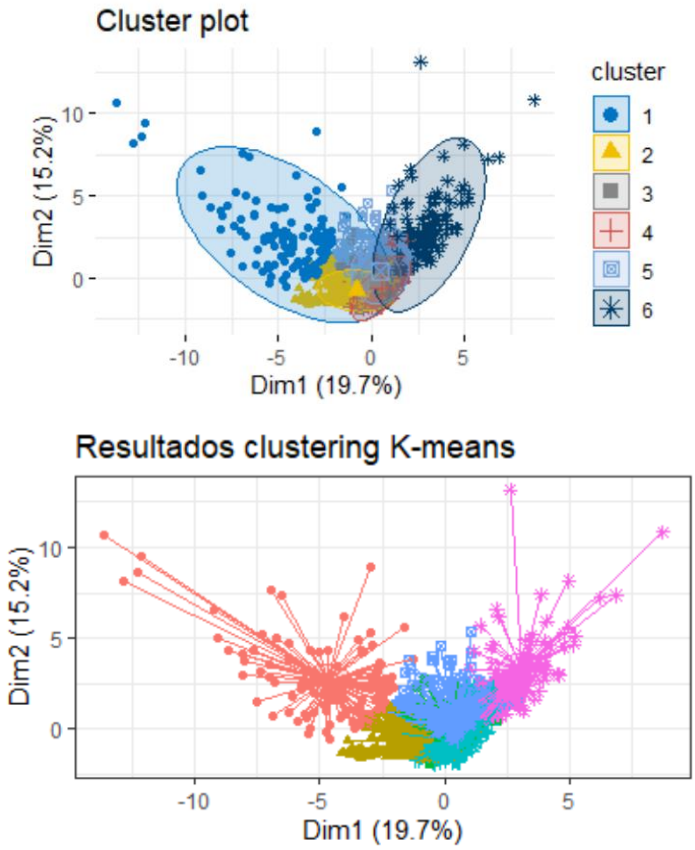
K-medias

Mediante el gráfico de codo se busca determinar el número óptimo de clusters.



En este caso se observa que en 6 podría presentarse un codo. Por lo que se define utilizar este número para clusterizar.

Gráfico de clusters y composición

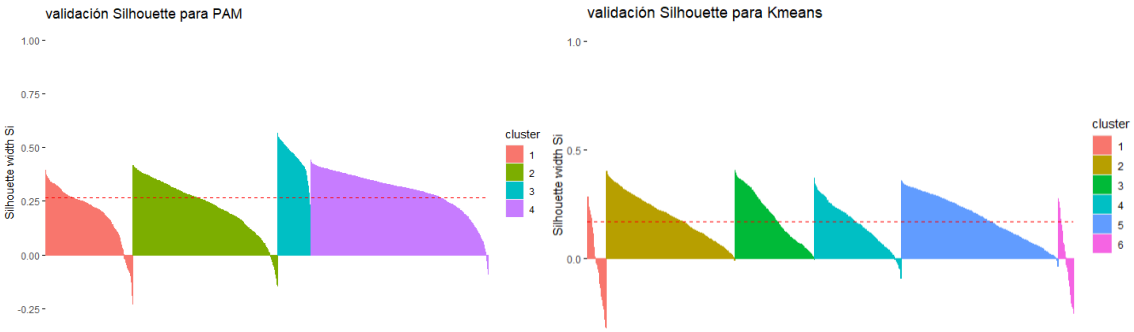


Con este modelo se obtienen la siguiente cantidad de individuos por cluster

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
118	795	492	536	972	97

Evaluación de modelo PAM y K-means

Se presenta a continuación una representación gráfica utilizando Silhouette como otro método para evaluar los clusters obtenidos por distintos modelos.



Cluster	PAM	K-Means
Cluster 1	0.20	-0.02
Cluster 2	0.23	0.19
Cluster 3	0.45	0.19
Cluster 4	0.30	0.15
Cluster 5		0.19
Cluster 6		0.00

Luego de realizar varios modelos se entiende que el conjunto de variables que más aportan para la segmentacion y definicion de los clusters son:

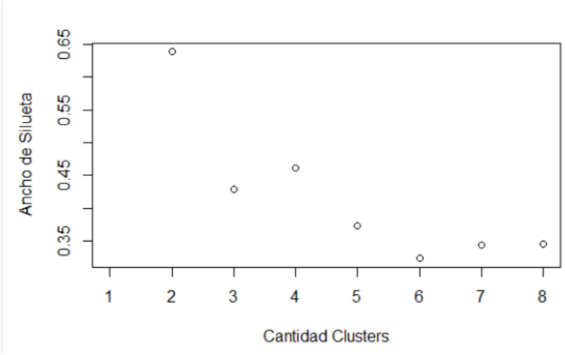
Variables finales para segmentacion:

- response_time_minute
- order_dias
- porcentaje_triggers
- porcentaje_voucher
- porcentaje_sessions
- porcentaje_ordenes_canceladas
- has_online_payment
- capacity_check

Los modelos finales obtenidos son:

PAM

Mediante el método de silueta se busca determinar el número óptimo de clusters.



Observando el gráfico se presenta en 2 el mayor valor para segmentar, sin embargo, este valor es un número muy bajo. En 4 luego se encuentra un nuevo máximo y se utilizará este valor para definir el número de clusters.

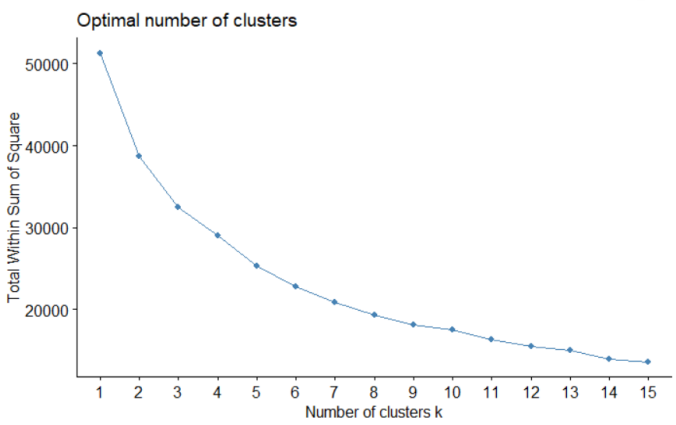
Con este modelo se obtienen la siguiente cantidad de individuos por cluster.

Cluster 1	Cluster 2	Cluster 3	Cluster 4
104	1516	1163	227

Se realizó una validación mediante Silhouette

K-medias

Mediante el gráfico del codo buscamos el número óptimo de clusters.



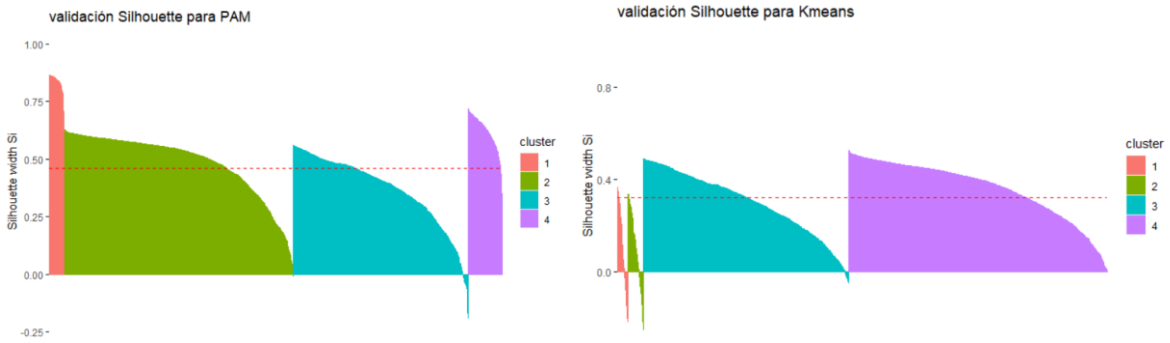
En este caso se observa que en 9 podría presentarse un codo, pero no es muy significativo. Realizar una segmentación por 9 clusters resulta un valor muy alto, por lo que se define utilizar 4 en donde se observa otro punto de inflexión y los clusters son algo más homogeneos.

Con este modelo se obtienen la siguiente cantidad de individuos por cluster.

Cluster 1	Cluster 2	Cluster 3	Cluster 4
65	95	1262	1588

Evaluación de modelo PAM y K-means

Se presenta a continuación una representación gráfica utilizando Silhouette como otro método para evaluar los clusters obtenidos por distintos modelos.



Cluster	PAM	K-Means
Cluster 1	0.80	0.08
Cluster 2	0.49	0.09
Cluster 3	0.37	0.30
Cluster 4	0.62	0.36

Se decide utilizar PAM como algoritmo de segmentación ya que se encuentran clusters más homogéneos. Adicionalmente se entiende que al ser PAM menos sensible a los outliers es un modelo más robusto que permite escalar de mejor forma.

Asignación de Outliers o nuevos individuos.

Una vez identificados los clusters se deben incorporar aquellos datos que fueron definidos como candidatos a Outliers. Para realizar esta tarea se escolarizan los valores cuantitativos utilizando la media y la desviación estándar del data set sin outliers, se calcula la distancia de Gower para las variables cualitativas y finalmente se asigna a cada uno de los clusters minimizando la distancia con cada uno de los medoides.

Cluster	Entrenamiento	C. Outliers	Total
Cluster 1	104	0	104
Cluster 2	1516	0	1516
Cluster 3	1163	6	1169
Cluster 4	227	23	250

Medoides resultantes de la segmentación para futuras clasificaciones.

partner_id	response_time_minute	order_dias	porcentaje_triggers	porcentaje_voucher	porcentaje_sessions	porcentaje_ordenes_canceladas	has_online_payment	capacity_check
12542	5,46093466	19,2312925	0,01450301	0,02971348	0	0	false	false
29888	6,8287037	1,96111111	0,0368272	0,06798867	0,03116147	0,03966006	true	false
11274	8,00769231	13,4722222	0,00453608	0,01319588	0,02639175	0,00989691	true	false
7964	14,7448718	37,5705882	0,00407077	0	0,04634414	0,0303742	true	true

V.III - Caracterización y descripción de los resultados obtenidos.

Análisis de Variables Cualitativas

Se buscará caracterizar en función de las variables cualitativas. Para esto, se medirá la frecuencia de cada una de ellas y así entender si se identifican valores predominantes de alguna de ellas dentro de los clusters identificados.

business_type_name - Tipo de negocio

Cluster	Kiosks	Market
1	2	102
2	736	780
3	574	595
4	0	250

- Se observa que en el cluster 1 y 4 son mayoritariamente Market.

delivery_time - Tiempo de envío

Cluster	Entre120Y 150	Entre150Y 180	Entre15Y 30	Entre30Y 45	Entre45Y 60	Entre60Y 90	Entre90Y 120	Horas12	Horas48
1	0	0	5	99	0	0	0	0	0
2	0	0	228	1209	78	1	0	0	0
3	0	0	184	936	47	2	0	0	0
4	0	0	44	60	18	128	0	0	0

- Cluster 4 con mayor tiempo estimado de entrega.

has_shipping_amount - Tiene costo de envío

Cluster	false	true	porcentaje TRUE
1	17	87	83,65%
2	1401	115	7,59%
3	1081	88	7,53%
4	225	25	10,00%

- En el cluster 1 mayormente tienen costo de envío.

has_online_payment - Tiene pago online

Cluster	false	true	porcentaje TRUE
1	104	0	0,00%
2	0	1516	100,00%
3	0	1169	100,00%
4	2	248	99,20%

- cluster 1 no tienen pago online y el resto sí.

has_discount - Tiene descuento

Cluster	false	true	porcentaje TRUE
1	104	0	0,00%
2	1515	1	0,07%
3	1168	1	0,09%
4	249	1	0,40%

- Esta variable se comporta igual en todos los clusters.

is_important_account - Es una cuenta importante

Cluster	false	true	porcentaje TRUE
1	103	1	0,96%
2	1258	258	17,02%
3	921	248	21,21%
4	14	236	94,40%

- Cluster 4 son mayormente cuentas importantes.

has_mov - Tiene valor mínimo

Cluster	false	true	porcentaje TRUE
1	99	5	4,81%
2	171	1345	88,72%
3	143	1026	87,77%
4	8	242	96,80%

- Cluster 1 mayormente no tiene minimo de compra.

accepts_vouchers - Acepta vouchers

Cluster	false	true	porcentaje TRUE
1	0	104	100,00%
2	96	1420	93,67%
3	104	1065	91,10%
4	11	239	95,60%

- Esta variable se comporta prácticamente igual en todos los clusters.

accepts_pre_order - Acepta pre órdenes

Cluster	false	true	porcentaje TRUE
1	0	104	100,00%
2	142	1374	90,63%
3	162	1007	86,14%
4	0	250	100,00%

- Esta variable se comporta prácticamente igual en todos los clusters y aceptan pre-ordenes

has_custom_photo_menu - Tiene fotos reales

Cluster	false	true	porcentaje TRUE
1	74	30	28,85%
2	1346	170	11,21%
3	1043	126	10,78%
4	240	10	4,00%

- Esta variable se comporta prácticamente igual en todos los clusters y no tienen fotos reales.

capacity_check - Tiene límites de órdenes

Cluster	false	true	porcentaje TRUE
1	104	0	0,00%
2	1516	0	0,00%
3	1169	0	0,00%
4	0	250	100,00%

- Cluster 4 es el único que tiene límite de capacidad.

is_logistic - Tiene delivery por PeYa

Cluster	false	true	porcentaje TRUE
1	0	104	100,00%
2	67	1449	95,58%
3	37	1132	96,83%
4	1	249	99,60%

- Esta variable se comporta prácticamente igual en todos los clusters y casi todos tienen delivery de PEYA

Analisis Variables cuantitativas

Se plantea ahora una caracterización utilizando las variables cuantitativas.

En este análisis se presenta cada una de las variables, su valor promedio en cada cluster, y el valor promedio del total de partners.

Se remarcan con Amarillo las celdas que estan por debajo del promedio y con verde aquellas que estan por encima del promedio general de los datos. De esta manera se puede identificar un comportamiento por cluster respecto a la media de la variable.

Variables	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Prom. Gral
qty_products	1614.952	470.092	511.695	6976.364	1060.506
response_time_minute	7.796	11.136	11.068	19.691	11.699
antigüedad	5.577	10.226	11.067	11.324	10.481
order_dias	23.124	6.446	23.824	77.810	19.573
porcentaje_triggers	0.031	0.097	0.068	0.014	0.077
porcentaje_voucher	0.068	0.134	0.071	0.022	0.098
porcentaje_ordenes_canc	0.004	0.039	0.010	0.025	0.025
chats_dias	0.002	0.062	0.069	1.593	0.189
porcentaje_sessions	0.004	0.047	0.029	0.071	0.041
NPS	0.131	0.219	0.230	0.210	0.219
porcentaje_fotos	0.175	0.691	0.705	0.872	0.694
porcentaje_orders_late (*)	0.097	0.446	0.351	0.420	0.395

(*) Calculado con menos observaciones, utilizando *NA omit*

Descripción de los clusters

Una vez analizadas las variables cualitativas y cuantitativas se describen cada uno de los clusters en lenguaje natural que permita una más fácil identificación por parte del negocio. Esta clasificación no es categórica, ni significa que todos los individuos la cumplan, sino que se trata de una descripción general con la que la mayoría de los elementos de un cluster pueden identificarse.

Cluster 1: El cluster 1 se caracteriza por agrupar a partners que son del tipo Market con menor antigüedad, y en su mayoría no son cuentas importantes. Como puntos fuertes: no tienen valor mínimo de compra, tienen la menor cantidad de órdenes tarde, el menor tiempo de respuesta, el menor porcentaje de órdenes canceladas, y la menor cantidad de chats y sesiones por órdenes.

Como puntos débiles: no cuentan con pago on-line, tienen costo de envío, tienen el menor NPS, y cuentan el menor número de fotos por producto en su catálogo.

Cluster 2: El cluster 2 está compuesto por partners de tipo de negocio Market y Kiosks en proporciones similares, y tienen la menor cantidad de productos en el histórico y con el mayor porcentaje de órdenes con voucher. Como puntos débiles se puede mencionar que son los que tienen mayor porcentaje de órdenes canceladas, mayor porcentaje de órdenes con llegadas tarde, mayor cantidad de triggers y menor cantidad de órdenes diarias. Este cluster es el que presenta mayores dificultades en prácticamente todas las áreas.

Cluster 3: El cluster 3 está compuesto por partners de tipo de negocio Market y Kiosks con el mayor NPS. Este cluster presenta a los partners con un comportamiento promedio o estándar.

Cluster 4: Los partners de este cluster pertenecen a Market, son los que tienen mayor antigüedad, en su mayoría son cuentas importantes y con la mayor cantidad de productos en el histórico. Como puntos fuertes, se puede mencionar que tienen la mayor cantidad de órdenes diarias, el menor porcentaje de triggers, la mayor cantidad de fotos por producto, y son los que menos utilizan el método de estímulo de ventas por vouchers. Como contrapartida, son los que tienen más chats y sesiones por órden, mayor tiempo de entrega, mayor tiempo de respuesta/confirmación de órdenes, y cuentan con control de límites de órdenes. Por el volumen de órdenes y la cantidad de productos en histórico, se entiende que en este cluster se encuentran las grandes cadenas de supermercados.

A continuación se muestra en formato tabla las características más relevantes de los diferentes clusters de acuerdo a las variables del modelo, identificando con verde aquellas que tienen una connotación positiva y con rojo las de connotación negativa. Para las variables antiguedad y porcentaje_vouchers, se indica en la tabla qué significan los colores verde y rojo, dado que la interpretación puede resultar ambigua.

variable	Cluster 1	Cluster 2	Cluster 3	Cluster 4
qty_products				
response_time_minute				
antiguedad	más nuevo			más antiguo
order_dias				
porcentaje_triggers				
porcentaje_voucher		más vouchers		menos vouchers
porcentaje_ordenes_canceladas				
chats_dias				
porcentaje_sessions				
NPS				
porcentaje_fotos				
porcentaje_orders_late				
tiempo de envío				
costo de envío				
pago on line				
cuentas importantes				
valor mínimo de compra				
límite de órdenes				

VI – Distribución

VI.I - Aplicación al problema de negocio

Se entregan los partners clasificados o caracterizados en grupos que tienen comportamientos similares. De esta forma, se podrán definir estrategias segmentadas para ayudarlos en los problemas identificados, o potenciar a los que tienen buenos indicadores desde distintos puntos de vista. Es importante medir periódicamente el impacto de estas acciones para evaluar su eficacia mediante los distintos indicadores que estén definidos.

Adicionalmente se entregan también los artefactos técnicos necesarios para que a medida que vayan ingresando nuevos partners, pasado el tiempo definido para que sean comparables, se puedan clasificar dentro de uno de los Clusters definidos. Esto permite anticiparse a los problemas o potenciar oportunidades, dado el comportamiento presentado en el cluster. Como ejemplo de código para incorporar nuevos Partners, se puede tomar como referencia el RMarkDown entregado para clasificar a los candidatos a outliers.

VI.II - Accionables

Área Operativa:

Cluster 2:

- Si bien son el cluster con menor cantidad de productos, son los que tienen la mayor cantidad de órdenes canceladas, lo que puede estar incidiendo en el bajo nivel de órdenes. Por tanto, se debería trabajar con los partners para mejorar el control de stock de sus productos.
- Dado que se producen entregas tardías al rider, recomendamos que se trabaje con los partners de este cluster para mejorar los tiempos de preparación de las órdenes para entregar al repartidor. Otra medida, para mitigar este problema podría ser hacer uso del límite de órdenes que no está siendo muy utilizado en este cluster.

Cluster 4:

- Si bien estos partners son los que tienen mayor cantidad de órdenes diarias, presentan la mayor demora en confirmar las órdenes; seguramente esto explique el alto volumen de chats y sesiones, recomendamos que se determinen estrategias conjuntas con el partner para mejorar los tiempos de confirmación de las órdenes.
- Dado que todos los partners de este cluster cuentan con límite de órdenes (capacity check), evaluar si los valores límite son correctos o se deben ajustar de modo de afrontar los picos de demanda y poder mejorar los tiempos de confirmación.

Área Comercial

Cluster 1:

- A pesar de tener un buen desempeño en el área operativa, los partners de este cluster cuentan con el menor índice de satisfacción de clientes (aunque el valor de NPS es aceptable). Lo que puede estar ocurriendo con el índice de satisfacción, es que éste se esté viendo afectado por no contar con pago online y además tener costo de envío. Por tanto se recomienda que se trabaje con estos partners de modo de implementar el pago online y no trasladar el costo de envío a los clientes.
- Al tratarse de partners nuevos, se recomienda que el área de Comercial los apoye para tener mayor visibilidad y utilizar las promociones (vouchers, descuentos).
- Trabajar con los partners para que agreguen fotos en sus productos, de modo de incrementar su NPS.

Cluster2:

- A pesar que en este cluster se encuentran muchas órdenes estimuladas con vouchers, operativamente están teniendo serios problemas (órdenes canceladas, llegadas tarde y mayor cantidad de triggers). Se recomienda disminuir este estímulo en tanto no se solucionen los problemas operativos.

Cluster3:

- A pesar de que este cluster no presenta problemas operativos y tiene el más alto NPS, no muestra el volumen más alto de órdenes. Dado que se observa que este cluster es de los que tiene la menor cantidad de productos, se recomienda trabajar con el partner para que aumente el número de órdenes mediante el aumento de variedad de productos disponibles.
- Para aumentar el nivel de órdenes, otra estrategia sería mediante campañas de marketing que pueden incluir mayor visibilidad de estos partners, hasta el uso de estímulos mediante vouchers y/o descuentos.

VI.III - Implementación del modelo

Una vez que se cuenta con los clusters y la definición de los medoides (para el caso clusterización PAM como fue el nuestro), se debe sistematizar la clasificación de nuevos partners dentro del cluster correspondiente, una vez transcurrida la ventana de tiempo definida para se pueda contar con la información suficiente y comparable. Para esto, en conjunto con el departamento de TI, se deben definir procesos automáticos que utilizando los medoides y el código de asignación a clusters que se realizó en el notebook entregado, permita de forma batch realizar esta tarea.

Se entiende que la incorporación de los nuevos partners en tiempo real no tiene ningún beneficio significativo respecto a incorporarlos en forma batch. Como quedó descrito en este análisis los Partners deben tener cierto tiempo de permanencia en la plataforma previo a poder incorporarlos a este tipo de análisis.

VI.IV - Evaluación del éxito del proyecto

Para evaluar el resultado de este trabajo se recomienda volver a relevar los datos de los partners 3 meses después de implementadas las medidas, y verificar cuál fue el desempeño de los partners en las distintas métricas en las que los analizamos.

VI.V - Oportunidades de mejora

A continuación se describen las principales limitantes encontradas y oportunidades de mejora.

Comprender cabalmente cada uno de los tipos de partners de modo de entender si clasifican dentro de la vertical de Groceries, podría ayudar a seleccionar correctamente el universo de datos.

En el análisis de outliers, no se contó con las herramientas adecuadas como para hacer un estudio multivariado, que sería lo recomendado en este caso.

El desconocimiento del negocio, nos puede haber llevado a tomar supuestos que podrían no ser válidos.

Para algunos partners, no se pudo utilizar la variable *cantidad de órdenes tarde entregadas al rider* para la clusterización debido a que existen un conjunto para los cuales no tenemos información, y que son aquellos que no cuentan con logística de PeYa. Se plantea la posibilidad de definir un mecanismo alternativo mediante el cual, aunque tenga delivery propio, se puedan registrar si hubo demoras en los tiempos de preparación.

Si se pudiera contar con la cobertura geográfica de los partners, permitiría hacer un mapa de oferta y demanda por zonas, distribuyendo a los riders de modo de evitar congestiones en los pedidos.

También falta el dato de facturación, lo cual es clave en términos de negocio para PeYa. Esto seguramente tendría un fuerte impacto en la clusterización y segmentación.

Hoy contamos con el tiempo de confirmación y si se entregó tarde, si pudiésemos contar con el tiempo de preparación del pedido, podremos comprender mejor cómo se explican las entregas tardías.

Otra alternativa de mejora sería poder contar con el horario en que opera el partner, de modo de optimizar el uso de los riders trabajando con partners que operen en distintos horarios.

El no contar con identificación de los clientes finales, no permite entender totalmente el comportamiento de los clientes de un partner e imposibilita realizar campañas de fidelización como una posible estrategia de aumento de ventas.