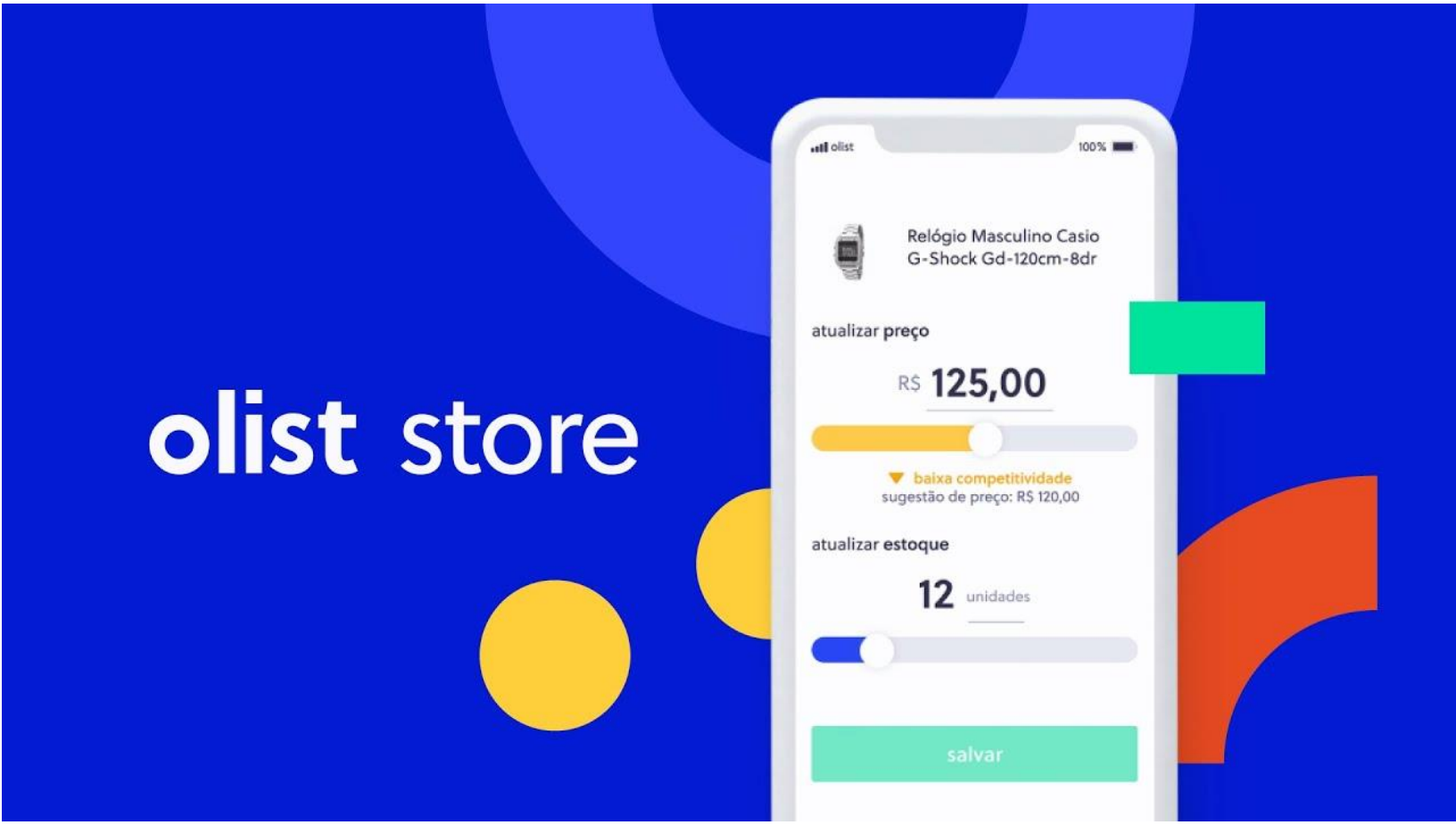
PROYECTO FINAL

Clasificación de clientes de E-Commerce



SEPTIEMBRE 9

CODERHOUSE
Integrantes del equipo: : Sánchez Wilmer – Rendón Luis - Domínguez Sofia – Reynoso Viviana

Contenido

Descripción del caso

Presentación de la empresa

Olist es una empresa de retail que ofrece servicios de E-commerce en México y Brasil. Su modelo de negocio permite apalancar a usuarios minoristas para realizar ventas a través de su plataforma online, ofreciendo además servicios de apoyo en logística de entregas, organización de catálogos, posicionamiento en sitios de Marketplace como Mercado Libre y asesoramiento para mantener la operación.

Objetivo de la investigación

Predecir una experiencia ***perfecta*** de compra para los clientes que adquieren productos a través del portal web, haciendo uso de la data colectada por la empresa entre 2016 y 2018

La experiencia de compra puede estar afectada por varios factores como ejemplo:

- Las fechas de entrega con retraso
- El precio de los productos
- El perfil del cliente

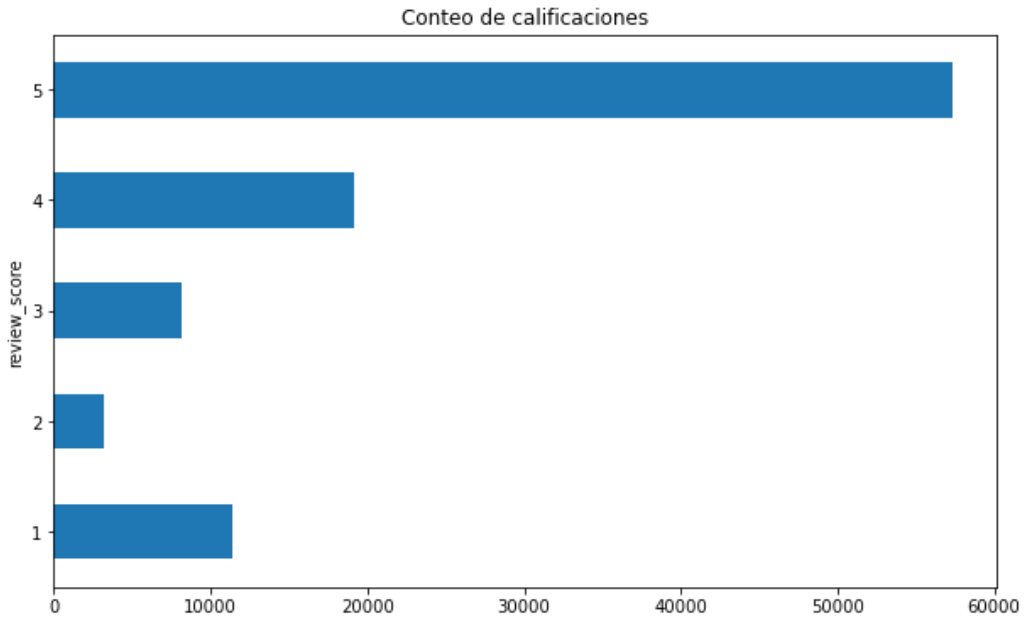
En términos generales los pasos que se van a usar para alcanzar el objetivo son:

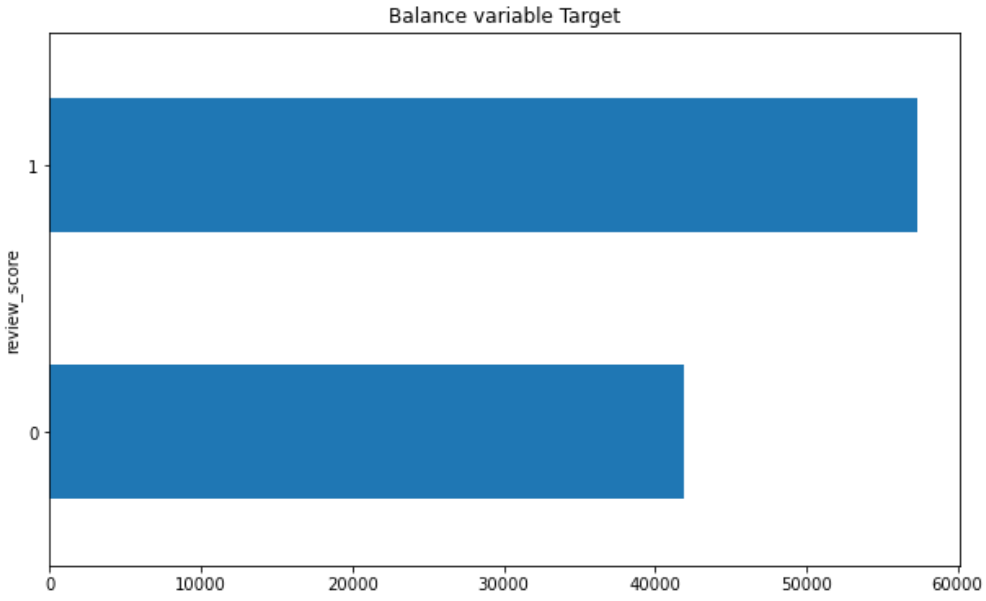
1. Analizar las variables que se consideran más relevantes para el objetivo
2. Filtrar y ajustar los datos
3. Uso de un algoritmo de clustering para perfilar los clientes y agregar los resultados al dataset
4. Entrenamiento de diversos modelos de clasificación
5. Comparar los resultados

Balance de la variable Target

Las calificaciones perfectas son aquellas de 5/5 y componen la mayoría de los datos, para ajustarse al objetivo de la investigación el dataset se balaceará al combinar calificaciones imperfectas 1,2,3 y 4, de la siguiente forma

- 1 si el score del cliente es 5 (Cliente satisfecho)
- 0 si el score del cliente es diferente de 5 (Cliente insatisfecho)


Conteo de calificaciones
review_score
5
4
3
2
1
0
10000
20000
30000
40000
50000
60000

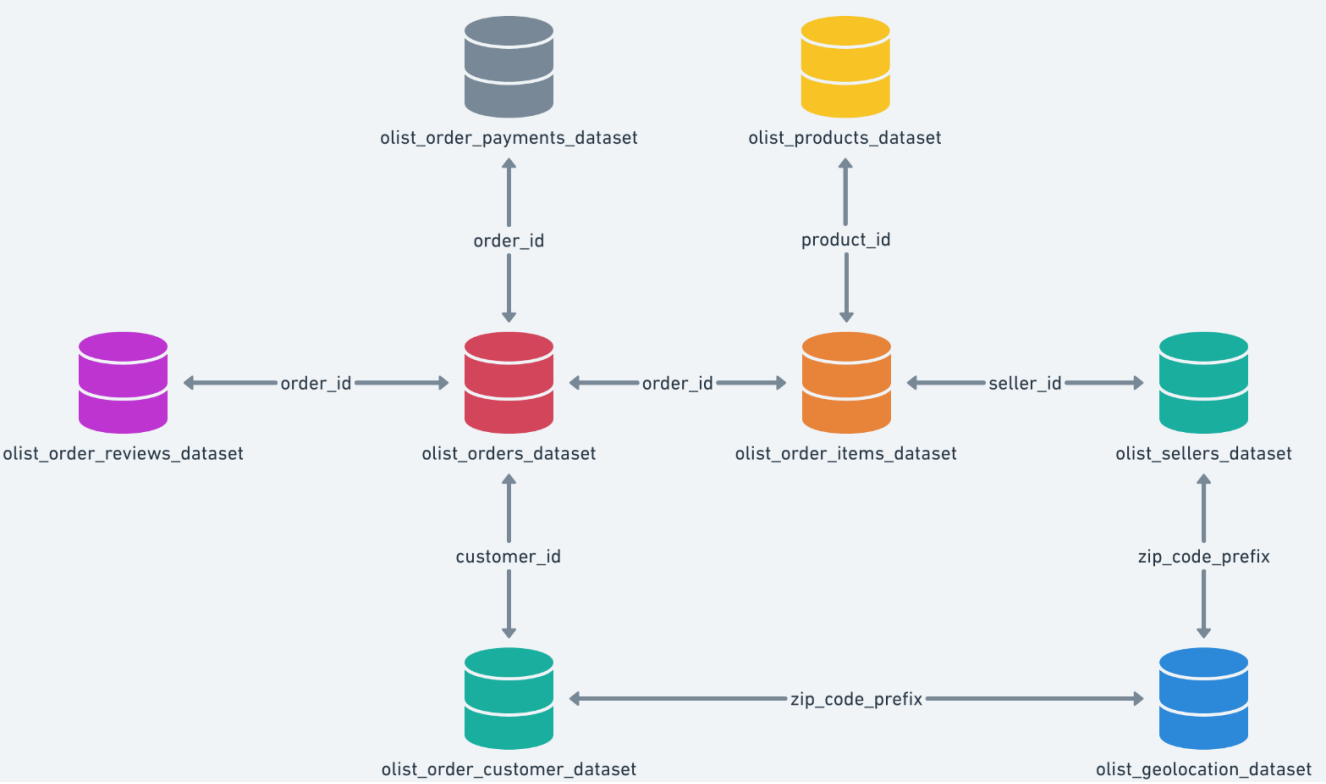

Balance variable Target
review_score
1
0
0
10000
20000
30000
40000
50000
60000

Descripción del Dataset

Es una base de datos relacional conformada por nueve archivos csv cada uno de los cuales contiene información detallada de un aspecto del negocio

1. Clientes
2. Ubicación
3. Artículos por pedido
4. Pagos
5. Pedidos
6. Calificaciones
7. Productos
8. Vendedores
9. Traducción de categorías

Los archivos están relacionados como se muestra en el siguiente diagrama:



Clientes (olist_customers_dataset)

Nombre de columna	*Descripción*
customer_id	Cada orden tiene un id único de cliente
customer_unique_id	Identificador único de cliente
customer_zip_code_prefix	Código postal
customer_city	Nombre de la ciudad donde está el cliente
customer_state	Nombre del estado donde está el cliente

Ubicación (olist_geolocation_dataset)

Nombre de columna	*Descripción*
geolocation_zip_code_prefix	Código postal
geolocation_lat	Latitud
geolocation_lng	Longitud
geolocation_city	Ciudad
geolocation_state	Estado

Artículos por pedido (olist_order_items_dataset)

Nombre de columna	*Descripción*
order_id	Identificador único de un pedido
order_item_id	Identificador de los artículos de un pedido
product_id	Identificador único de un producto en venta
seller_id	Identificador único de un vendedor
shipping_limit_date	Fecha límite de entrega del vendedor para manejo y empaque
price	Valor del artículo
freight_value	Valor del flete. En un pedido con varios artículos este valor se divide entre el número de artículos

Pagos (olist_order_payments_dataset)

Nombre de columna	*Descripción*
order_id	Identificador único de un pedido
payment_sequential	Número secuencial por cada método de pago usado por un cliente
payment_type	Método de pago seleccionado • credit_card • boleto • voucher • debit_card • not_defined
payment_installments	Cantidad de cuotas seleccionadas para diferir el pago
payment_value	Valor total pagado por el cliente

Pedidos (olist_orders_dataset)

Nombre de columna	*Descripción*
order_id	Identificador único de un pedido
customer_id	Identificador único de cliente
order_status	Estatus de la orden • delivered • shipped • canceled • unavailable • invoiced • processing • created • approved
order_purchase_timestamp	Fecha la compra de cada pedido
order_approved_at	Fecha de aprobación del pedido
order_delivered_carrier_date	Fecha de entrega al carrier
order_delivered_customer_date	Fecha de entrega al cliente
order_estimated_delivery_date	Fecha estimada de entrega al hacer la compra

Calificaciones (olist_order_reviews_dataset)

Nombre de columna	*Descripción*
review_id	Identificador único de la calificación del cliente
order_id	Identificador único del pedido calificado
review_score	Calificación, valor de 1 a 5
review_comment_title	Título del comentario del cliente (En portugués)
review_comment_message	Comentario del cliente (En portugués)
review_creation_date	Fecha en la que se le envió la encuesta al cliente
review_answer_timestamp	Fecha en la que el cliente contesto la encuesta

Productos (olist_products_dataset)

Nombre de columna	*Descripción*
product_id	Identificador único del producto
product_category_name	Categoría del producto (73 categorías)
product_name_lenght	Longitud del nombre del producto
product_description_lenght	Longitud del texto que describe el producto
product_photos_qty	Cantidad de fotos de la publicación
product_weight_g	Peso en gramos del producto
product_length_cm	Longitud del producto en centímetros
product_height_cm	Altura del producto en centímetros
product_width_cm	Ancho del producto en centímetros

Vendedores (olist_sellers_dataset)

Nombre de columna	*Descripción*
seller_id	Identificador único del vendedor
seller_zip_code_prefix	Código postal de la ubicación del vendedor
seller_city	Ciudad a la que pertenece el vendedor
seller_state	Estado al que pertenece el vendedor

Traducción nombre de categorías (product_category_name_translation)

Nombre de columna	*Descripción*
product_category_name	Nombre de categorías en portugués
product_category_name_english	Nombre de categorías en inglés

Análisis Exploratorio de Datos (EDA)

Datos del negocio

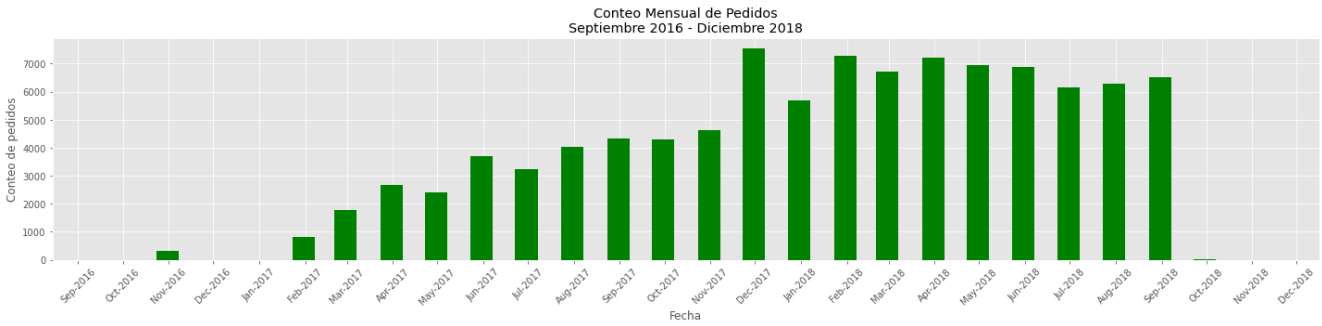
El csv que contiene la variable target es el de Calificaciones, este contiene 99 mil registros lo cual da una referencia del volumen de datos que tendremos disponibles para trabajar. La cantidad de clientes únicos es de 96 mil lo que significa que hay clientes que hacen varias compras y por registran más calificaciones.

Olist no vende directamente los productos solo ofrece servicios de logística por lo que también existe un alto volumen de vendedores, se identificó que hay una relación de 1 vendedor por cada 30 clientes.

En total el E-Commerce se han registrado 33 mil productos en 73 diferentes categorías

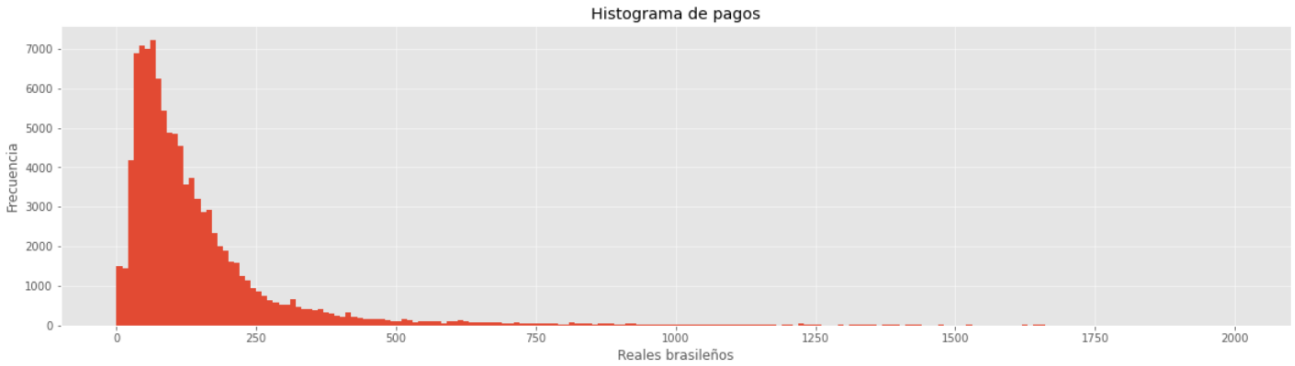
Periodo de recolección de los datos

Analizando la columna de "fecha de despacho" para los pedidos, se tiene que los datos fueron colectados con mayor consistencia entre febrero de 2017 y septiembre de 2018. El número mayor de compras fue en el mes de diciembre de 2017



Análisis de los pagos por orden

La gran mayoría de pagos son de montos que varían entre 0 y 500 reales brasileños



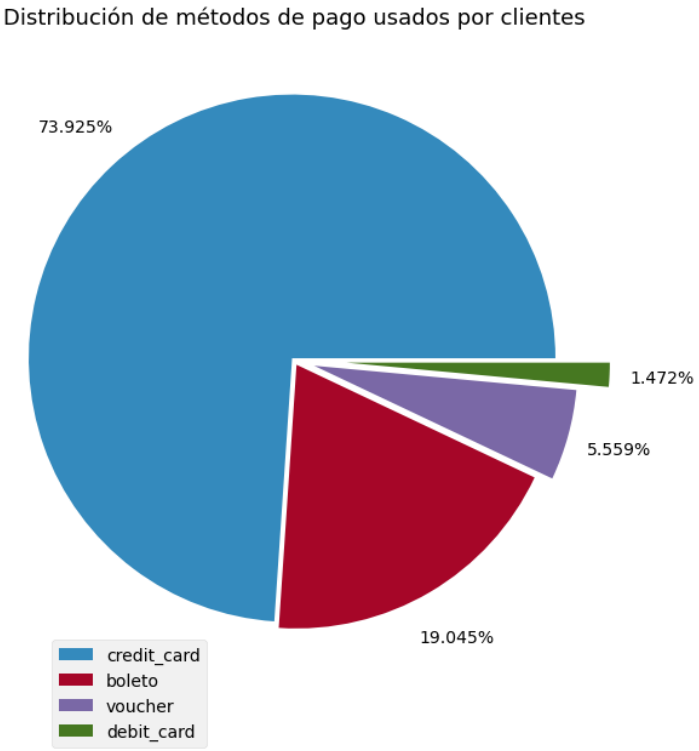
Análisis temporal del estado de los pedidos

La proporción de pedidos en estado de “entregado” es mucho mayor en el periodo febrero 2017-septiembre 2018, esto lo que significa es la variable order_status no tiene poder de clasificación por lo tanto será descartada en los modelos de clasificación

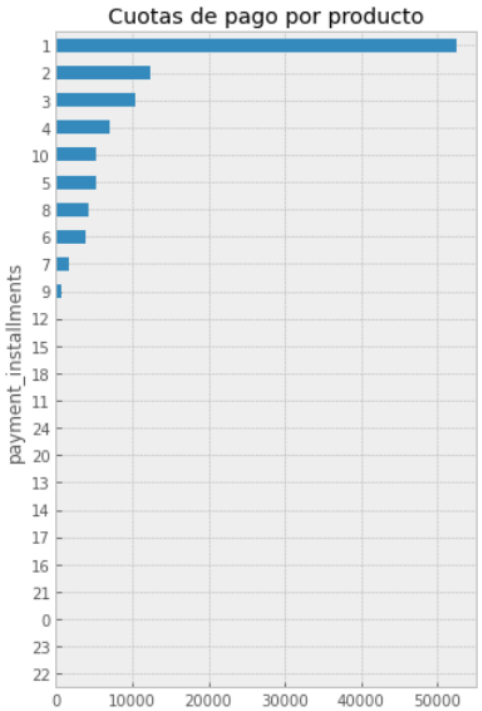
order_purchase_timestamp	delivered	invoiced	shipped	processing	unavailable	canceled	created	approved
2016-09-30 00:00:00	1	0	1	0	0	2	0	0
2016-10-31 00:00:00	265	18	8	2	7	24	0	0
2016-11-30 00:00:00	0	0	0	0	0	0	0	0
2016-12-31 00:00:00	1	0	0	0	0	0	0	0
2017-01-31 00:00:00	750	12	16	9	10	3	0	0
2017-02-28 00:00:00	1653	11	21	32	45	17	0	1
2017-03-31 00:00:00	2546	3	45	23	32	33	0	0
2017-04-30 00:00:00	2303	14	49	10	9	18	0	1
2017-05-31 00:00:00	3546	16	55	23	31	29	0	0
2017-06-30 00:00:00	3135	11	47	12	24	16	0	0
2017-07-31 00:00:00	3872	7	56	11	52	28	0	0
2017-08-31 00:00:00	4193	20	41	18	32	27	0	0
2017-09-30 00:00:00	4150	17	38	22	38	20	0	0
2017-10-31 00:00:00	4478	16	33	20	58	26	0	0
2017-11-30 00:00:00	7289	35	72	25	84	37	2	0
2017-12-31 00:00:00	5513	13	57	35	42	11	2	0
2018-01-31 00:00:00	7069	15	74	29	48	34	0	0
2018-02-28 00:00:00	6555	6	57	6	30	73	1	0
2018-03-31 00:00:00	7003	23	133	9	17	26	0	0
2018-04-30 00:00:00	6798	14	99	8	5	15	0	0
2018-05-31 00:00:00	6749	24	54	6	16	24	0	0
2018-06-30 00:00:00	6099	3	43	0	4	18	0	0
2018-07-31 00:00:00	6159	13	60	1	18	41	0	0
2018-08-31 00:00:00	6351	23	47	0	7	84	0	0
2018-09-30 00:00:00	0	0	1	0	0	15	0	0
2018-10-31 00:00:00	0	0	0	0	0	4	0	0

Análisis de los tipos de pago

El método de pago mas popular entre los brasileños es la tarjeta de crédito con un **73.92 %** de uso, seguido por boleto que es un medio de pago en efectivo a través de un ticket válido en Brasil y que se usa para realizar compras online, en Olist este ticket representó el **19.04%** de los pagos, el tercer método es el voucher bancario con un **5.55%** y finalmente el pago con tarjeta de debito que es de tan solo **1.42%.**



A pesar de que la gran mayoría de los clientes usan tarjetas de crédito solo el **50.55%** difieren sus pagos en cuotas, esto lo que puede significar es que uno de los motivos de uso de este medio de pago es la protección contra los fraudes y copia de datos.

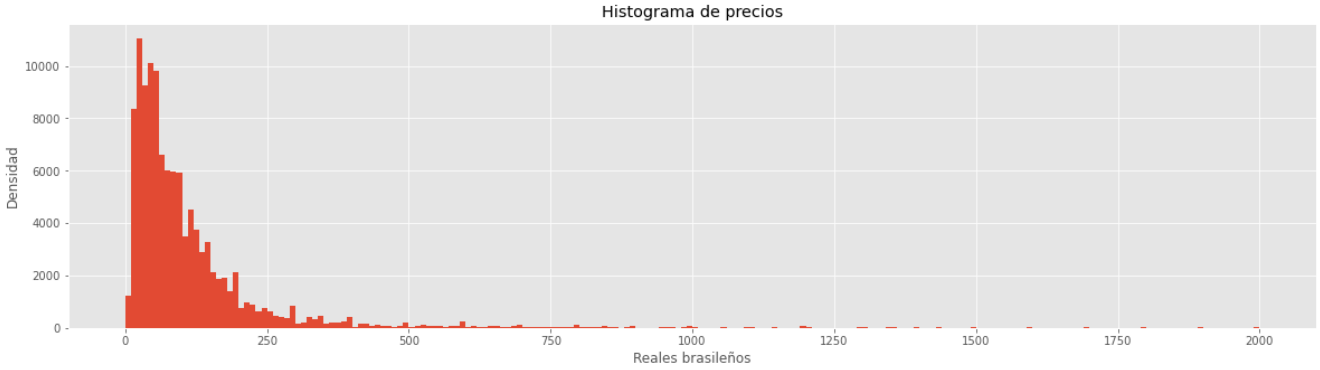


Análisis de categorías

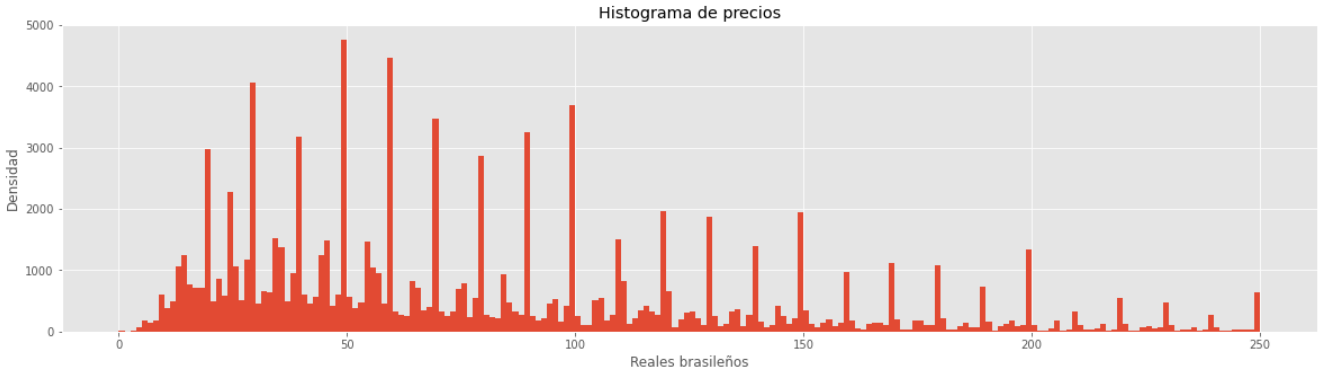
Las 7 categorías marcadas en verde son las que representan el mayor potencial de ingresos, superan los 6000 productos vendidos, así como una conversión de 1 millón o más Reales

Análisis de precios

El siguiente histograma muestra que los productos comercializados en el portal web son en su mayor proporción de un costo bajo, entre 0 y 250 reales.



Probablemente las barras más pronunciadas pertenezcan a productos de alta demanda.



Limpieza de datos

La base de datos original se compone de 9 archivos CSV cada uno de los cuales contiene diferente información. Para facilitar la manipulación de los datos se hizo un merge entre todos estos archivos a través de sus Id, posteriormente se eliminaron las columnas innecesarias

Se eliminan todos los ID ya que solo son útiles para el merge. El único ID que se mantiene es el del cliente único ya que será usado en una sección posterior

- order_id
- customer_id
- review_id
- product_id
- seller_id
- order_item_id

Se eliminan datos de la calificación que no se van a usar

- review_comment_title
- review_comment_message
- review_answer_timestamp

Se eliminan las ciudades ya que posteriormente solo es requerida el estado al que pertenece dicha ciudad

- customer_city
- seller_city

Se considera que la secuencia de pago no es un dato relevante para la calificación

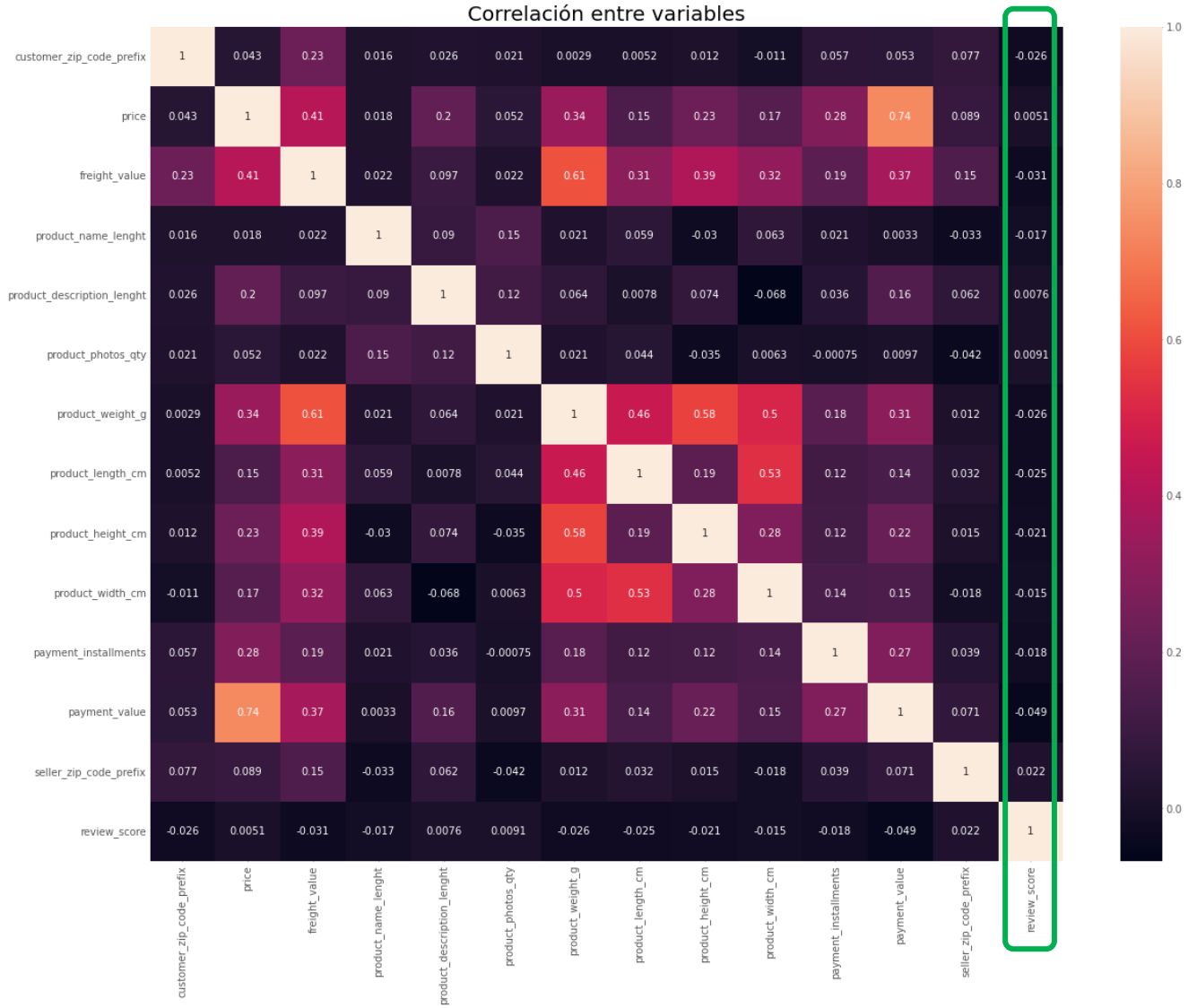
- payment_sequentia

Conforme se avanza en la ingeniería de datos se eliminarán más variables.

Matriz de correlación

No se muestran correlaciones relevantes para la variable target, el dataset original muestra algunas correlaciones obvias como por ejemplo:

1. El precio del producto y el pago realizado por el cliente
2. El precio de envío y las dimensiones del producto enviado
3. La descripción del producto y cantidad de fotos puestas por el vendedor


Correlación entre variables
customer_zip_code_prefix 1 0.043 0.23 0.016 0.026 0.021 0.0029 0.0052 0.012 -0.011 0.057 0.053 0.077 -0.026
price 0.043 1 0.41 0.018 0.2 0.052 0.34 0.15 0.23 0.17 0.28 0.74 0.089 0.0051
freight_value 0.23 0.41 1 0.022 0.097 0.022 0.61 0.31 0.39 0.32 0.19 0.37 0.15 -0.031
product_name_lenght 0.016 0.018 0.022 1 0.09 0.15 0.021 0.059 -0.03 0.063 0.021 0.0033 -0.033 -0.017
product_description_lenght 0.026 0.2 0.097 0.09 1 0.12 0.064 0.0078 0.074 -0.068 0.036 0.16 0.062 0.0076
product_photos_qty 0.021 0.052 0.022 0.15 0.12 1 0.021 0.044 -0.035 0.0063 -0.00075 0.0097 -0.042 0.0091
product_weight_g 0.0029 0.34 0.61 0.021 0.064 0.021 1 0.46 0.58 0.5 0.18 0.31 0.012 -0.026
product_length_cm 0.0052 0.15 0.31 0.059 0.0078 0.044 0.46 1 0.19 0.53 0.12 0.14 0.032 -0.025
product_height_cm 0.012 0.23 0.39 -0.03 0.074 -0.035 0.58 0.19 1 0.28 0.12 0.22 0.015 -0.021
product_width_cm -0.011 0.17 0.32 0.063 -0.068 0.0063 0.5 0.53 0.28 1 0.14 0.15 -0.018 -0.015
payment_installments 0.057 0.28 0.19 0.021 0.036 -0.00075 0.18 0.12 0.12 0.14 1 0.27 0.039 -0.018
payment_value 0.053 0.74 0.37 0.0033 0.16 0.0097 0.31 0.14 0.22 0.15 0.27 1 0.071 -0.049
seller_zip_code_prefix 0.077 0.089 0.15 -0.033 0.062 -0.042 0.012 0.032 0.015 -0.018 0.039 0.071 1 0.022
review_score -0.026 0.0051 -0.031 -0.017 0.0076 0.0091 -0.026 -0.025 -0.021 -0.015 -0.018 -0.049 0.022 1
customer_zip_code_prefix
price
freight_value
product_name_lenght
product_description_lenght
product_photos_qty
product_weight_g
product_length_cm
product_height_cm
product_width_cm
payment_installments
payment_value
seller_zip_code_prefix
review_score
1.0
0.8
0.6
0.4
0.2
0.0

Ingeniería de datos

En el análisis anterior se puede observar que no existen variables influyentes para nuestro target por lo que se tomó la decisión de crear nuevas variables que permitieran tener mayor correlación.

Conversión de fechas a días

Se crean 7 variables relacionadas restando diferentes fechas disponibles en el dataset original y traduciendo ese valor un valor numérico.

1. Días entre la entrega y la fecha de compra
2. Días entre la promesa de entrega de Olist y la entrega real
3. Días entre la entrega al cliente y la calificación en la página
4. Días entre la entrega al cliente y el tiempo límite para la entrega del carrier
5. Días entre la entrega al carrier por parte del vendedor y la fecha límite de despacho
6. Días entre la compra del cliente y la aprobación del pedido

Conversión de estados en regiones

Para reducir los 27 estados de Brasil a solo 5 regiones se usó la siguiente tabla:

Region_Nro	Region	Estados	Capitales	REF
1	Norte	Acre	Rio Branco	AC
1	Norte	Amapá	Macapá	AP
1	Norte	Amazonas	Manaus	AM
1	Norte	Pará	Belém	PA
1	Norte	Rondônia	Porto Velho	RO
1	Norte	Roraima	Boa Vista	RR
1	Norte	Tocantins	Palmas	TO
2	Nordeste	Alagoas	Maceió	AL
2	Nordeste	Bahia	Salvador	BA
2	Nordeste	Ceará	Fortaleza	CE
2	Nordeste	Maranhão	São Luís	MA
2	Nordeste	Paraíba	João Pessoa	PB
2	Nordeste	Pernambuco	Recife	PE
2	Nordeste	Piauí	Teresina	PI
2	Nordeste	Rio Grande do Norte	Natal	RN
2	Nordeste	Sergipe	Aracaju	SE
3	Centro -oeste	Goiás	Goiânia	GO
3	Centro -oeste	Mato Grosso	Cuiabá	MT
3	Centro -oeste	Mato Grosso do Sul	Campo Grande	MS
3	Centro -oeste	Distrito Federal	Brasília	DF
4	Sudeste	Espírito Santo	Vitória	ES
4	Sudeste	Minas Gerais	Belo Horizonte	MG

	4	Sudeste	Rio de Janeiro	Rio de Janeiro	RJ
	4	Sudeste	São Paulo	São Paulo	SP
	5	Sur	Paraná	Curitiba	PR
	5	Sur	Rio Grande do Sul	Porto Alegre	RS
	5	Sur	Santa Catarina	Florianópolis	SC

Conversión de categorías

Como se vio anteriormente en el análisis EDA se tienen 73 categorías de productos, solo 7 superaron el millón de reales en ventas y un volumen de 6000 productos, a estas categorías se les clasificó como Top asignándoles el número 1

1. bed_bath_table
2. health_beauty
3. computers_accessories
4. furniture_decor
5. watches_gifts
6. sports_leisure
7. housewares

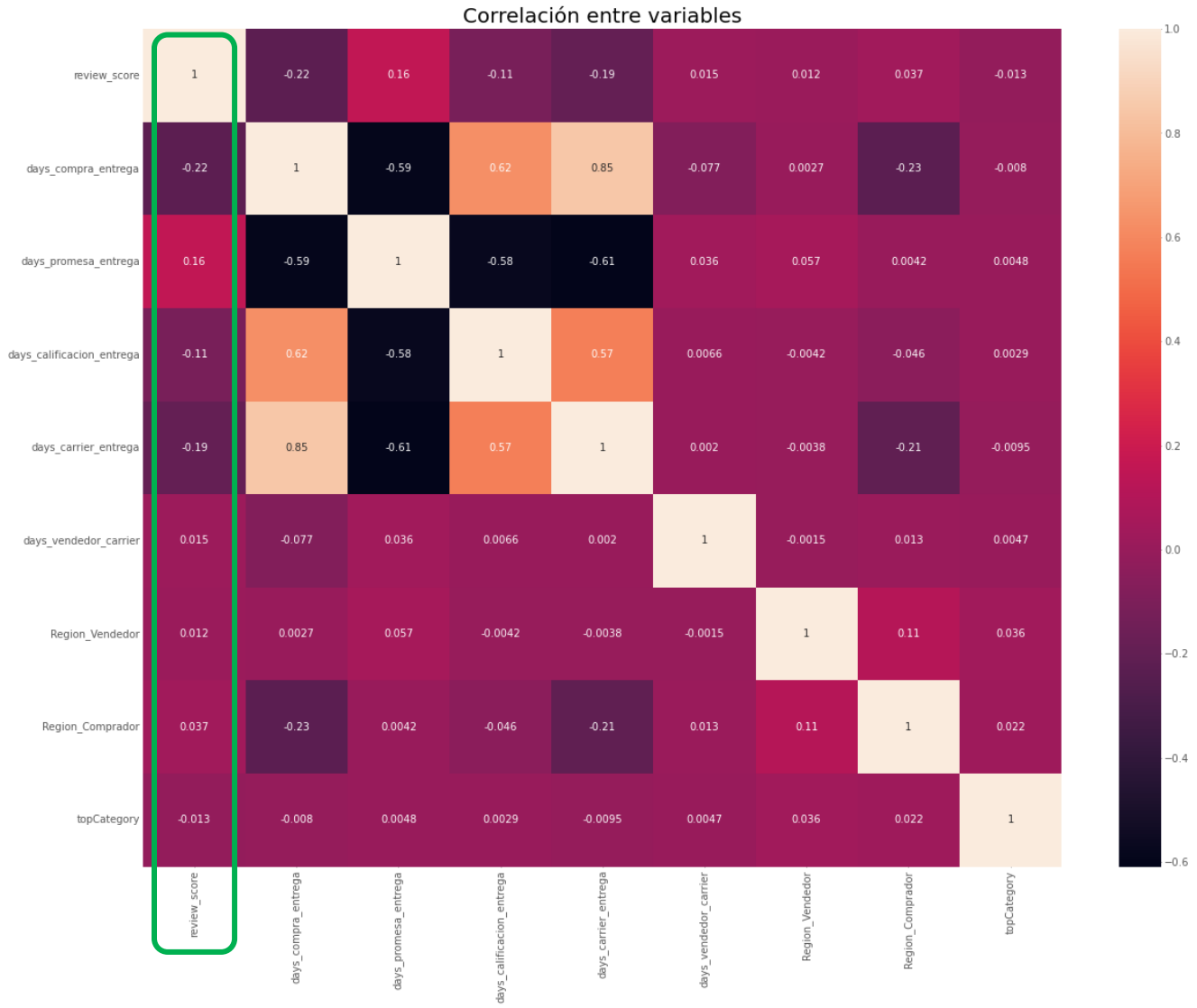
Al resto de categorías se les asigno el número 0, y de esta forma creamos una nueva variable para reducir la dimensionalidad.

Conversión de precios a rangos de precios

El precio es una variable continua que no podemos usar en nuestro análisis, sin embargo, para darle un uso se agruparon los precios de los productos con base en el análisis EDA en los siguientes rangos:

- Menor que 50
- Menor que 100
- Menor que 150
- Menor que 200
- Mayor igual que 200

Nueva matriz de correlación



Como se puede observar la nueva matriz de correlación tiene algunas características con mayor peso, esto ayudará a que el modelo tenga mayores opciones de predecir correctamente.

K-Means

Para enriquecer el dataset con una variable adicional se usó el algoritmo de K-Means en conjunto con la teoría de marketing RFM, esto con el objetivo de clasificar todos los clientes del dataset en clusters y así tener una característica más para las pruebas en algoritmos de clasificación

Perfiles de clientes (RFM)

RFM Es una técnica para segmentar grupos de clientes, diferenciando clientes con mayor valor para la compañía y permitiendo tomar decisiones de negocio como la inversión para fidelizar, ofrecer beneficios, incrementar ventas, diferenciar necesidades, entre otros

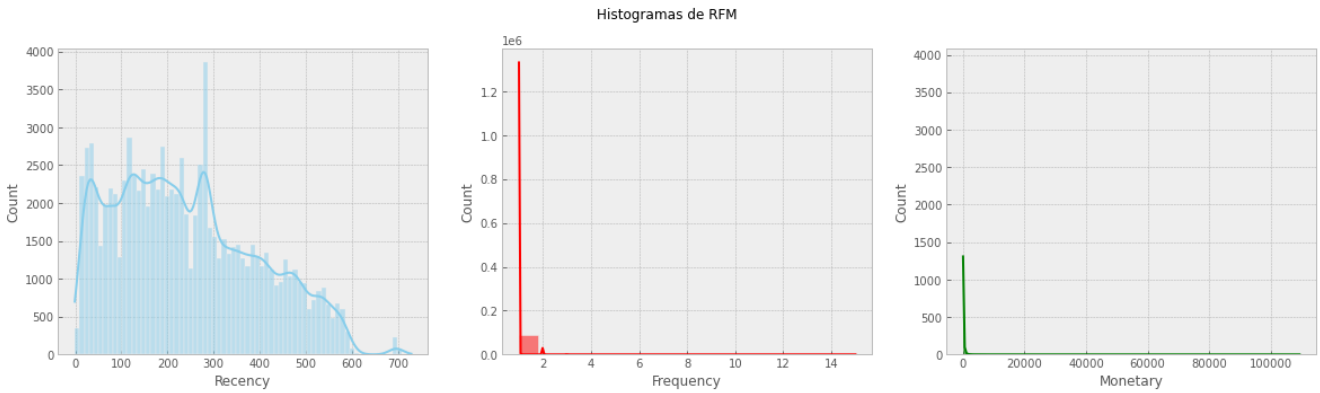
Para aplicar la técnica RFM hay que tener en cuenta las tres principales variables

- *Recencia*: el cliente que hace una compra en un sitio web está más predispuesto que otro a repetir. Por lo que la primera pregunta debe ser: ¿Cuándo fue la última vez que el cliente hizo una compra?
- *Frecuencia*: define el número de interacciones del cliente con la marca en un espacio determinado de tiempo.
- *Valor monetario*: refleja la cantidad que se ha gastado el cliente en las compras hechas en ese espacio de tiempo.

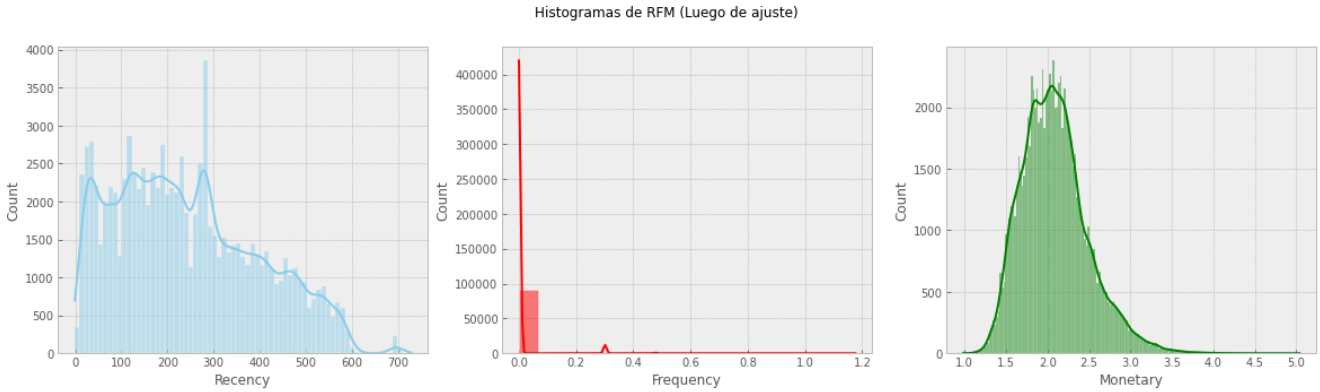
El análisis RFM aplicado al marketing, incrementa el grado de engagement y permite a la marca dirigir promociones específicas a cada grupo de clientes, mejorando las tasas de conversión y logrando un mejor resultado en las campañas.

Ajuste de variables RFM

La siguiente figura muestra los histogramas de las variables calculadas, se puede observar que existe una alta dispersión en la frecuencia de compra y en el valor monetario que los clientes ofrecen a la empresa.



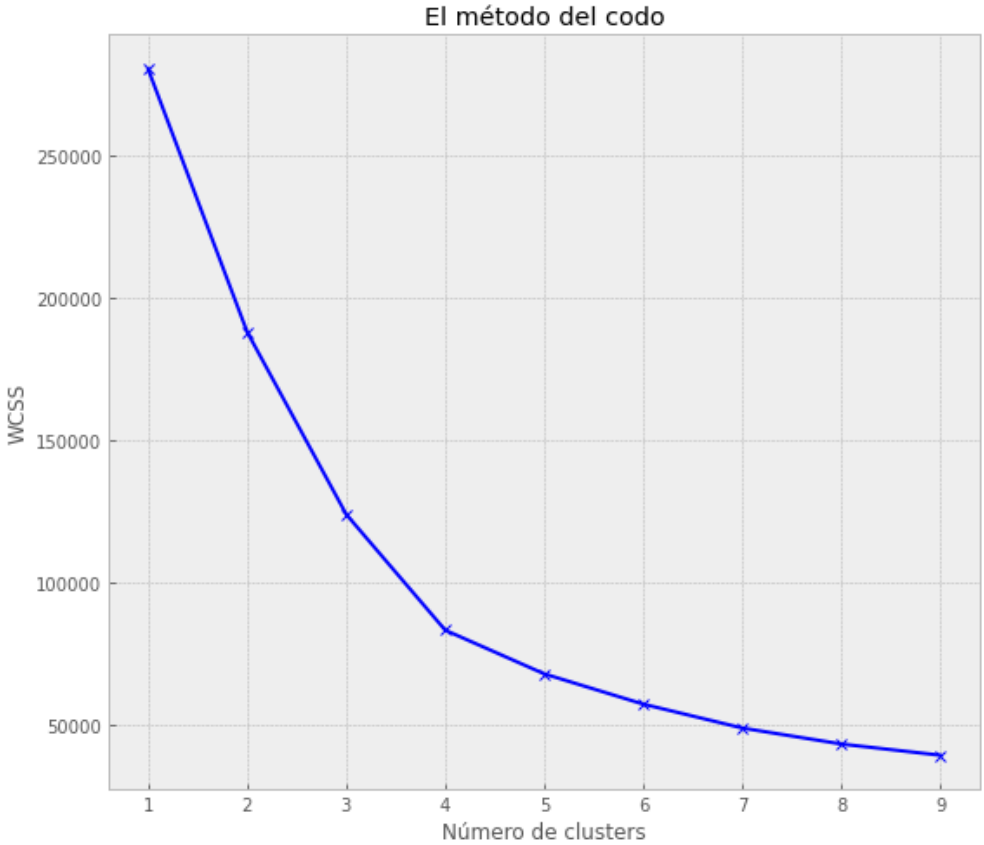
Se sabe que K-Means es altamente sensible a outliers por lo que para disminuir los efectos de este sesgo se aplicó una escalación logarítmica a estas variables



Se observa una mejora en la dispersión principalmente para el valor monetario, en el caso de la frecuencia de compra se sabe desde el análisis inicial EDA que la mayoría de los clientes no repiten compras por lo que esta variable pierde influencia en la segmentación.

Selección de K

Haciendo uso del método del codo se obtiene que el valor de K puede ser de 4 o 5 clusters. Debido a que el valor de las distancias es menor en 5 se tomó la decisión de asignar este valor a K.

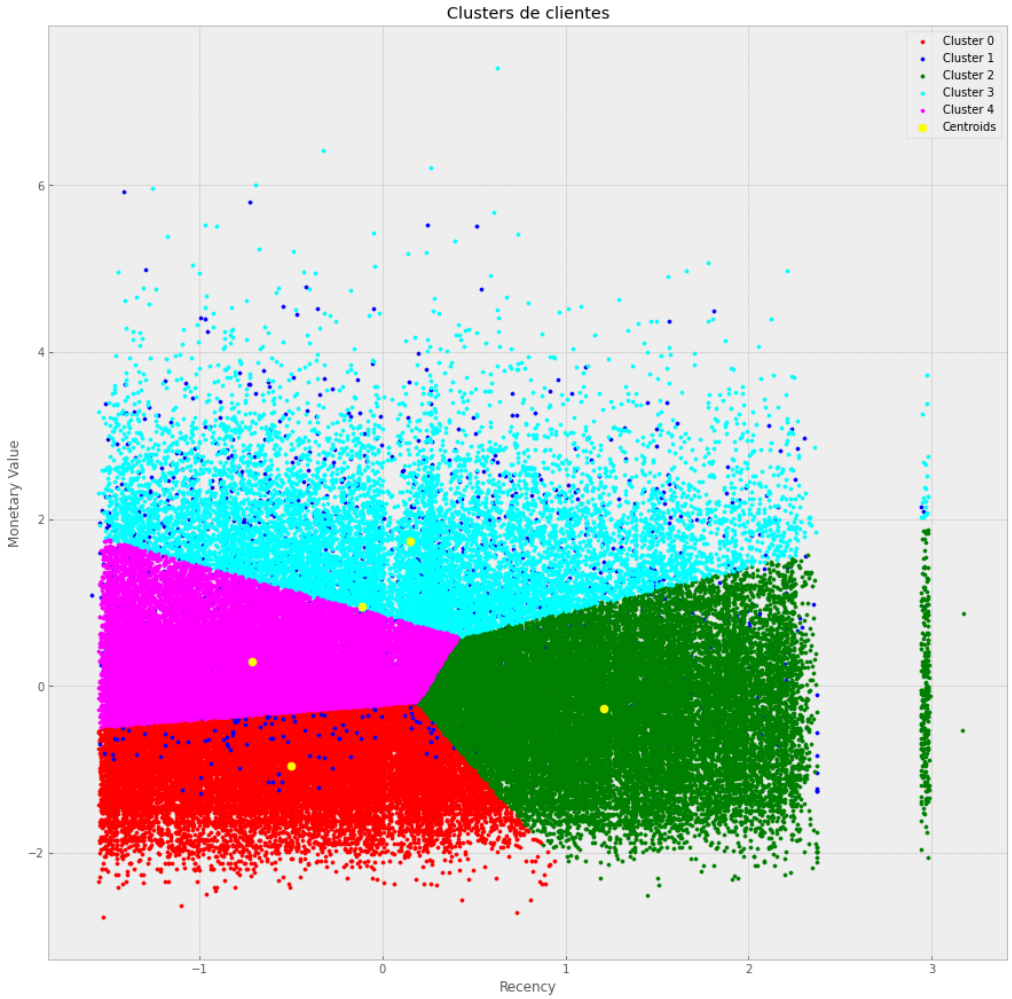


Resultados de K-Means

Para visualizar el efecto de las 3 variables de RFM se realizó la siguiente tabla donde se puede observar el promedio de cada valor en cada cluster resultante de K-Means.

	Recency	Frequency	Monetary	
	mean	mean	mean	count
Cluster				
0	166.0	1.0	55.0	25393
1	225.0	2.0	488.0	2807
2	427.0	1.0	114.0	25626
3	266.0	1.0	814.0	11496
4	133.0	1.0	174.0	28074

Se realiza una visualización 2D de los cluster usando las variables “Recency” y “Monetary” ya que la frecuencia como ya se comento es la variable menos influyente.



Interpretación de resultados

Se tienen 5 grupo de clientes:

- Cluster 0: Clientes que compraron recientemente y gastaron poco
- Cluster 1: Clientes más leales, con más compras de valor moderado
- Cluster 2: Clientes que compraron alguna vez pero que no repitieron la compra
- Cluster 3: Clientes valiosos que realizan compras de mayor valor
- Cluster 4: Clientes potenciales, que compraron recientemente con un valor más alto en sus compras

En una campaña de marketing esta segmentación puede ayudar a definir estrategias de campañas o a tomar decisiones dirigidas a cada grupo de clientes, en nuestro estudio este resultado será una variable adicional.

Algoritmos de clasificación

Todos los datos trabajados hasta el momento tienen como objetivo mejorar las posibilidades de clasificación de los algoritmos

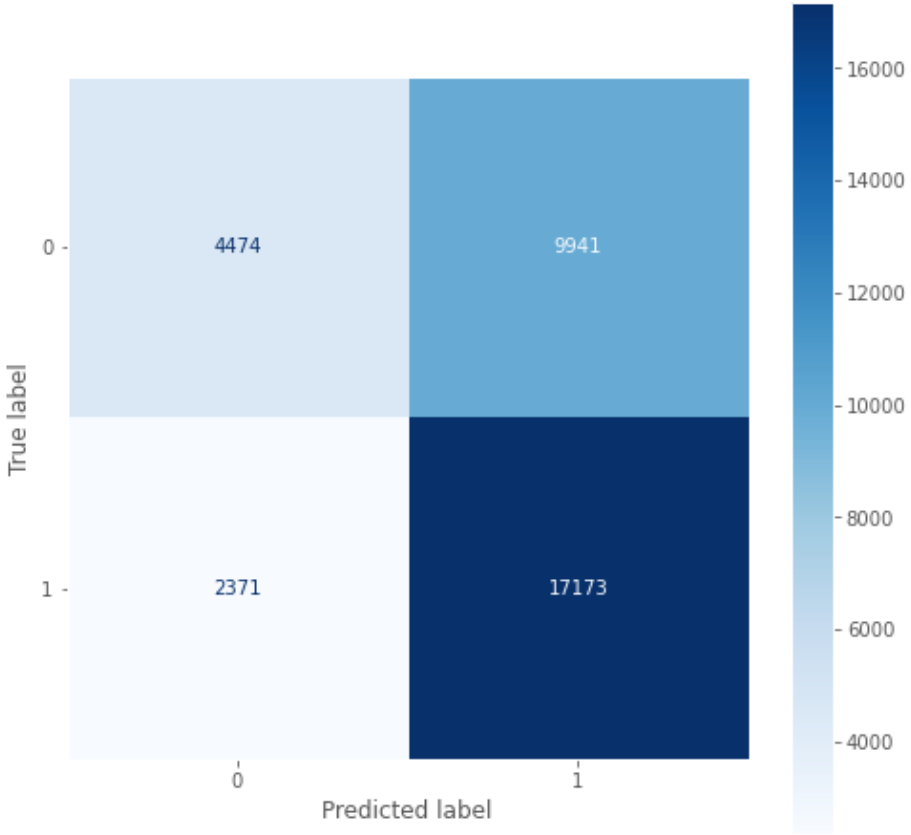
Recordando que:

1 => El cliente dio una calificación perfecta, cliente satisfecho
0 => El cliente no dio una calificación perfecta, cliente insatisfecho

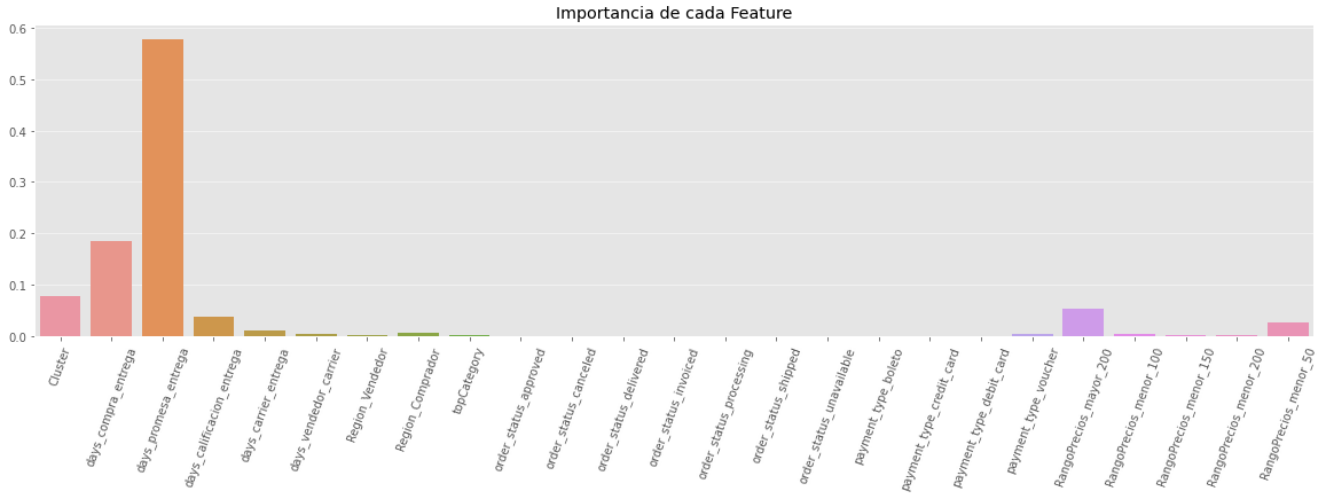
El error más costoso en este problema es el **falso positivo** ya que afirmar que un cliente está satisfecho cuando en realidad no lo está afecta más a un E-Commerce, por lo tanto, nuestro criterio de selección tendrá mas peso en la precisión.

Decision Tree

La matriz de confusión muestra que para este modelo se tiene un alto número de verdaderos positivos este valor puede estar un poco sesgado ya que el dataset no está perfectamente balanceado. Nuestra variable de interés el falso positivo supera por mucho su contraparte el falso negativo

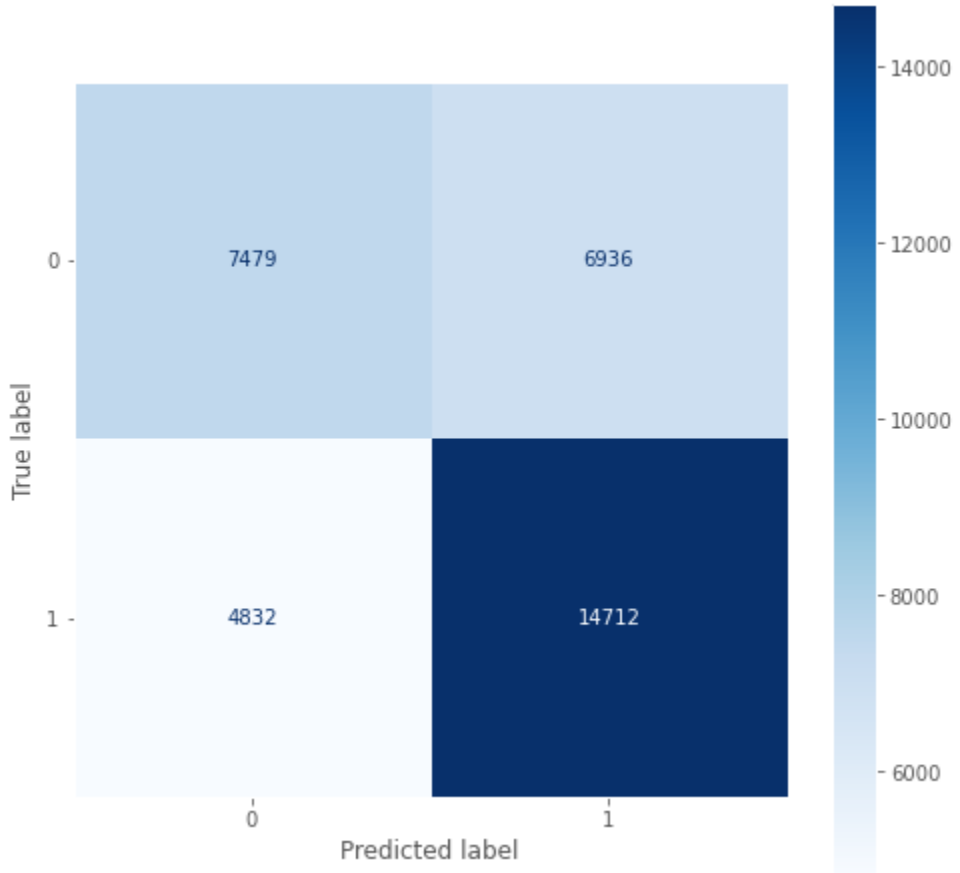


En este modelo la variable más influyente es la diferencia en días entre la promesa de entrega y la fecha real de entrega, esto tienen sentido ya que los clientes esperan su pedido a tiempo

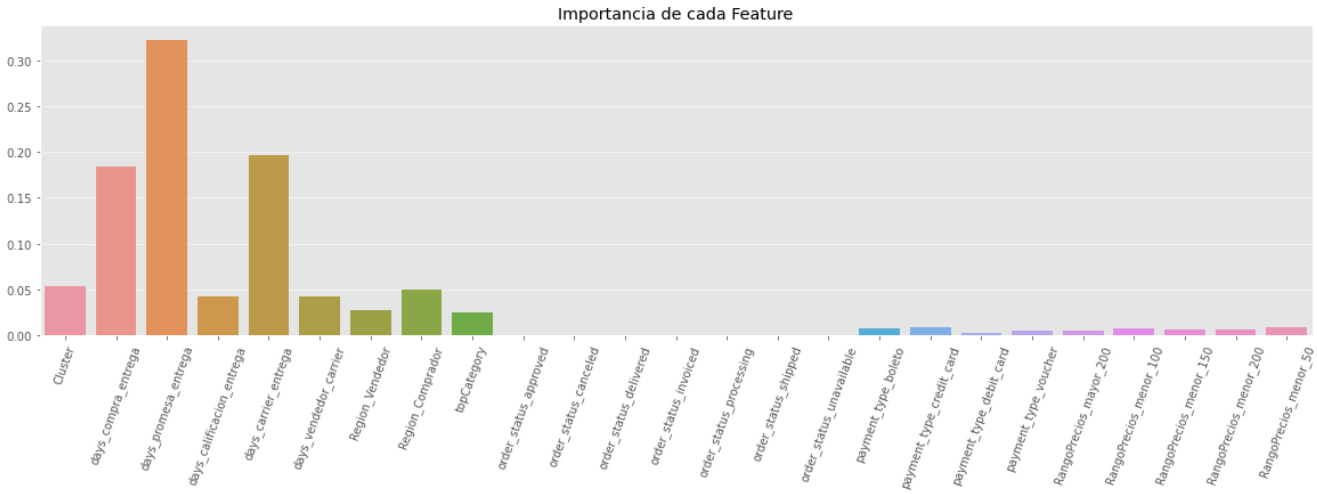


Random Forest

Para este modelo se tiene un mejor resultado ya que se disminuye el falso positivo

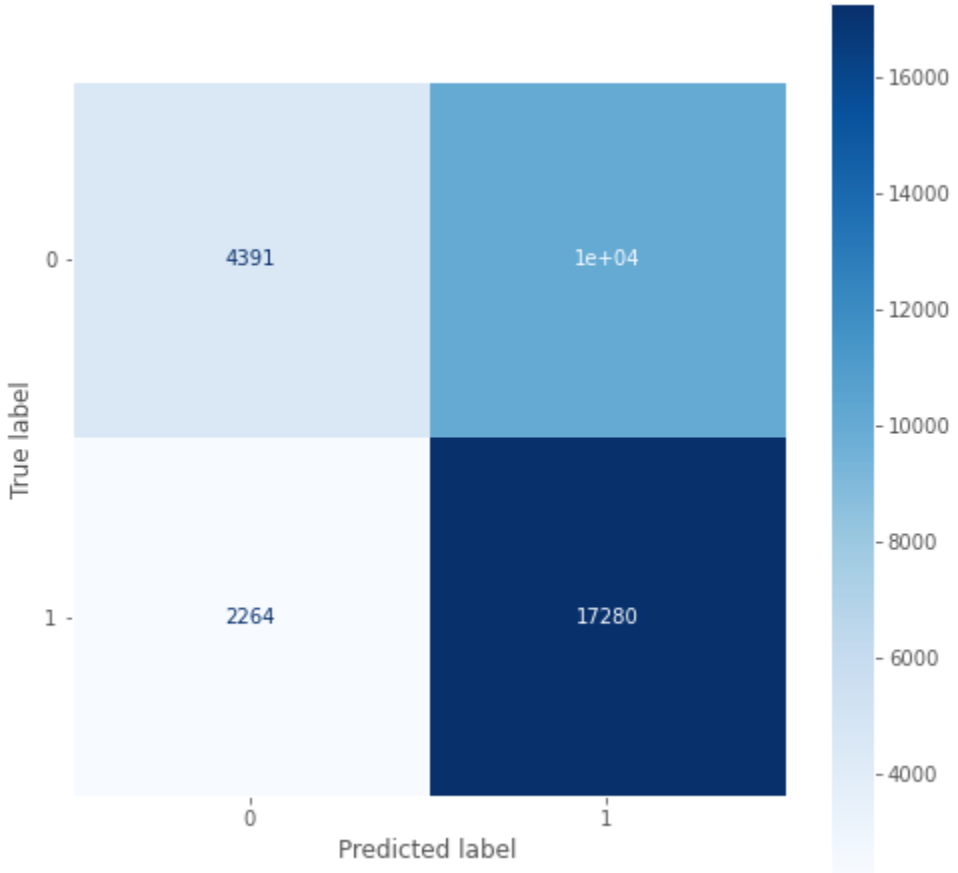


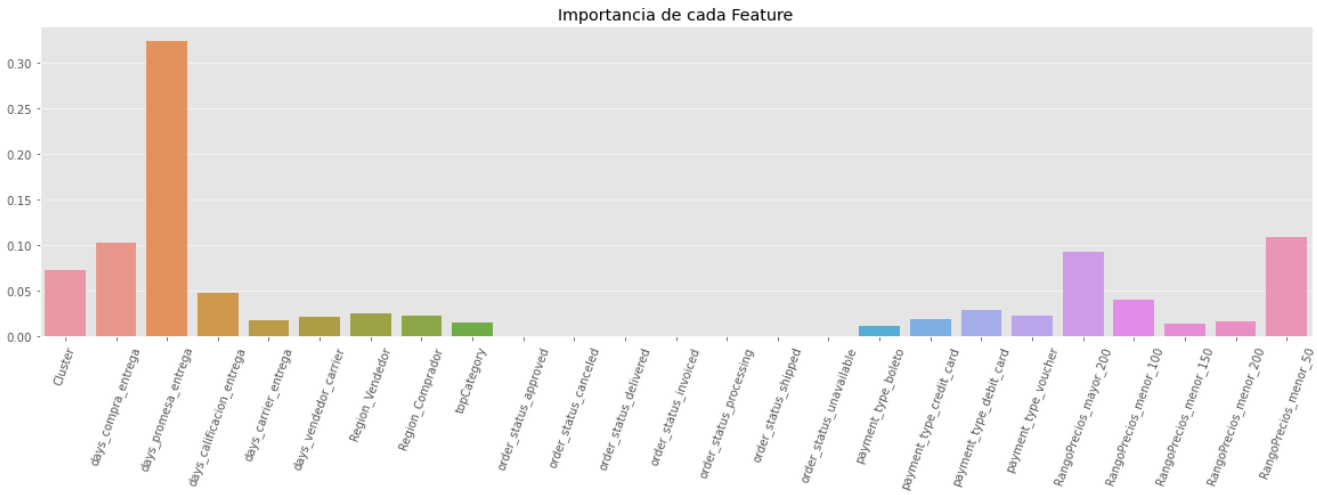
La variable más influyente continúa siendo los días entre la entrega y la promesa de entrega



XGBoost

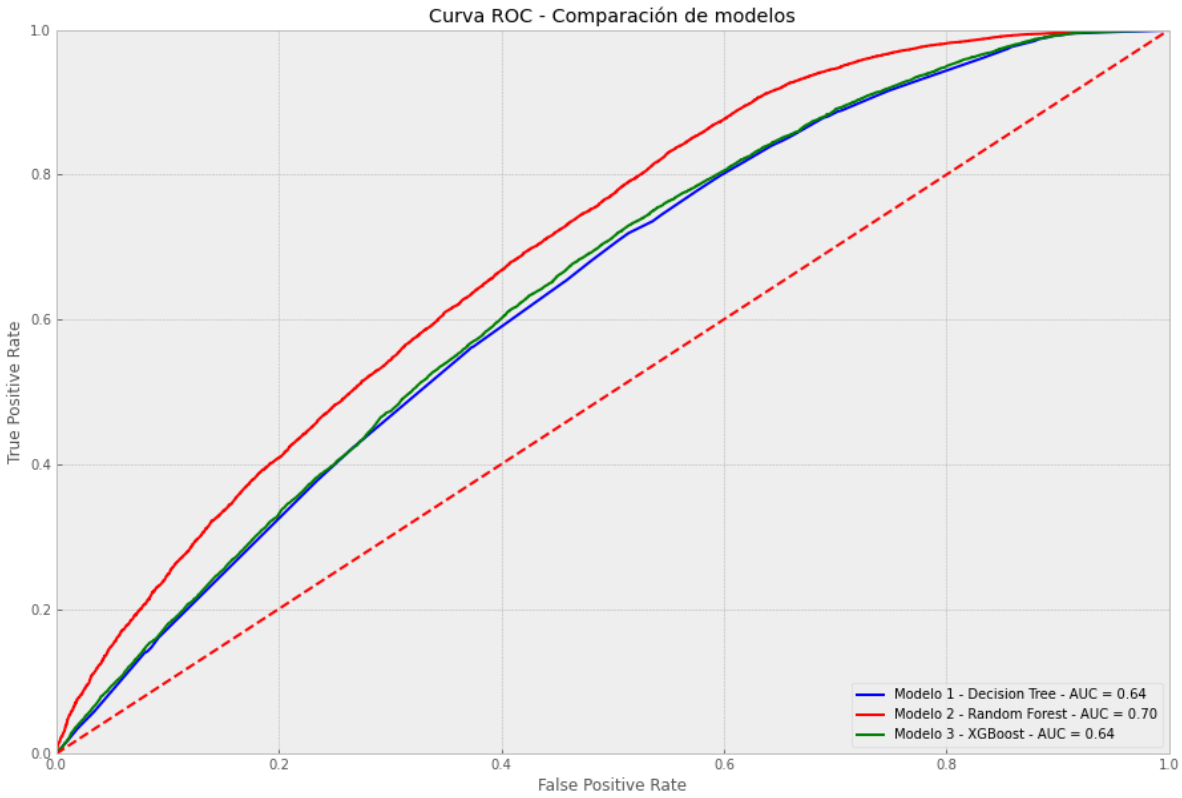
Este modelo genero el peor resultado ya que el falso positivo incrementó, el random forest generó más aciertos y con menos errores





Comparación de modelos

La curva ROC nos muestra que el modelo con mayor área bajo la curva es el Random Forest



En la siguiente tabla se hace una comparación de las métricas de evaluación de cada algoritmo, el mejor resultado para nuestro objetivo es la precisión del **67,9%** la cual fe conseguida por el random forest

	DecisionTree	RandomForest	XGBoost
accuracy_score	0.637445	**0.653464**	0.638152
precision	0.633363	**0.679601**	0.632874
recall	**0.878684**	0.752763	**0.884159**
f1	**0.736122**	0.714313	**0.737705**

Conclusiones

- Se selecciona el Random Forest por tener mejor precisión y mayor área bajo la curva.
- El modelo seleccionado dio un 67% de precisión
- El modelo no estaba completamente balanceado por lo que el accuracy de 65% puede ser una métrica engañosa.
- La variable mas influyente siempre fueron los días de retraso en la entrega de los paquetes por lo que al negocio le interesaría mucho mejorar la logística
- El algoritmo de K-Means generó 5 agrupaciones de clientes esta variable se mantuvo en todos los casos en el top 5 de variables influyentes para los algoritmos de clasificación
- El resultado del algoritmo de K-Means le puede dar al negocio información para generar estrategias adecuadas a cada clúster, por ejemplo:
    - Brindar beneficios a los clientes de los clústeres que representan más ingresos
    - Incentivos para los fidelizar clientes nuevos
    - Premiar la frecuencia de compra de los clientes que tengan más de una compra

Futuras líneas de investigación

- Hacer un estudio más profundo de las variables que afecten los retrasos en la entrega
- Investigar que productos tienen más problemas en la entrega, así como las características de estos.
- Estudiar la relación entre las entregas y los vendedores.
- Probar más algoritmos como SVM y KNN.
- Hacer comparaciones entre K-Means y HDBSCAN
- Profundizar en el estudio de los hiperparametros de cada modelo e incrementar las variaciones en cada uno

Referencias

https://olist.com/es-mx/
https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce
https://en.wikipedia.org/wiki/RFM_(market_research)
https://www.unir.net/marketing-comunicacion/revista/analisis-rfm/
https://www.rapyd.net/blog/what-is-boleto-everything-you-need-to-know/