IT ACADEMY

Ciencia de Datos - Proyecto final

Análisis de datos y Machine Learning aplicados a la gestión del e-commerce

Marc Hernández de Marcos

31 de marzo de 2023

Abstract

Estamos viviendo una nueva era del comercio electrónico. La pandemia supuso un cambio en los patrones de consumo e impulsó al comercio electrónico a aquellos negocios que se vieron obligados a cerrar sus establecimientos.

Actualmente los negocios tradicionales han pasado a ser negocios onmicanales, incluyendo los canales digitales y redes sociales.

El análisis de datos y el machine learning son una herramienta imprescindible para conocer a los usuarios y ofrecerles una experiencia personalizada con el objetivo de optimizar recursos e inversiones y maximizar las ventas.

1. Introducción

El estudio de E-commerce 2022 en nuestro país, de la compañía IAB Spain, en su octava edición, revela que el 94 % de la población entre los 16 y los 70 años es internauta, y de estos, un 78 % son compradores online, una cifra de 24,7 millones.

El mismo estudio recoge el aumento de usuarios que compran en tiendas físicas y online respecto al año 2021, comportamiento que aumentará en el 2023. Baja el número de los compradores polarizados que solo compra online o solo compra en establecimientos.

Internet sigue siendo el canal de búsqueda de información principal para el 93 % de usuarios, antes de tomar la decisión de compra, y son un 80 % los que realizan la compra a través de este mismo canal.

Los futuros clientes pueden acceder con facilidad a la información y comprar desde cualquier lugar. La competencia es enorme y es evidente la importancia de las acciones de marketing a la hora de trabajar con e-commerce, con el objetivo de captar leads y convertirlos en clientes.

Se conoce como Customer Jorney la relación del cliente con la empresa desde la primera toma de contacto hasta la postventa. Este recorrido se diseña para atraer al cliente, educarle y comprometerle con la marca. Las empresas ofrecen valor compartiendo contenido relacionado con los intereses de su audiencia, ayudándoles a descubrir sus necesidades y ofreciéndoles soluciones que se transformarán en ventas.

El Coste de Adquisición de un Cliente (CAC) es la inversión realizada para incorporar a ese cliente. El cálculo es CAC se realiza en función del canal donde se captan los nuevos clientes, consiste en dividir el presupuesto de marketing por el número de clientes atraídos.

El Customer Jorney no acaba con la primera venta, la inversión se rentabiliza con la última e importante etapa: la fidelización.

El Valor de la Vida Útil de un Cliente (CLV) es el valor monetario del cliente durante el tiempo que está en la plataforma, antes de lo que se conoce como abandono o Churn Customer.

Las acciones postventa logran que cliente se comprometa con la marca. Un cliente comprometido es un promotor de la marca, participa más en redes sociales y es un cliente recurrente a lo largo del tiempo. Así aumenta la autoridad de la marca que, junto el boca a boca, hace que atraer a nuevos clientes tenga menos coste.

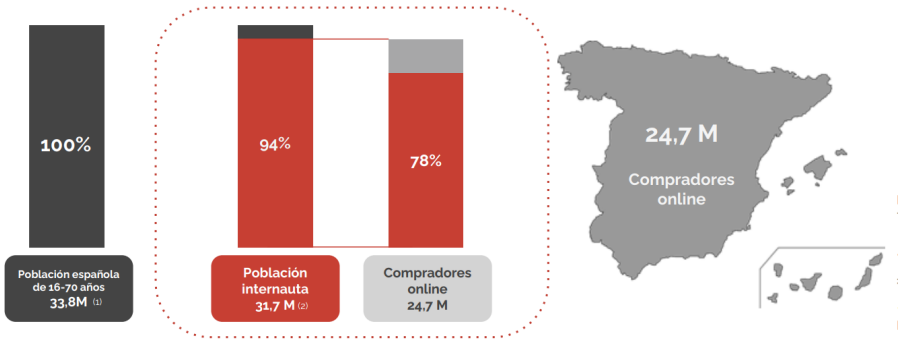


Figura 1: Clientes online año 2022.

2. Machine Learning orientado a la fidelización del cliente y la previsión de ventas

El valor del CAC no debe ser mayor del CLV para que las inversiones en marketing sean rentables. Aquí radica la importancia de conocer a la audiencia

de un e-commerce, permite tener una adecuada planificación de las campañas de marketing, saber a qué clientes dirigirlas para obtener el máximo beneficio y prever qué presupuesto se debe invertir.

Los objetivos de este proyecto son, a través del análisis de datos y del machine learning, segmentar la audiencia de un e-commerce, calcular la vida útil de un cliente y hacer una previsión de las ventas.

2.1. Segmentación de clientes.

Los clientes aportan diferentes valores al e-commerce, algunos crean más valor con compras recurrentes siendo fieles a la marca y otras son compradores puntuales.

Segmentamos a los clientes para hacer campañas de marketing orientadas específicamente a cada segmento, ya sea hacia los clientes de mayor valor con recompensas, por ejemplo, o a los de menos valor con promociones.

Existen muchas posibles segmentaciones, en este proyecto la segmentación se basaré en el análisis RFM, ."Actualidad"(días desde la última compra de un cliente hasta el día de la segmentación) o Recency, Frecuencia (periodos entre compras) o Frequency y Valor Monetario (gasto total del cliente en la plataforma) o Monetary. Los grupos generados serán los siguientes:

- Valor bajo : clientes poco frecuentes y que generan ingresos muy bajos.
- Valor medio : cliente frecuente con ingresos moderados.
- High Value : cliente muy activo que genera altos ingresos.

2.2. Cálculo del valor de la Vida útil de un cliente.

El cálculo del valor del cliente o Customer Lifetime Value (CLV) en este proyecto tiene un enfoque predictivo basado en un modelo probabilístico. Este modelo ajusta una distribución de probabilidad a los datos y con esa información estima otros parámetros como el número de transacciones futuras o el valor monetario futuro. Para desarrollar este modelo trabajamos con los datos transaccionales históricos del negocio.

Los modelos utilizados son los siguientes:

- El modelo BG/NBD, Distribución Beta Geométrica/Binomial Negativa predice el número de compras futuras a través la distribución de los comportamientos de compra de cada cliente ajustados a los datos históricos. También predice la tasa de abandono.
- El submodelo Gamma-Gamma se combina con el anterior en el cálculo del CLV agregando el valor económico a través de la distribución de ingresos; predecirá la ganancia promedio para cada cliente.

Los modelos usados pertenecen al paquete de Python Lifetimes, que se orienta principalmente al cálculo del CLV y la tasa de abandono de cada cliente.

2.3. Previsión de ventas con series temporales.

Es desarrollo del apartado de previsión de ventas en este proyecto se ha basado en el paquete Prophet. Esta biblioteca es de código abierto y está diseñada para hacer pronósticos para conjuntos de datos de series temporales univariadas, esto es, se considera una única variable endógena.

Prophet implementa un modelo de pronóstico de series de tiempo aditivolas, donde las tendencias no lineales se ajustan a la estacionalidad anual, semanal y diaria, además de los efectos de las vacaciones. Está diseñado para ser fácil y completamente automático.

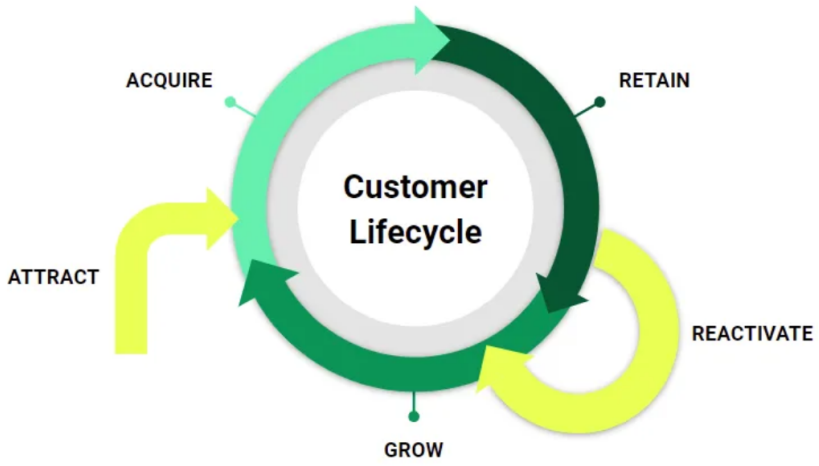


Figura 2: Ciclo vital de un cliente en ecommerce.

3. Conjunto de datos trabajados en el proyecto

El conjunto de datos ha sido facilitado por la empresa Olist, el mayor especialista de plataformas online o marketplaces de Brazil.

Un marketplace es una plataforma digital que agrupa diferentes tiendas online que venden productos o servicios de terceros. Es una web donde diferentes vendedores ofrecen sus productos de una o varias temáticas. La plataforma constituye el intermediario por medio del cual se realiza la transacción de compraventa entre un comprador y un vendedor digital.

Olist ofrece a pequeños negocios de todo Brasil vender en sus marketplaces y enviar también sus productos a través de su red de logística de una forma fácil y con un contrato sencillo. Después de recibir el producto se le envía al cliente una encuesta de satisfacción.

Los datos comerciales son reales y se han anonimizado, así como las referencias a las compañías y los socios.

Este conjunto de datos está disponible en el siguiente link:
https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce?
datasetId=55151sortBy=voteCountselect=olist$_o rder_i tems_d ataset.csv$

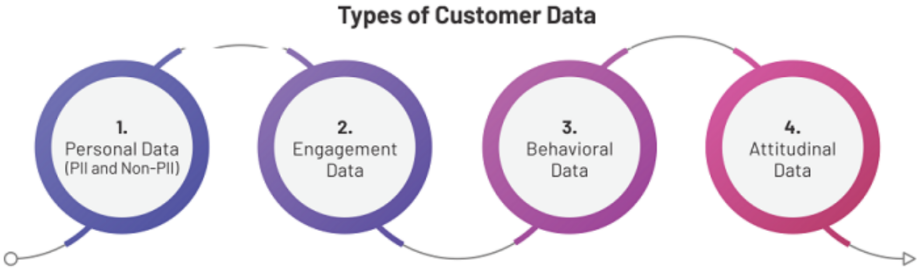


Figura 3: Tipos de datos del cliente online.

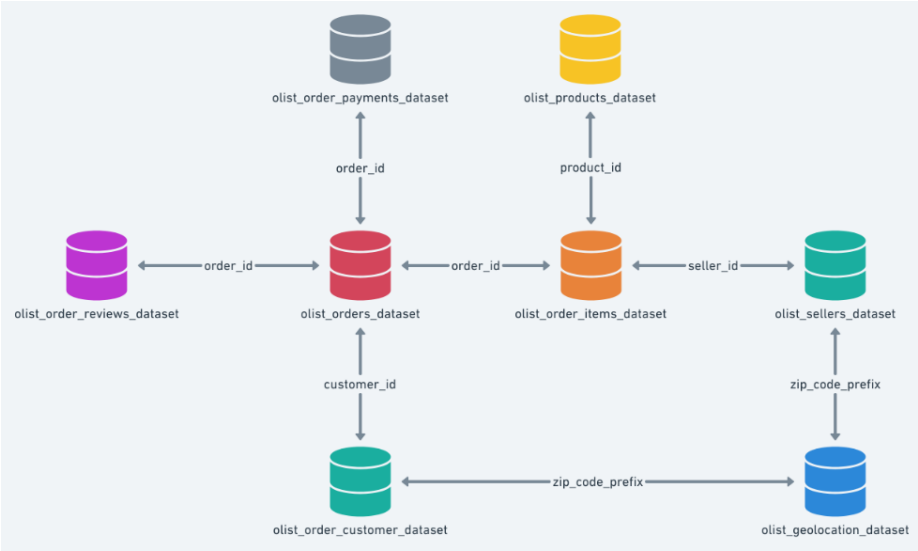
El dataset tiene 99441 pedidos realizados desde 2016 a 2018, con la posibilidad de haber varios artículos por pedido gestionados por distintos vendedores.

La información relativa a los pedidos como el estado, el precio, el pago o el envío, así como la información relacionada con los artículos vendidos, localización de los clientes y sus opiniones sobre el servicio, se organiza en diferentes archivos:

- olist$_{c}ustomers_{d}ataset.csv$
- olist$_{g}eolocation_{d}ataset.csv$
- olist$_{o}rder_{i}tems_{d}ataset.csv$
- olist$_{o}rder_{p}ayments_{d}ataset.csv$
- olist$_{o}rder_{r}eviews_{d}ataset.csv$
- olist$_{o}rders_{d}ataset.csv$
- olist$_{p}roducts_{d}ataset.csv$
- olist$_{s}ellers_{d}ataset.csv$
- product$_{c}ategory_{n}ame_{t}ranslation.csv$

El dataset ocupa 15.44 MB, tiene asignado el DOI 10.34740/kaggle/dsv/195341 y se distribuye bajo la licencia Creative Commons CC BY-NC-SA 4.0.

Nuestro conjunto de datos tiene múltiples variables. Hay variables que son identificadores únicos de clientes, pedidos, vendedores, artículos, opiniones, otras categóricas relacionadas con las categorías de los artículos o el tipo de pago, otras nominativas, con la localización de los clientes o el destino de los pedidos, otras numéricas con el precio o dimensiones de los artículos.



Estrutura de las relaciones entre los diferentes conjuntos de datos.

4. Metodología y objetivos

La metodología utilizada para el desarrollo de este proyecto ha sido la siguiente:

1. Importación de las librerías.
2. Carga de los conjuntos de datos y extracción de los datos necesarios a un dataframe.
3. Limpieza de los datos.
4. Análisis de los datos obtenidos. Explorar los datos para conocer el comportamiento de los clientes.
5. Visualización de los datos.
6. Segmentación de clientes. Los clientes quedaran segmentados por grupos con una cluterización basada el modelo K-means y la metodología RFM.
7. Cálculo del valor de la vida útil de un cliente CLV. Con los modelos BG/NBD y Gamma-Gamma combinados.
8. Aplicación de la librería Lifetimes. Cálculo de la vida útil de un cliente CLV. Modelos BG/NBD y Gamma-Gamma. Cálculo de la probabilidad de abandono de los clientes.
9. Aplicación de la librería Prophet. Previsión de ventas con series temporales univariantes.

5. Resultados

Los datos obtenidos son los siguientes:

1. La serie temporal del dataset empieza el 04/09/2016 y acaba el 03-09-2018. Vemos que en septiembre, noviembre y diciembre de 2016 no se aprecian ventas, tampoco en septiembre de 2018. Hay picos de venta marcados que corresponden a Pascua, 09-04, Día de la madre 2018, 13-05 y Black Friday 2017, 24-11.

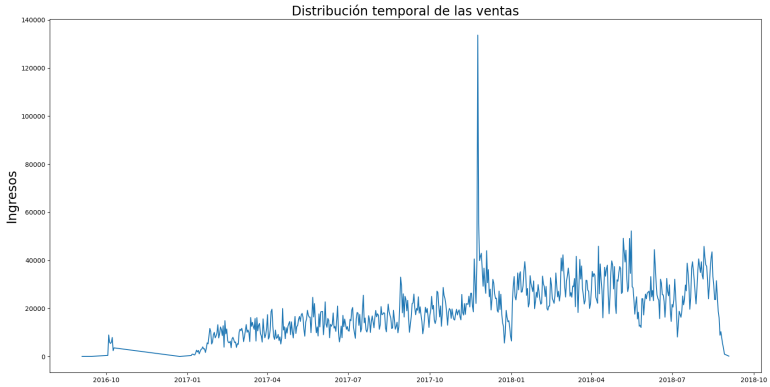


Figura 4: Distribución de las ventas

2. Análisis de los artículos y sus precios:

   - La media de artículos comprados es de 1.19 artículos, el máximo número en una compra es de 23 artículos. Vamos que más del 75 % de las compras son de un sólo artículo. - El precio medio de los artículos es de 120.65 reales brasileños. Sin embargo solo el 25 % de los artículos que está por encima de esta media.

3. Análisis de los pedidos:

   Los pedidos cancelados representan sólo un 0,5 % son 579 pedidos cancelados o erróneos. Hay 3216 pedidos de clientes recurrentes, lo que representa solo un 3,2 % del total.

4. Análisis de los clientes: El gasto medio por cliente es de 129 reales brasileños, pero sólo el 25 % de los clientes gastan más de 144 reales brasileños por compra. La compra mínima es de 0,85 reales brasileños y la máxima de 7388 reales brasileños. Un 25 % de los clientes, aproximadamente, llevan más de un año sin comprar, más del 25 % lleva tres meses. La media es de 243 días. Sólo un 3 % de los clientes es recurrente, y de ellos la mayoría hace 3 compras o menos. Los cinco clientes que más compran harían entre 0,14 y 0,65 ventas en un mes. La alta mayoría de los clientes compran sólo una vez.

5. La segmentación de clientes da como resultado que sólo 84 clientes tienen un valor alto, no llega a un 1 cliente por cada mil. Un 43 % tiene un valor

medio y el resto un valor bajo. Los clientes con el mejor valor se sitúan en la zona amarilla, donde el Recency es mayor, lo que significa que el cliente ha vuelto a comprar entre 600 y 700 días después de hacer su primera compra; y que la frecuencia de estas compras es superior a 10 compras.

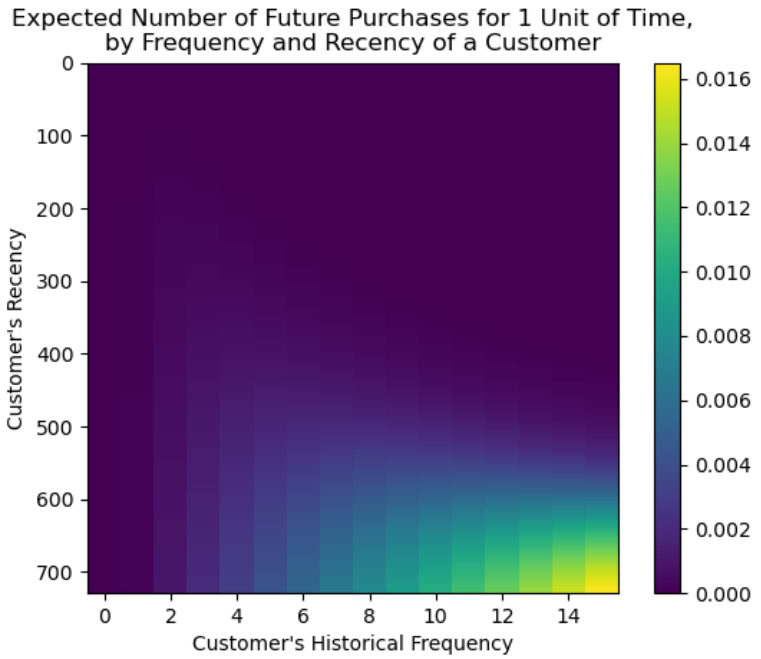
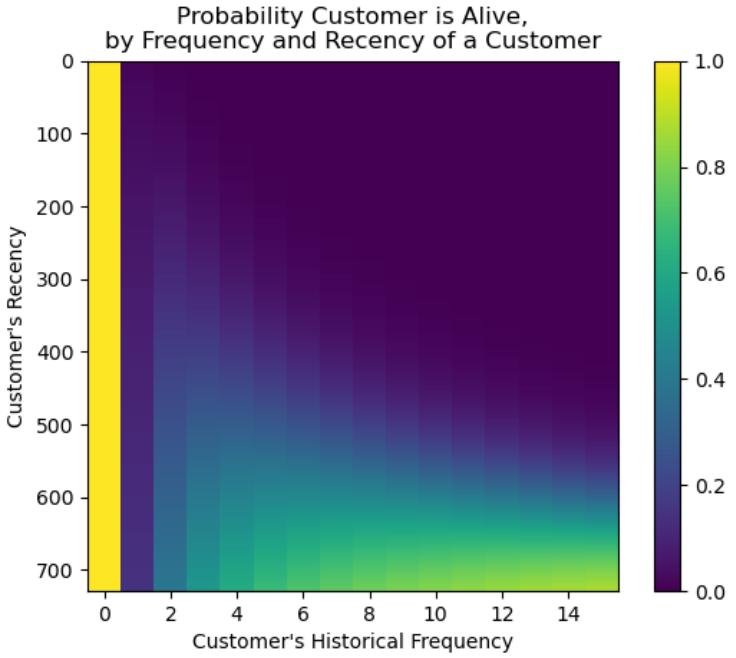


Figura 5: Matriz de predicción de ventas

Los clientes con más probabilidades de estar activos se sitúan en la zona verde-amarilla de la matriz, las compras son frecuentes y lleva tiempo en el marketplace. Los clientes que aún no han hecho una compra se les considera vivos.

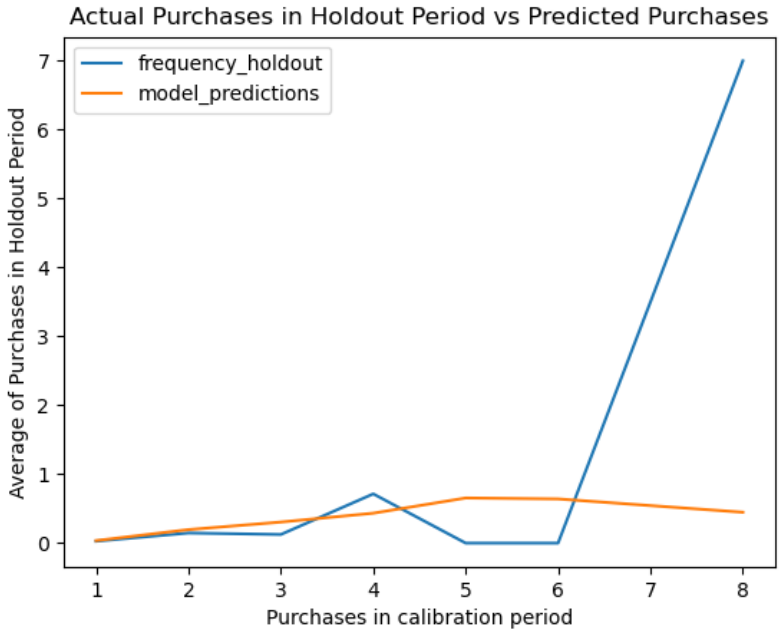


Matriz de probabilidad de retención de un cliente

En este dataset tanto el valor de los clientes como su retención son muy bajos.

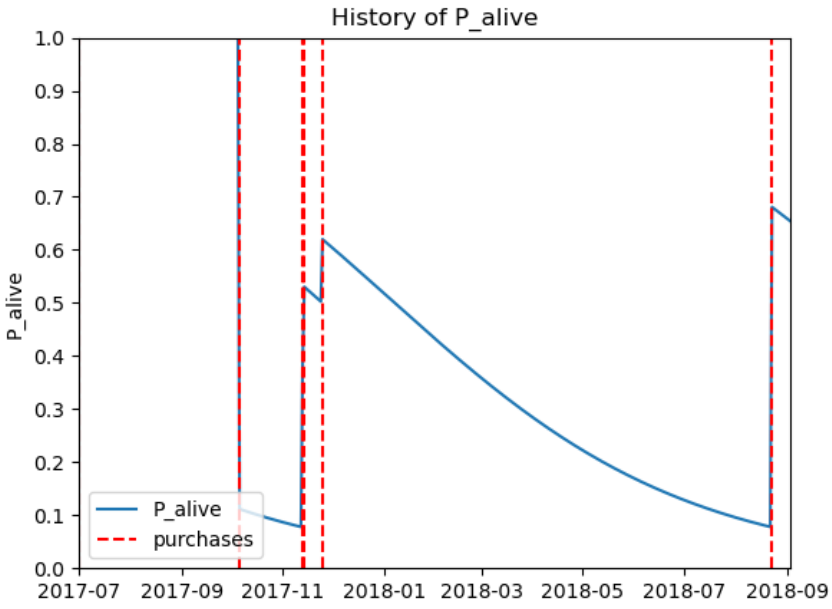
6. Cálculo del valor de la vida útil de un cliente: Modelo BetaGeoFitter, predice las compras y la tasa de abandono.

   La predicción del modelo es buena para una compra y para dos, empeora en la tercera compra y a partir de aquí va aumentando el error con cada compra añadida. Es un comportamiento esperable debido a la falta de muestras de clientes recurrentes.



Ventas reales vs predicción de ventas

7. La probabilidad de que un cliente se mantenga activo y fiel a la marca aumenta con cada compra y disminuye drásticamente en los periodos entre compra y compra.



Evolución de la probabilidad de retención de un cliente

Modelo Gamma-Gamma, predice la ganancia promedio de cada cliente.

La media de las ganancias por cliente es de 109,68 reales brasileños, y el modelo predice 110,07 reales brasileños.

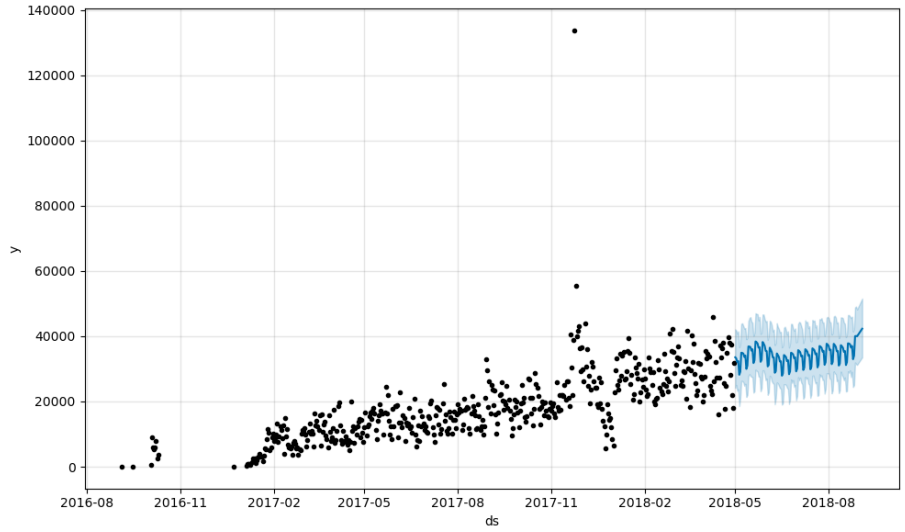
Ambos modelos se combinan para obtener el CLV en un periodo de 30 días.

El modelo prevee que un cliente generará en 30 días una media de 11,58 reales brasileños. El 50 % de los clientes generarán un valor de por debajo de los 6 reales brasileños, un 25 % por debajo de los 12,73 y del otro 25 % el valor máximo sería de 277,21 reales brasileños.

Predicción de tasa de abandono (Churn Prediction).

Esta métrica nos confirma que en este dataset no hay clientes recurrentes, la tasa de abandono es de un 94 % de media.

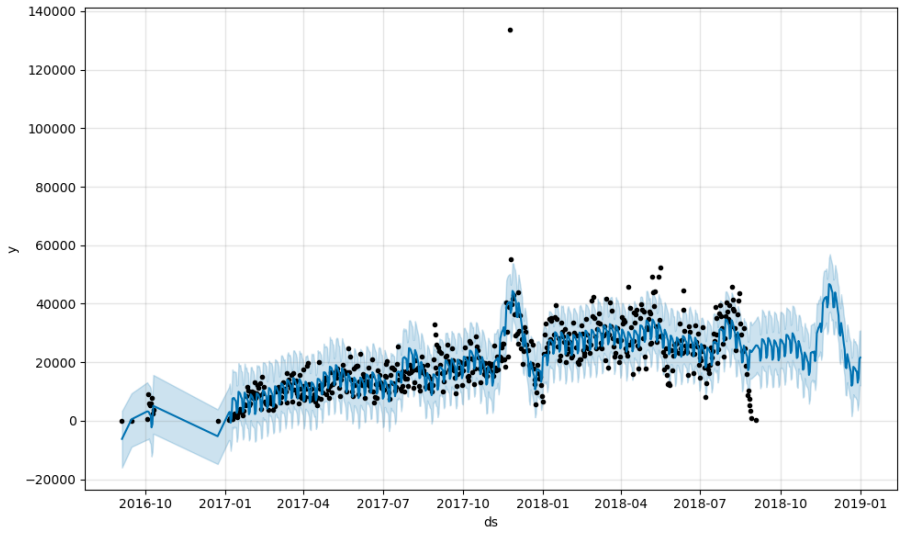
8. Predicción de ventas (Sales forecasting). Modelo Prophet. El MAE es alto, el modelo no ajusta bien por defecto. Deberemos incluir la tendencia y la estacionalidad.



Predicción de ventas dentro del dataset

Predicción del modelo.

El modelo ajusta mejor en la predicción de estos tres meses al tener las muestras del dataset entero en las que basarse.



Predicción de ventas durante tres meses

Descomposición de la predicción.

Al ser Prophet un modelo de agregación podemos descomponerlo en diferentes métrica, tendencia, distribución semanal, anual y diaria.

Descomposición de la serie temporal

La serie temporal tiene dos tendencias diferenciadas, una ascendente y otra plana.

Con la distribución anual se observan ver los eventos que más influyen en las ventas, como el Black Friday en noviembre.

Durante la semana se compra algo más el lunes, pero las compras se reparten entre lunes y viernes y caen los fines de semana.

Durante el día hay una clara tendencia a hacer las compras por la noche.

Con estos datos se ajustaár mejor el modelo en la siguiente previsión.

6. Conclusiones

El análisis de los datos y predicciones de los modelos usados han dado como resultado datos muy útiles para conocer a los clientes de este e-commerce.

Estos datos revelan comportamientos sobre el proceso de compra que serán valiosos a la hora de diseñar campañas de marketing y estrategias de captación y retención de clientes.

El tamaño del dataset ha supuesto una limitación en la predicción de las ventas por falta de muestras a lo largo de más años.

7. Referencias

https://iabspain.es/estudio/estudio-e-commerce-2022/ https://rockcontent.com/es/blog/ecommerce-marketing/ https://medium.com/swlh/customer-analytics-techniques-to-unveil-customer-insights-7d937dd88ff9 https://www.analyticsvidhya.com/blog/2020/10/a-definitive-guide-for-predicting-customer-lifetime-value-clv/ https://clevertap.com/blog/rfm-analysis/

https://lifetimes.readthedocs.io/en/latest/index.html

https://towardsdatascience.com/predicting-customer-lifetime-value-with-buy-til-you-die-probabilistic-models-in-python-f5cac78758d9

https://productcoalition.com/customer-lifetime-value-explained-9c87c582c4a0

https://www.kaggle.com/code/rajivaiml/brazilian-ecommerce-eda-rfm-nmfLets-do-the-RFM-Analysis-for-the-Sellers

https://facebook.github.io/prophet/

https://www.modeldifferently.com/2022/04/analisis$_p rediccion_t s_p rophet$/21$-$ $prophet-en-profundidad-las-componentes-de-tendencia-estacionalidad-festivos-y-regresores$

https://machinelearningmastery.com/time-series-forecasting-with-prophet-in-python/

https://www.ucm.es/data/cont/docs/518-2016-09-22-Tema3$_S eries$