

## Descripción del caso:

El portal [losdelamoto.com](https://losdelamoto.com) es un sitio web especializado en venta de accesorios para motociclistas orientado a un público joven y de alta capacidad de compra, por esta razón, ofrece artículos de marcas reconocidas así como algunas piezas únicas y de colección.

El portal tiene una fuerte presencia en instagram con más de 2.000 seguidores, hace campañas en motores de búsqueda y ha establecido varias alianzas con concesionarios de motos para poner su marca en las tiendas.

Los administradores del portal quieren mejorar las ventas usando las mejores técnicas de Data Science:

1. ¿Qué ideas se te ocurren? (Para cada idea una describe la acción, las herramientas / modelos y las métricas que usarías para medir su impacto)
  - a. Partiendo de una revisión exhaustiva de la información con la que se cuenta ya sea información transaccional o demográfica, se recomienda construir un perfilamiento (segmentación) de clientes para de esta forma identificar grupos de clientes con diferentes características identificando sus intereses con respecto a los productos que ofrece la compañía y de esta manera poder realizar campañas de marketing a perfiles de clientes en específico.
  - b. De forma complementaria al ya tener identificados los productos más vendidos, marcas, precios y costos, se podrían armar categorías de productos con similares características y ofrecerlos mediante publicidad.
  - c. También la empresa podría adoptar una opción de créditos, observando el comportamiento financiero de los clientes y así ofrecer diferentes formas de pago para la adquisición de productos.
2. Tenemos acceso a los siguientes 4 archivos de datos pero cada uno tiene un costo por lo que debes escoger solamente 2:
  - a. Datos de visitas al sitio tomados de Google Analytics.
  - b. Datos de ventas.
  - c. Catálogo de productos.
  - d. Estadísticas del sector motos offline.

Cuáles escoges y por qué:

Teniendo en cuenta el objetivo de aumentar las ventas, escogería :

- Datos de visitas al sitio tomados de Google Analytics
- Datos de ventas

Dado que el portal es una iniciativa web aporta bastante la información de visitas al sitio, ya que uno, la iniciativa online puede que tenga un comportamiento significativamente diferente al offline y dos, teniendo en cuenta que el grupo objetivo son personas jóvenes y ellos están más adaptados a lo que ofrece la tecnología hoy en día al alcance de un celular o un computador. Y las ventas, son necesarias para a

partir de estas identificar los productos que más se venden y generar campañas de marketing de los productos que más se venden.

3. Adjunto encontrarás una muestra de los archivos de ventas, productos y visitas.  
¿Qué información puedes encontrar en ellos?
  - a.
  - b.
4. ¿Qué tipo de reportes y gráficas harías para comunicar tus conclusiones?
  - a.
  - b.

## Preguntas abiertas

1. Data wrangling. Crea un script para transformar los siguientes datos.

Data original de ventas:

	2019	2018	2017
Cascos	17.000 un.	15.700 un	14.301 un
Guantes	8.100 un	5.400 un	2.100 un

Convertir a:

Producto	Año	Unidades
Cascos	2019	17.000
Cascos	2018	15.700
Cascos	2017	14.301
Guantes	2019	8.100
Guantes	2018	5.400
Guantes	2017	2.100

2. La tabla session almacena las visitas de los usuarios (autenticados o no) al sitio web:

Column_name	Datatype
session_id	autonumeric long
start_time	datetime NOT NULL
end_time	datetime NULL
user_id	long NULL
ip	text

Hacer una consulta en SQL que retorne la lista de usuarios y cuánto tiempo pasó entre su última y penúltima visita.

3. Cada cuánto se debe entrenar un modelo y cómo evaluar el valor decreciente de más información vs el costo de adquirir, procesar y almacenar la información.

Respecto a cada cuanto se debe entrenar un modelo, depende del tipo de modelo y el contexto en que se utilice. Hay modelos más estables y modelos en donde puede haber sobreajuste y es necesario calibrar más seguido para evitar que con nueva data el modelo ya no funcione de igual manera. Una forma de evitar calibrar un modelo muy seguido es realizar una prueba de estabilidad, es decir, dado una ventana de tiempo se pueden escoger periodos de tiempo no necesariamente continuos y desarrollar el modelo con estos y entrenar con los restantes. De manera similar se puede observar la estabilidad de las variables que incluye el modelo e indicadores de desempeño a través del tiempo.

El valor decreciente debería ser planeado desde la construcción del modelo, el costo de variables que entran al modelo, su implementación y sus posibles variaciones.

4. Si tienes una base para detectar fraudes y tienes una gran base de transacciones (incluyendo fraudulentas) que tipo de modelo sugieres?

Si existe dentro de la base una marca con la que se puede identificar desde un inicio si es fraude o no, se pueden utilizar metodologías como regresión logísticas, máquinas de vectores soporte, xgboosts en caso de que no esté dicha marca se podría utilizar un algoritmo de clasificación.