DD360

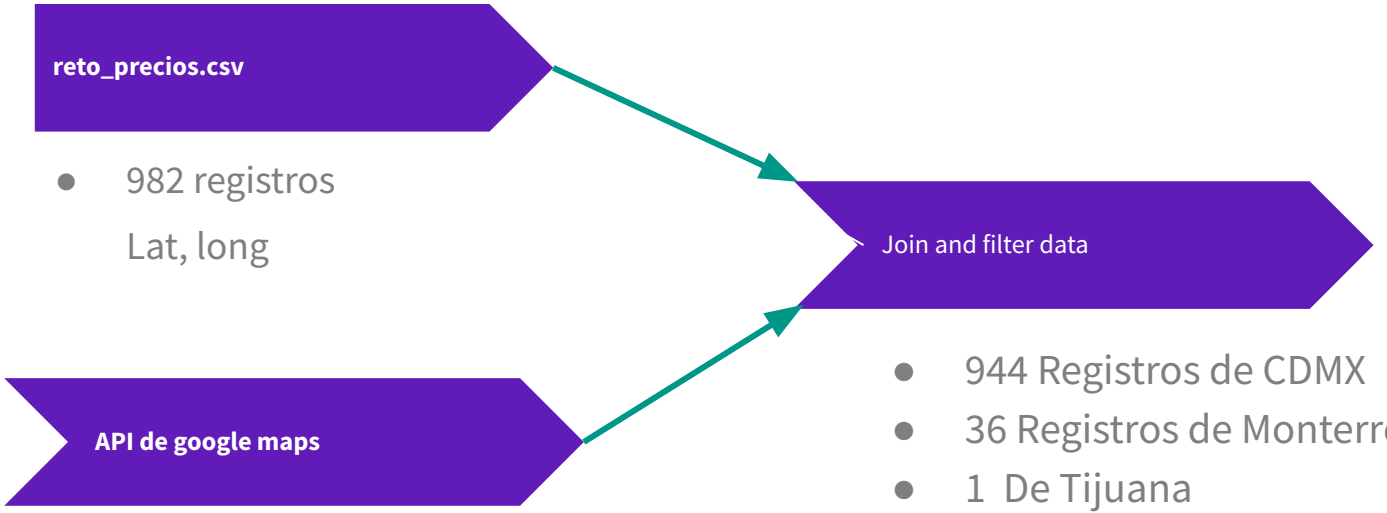
Business case

Deptos. y m2

Problem statement

Identificar los factores que influyen en el precio por metro cuadrado de cada vivienda en un conjunto de departamentos en CDMX.

Compresión de los datos



Modelado

Utilizando un modelo lineal, los factores que se encontraron que determinan el precio por m2 de los departamentos son:

- **La calle en donde se encuentra ubicado**
- Número de cajones de estacionamiento
- Número de habitaciones
- Número de baños
- M2 del departamento

Comentarios

Problem statement

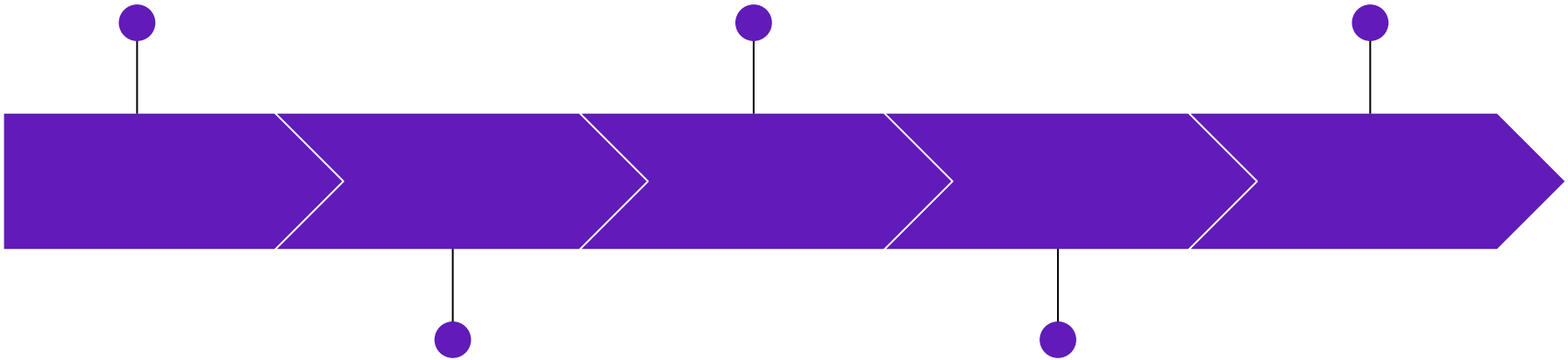
Solicitamos que resumas los principales temas que mencionan los comensales al momento de dejar una reseña.

Preparación de los datos

Se removió puntuación y stopwords de los comentarios

Se descartan palabras poco comunes

Resumen de palabras de los tópicos



Se aplica algoritmo de porter:

Niños, niño -> niño

Selección número de tópicos

Interpretación

Se seleccionaron 4 tópicos resumidos por las palabras:

1. Lugar, excelente, atención, rico , buen, precio, comida, platillo, variedad y sazón.
2. Mejor, taco, caro, mejor, comer, estar, carne y recomendar.
3. Delicioso, sabor, birria, restaurante y bastante.
4. Atención, calidad, mexicana y limpio.

Estimación oferta de carros usados

Contexto de negocio

Clikauto es un marketplace digital de compra y venta de autos usados.

El core de su funnel es la oferta que se hace en su sitio web a sus leads para la compra de autos usados.

Problem statement

Dar en página web una oferta al usuario considerando el año, marca, modelo, versión, subversión y kilometraje de su auto.

Estimación oferta de carros usados

Principales retos

- Data “sucia”. La principal fuente de datos, mercado libre, requiere mucha limpieza.
- El proceso bootstrap requirió implementarse en paralelo para ser eficiente.
- Comunicación efectiva con el equipo de operaciones para implementarse y tener feedback para mejoras posteriores del modelo.

Conjunto de datos

- Input: Scrawler de mercado libre
- Rows: combinaciones de año , marca, modelo, versión, subersión, km, precio
- # rows aprox. 100,000 + delta de 1000 semanal (después de ETL de limpieza y bootstrap)

Resumen del proyecto

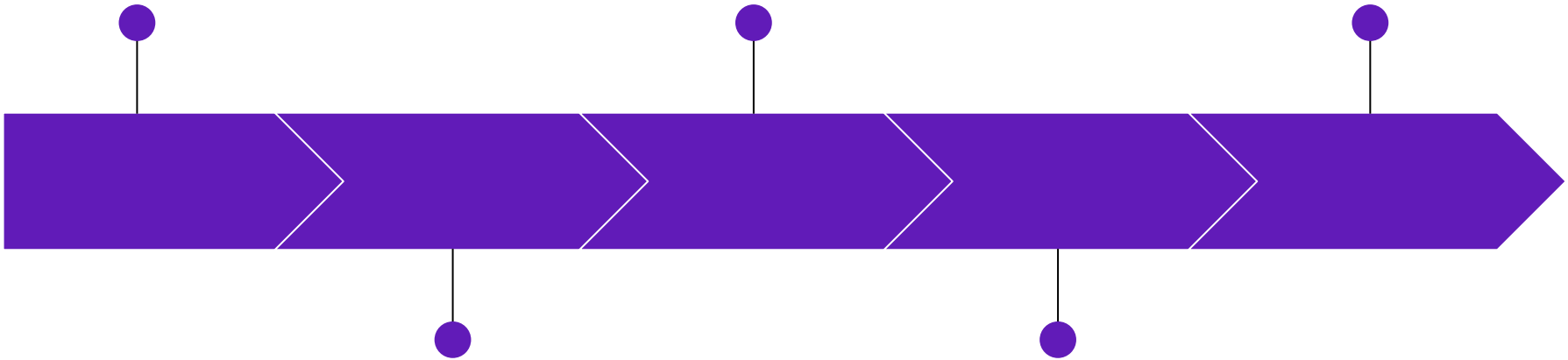
Extracción de data de mercado libre: Proceso dockerizado que se ejecuta diario.

Lectura de S3 de data de mercado y libre(dump) y escritura de data limpia solo de (Año,marca, modelo, versión, subversión, kilometraje, color y precio)

Muestreo bootstrap para eliminar outliers en los precios

Modelo ML de regresión para estimar el precio de los autos

Despliegue del modelo en un contenedor y feedback del modelo en un sheets de google



Aprendizajes

- Una hoja de cálculo es un producto de datos que puede generar mucho valor
- El método bootstrap y la desigualdad de Chebyshev son geniales para detectar outliers.
- El feeling de negocio dio origen a los sistemas expertos.
- El uso de servicios de Glue y Sagemaker en AWS.