

Primera entrega

Integrantes:

- Lukas Guerra
- Jhon Alexander Longas
- Mateo Caicedo

1. Descripción del problema

La predicción de las cantidades que se han de vender en un lugar ya sea para un determinado producto o para un conjunto de productos similares, ha atraído la atención de las diferentes comercializadoras pues esta nueva herramienta permite la disminución de costos en sus inventarios al optimizar su almacenamiento o incluso las pérdidas por daño o perecimiento de productos. Por esto se pretende realizar una predicción sobre las ventas de diferentes *familias* de productos en la comercializadora Favorita con la información de las ventas de los distintos productos en las diferentes tiendas del país.

2. Data

El [dataset](#) que se utilizará proviene de una competencia de la plataforma *kaggle* la cual consta de 5 archivos que contienen información acerca de:

- *Holiday_events*: Este archivo contiene información de los días de festividad en el país de Ecuador, por cada día proporciona el lugar en que se celebra, la fecha, el tipo y si es local o regional
- *Oil*: Aquí se muestra información del precio de la gasolina en el país en las diferentes fechas
- *Stores*: Contiene la información de las tiendas, la ciudad, el estado, el tipo
- *Train*: Datos de entrenamiento del modelo donde se guardan las ventas, el tipo de producto, la fecha y si se encontraba en promoción
- *Test*: Datos de prueba del modelo, contiene las mismas columnas de *train* excepto por las ventas
- *Transactions*: transacciones realizadas en las tiendas por fechas

3. Métrica

La métrica de evaluación principal para el modelo será el error logarítmico medio cuadrático (RMSLE), el cual se calcula mediante la siguiente expresión:

$$\epsilon = \frac{\sqrt{\sum_{i=1}^n (\log(p_i + 1) - \log(a_i + 1))^2}}{n}$$

Donde a_i y p_i corresponden a los valores reales y los valores de las predicciones respectivamente y n el número total de observaciones.

- 4.** Para una predicción como estas en una empresa con tantos puntos de venta, donde los inventarios de cada sucursal son considerablemente grandes, se espera que el nivel de predicción sea bastante alto y mas que todo en esos productos de fácil perecimiento. Se pretende con un modelo como este reducir el lo que mas se pueda ese costo que genera la mercancía guardada, valga la redundancia, reduciendo la cantidad de inventario y además reducir los costos de perdida que producen los productos que son más perecederos.