PI Data Scientist Challenge

El problema

Se proporciona un dataset con datos transaccionales referidos a ventas de distintos productos. El objetivo es construir un modelo de regresión simple para predecir las ventas por producto de una tienda en particular, que forma parte de una cadena de tiendas, y descubrir cuáles son los principales factores que influencian dicha predicción.

Asunciones

- La categoría del producto podría tener cierto impacto en las ventas: productos de consumo general se venden más que bebidas alcohólicas.
- El tipo de tienda y su ubicación es importante para las ventas.
- El tamaño de la tienda puede ser importante (¿la gente va a las tiendas grandes para comprar todo lo que necesita de una sola vez y conseguir mejores precios, o prefiere visitar tiendas pequeñas?)

Su tarea es comprender los datos y mostrarnos el proceso de razonamiento utilizado para llegar a su resolución.

Para ello, deberás:

- Descargar y analizar los datos
  - Limpieza de datos
  - Gráficos
  - Tablas
- Procesamiento de datos:
  - Seleccionar variables que consideres importante para el modelo
  - Feature engineering para variables numéricas, categóricas y datetimes
- Modelado:
  - Entrenamiento
  - Cálculo de métricas apropiadas
  - ¿Cuáles son las principales variables que utiliza el modelo?
- Test
  - Aplicación de modelo en test dataset.

Tener en cuenta que la solución deberá ser legible, reproducible y eficiente para la resolución del challenge. Se libre de elegir cualquier tipo de gráfico o tabla para mostrar la información del modo que creas conveniente.

Santa Rosa 320 Piso 5 A – Teléfono +54(351)5693075
Email: info@piconsulting.com.ar
www.piconsulting.com.ar

Los datos

Se proveen dos archivos train y test con la misma estructura:

- **Item_Identifier:** nombre o identificador del producto
- **Item_Weight:** peso del producto en gramos
- **Item_Fat_Content:** clasificación del producto en términos de grasas contenidas en él.
- **Item_Visibility:** scoring de visibilidad del producto: medida que hace referencia al conocimiento del producto en el consumidor. ¿Qué tan fácil puede ser encontrado el producto?
- **Item_Type:** tipo de producto
- **Item_MRP:** máximum retailed price. Precio calculado por el fabricante que indica el precio más alto que se puede cobrar por el producto.
- **Outlet_Identifier:** identificador de la tienda
- **Outlet_Establishment_Year:** año de lanzamiento de la tienda
- **Outlet_Size:** tamaño de la tienda
- **Outlet_Location_Type:** clasificación de las tiendas según ubicación
- **Outlet_Type:** tipo de tienda
- **Item_Outlet_Sales:** ventas del producto en cada observacion

Presentación de resultados

- El entregable principal es un Informe con tus respuestas: puede ser un jupyter notebook, una presentación en PPT o similar, un HTML, o como lo prefieras. Queremos conocer detalles como los criterios utilizados, dificultades encontradas en el camino y resultados parciales.
    - ¿Qué otras fuentes de datos o información crees que serían interesantes para mejorar tu análisis?
- Podés usar cualquier lenguaje, herramientas y servicios en la nube que desees.
    - Nos gustaría ver tu código o implementación.
- Tendrás una presentación de 30 minutos para mostrar los resultados.
- Esperamos recibir tu trabajo en alrededor de 1 semana posterior al envío del challenge.

Si tenés alguna duda, no dudes en contactarnos en recursos.humanos@piconsulting.com.ar