Universidad de Santiago

Facultad de Administración y Economía

Diplomado en Data Mining

Pauta informe módulo 7

**Fecha de entrega: Jueves 29 de Agosto**

**Contexto**

La empresa ROMA es un retailer chileno que vende productos de consumo masivo, y necesita determinar qué productos deben seguir vendiendo y qué productos deben remover de su stock, con el objetivo de reducir costos de inventario.

Los costos asociados a inventario se entienden como la suma del costo de mantener un producto que no se venderá y el costo de oportunidad de tener quiebres en stock en productos que tienen una alta rotación.

La empresa tiene información histórica que podría utilizar para identificar qué productos deben mantenerse y cuáles pueden ser removidos. Por lo tanto, ROMA lo contrata como consultor para que pueda ayudarle a reducir costos de inventario usando Analytics, específicamente modelos predictivos. Para tal labor, Roma le entrega un archivo consolidado que contiene datos históricos: Niveles inventarios, SKU, precios promedios, mínimo, y otra información relevante, además incluye una variable que indica si los productos tuvieron una venta lo suficientemente alta para cubrir los costos de inventarios, durante los últimos 6 meses.

Luego de estudiar los datos, usted se da cuenta que es posible que es desarrollar un modelo predictivo para identificar que productos tendrán un nivel de ventas alto y cuales tendrán un nivel de ventas bajo, luego de proponerle la metodología a ROMA, obtiene la aprobación y se le encarga construir el modelo.

**Problema**

Su labor consistirá en construir el mejor modelo predictivo que indique que producto tendrá nivel de venta esperado alto. Esto puede realizarse mediante un modelo de clasificación binaria:

- 1: Alto nivel de venta esperado -> No debe removerse del stock
- 0: Bajo nivel de venta esperado -> Debe removerse del stock

El threshold del modelo que construya debe considerar que el costo de clasificar mal un producto con bajo nivel de ventas esperadas (predecir 1 cuando era 0) corresponde a 10 UM, mientras que el costo de clasificar mal a un producto con alto nivel de ventas (predecir 0 cuando era 1 es 15 UM.

**Instrucciones**

Se le proporcionaron tres archivos:

- El archivo TRAIN contiene los datos necesarios para que usted pueda construir el mejor modelo posible.
- El archivo TEST, tiene el mismo formato del archivo TEST, pero debe utilizarlo para generar una predicción a nivel de sku.
- El archivo SUBMIT, es el archivo que debe completar con la predicción para cada sku.

La estructura de los archivos TRAIN y TEST es la siguiente:

| **CAMPO** | **DEFINICIÓN** |
|---|---|
| ORDEN | Correlativo asociado al SKU |
| SKU | Identificador único del SKU |
| VENTA_ALTA | Variable dependiente que indica si las ventas fueron altas (1) o bajas (0) (TARGET) |
| MARKETING | Indica si el producto se comercializa en tienda (D), online + tienda (S) |
| ACTUALIZACIONES | Veces que se han ordenado al proveedor nuevos lotes del producto |
| HA_ACTUALIZADO | Si se ha ordenado un nuevo lote al proveedor |
| VENTAS_ACUMULADAS | Ventas acumuladas del producto, durante 6 meses |
| PRICE | Precio promedio del producto durante los 6 meses |
| LANZAMIENTO | Año de lanzamiento del producto |
| CANTIDAD | Cantidad (miles) de unidades en stock al último mes observado |
| PRECIO_MIN | Precio mínimo en los 6 meses |
| SD_PRECIO | Desviación estándar del precio en los 6 meses |

La estructura del archivo SUBMIT es la siguiente:

| **CAMPO** | **DEFINICIÓN** |
|---|---|
| SKU | Correlativo asociado al SKU |
| PRED | Campo que debe completar con la predicción continua de su modelo (valores entre 0 y 1) |
| PRED_CLASS | Campo que debe completar con la predicción binaria de su modelo al seleccionar un threshold en particular |

**Importante**: Los campos en amarillo deben ser completados por usted.

El trabajo se calificará de la siguiente forma: Debe enviar un correo electrónico a nicolastaglelucero@gmail.com incluyendo 2 archivos:

1) Archivo SUBMIT, con la siguiente estructura:

    - Nombre del archivo: SUBMIT_GRUPO.txt
    - Tipo: txt separado por tabulador "\t"
    - Decimales: Punto "."

2) Informe (1 plana) con el siguiente contenido:

    - ¿Cuántas variables utilizó en su modelo? ¿Construyó variables adicionales? ¿Cuál fue la motivación?
    - ¿Qué metodología de modelamiento utilizó (Ej Random Forest)? ¿Por qué?
    - ¿Cuál es el AUC del modelo seleccionado?
    - Matriz de confusión del modelo seleccionado

Preguntas pueden realizarse al mismo correo indicado anteriormente.