Challenge Analítica Avanzada e Inteligencia Artificial

Centro de Investigación Coppel

Problema UNO

Considera el archivo **20210513_Challenge_AA.csv**. Este documento contiene información sociodemográfica y transaccional sobre algunos clientes sobre los cuales se busca hacer una clasificación de acuerdo a la bandera de la columna *Target*.

El problema del negocio a resolver se trata de un modelo de contactabilidad, se busca asociar una probabilidad de contacto a cada cliente en función a ciertos segmentos horarios, a saber: mañana, mediodía, tarde y noche. Este detalle se encuentra en la columna *Horario.*

En miras en resolver el problema de clasificación planteado, a continuación, realiza la siguiente rutina:

1. Carga los datos del archivo .csv
2. Selecciona el horario de tu preferencia, para en adelante trabajar con él para realizar el modelo, realiza el filtro adecuado.
3. Indica la dimensión del nuevo fichero, número de filas y columnas, nombres de las variables y descripción de los datos.
4. Obtén el número de observaciones, valores perdidos, valores NA.
5. Expresa los detalles estadísticos básicos: promedio, percentiles, desviaciones estándar, valores mínimos y máximos, mediana, etc.
6. Elabora un programa para crear un gráfico para obtener una estadística general de los datos, para todos los datos.
7. Elabora un programa para crear un gráfico de barras para obtener la frecuencia de las categorías de los datos, para las variables continuas.
8. Realiza un análisis de correlaciones para hallar aquellas variables que son linealmente dependientes y establece un criterio para determinar niveles altos y bajos.
9. Establece un criterio para determinar aquellos factores influyentes con respecto al target. Enumera de mayor a menor según el nivel de influencia sobre la variable objetivo. (Hint: Information value, weight of evidence)
10. Con base en los resultados de los puntos anteriores, elabora al menos cinco nuevas variables que de acuerdo a los resultados estadísticos posean buen poder predictivo con respecto a la variable objetivo. Explica la construcción de las variables.
11. Prepara un conjunto de entrenamiento y validación para realizar las determinaciones correspondientes. Justifica el porcentaje usado para cada conjunto.
12. Elabora al menos tres propuestas de modelos de clasificación con las variables de tu preferencia. Justifica las propuestas.

13. Sobre las propuestas realizadas, realiza la valoración de las métricas correspondientes sobre las muestras de entrenamiento y validación. Selecciona un modelo ganador. Justifica el uso de las métricas utilizadas y la selección del modelo que prefieres.
14. Explica cómo visualizas la implementación del modelo ganador, este podrá ser un punto a considerar dentro de tu modelo preferido.
15. Prepara un informe ejecutivo que detalle cada uno de los puntos de esta rúbrica, desde la carga hasta la presentación de resultados y su implementación. El entregable de este challenge será un archivo con el código y el reporte ejecutivo en formato PDF.

Problema DOS

Considera los archivos adjuntos en formato csv. Estos archivos contienen lo siguiente:

1. CostSKUCouches.csv los sku y su costo
2. Festivals - Pricing.xlsx contiene las semanas en las cuales incide esta fecha de compra de los productos (es por ello que las variables son binarias)
3. NationalGlobalSalesWeightedBasePriceSkuStoreMattressesMatrix.csv contiene el precio base de los sku en cada una de las semanas
4. NationalGlobalSalesWeightedDiscountedPriceSkuStoreMattressesMatrix.csv contiene el precio de descuento de los sku en cada una de las semanas
5. NationalUnitSalesSkuStoreMattressesMatrix.csv contiene el número de unidades de cada sku en cada una de las semanas
6. NationTotalUnitsSoldCouches.csv el número total de unidades vendidas en cada semana
7. SKUInformation_Couches.csv contiene los atributos de cada sku

El problema a resolver es dar un pronóstico de las unidades a vender en la 3er semana a partir de la última semana que se tiene. En este caso la última semana con la que cuentas es 2021.07 (semana 7 del 2021) y se busca el forecast para la semana 2021.10. La restricción que se tiene es que no se puede usar data de las últimas dos semanas, en este caso 2021.06 y 2021.05.

Realiza los siguientes puntos:

1. Carga los archivos con los siguientes nombres de acuerdo al orden proporcionado al inicio
    a. df_cost
    b. df_festival
    c. df_base_price
    d. df_discounted_price
    e. df_units
    f. df_total_units
    g. df_infor
2. Una vez cargados cada uno de los archivos la idea es tener tener un dataframe en el cual están los sku con cada uno de estos atributos pero ordenados por las semanas, por ejemplo:

semana	sku	demas columnas de los dataframes
2019.45	2	
2019.46	2	
...		
2018.21	5	
2018.22	5	

3. Determina cuales son los skus más vendidos (90%) de acuerdo número de unidades vendidas en el último año, y filtra el dataframe del punto anterior de acuerdo a esta lista.
4. Gráfica la serie de tiempo de cada sku de sus unidades vendidas, su precio base y su precio de descuento de los últimos dos años.
5. Da los estadísticos descriptivos que consideres importantes para el análisis de cada sku
6. Crea las variables que consideres necesarias para tu modelo
7. Selecciona las variables que consideres importantes
8. Crea tu modelo para realizar el pronóstico de cada sku