Hackathon Data Science

Reparto a Domicilio

Autor: **Eduardo Baffi**

JUMP2DIGITAL

Barcelona Digital Talent

ÍNDICE

1.- RETO A DESARROLLAR

Akadelivers es una empresa de reparto a domicilio especializada en la entrega de paquetes en menos de 1 hora, lo que se denomina (Q-commerce = Quick commerce) Esta empresa tiene una aplicación móvil con la que sus usuarios pueden elegir entre un catálogo de productos de tiendas locales de su ciudad y que les sean entregados en menos de 10 minutos a la dirección que deseen.

Cuando un usuario pide un pedido a través de Akadelivers se le cobra directamente el coste total (coste del producto + gastos de servicio + gastos de envío). Una vez el usuario ha pagado un producto, el repartidor que se encuentre más próximo a la tienda que tiene el producto se acerca a esta, paga el producto, lo recoge y lo lleva a la dirección que el usuario ha elegido. Akadelivers se lo llevara a la dirección indicada.

Datasets

Variables del dataset:

order_id: Número de identificación del pedido.

local_time: Hora local a la que se realiza el pedido.

country_code: Código del pais en el que se realiza el pedido.

store_address: Número de tienda en la que se realiza el pedido.

payment_status: Estado del pedido.

n_of_products: Número de productos que se han comprado en ese pedido.

products_total: Cantidad en Euros que el usuario ha comprado en la app.

final_status: Estado final del pedido (este será la variable 'target' a predecir) que indicará si el pedido será finalmente entregado o cancelado. Hay dos tipos de estado:

- *CanceledStatus: La entrega se ha cancelado.*
- *DeliveredStatus: La entrega se ha realizado correctamente.*

Archivos:

***train.csv:** Este dataset contiene tanto las variables predictoras como el estado del pedido (TARGET).*

***test_X.csv:** Este dataset contiene las variables predictoras con las que se tendrá que predecir el estado de un pedido.*

***ejemplo_predicciones.csv:** Este dataset es un ejemplo de cómo se debera entregar las predicciones del objetivo número 6 de las tareas a relizar.*

2.- OBJETIVO DEL ESTUDIO.

El objetivo del trabajo es contestar 5 preguntas respecto a los datasets. Para las preguntas 1, 2, 3 y 4 fue necesario emplear el dataset 'train.csv'. Para la pregunta 5 fue empleado el dataset 'train.csv' y 'test_X.csv'.

1. *¿Cuáles son los 3 países en los que más pedidos se realizan?*
2. *¿Cuáles son las horas en las que se realizan más pedidos en España?*
3. *. ¿Cuál es el precio medio por pedido en la tienda con ID 12513?*
4. *Teniendo en cuenta los picos de demanda en España, si los repartidores trabajan en turnos de 8horas.*

   Turno 1 (00:00-08:00)

   Turno 2 (08:00-16:00)

   Turno 3 (16:00-00:00)

   Qué porcentaje de repartidores pondrías por cada turno para que sean capaces de hacer frente a los picos de demanda. (ej: Turno 1 el 30%, Turno 2 el 10% y Turno 3 el 60%).

5. *Realiza un modelo predictivo de machine learning a partir del dataset 'train.csv' en el cual a partir de las variables predictoras que se entregan en el dataset 'test_X' se pueda predecir si el pedido se cancelará o no (columna 'final_status').*

   Siendo:

   Para simplificar, podeis asignar los valores 'CanceledStatus' a 0 y los valores 'DeliveredStatus' a 1.

   0 = CanceledStatus

   1 = DeliveredStatus

   Entrega las predicciones en un csv a parte. Tal y como puede verse en el ejemplo de 'ejemplo_predicciones'. La calidad de la predicción se medira a partir del f1-score(macro).

3.- RESULTADOS

El código de trabajo está en el notebook **AkaDelivers_EduardoBaffi.ipynb**

3.1. Limpieza y preparación de los datos obtenidos

El proceso de limpieza y preparación de los datos se ha llevado a cabo utilizando “Numpy”, que es una librería para el lenguaje de programación Python que da soporte para crear vectores y matrices grandes multidimensionales, junto con una gran colección de funciones matemáticas de alto nivel para operar con ellas. Se ha utilizado también “Pandas”, que es una biblioteca de software escrita como extensión de NumPy para manipulación y análisis de datos para el lenguaje de programación Python. En particular, ofrece estructuras de datos y operaciones para manipular tablas numéricas y series temporales.

Principales librerías utilizadas: *numpy, pandas, matplotlib, seaborn*, entre otras.

3.2. Question 1

“¿Cuáles son los 3 países en los que más pedidos se realizan?”

Los tres países en los que más pedidos se realizan son: Argentina, España y Turquía. El dataset tiene 54330 pedidos diferentes. Los resultados para cada uno de los tres países son:

País	Total de pedidos	Percentage
Argentina	11854	21,82%
España	11554	21,27%
Turquía	5696	10,48%

Tabla 1: Los 3 países con más entregas

3.3. Question 2

“¿Cuáles son las horas en las que se realizan más pedidos en España?”

Las cinco horas exactas con más entregas en España están en la próxima tabla.

Hora Completa	Total de Pedidos
20h	1716
21h	1155
19h	1128
13h	1047
14h	956

Tabla 2: Las 5 horas completas con más pedidos en España

Las cinco horas exactas con más entregas son:

Hora Exacta	Total de Pedidos
10:03:06	17
10:03:07	17
21:03:01	16
10:03:05	15
10:03:08	15

Tabla 3: Las 5 horas exactas con más pedidos en España

3.4. Question 3

"¿Cuál es el precio medio por pedido en la tienda con ID 12513?"

El precio medio por pedido de la tienda con ID 12513 es de 17,39 Euros.

3.5. Question 4

"¿Qué porcentaje de repartidores pondrías por cada turno para que sean capaces de hacer frente a los picos de demanda.?"

El porcentaje necesario de repartidores para cada turno es:

- Turno 1 (00:00:00-07:59:59): 0.1%
- Turno 2 (08:00:00-15:59:59): 38.9%
- Turno 3 (16:00:00-23:59:59): 61.0%

El número de repartidores por turno debe ser redondeado a depender del número total de repartidores, ya que el número total de repartidores debe de ser un entero (int).

3.6.Pregunta 5

Resultado del mejor modelo predictivo:

Random Forest Classifier:

F1 Score: 0.9046054198218794
Accuracy: 0.901727014031989
Confusion Matrix:
[[8882 1366]
 [637 9497]]
Classification Error: 0.09827298596801104
Sensitivity: 0.9371422932701796
Specificity: 0.8667056986729118
False Positive Rate: 0.1332943013270882
Precision: 0.8742520482371352

El resultado de la predicción está en el archivo .CSV a parte llamado ‘predition_EduardoBaffi.csv’