92.26 Industrias X.0

TP – EXPLORACIÓN, VISUALIZACIÓN DE DATOS Y MACHINE LEARNING

Fecha de entrega limite: 14/12 a las 19.00 hs

El presente trabajo se deberá realizar en Python.

Las consignas son las siguientes:

1) *Explorar* el dataset *"Bank Marketing"*, el cuál será utilizado para predecir si un cliente del banco pedirá un plazo fijo o no. Explicar la cantidad de filas y columnas que tiene el mismo, cuantas son numéricas y cuantas categóricas. ¿Existen valores vacíos?
   Responder las siguientes preguntas:
   - ¿Cuánta gente está casada, soltera y divorciada?
   - ¿Cuánta gente tiene educación primaria, secundaria y terciaria?
   - ¿Cuánta gente se suscribió al plazo fijo y cuantas no?

2) *Proceso de limpieza*. Todo dataset debe ser limpiado y preparado para realizar un análisis. Por lo general existen columnas vacías, filas con datos faltantes, columnas de mayor interés que otras. ¿Que trabajo realizarías? ¿Existen valores faltantes? Completar los "vacíos", removerlos o en caso de incluirlos explicar por qué se dejan en el dataset. Por otro lado, seleccionar las columnas que pueden ser de interés para el modelo de predicción.

3) *Visualización y análisis de los datos*. ¿Existe correlación entre variables? ¿Hay mejores variables para predecir si un cliente pedirá un plazo fijo?. Realizar al menos 3 gráficos que sirvan para entender mejor la base de datos y que permitan luego realizar un análisis más en detalle (¿Te animás a mejorar el formato del gráfico?)

4) **Machine Learning.** El banco europeo nos contrató como ingenieros industriales y data scientists para predecir si los clientes del banco van a pedir un plazo fijo ("*term deposit*") o no, ya que se está buscando realizar una campaña de marketing para atraer a los clientes a pedirlo. La campaña se realizará con todos aquellos que el modelo prediga que pedirán el plazo. De esta forma los ingresos y costos por la campaña de marketing se traducen de la siguiente forma:

Cliente pide Plazo Fijo?		Predicción	
		No	Si
Realidad	No	0 euros	.- 50 euros
	Si	.- 25 euros	250 euros

(Lease por ejemplo que si el modelo predice que la persona si pedirá un plazo fijo, se hará campaña de marketing con esa persona. Si luego realmente pide el plazo, significará una ganancia de 250 euros).

Proponer al menos ***3 modelos de Machine Learning*** que busquen maximizar las ganancias para el banco con la campaña de marketing y elegir el mejor para hacerle la propuesta final al banco.

El trabajo práctico se deberá entregar con:

- Un informe o presentación (PDF o PPT) de lo realizado. Ser concreto pero incluir todo el análisis realizado.
- El código realizado en Python. Se tiene que poder correr para probar el mejor modelo propuesto.

Comentarios:

- No quedarse solo con las consignas presentadas. Todo análisis extra que se realice será tenido en cuenta.

- La variable respuesta Y tiene los valores 2: Si y 1: No. Modificarla para que los valores sean 1:Si y 0:No

- El beneficio máximo que obtengan puede ser menor que 0

-Recordar agregar en las funciones que realizan transformaciones y en los modelos el parámetro random_state para asegurar la reproductibilidad del código.

- Vamos a evaluar a los grupos en función del beneficio final obtenido. HAY PREMIO PARA EL GRUPO QUE OBTENGA UN MAYOR BENEFICIO!!! *(Salvo que hagan trampa, en ese caso los recursamos 😄)*

Explicación de las variables

- bank client data:

1 - age (numeric)

2 - job : type of job (categorical: "admin.","unknown","unemployed","management","housemaid","entrepreneur", "student","blue-collar","self-employed","retired","technician","services")

3 - marital : marital status (categorical: "married","divorced","single"; note: "divorced" means divorced or widowed)

4 - education (categorical: "unknown","secondary","primary","tertiary")

5 - default: has credit in default? (binary: "yes","no")

6 - balance: average yearly balance, in euros (numeric)

7 - housing: has housing loan? (binary: "yes","no")

8 - loan: has personal loan? (binary: "yes","no")

- related with the last contact of the current campaign:

9 - contact: contact communication type (categorical: "unknown","telephone","cellular")

10 - day: last contact day of the month (numeric)

11 - month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")

12 - duration: last contact duration, in seconds (numeric)

- other attributes:

13 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

14 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)

15 - previous: number of contacts performed before this campaign and for this client (numeric)

16 - poutcome: outcome of the previous marketing campaign (categorical: "unknown","other","failure","success")

- output variable (desired target):

17 - y - has the client subscribed a term deposit? (binary: "yes","no")