Trabajo Final MD004

PARTE I

Imaginemos que trabajas en una empresa de análisis de datos para una compañía de transporte público que opera una flota de autobuses en Barcelona. Debes desarrollar un modelo predictivo para estimar el consumo de combustible de los autobuses en función de diferentes variables. Dispones de un conjunto de datos que incluye 35 variables explicativas (31 continuas y 4 categóricas)

El dataset que te hacen llegar contiene información recopilada de sensores instalados en los autobuses, así como datos operativos y ambientales. A continuación, se presenta un ejemplo de las variables presentes en el conjunto de datos

Ejemplo variables continuas:

- Distancia recorrida desde el último repostaje (en kilómetros)
- Velocidad media del autobús durante el trayecto (en kilómetros por hora)
- Carga promedio de pasajeros a bordo
- Temperatura exterior durante el trayecto (en grados Celsius)
- Año de fabricación del autobús

Variables categóricas:

- Tipo de motor del autobús (convencional, eléctrico, híbrido)
- Ruta del autobús (urbana, interurbana)
- Tramo de la semana (fin_de_semana, no_fin_de_semana)
- Condiciones climáticas (soleado, nublado, lluvioso, nevado)

¿Podrías describir que estrategia seguirías para desarrollar un modelo predictivo utilizando este conjunto de datos? (solo menciona pasos y técnicas) (2pts)

PARTE II

Se dispone del siguiente dataset que contiene datos cualitativos y cuantitativos de clientes de una empresa de telecomunicaciones india en la que se detallan aspectos de los clientes de la empresa. El objetivo del presente dataset es encontrar acciones concretas que nos ayuden a prevenir que un cliente haga churn:

- device user's – device brand (Categorical)
- first_payment_amount – user's first payment amount(Numeric)
- age – user's age(Numeric or categorical?)
- city – user's city(Categorical)
- number_of_cards – #of cards user owns
- payments_initiated – #of payments initiated by user

- payments_failed – #of payments failed
- payments_completed – #of payments completed
- payments_completed_amount_first_7days – amt of payment completed in first 7 days of joining
- reward_purchase_count_first_7days – #of rewards claimed in first 7 days
- coins_redeemed_first_7days – coins redeemed in first 7 days
- is_referral – is user a referred user
- visits_feature_1 – #of visits made by user to product feature 1
- visits_feature_2 – #of visits made by user to product feature 2
- given_permission_1 – has user given permission 1
- given_permission_2 – has user given permission 2
- user_id – user identifier
- is_churned – whether user churned

Data: MD004_ACFinal_customer_churn_data.csv

Se pide:

1. **Análisis exploratorio de los datos(2pts)**
- análisis descriptivo de la variable objetivo (métricas+gráficos) comentando los resultados
- visualizaciones que ayuden a entender la relación entre los atributos y la variable objetivo is_churned (métricas+gráficos) comentando los resultados
2. **Selección de variables, mediante el uso de técnicas estadísticas (usad al menos 2: Correlación, PCA, ANOVA, Información Mutua), para el desarrollo de un modelo de regresión logística (2p)**
- justificad la elección de vuestras variables e interpretad los resultados de las técnicas usadas
3. **Desarrollo del modelo de regresión logística (3p)**
- Selección y justificación de la métrica de optimización del modelo
- Desarrollo de al menos 2 modelos y comparación de resultados (Matriz de confusión)
4. **Conclusiones y vías abiertas (1p)** ¿Qué recomendaciones le daríais a esta empresa para reducir churn?, ¿Cómo os ayuda el modelo que habéis calculado a llegar a estas conclusiones?

Recordad: este ejercicio no tiene una solución única, se espera que se haga una interpretación de los datos obtenidos en todos los puntos y **sobre todo que seáis capaces de sacar recomendaciones que ayuden a reducir el customer churn.**

Entrega:

- fichero Jupyter Notebook con MD004NombreApellidosACFinal.ipynb
- fichero .pdf con MD004NombreApellidosACFinal.pdf
- Plazo: 21/02/2024 19:00