Miriam Guadalupe Reyes Mendoza

Reporte de actividades

07/06/2024

REPORTE DE ACTIVIDADES

ACTIVIDADES REALIZADAS

Repositorio de trabajo:
https://github.com/miriamyi01/Data_Challenge

1. DATA ENGINEERING Y VISUALIZACIÓN

Objetivo: Usando JavaScript o Python y la API (INEGI) realizar la extracción de datos y el manejo de datos necesario para generar un dashboard en la plataforma de tu preferencia resumiendo los puntos principales de la población incluyendo comentarios relevantes.

Solución:

- Dashboard: https://datachallenge-exercise1.streamlit.app/

- Este código en Python extrae y procesa datos demográficos de la API pública del Instituto Nacional de Estadística y Geografía (INEGI) de México. Los datos incluyen información sobre población, género, religión, natalidad, mortalidad, nupcialidad, edad y migración. La aplicación web, construida con Streamlit, permite visualizar estos datos interactivamente, seleccionando un año específico.

- **Datos:**

- Se definen URLs que apuntan a la API de INEGI para obtener datos específicos.
- Se envían solicitudes GET a estas URLs y se procesan las respuestas JSON.
- Los datos extraídos se almacenan en un DataFrame de pandas.
- Se realizan cálculos específicos como población total por género, grupos de edad, y diferentes indicadores demográficos.

- **Visualización:**
  - Se crean gráficos interactivos usando Plotly para representar datos demográficos y nupciales.
  - Se generan gráficos de barras y donuts para mostrar diferentes métricas.

- **Dashboard:**
  - Se estructura la interfaz del dashboard con títulos y secciones para cada conjunto de datos.
  - Se muestra el gráfico correspondiente o una advertencia si no hay datos disponibles para el año seleccionado.

- **Notas:**
  - Streamlit permite crear aplicaciones web interactivas de manera sencilla y rápida, lo que lo hace ideal para prototipos y el despliegue de dashboards.
  - En algunos años, los datos no están disponibles. En esos casos, se muestran mensajes de advertencia personalizados en lugar de los gráficos.

2. ANÁLISIS EXPLORATORIO Y MACHINE LEARNING

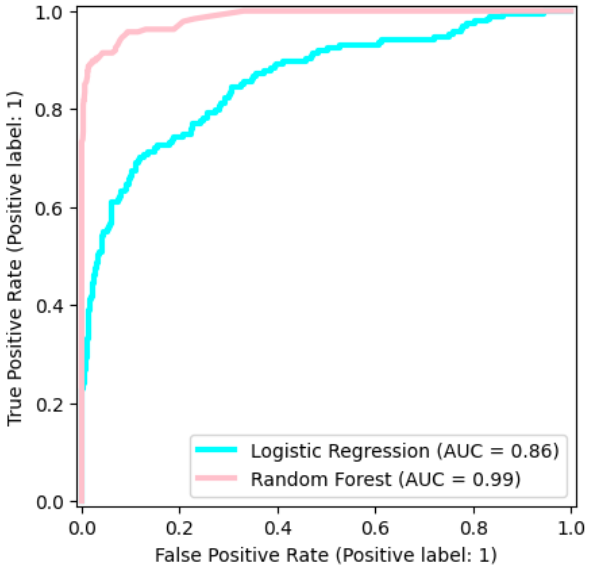
Objetivo: Implementar modelos de regresión para entender impacto de variables explicativas en el churn usando (E Commerce Dataset), explicar conclusiones sobre el impacto de las variables de acuerdo con el desempeño del mejor modelo implementado.

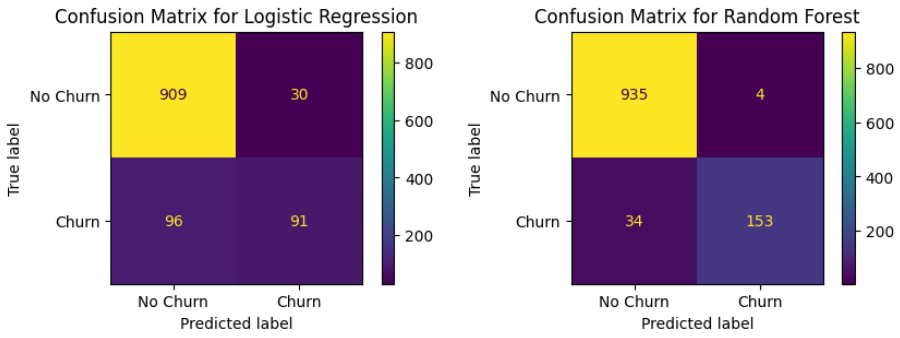
Solución:

- **Datos:**
  - Se cargan los datos y se identifican características numéricas y categóricas.
  - Se estandarizan las características numéricas y se codifican las categóricas usando OneHotEncoder.

- **Modelos:**
  - Se definen dos modelos, uno de regresión logística y otro de bosque aleatorio.
  - Se dividen los datos en conjuntos de entrenamiento y prueba.
  - Ambos modelos se ajustan a los datos de entrenamiento y se evalúan utilizando el conjunto de prueba.

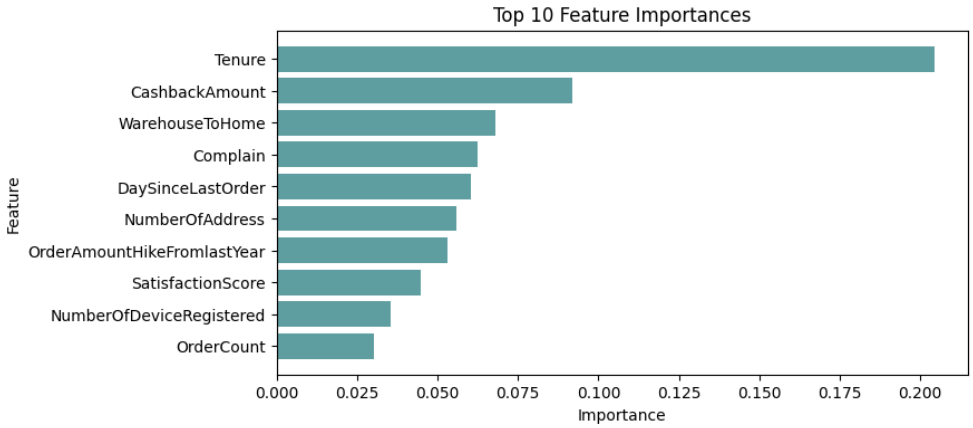
  - **Desempeño:**
    - La regresión logística obtiene una precisión de 0.83 y un AUC de 0.86.
    - El bosque aleatorio obtiene una precisión de 0.97 y un AUC de 0.99, siendo el mejor modelo.





    - En la gráfica de rendimiento, se observa que el modelo de bosque aleatorio supera claramente a la regresión logística. La matriz de confusión muestra que el bosque aleatorio tiene menos falsos negativos y falsos positivos, indicando un mejor rendimiento en la clasificación.

- **Variables:**
  - Según el modelo de bosque aleatorio, las características más importantes que afectan el churn son la tenencia, el monto del cashback, y la distancia del almacén a casa, entre otras.
  - Variables como la satisfacción del cliente y el número de dispositivos registrados también tienen un impacto significativo.



- **Conclusión:** Basándonos en las 10 primeras características más importantes, podemos concluir que:

  - **Tenure:** El tiempo que un cliente ha estado con la empresa es la característica más importante para el modelo. Esto puede indicar que los clientes que han estado con la empresa durante más tiempo tienen comportamientos de compra más predecibles.

  - **CashbackAmount:** La cantidad de reembolso que un cliente recibe también es una característica importante. Esto puede sugerir que los clientes que reciben mayores cantidades de reembolso tienden a comprar más.

  - **WarehouseToHome:** La distancia desde el almacén hasta el hogar del cliente también es una característica importante. Esto puede indicar que la distancia afecta la frecuencia o la cantidad de compras de un cliente.

  - **Complain:** Si un cliente se ha quejado o no también es una característica importante. Esto puede sugerir que los clientes que se han quejado en el pasado pueden tener un comportamiento de compra diferente.

- **DaySinceLastOrder:** El número de días desde la última orden de un cliente también es una característica importante. Esto puede indicar que los clientes que han hecho pedidos recientemente pueden tener más probabilidades de hacer pedidos en el futuro.

- **NumberOfAddress:** El número de direcciones que un cliente tiene registradas es una característica importante. Esto puede sugerir que los clientes con más direcciones pueden hacer más pedidos.

- **OrderAmountHikeFromlastYear:** El aumento en la cantidad de la orden desde el año pasado también es una característica importante. Esto puede indicar que los clientes cuyas cantidades de pedidos han aumentado son más propensos a hacer pedidos en el futuro.

- **SatisfactionScore:** La puntuación de satisfacción del cliente también es una característica importante. Esto puede sugerir que los clientes más satisfechos son más propensos a hacer pedidos.

- **NumberOfDeviceRegistered:** El número de dispositivos registrados por un cliente también es una característica importante. Esto puede indicar que los clientes que tienen más dispositivos registrados pueden hacer más pedidos.

- **OrderCount:** El número de pedidos que un cliente ha hecho también es una característica importante. Esto puede indicar que los clientes que han hecho más pedidos en el pasado son más propensos a hacer pedidos en el futuro.

Estas características pueden ser útiles para entender el comportamiento de compra de los clientes y para hacer predicciones sobre futuras compras.

3. SQL

Objetivo: A partir de los datos adjuntos (SQL_TEST.xlsx), contesta lo siguiente:

Solución:

- Carga el archivo de Excel llamado SQL_TEST.xlsx y establece una conexión a una base de datos SQLite en memoria.

- Se definen rangos para tres tablas (Customer, Product, Station) que se encuentran en el archivo de Excel. Los rangos incluyen las filas donde comienzan y terminan las tablas y las columnas que se deben leer.
- Para cada tabla, carga el nombre de la tabla y los nombres de las columnas del archivo de Excel, y luego carga los datos de la tabla en un DataFrame de pandas.
- Importa los datos del DataFrame a la tabla correspondiente en la base de datos SQLite.
- Define y ejecuta tres consultas SQL.
- Finalmente, imprime los resultados de las tres consultas.

1. Query con los campos y el resultado de lo siguiente: Name y LastName que representan el top 1 de clientes con más compras en cada región MX y USA

   Solución:

```
Query 1 Result:
           Name        LastName Region
0  MARIA LUISA            MENA     MX
1     ANAPAOLA  MUNOZ ARELLANO     US
```

   - El resultado muestra que María Luisa Mena es la cliente con más compras en la región MX y Ana Paola Muñoz Arellano es la cliente con más compras en la región US.

2. ¿Cuáles son los emails de los clientes mujeres con valor de productos comprados mayor a $100? Compartir Query y resultado.

   Solución:

```
Query 2 Result:
      Email
0    email2
1    email3
2    email5
3    email9
4   email10
```

- Los correos electrónicos de las clientes mujeres que han comprado productos con un valor superior a 100 fueron el email2, email3, email5, email9 y email10.

3. ¿Cuál es el número de productos, número de clientes y amount total por región? Compartir Query y resultado.

   Solución:

```
Query 3 Result:
   Region  NumberOfProducts  NumberOfCustomers  TotalAmount
0     MX                 7                  7         4170
1     US                 6                  6          252
```

- En la región MX, hay 7 productos únicos, 7 clientes únicos y el total de ventas es de 4170. En la región US, hay 6 productos únicos, 6 clientes únicos y el total de ventas es de 252.

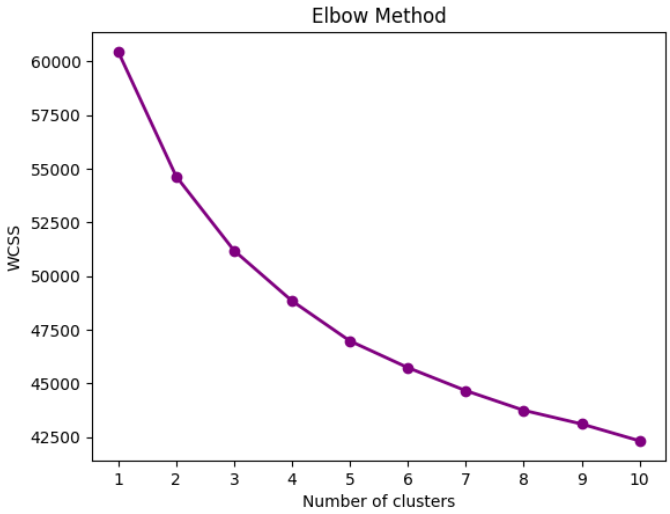
4. CLUSTERIZACIÓN Y EXPLICACIÓN DE SEGMENTOS

Objetivo: Realiza un modelo de clusterización para clientes basado en sus características de compra, incluye insights de cada clúster basados en su comportamiento (E Commerce Dataset).

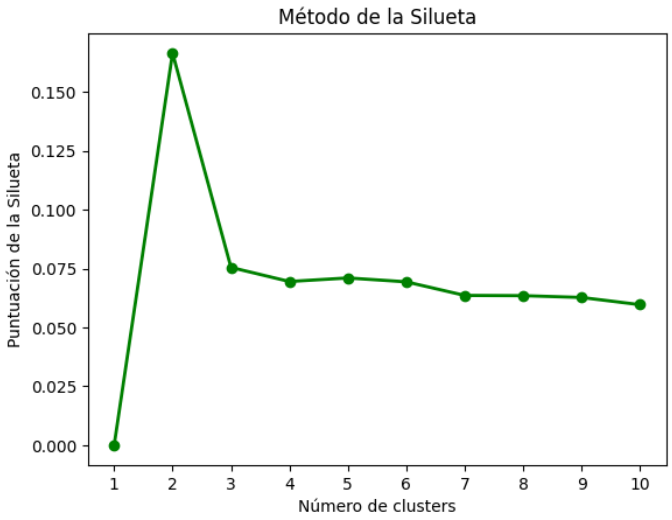
Solución:

- **Datos:**
  - Se cargan los datos desde un archivo Excel.
  - Se definen las características numéricas y categóricas.
  - Se eliminan los valores faltantes y se realiza el preprocesamiento de los datos utilizando StandardScaler para las características numéricas y OneHotEncoder para las características categóricas.

- **Clusters:**
  - *Método del codo:* Se evalúa la suma de las distancias al cuadrado dentro de los clusters (WCSS) para diferentes números de clusters y se grafica el resultado para identificar el "codo" en el gráfico, que indica el número óptimo de clusters.



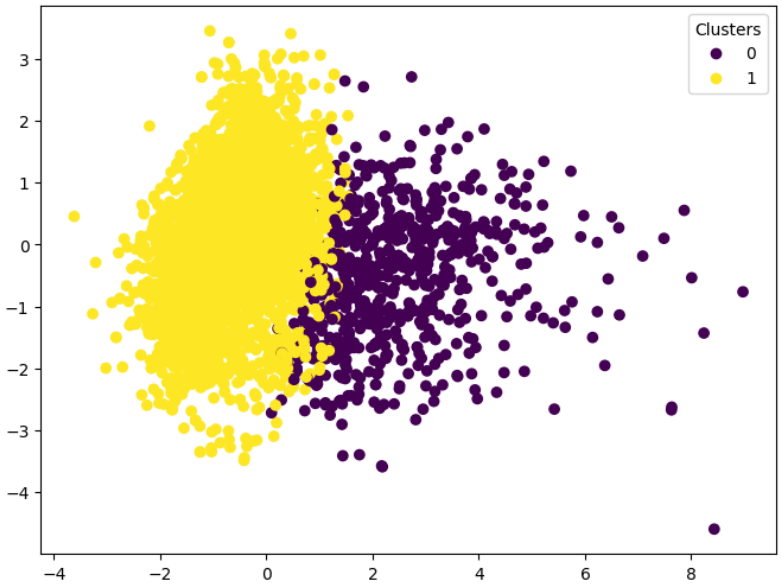
- *Método de la silueta:* Se calcula la puntuación de la silueta para diferentes números de clusters (de 2 a 10) y se grafica el resultado para corroborar el número óptimo de clusters.



- **Algoritmo K-means:**
  - Se aplica el algoritmo de K-means con el número óptimo de clusters determinado previamente.
  - Se asignan los clusters a los datos originales.

  - **Análisis:**
    - Se agregan los clusters a los datos originales.

- Se visualizan los clusters generados utilizando la técnica de reducción de dimensionalidad PCA (Análisis de Componentes Principales) para reducir los datos a 2 dimensiones y graficar los clusters.



- **Insights:**
  - Se generan insights sobre cada cluster analizando las características promedio de cada cluster, proporcionando una visión general del comportamiento de los clientes en cada grupo.

Cluster	Tenure	CityTier	WarehouseToHome	HourSpendOnApp	NumberOfDeviceRegistered
0	10.529326	1.790323	15.708211	3.000000	3.725806
1	8.390362	1.689521	15.751617	2.977038	3.760349

Cluster	SatisfactionScore	NumberOfAddress	Complain	OrderAmountHikeFromlastYear	CouponUsed
0	3.057185	3.618768	0.280059	15.822581	4.145161
1	3.056274	4.348318	0.282665	15.706662	1.184670

Cluster	OrderCount	DaySinceLastOrder	OrderCount	DaySinceLastOrder	CashbackAmount
0	7.189150	8.340176	7.189150	8.340176	176.539179
1	1.862872	3.684994	1.862872	3.684994	161.494945

Cluster	CashbackAmount	PreferredLoginDevice	PreferredPaymentMode	Gender
0	176.539179	Mobile Phone	Debit Card	Male
1	161.494945	Mobile Phone	Debit Card	Male

Cluster	PreferedOrderCat	MaritalStatus	PreferredLoginDevice	PreferredPaymentMode
0	Laptop & Accessory	Married	Mobile Phone	Debit Card
1	Laptop & Accessory	Married	Mobile Phone	Debit Card

- **Conclusión:** El análisis de clustering revela diferencias significativas entre dos grupos de clientes.

  - *Cluster 0:*
    - Tiempo de relación con la empresa más largo (10.53 años en promedio), sugiriendo una mayor lealtad.
    - Mayor tiempo dedicado a la aplicación (3 horas en promedio) y mayor actividad en términos de uso de cupones y número de pedidos.
    - Mayor cantidad de pedidos y montos de reembolso más altos, lo que indica un mayor valor para la empresa.
    - Niveles de satisfacción similares a cluster 1, pero con una proporción de quejas del 28%.
    - Características demográficas y de preferencia similares a cluster 1.

  - *Cluster 1:*
    - Menor tiempo de relación con la empresa (8.39 años en promedio).

- Menor tiempo dedicado a la aplicación (2.98 horas en promedio) y menos actividad en términos de uso de cupones y número de pedidos.
- Menor cantidad de pedidos y montos de reembolso más bajos en comparación con cluster 0.
- Niveles de satisfacción similares a cluster 0, pero con una proporción de quejas ligeramente más alta (28.27%).
- Características demográficas y de preferencia similares a cluster 0.

Podríamos decir entonces que el cluster 0 está formado por clientes más antiguos y activos, con una mayor interacción con la plataforma y una lealtad más alta, mientras que el cluster 1 está compuesto por clientes más nuevos y menos interactivos. Aunque ambos clusters muestran niveles de satisfacción y características demográficas similares pero las diferencias en el comportamiento de compra sugieren la necesidad de estrategias diferenciadas para retener y satisfacer a cada grupo de clientes.