

Documentación del Proyecto ETL de Customer Churn

Descripción del Problema

El abandono de clientes (churn) es uno de los desafíos más significativos que enfrentan las empresas basadas en suscripciones, como las compañías de telecomunicaciones. El churn ocurre cuando los clientes cancelan sus servicios, lo que lleva a una pérdida de ingresos recurrentes. Comprender los factores que contribuyen al churn puede ayudar a las empresas a tomar medidas proactivas para mejorar la retención de clientes.

Objetivo:

Este proyecto tiene como objetivo construir un pipeline ETL (Extract, Transform, Load) para analizar datos de clientes, identificar los factores que contribuyen al churn y extraer información valiosa para reducir el abandono. El resultado final incluirá un panel de control para visualizar métricas clave de churn y perspectivas sobre los clientes.

Contexto

Las compañías de telecomunicaciones suelen ofrecer múltiples servicios, como teléfono, internet y suscripciones de TV. Factores como la duración del contrato, la calidad del servicio, los precios y el soporte al cliente pueden influir en la decisión de un cliente de quedarse o abandonar el servicio. Identificar señales tempranas de churn puede ayudar a las empresas a implementar intervenciones específicas para mejorar la satisfacción y la retención de clientes.

Descripción del Conjunto de Datos

El conjunto de datos principal para este proyecto es el **Telco Customer Churn Dataset** obtenido de [Kaggle](#). Este conjunto de datos contiene información detallada sobre los clientes, incluidos los tipos de servicio, detalles del contrato, métodos de pago y un indicador de churn.

Principales Columnas del Dataset:

- **CustomerID:** Identificador único del cliente.
- **Gender:** Género del cliente (Male/Female).
- **SeniorCitizen:** Indica si el cliente es un adulto mayor (1) o no (0).

- **Partner:** Si el cliente tiene pareja (Yes/No).
- **Dependents:** Si el cliente tiene dependientes (Yes/No).
- **Tenure:** Tiempo en meses que el cliente ha estado con la empresa.
- **PhoneService:** Si el cliente tiene servicio de teléfono (Yes/No).
- **MultipleLines:** Si el cliente tiene múltiples líneas telefónicas (Yes/No/No phone service).
- **InternetService:** Tipo de conexión a internet (DSL/Fiber optic/No).
- **OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies:** Servicios adicionales contratados (Yes/No/No internet service).
- **Contract:** Tipo de contrato (Month-to-month/One year/Two year).
- **PaperlessBilling:** Si el cliente tiene facturación sin papel (Yes/No).
- **PaymentMethod:** Método de pago (Electronic check, Mailed check, Bank transfer, Credit card).
- **MonthlyCharges:** Monto mensual que paga el cliente.
- **TotalCharges:** Monto total pagado por el cliente.
- **Churn:** Variable objetivo que indica si el cliente ha abandonado el servicio (Yes/No).

Tamaño del Conjunto de Datos:

- Filas: 7,043
- Columnas: 21

Proceso

1. Fuentes de Datos

- **Fuente:** Kaggle (archivo CSV)

2. Extracción de Datos

- **Configuración del Ambiente de Trabajo**
Para garantizar la reproducibilidad y estabilidad del proceso, se configuró un ambiente virtual en Python donde se instalaron las librerías necesarias. Entre estas se encuentran pandas para la manipulación de datos, kaggle

para la descarga del dataset, mysql-connector-python para la conexión con MySQL y python-dotenv para la gestión segura de credenciales.

- **Descarga de Datos desde Kaggle**

El dataset *Telco Customer Churn*, disponible en Kaggle, fue descargado mediante la API de la plataforma. Para ello, se configuró la autenticación con una clave de acceso, lo que permitió la descarga automatizada del archivo CSV con los datos de los clientes.

- **Carga y Validación de los Datos**

Una vez descargado, el dataset fue cargado en un DataFrame de pandas, donde se realizó una primera inspección para verificar su integridad. Se revisaron aspectos como la cantidad de registros y columnas, los tipos de datos y la presencia de valores nulos o inconsistentes.

- **Limpieza y Preparación Inicial**

Se aplicaron algunas transformaciones básicas para garantizar la calidad de los datos antes de su almacenamiento. Entre ellas, la conversión de ciertos valores al formato numérico, la eliminación de espacios en blanco y la gestión de valores vacíos.

- **Completar registros:** Debido a que el dataset tiene menos registros que los solicitados se completo con datos aleatorios generados, se garantizo que se mantengan los **CustomerID diferentes para cada registro**

- **Almacenamiento en MySQL**

Para facilitar su acceso y procesamiento posterior, los datos fueron almacenados en una base de datos MySQL. Se creó una estructura de tabla acorde a las características del dataset, asegurando que cada columna tuviera un tipo de dato apropiado. Luego, se insertaron los registros extraídos, verificando que la carga se realizara correctamente.

- **Validación Final**

Finalmente, se realizaron consultas sobre la base de datos para verificar que los datos fueron cargados correctamente. Se revisó la cantidad de registros almacenados y se confirmó que las estructuras de datos coincidieran con la fuente original.

3. Transformación de Datos (Posibles transformaciones)

3.1 Limpieza de datos

- **Manejo de valores nulos:** eliminar filas con demasiados valores faltantes o imputar valores con la media/moda.
- **Corrección de formatos** (ej., convertir fechas si hay alguna columna con fechas).
- **Eliminación de duplicados.**

3.2 Conversión de variables categóricas

- **Convertir variables como gender, Contract, PaymentMethod en variables dummy (one-hot encoding) o en valores numéricos.**
- **Transformar SeniorCitizen (0/1) en una variable categórica.**

3.3 Creación de nuevas variables

- **TotalCharges: Convertir a numérico si está en formato string.**
- **Tenure: Transformarlo en rangos de permanencia (ej., 0-12 meses, 13-24 meses...).**
- **MonthlyCharges y TotalCharges: Crear una nueva variable $\text{AvgMonthlySpend} = \text{TotalCharges} / \text{tenure}$ para ver patrones de gasto.**

4. Carga de Datos

- Cargar los datos transformados en una base de datos final para su análisis.
- Asegurar la integridad de los datos e indexación para una recuperación optimizada.

5. Creación de Dashboard

- Usar una herramienta de visualización (por ejemplo, Power BI o Looker Studio) para crear paneles interactivos que muestren:
 - Tasas de churn por demografía y tipo de servicio.
 - Tendencias de churn a lo largo del tiempo.
 - Factores clave que contribuyen al churn.

Evidencias del Proceso ETL

1. Extracción de Datos

- **Extraccion_muestra.csv: Se guarda una muestra de los primeros 10 registros extraídos del dataset, para verificar que los datos fueron descargados correctamente.**

```
__init__.py  main.py  extraccion_muestra.csv  extract.py 3  .env  transform.py  load.py  README.md  .gitig  ...
evidencias > extraccion_muestra.csv
1  customerID,gender,SeniorCitizen,Partner,Dependents,tenure,PhoneService,MultipleLines,InternetService,OnlineSecurity,OnlineBackup,De
2  7590-VHVEG,Female,0,Yes,No,1,No,No phone service,DSL,No,Yes,No,No,No,Month-to-month,Yes,Electronic check,29.85,29.85,No
3  5575-GNVEDE,Male,0,No,No,34,Yes,No,DSL,Yes,No,Yes,No,No,No,One year,No,Mailed check,56.95,1889.5,No
4  3668-QPYBK,Male,0,No,No,2,Yes,No,DSL,Yes,Yes,No,No,No,Month-to-month,Yes,Mailed check,53.85,108.15,Yes
5  7795-CFOQW,Male,0,No,No,45,No,No phone service,DSL,Yes,No,Yes,Yes,No,No,One year,No,Bank transfer (automatic),42.3,1840.75,No
6  9237-HQITU,Female,0,No,No,2,Yes,No,Fiber optic,No,No,No,No,No,Month-to-month,Yes,Electronic check,70.7,151.65,Yes
7  9305-CDSKC,Female,0,No,No,8,Yes,Yes,Fiber optic,No,No,Yes,No,Yes,Yes,Month-to-month,Yes,Electronic check,99.65,820.5,Yes
8  1452-KIOVK,Male,0,No,Yes,22,Yes,Yes,Fiber optic,No,Yes,No,No,Yes,No,Month-to-month,Yes,Credit card (automatic),89.1,1949.4,No
9  6713-OKOMC,Female,0,No,No,10,No,No phone service,DSL,Yes,No,No,No,No,Month-to-month,No,Mailed check,29.75,301.9,No
10 7892-POOKP,Female,0,Yes,No,28,Yes,Yes,Fiber optic,No,No,Yes,Yes,Yes,Yes,Month-to-month,Yes,Electronic check,104.8,3046.05,Yes
11 6388-TABGU,Male,0,No,Yes,62,Yes,No,DSL,Yes,Yes,No,No,No,No,One year,No,Bank transfer (automatic),56.15,3487.95,No
12
```

- **Datos_limpiados_valores_nulos.csv,**
datos_limpiados_muestra.csv y
datos_limpiados_estadisticas.csv: Proceso de validación y
limpieza de datos.

```
__init__.py  main.py  datos_limpiados_valores_nulos.csv  extract.py 3  .env  transform.py  load.py  README.md  ...
evidencias > datos_limpiados_valores_nulos.csv
1  ,0
2  customerID,0
3  gender,0
4  SeniorCitizen,0
5  Partner,0
6  Dependents,0
7  tenure,0
8  PhoneService,0
9  MultipleLines,0
10 InternetService,0
11 OnlineSecurity,0
12 OnlineBackup,0
13 DeviceProtection,0
14 TechSupport,0
15 StreamingTV,0
16 StreamingMovies,0
17 Contract,0
18 PaperlessBilling,0
19 PaymentMethod,0
20 MonthlyCharges,0
21 TotalCharges,11
22 Churn,0
23
```

```
__init__.py  main.py  datos_limpiados_muestra.csv  extract.py 3  .env  transform.py  load.py  README.md  ...
evidencias > datos_limpiados_muestra.csv
1  customerID,gender,SeniorCitizen,Partner,Dependents,tenure,PhoneService,MultipleLines,InternetService,OnlineSecurity,OnlineBackup,De
2  7590-VHVEG,Female,0,Yes,No,1,No,No phone service,DSL,No,Yes,No,No,No,Month-to-month,Yes,Electronic check,29.85,29.85,No
3  5575-GNVEDE,Male,0,No,No,34,Yes,No,DSL,Yes,No,Yes,No,No,No,One year,No,Mailed check,56.95,1889.5,No
4  3668-QPYBK,Male,0,No,No,2,Yes,No,DSL,Yes,Yes,No,No,No,Month-to-month,Yes,Mailed check,53.85,108.15,Yes
5  7795-CFOQW,Male,0,No,No,45,No,No phone service,DSL,Yes,No,Yes,Yes,No,No,One year,No,Bank transfer (automatic),42.3,1840.75,No
6  9237-HQITU,Female,0,No,No,2,Yes,No,Fiber optic,No,No,No,No,No,Month-to-month,Yes,Electronic check,70.7,151.65,Yes
7  9305-CDSKC,Female,0,No,No,8,Yes,Yes,Fiber optic,No,No,Yes,No,Yes,Yes,Month-to-month,Yes,Electronic check,99.65,820.5,Yes
8  1452-KIOVK,Male,0,No,Yes,22,Yes,Yes,Fiber optic,No,Yes,No,No,Yes,No,Month-to-month,Yes,Credit card (automatic),89.1,1949.4,No
9  6713-OKOMC,Female,0,No,No,10,No,No phone service,DSL,Yes,No,No,No,No,Month-to-month,No,Mailed check,29.75,301.9,No
10 7892-POOKP,Female,0,Yes,No,28,Yes,Yes,Fiber optic,No,No,Yes,Yes,Yes,Yes,Month-to-month,Yes,Electronic check,104.8,3046.05,Yes
11 6388-TABGU,Male,0,No,Yes,62,Yes,No,DSL,Yes,Yes,No,No,No,No,One year,No,Bank transfer (automatic),56.15,3487.95,No
12
```

```
_init_.py  main.py  datos_limpiados_estadisticas.csv  extract.py 3  .env  transform.py  load.py  README.md  ...
evidencias > datos_limpiados_estadisticas.csv
1  ,SeniorCitizen,tenure,MonthlyCharges,TotalCharges
2  count,7043.0,7043.0,7043.0,7032.0
3  mean,0.1621468124378816,32.37114865824223,64.76169246059918,2283.3004408418656
4  std,0.3686116056100131,24.55948102309446,30.090047097678493,2266.771361883145
5  min,0.0,0.0,18.25,18.8
6  25%,0.0,9.0,35.5,401.45
7  50%,0.0,29.0,70.35,1397.475
8  75%,0.0,55.0,89.85,3794.7375
9  max,1.0,72.0,118.75,8684.8
10
```

2. Carga de Datos en MySQL

- **Consulta SQL:** Se ejecuta esta consulta para verificar la creación de la base de datos, la tabla de datos extraídos y cuántos registros han sido insertados en la base de datos.

```
mysql> Show databases
-> ;
+-----+
| Database |
+-----+
| customer_churn |
| ejec1_etl |
| etl_db |
| information_schema |
| modelo_uao |
| mysql |
| performance_schema |
| sys |
+-----+
8 rows in set (0.00 sec)

mysql>
```

```
mysql> Show tables;
+-----+
| Tables_in_customer_churn |
+-----+
| customer_churn_extract |
+-----+
1 row in set (0.00 sec)

mysql> |
```

```
mysql> SELECT COUNT(*) AS total_registros FROM customer_churn_extract;
+-----+
| total_registros |
+-----+
| 12043 |
+-----+
1 row in set (0.00 sec)

mysql> |
```

3. Respuesta en Terminal: Cadena de respuestas de la terminal a lo largo del proceso de extracción para validar el funcionamiento de la lógica.

```
(customer-churn-etl-py3.12) PS C:\Users\juanm\OneDrive\Documentos\ETL\Proyecto\customer-churn-etl> poetry run main.py

🔥 Iniciando proceso ETL...

📦 Extrayendo datos desde Kaggle...
Dataset URL: https://www.kaggle.com/datasets/blastchar/telco-customer-churn
📁 Evidencias guardadas en evidencias/
✅ Extracción completada.
📁 Evidencia de extracción guardada en evidencias/.

📦 Cargando datos en MySQL...
📦 Cargando datos en MySQL...
📦 Cargando datos en MySQL...
📦 Cargando datos en MySQL...
📦 Cargando datos en MySQL...
📦 Cargando datos en MySQL...
✅ Base de datos y tabla verificadas.
📦 Cargando datos en MySQL...
📦 Cargando datos en MySQL...
📦 Cargando datos en MySQL...
✅ Base de datos y tabla verificadas.
📦 Cargando datos en MySQL...
📦 Cargando datos en MySQL...
📦 Cargando datos en MySQL...
📦 Cargando datos en MySQL...
📦 Cargando datos en MySQL...
📦 Cargando datos en MySQL...
✅ Base de datos y tabla verificadas.
📦 Cargando datos en MySQL...
✅ 12043 registros cargados en MySQL.
📁 Evidencia de carga guardada en evidencias/log_carga.txt.
✅ Carga en base de datos completada.

🔥 ETL finalizado exitosamente. 🔥
(customer-churn-etl-py3.12) PS C:\Users\juanm\OneDrive\Documentos\ETL\Proyecto\customer-churn-etl>
```