Examen Práctico - Minería de Datos

Universidad San Francisco de Quito

28 de febrero de 2025

1. Contexto de Negocio

La empresa **TelcoAndes** proporciona servicios de:

- Mensajería (SMS).
- Llamadas de voz.
- Internet móvil (datos).

Se cuenta con la siguiente información en tres tablas principales (además de las relaciones intermedias necesarias):

1. **Tabla de Usuarios**:
   - `user_id`: Identificador único del usuario.
   - `nombre`, `apellido`: Datos personales.
   - `edad`, `genero`, `region`: Información demográfica.
   - `fecha_alta`: Fecha en la que el usuario se suscribió a TelcoAndes.
   - `fecha_abandono`: Fecha en la que el usuario abandonó a TelcoAndes.
   - `plan_id`: Identificador del plan al cual está suscrito.
2. **Tabla de Planes**:
   - `plan_id`: Identificador único del plan.
   - `nombre_plan`: Nombre comercial del plan.
   - `costo_mensual`: Tarifa base que paga el usuario cada mes.
   - `minutos_incluidos`, `sms_incluidos`, `datos_incluidos`: Cantidades mensuales incluidas sin costo adicional.
3. **Tablas de Consumo (Ej. Llamadas, SMS, Datos)**:
   - `user_id`: Para enlazar con la tabla de usuarios.
   - `fecha_consumo`: Fecha y hora de la transacción (llamada, SMS, uso de datos).
   - `duracion_llamada` (en la tabla de llamadas).
   - `cantidad_sms` (en la tabla de SMS).
   - `mb_consumidos` (en la tabla de datos).

1.1. Problema de Negocio

En este examen, asumimos que **TelcoAndes** desea:

Predecir el Monto Total (Facturación) que un Usuario Pagará en el Próximo Mes

La predicción contempla tanto la tarifa fija del plan (`costo_mensual`) como los costos extra (llamadas, SMS y datos que exceden lo incluido en el plan).

2. Instrucciones Generales

Se solicita aplicar un proceso de minería de datos completo, abordando los siguientes pasos:

1. **Análisis Exploratorio de Datos (EDA)**
   - Describir la calidad de los datos, tipos de variables, rangos y posibles outliers.
   - Visualizaciones básicas que muestren distribuciones (histogramas, boxplots, etc.) para la variable *costo total* (o variable objetivo) y sus posibles relaciones con características demográficas (`edad`, `region`, `genero`, etc.).
2. **Data Wrangling**
   - Manejo de valores faltantes en cualquiera de las tablas.
   - Identificación y tratamiento de outliers en consumo (llamadas extremas, SMS masivos, sobreuso de datos).
   - Unificación de las tablas para crear un dataset analítico que permita entrenar un modelo (`JOINs` apropiados, cuidado con la duplicación de registros).
3. **Modelado de Datos (Snowflake o Star Schema)**
   - Explicar el modelo de datos diseñado para el almacenamiento y la consulta (dimensiones y hechos).
   - Indicar cómo las tablas de usuarios, planes y consumos se relacionan para conformar una tabla de *hechos* que contenga el costo total por usuario y periodo.
4. **Feature Engineering**
   - Crear variables relevantes (e.g., `total_llamadas`, `total_sms`, `total_mb`, `uso_extra`, etc.).
   - Considerar transformaciones de las variables originales (log-transform, escalados, etc.).
   - Incluir variables demográficas (edad, región, etc) si las consideran importantes.
5. **Entrenamiento y Evaluación de Modelos de Regresión**
   - Separar el dataset en entrenamiento y prueba.

- Entrenar varios modelos de regresión (Regresión Lineal, Gradient Descent, Stocastic GD, etc).
- Evaluar métricas de error (`RMSE`, `MAE`, `R^2`) y compararlas.
- Seleccionar el mejor modelo para predecir el costo total esperado del usuario en el próximo mes.

6. **Conclusiones y Recomendaciones de Negocio**
   - Explicar con claridad el modelo seleccionado y su precisión.
   - Comentar cómo estos resultados pueden usarse para definir estrategias de precios, promociones o paquetes de servicios personalizados.
   - Proponer mejoras a futuro (más datos, otras variables o técnicas).

3. Estructura de Repositorio y Entrega

Cada estudiante (o equipo) debe crear un repositorio en GitHub con la siguiente estructura de carpetas:

```
.
 data
    raw/
    clean/
    ml/
 notebooks
    1_EDA.ipynb
    2_Data_Wrangling.ipynb
    3_Data_Modeling.ipynb
    4_Feature_Engineering.ipynb
    5_Training_Evaluation.ipynb
    6_Conclusions.ipynb
 README.md
 requirements.txt
```

- **data/raw**: Dataset original.
- **data/clean**: Dataset después de limpiezas y modelamiento.
- **data/ml**: Dataset despues del feature engineering.
- **notebooks**:
  - `1_EDA.ipynb`: Análisis Exploratorio de Datos.
  - `2_Data_Wrangling.ipynb`: Limpieza y unificación de datos.
  - `3_Data_Modeling.ipynb`: Diseño del esquema (Snowflake o Star).

- `4_Feature_Engineering.ipynb`: Creación y selección de variables.
- `5_Training_Evaluation.ipynb`: Entrenamiento y evaluación de modelos.
- `6_Conclusions.ipynb`: Conclusiones y recomendaciones de negocio.

- **README.md**: Descripción del proyecto, cómo ejecutarlo y cualquier información relevante.
- **requirements.txt**: Lista de librerías necesarias (`pandas`, `numpy`, `sklearn`, `matplotlib`, `seaborn`, etc.).

Al finalizar, entregar:

- Enlace al **repositorio de GitHub** con toda la estructura solicitada.

4. Criterios de Evaluación

1. **Documentación del EDA y Data Wrangling**.
2. **Modelado de Datos (Snowflake o Star) y Justificación**.
3. **Calidad de la Feature Engineering y Transformaciones**.
4. **Entrenamiento y Evaluación de los Modelos** (uso de métricas apropiadas).
5. **Conclusiones de Negocio Claras y Relevantes**.
6. **Estructura de Carpeta y Organización del Proyecto**.

¡Éxitos en el Examen Práctico y a demostrar sus habilidades en Minería de Datos!