Data Scientist Challenge – Entel

Esta prueba técnica tiene por objetivo evaluar tus habilidades en programación y manejo de datos. También se evalúa el uso de Git como sistema de control de versiones y el orden en el código para garantizar su reproducibilidad. Se recomienda guardar el desarrollo en tu plataforma de almacenamiento de repositorios favorita (Github, BitBucket, GitLab, etc) en un repositorio privado y darle acceso al mail que te envió la prueba.

Si no conoces GIT, siempre se puede aprender. Entréganos tu proyecto (o repositorio) de la forma que te acomode como respuesta al mail del que te enviamos la prueba.

Instrucciones Git (recomendado)

1) Crear un repositorio privado en la plataforma de Git que más te acomode
2) Trabajar con una rama principal y otra de desarrollo

Instrucciones del desafío

1) Debes responder el mail con asunto 'Challenge Data Scientist - [Nombre][Apellido]', ejemplo: Challenge Data Scientist – Juan Soto. Debes adjuntar el link al repositorio.
2) Se aceptarán cambios en el repositorio hasta la fecha definida en el mail.
3) El repositorio debe tener las instrucciones necesarias para correr tu solución en Python, R o similar en forma reproducible.
4) Dentro del repositorio deben estar todos los archivos necesarios para que los evaluadores puedan clonar y luego ejecutar tu código sin problemas.

Instrucciones prueba

En este desafío te entregamos un archivo llamado **dataset_prueba.csv** con X variables e Y columnas, donde encontrarás datos de una TELCO sobre el uso telefónico de ciertos usuarios.

Este dataset tiene datos de 99.999 clientes (mobile_number) en un período de 3 meses. Contiene información como el promedio de recargas, uso de la red 2G y 3G, número de llamadas emitidas y recibidas, etc. Existen 2 columnas que no son variables: **mobile_number y last_date_of_month**, y una columna target: **churn.**

Si lo necesitas, puedes encontrar las descripciones de los acrónimos de las columnas en el archivo **diccionario_datos.csv** adjuntado en el mail.

Un problema muy común en la industria es intentar predecir cuáles serán los clientes que se irán o no de la compañía, conocido como tasa de fuga o **churn** (en adelante la llamaremos así). Considera que perder un cliente es muy costoso para cualquier empresa, por lo que se necesitan medidas de mitigación focalizada a clientes propensos a fugarse.

El objetivo de esta prueba es analizar el **churn** de los clientes con respecto a otras variables e intentar predecirlo en el último mes disponible.

Primera parte: Análisis exploratorio (Estimamos que esto debería tomar 90 minutos de desarrollo)

Se recomienda especificar dentro de tu código la pregunta que estás respondiendo para mantener un orden coherente.

Primero, debes separar el dataset **dataset_prueba.csv** en dos conjuntos, entrenamiento y test. El conjunto de entrenamiento contiene los registros con fechas (columna last_date_of_month) 7/31/2014 y 6/30/2014. Mientras que el conjunto de test contiene los registros con fecha 8/31/2014. **Las preguntas 1 a 5 consisten en explorar y editar los datos del conjunto de entrenamiento.**

Para las preguntas número 1 a la 4, necesitamos que entregues tu respuesta en el archivo de texto adjunto **respuestas.txt**, editando el valor por defecto que te entregamos. Ten en cuenta que este archivo se evalúa de forma automática, solo edita los valores númericos.

1. Crea una columna a partir de la columna ARPU (average revenue per user) que clasifique a los clientes en tres categorías. Llama a esta columna *clasificacion_clientes_revenue*.
    a. El decil con mayor valor, tendrá la clasificación **platino**
    b. El siguiente decil, tendrá la clasificación **gold**
    c. El resto, son de categoría **normal**
    d. Anota tu respuesta en el archivo de texto con la cantidad de registros según clasificación #platino #gold #otros ej: 20000 25000 150000
2. Crea una columna binaria con nombre *flag_recarga*, si es que hizo (1) o no (0) una recarga durante el mes, esto proviene de la columna *total_rech_num* (cantidad de recargas totales).
    a. Anota tu respuesta en el archivo de texto con la cantidad de valores #1 #0, ej: 15000 185000
3. ¿Cuál es la proporción de churn de cada mes (junio y julio) como 1s sobre el total?
    a. Anota tu respuesta en el archivo de texto con los 2 porcentajes por mes; 0.053 0.038
4. Elimina las columnas que contienen sobre 70% de valores nulos, adjuntar número de columnas restantes. ¿Hay columnas que no deberían tener valores nulos?
    a. Anota tu respuesta en el archivo de texto con la cantidad de columnas que fueron borradas, ej: 13
    b. (Bonus) ¿Hay más columnas que borrarías? ¿Por qué?
5. Si un cliente hace **churn** o no, ¿se observan diferencias en la distribución de la variable total_rech_num? Hint: graficar para comparar te puede ayudar a responder esta pregunta. **No es necesario adjuntar respuesta en el archivo respuestas.txt**

Segunda parte: Modelo (estimamos que te tomará 90 minutos)

Tienes total libertad para hacer lo que se te ocurra, pero esperamos una explicación de por qué tomaste tal decisión. Recuerda que en esta sección puedes utilizar el conjunto de test. Considera que es más importante el análisis y las decisiones que las métricas de rendimiento. **Nosotros no sabemos lo que estás pensando, anota lo que sea relevante para entender tu razonamiento y decisiones.**

6. Usa uno o más modelos predictivos, que logren predecir si un cliente realizará o no **churn** en el último mes del dataset. Explica por qué elegiste tal modelo.
    a. ¿Qué métricas usaste para evaluar el desempeño del modelo? ¿Por qué?
    b. ¿Cómo podrías mejorar la performance del modelo?
    c. ¿Por qué elegiste este modelo?

No adjuntar respuestas de esta sección en el archivo respuestas.txt