Prueba Técnica Data Science Engineer

El presente documento describe la prueba técnica que debe ser realizada por los candidatos para cargo de DSE en el equipo de Machine Learning de PayU. La prueba consiste en la realización de un proceso de entrenamiento de modelos de clasificación desde la preparación de sus datos hasta la evaluación de los mismos.
La ponderación de la evaluación de esta prueba es de 80% para el proceso y 20% para los resultados obtenidos en los modelos entregados por el candidato, por lo que se sugiere prestar especial atención a la presentación del informe.
Las actividades descritas deben ser realizadas usando los lenguajes de programación *R* o *Python* y la entrega debe hacerse usando *RMarkdown*, *Jupyter Notebook* o similar.

La prueba consiste de 5 actividades: *Preparación de los datos, Análisis Exploratorio, Feature Engineering, Entrenamiento de modelos y Análisis de resultados*.

DATA SET:

En el archivo ***dataset.zip*** se encuentran 9 archivos en formato *CSV* con información de transacciones procesadas por PayU y realizadas entre octubre 2016 y mayo 2017 por un negocio dedicado al transporte de pasajeros.

Existen dos tipos de archivos en el *dataset*; los archivos de **Transacciones** tienen nombre *transaction[añomes].csv* donde *[añomes]* es un indicador del mes en el que fueron realizadas las transacciones. Por ejemplo, el archivo *transaction201610.csv* contiene las transacciones correspondientes al mes de octubre de 2016. El archivo de **Listas** tiene nombre *lists201610201705.csv* y contiene información sobre listas negras y listas blancas relacionadas con las transacciones.

Los archivos de transacciones mensuales contienen la siguiente información:

- **creation_date (*datetime*):** La fecha y hora en la que se realizó la transacción.
- **email (*string*)** : El correo electrónico del comprador
- **chargeback (*boolean*):** true si la transacción fue marcada como *contracargo**.
- **status (*categorical, integer*)**: Denota el estado que la transacción tiene dentro del sistema
    - *4* : La transacción fue rechazada por el procesador de pagos
    - *6* : La transacción fue rechazada por una entidad bancaria
    - *8* : La transacción fue aprobada por un banco
    - *11* : La transacción resultó en un fraude confirmado
    - *12* : La transacción fue identificada como un intento de fraude
- **card (string):** Un hash que representa la tarjeta de crédito
- **uuid (integer):** Identificador de la transacción
- **amount:** Monto de la transacción

**Un contracargo implica que un comprador realizó un proceso de desconocimiento de compra. En términos prácticos, para este ejercicio se puede tomar como un indicador de fraude confirmado.*

El archivo de evaluaciones contiene:

- **eval_uuid (integer):** Id de la evaluación
- **uuid (integer):** El id de la transacción a la cual corresponde esta evaluación
- **blacklist (boolean):** true si la transacción ha sido incluida en una lista negra
- **whitelist (boolean):** true si la transacción ha sido incluida en una lista blanca

I – ANÁLISIS EXPLORATORIO

Haga un análisis exploratorio de los datos y reporte lo que considere importante.

II - PREPARACIÓN DE LOS DATOS

Con el fin de generar modelos que ayuden a determinar si una nueva transacción debe ser rechazada o no, es necesario etiquetar las transacciones en el dataset como legítimas (legit) o fraudulentas (fraud). Las normas para asignar estas etiquetas son las siguientes:

1 - Una transacción se considera fraudulenta si cualquiera de las siguientes condiciones se cumple:

- Su estado es 11 o 12
- Ha sido marcada como contracargo
- Ha sido incluida en una lista negra
- El email o la tarjeta ha sido usado en una transacción fraudulenta. Este proceso se conoce como expansión de fraudes y se debe continuar haciendo mientras sea posible.

2 - Las transacciones con estado 8 que no fueron marcadas por la norma 1 como fraudulentas, son consideradas legítimas.

3 - Todas las transacciones que no hayan sido tenidas en cuenta por las normas 1 y 2 deben ser eliminadas del *dataset*.

Usando las normas descritas, modifique el dataset para generar el conjunto de entrenamiento que será usado para construir modelos de clasificación. Aplique cualquier otra transformación que considere necesaria.

III - FEATURE ENGINEERING

Agregue las características que considere que pueden ser útiles en la generación de modelos de clasificación para identificar transacciones fraudulentas. Adicionalmente, agregue:

- Al menos dos características que analicen el texto del correo electrónico (pueden ser características simples como la longitud de la cadena, o el número de caracteres numéricos)
- Una característica que contenga el número de veces que el email de la transacción ha sido visto en el pasado (antes de su fecha de creación).

IV – ENTRENAMIENTO DE MODELOS

Elija dos modelos de clasificación (los que desee) y entrénelos con los datos que preparó. Describa el proceso que llevó a cabo. ***Ej***: ¿Cómo seleccionó los parámetros de configuración? ¿Cómo separó el conjunto de entrenamiento?

V - ANÁLISIS DE LOS RESULTADOS

Presente las métricas de desempeño que considere necesarias y haga un análisis de los resultados que obtuvo con ambos modelos.

- ¿Cuáles fueron las variables más importantes para los clasificadores que construyó?
- ¿Tuvo limitaciones para la creación de los modelos? ¿Cuáles?
- ¿Qué mejoraría en los datos para obtener resultados superiores?
- ¿Cómo pondría en operación los modelos que construyó? ¿Cómo seleccionaría los umbrales de decisión para clasificar una transacción nueva como legítima o fraudulenta?