2do Examen Parcial

Rolando Oviedo Quezada

July 30, 2022

1 Introducción

El examen constará de una única fase, la cuál tiene como objetivo retarlos y evaluar el conocimiento adquirido en el módulo, con datos que tengan una distribución parecida a lo que seguramente encontrarán en el mundo real. De esta forma, esta evaluación final trata de ponerlos a prueba, de llevarlos a esa situación de trabajo como Científico de Datos, con un desbalanceo radical, datos grandes, y varias fuentes de datos.

1.1 Instrucciones:

1. El examen se debe realizar de forma individual; en caso de que los resultados y el código tenga una relación de desarrollo y lógica igual para dos o más alumnos, el examen les será anulado.
2. Cuentan con un total de 1 semana para realizarlo, por lo que la entrega será el próximo sábado 5 de Agosto.

2 Evaluación Práctica

Usted es científico de Datos de una institución Financiera, que cuenta con un universo total de 2,000 clientes. Los cuáles pueden no tener producto activo, tener uno o tener varios. Debido al avance tecnológico, se han presentado diversos fraudes, esto es un conflicto para la institución Financiera, pues representa una pérdida de dinero, ya que la institución absorbe la deuda, y una insatisfacción para el cliente, debido a que este tiene una mala experiencia.

Se le solicita al Científico de Datos, proponer una solución al problema, como tal el objetivo puntual es: lograr la identificación efectiva de las transacciones fraudulentas. El cómo, dependerá de la solución que usted, como científico de datos, diseñe. Como esto es un proyecto de alta urgencia, se solicita que dicha solución sea propuesta en una semana, dónde esperaremos contar con lo siguiente como entrega:

1. Presentación en formato pdf con el proceso, y resultados.
2. Diagrama de la solución (puede ser dibujado en papel, y subir foto) El diagrama deberá explicar el funcionamiento de la solución, y porque diseñó ese flujo.
3. Código de la solución.
4. Modelo o Modelos generados.

Adicionalmente, a esta solicitud, se incluye un diccionario de datos, de forma que el científico conozca las fuentes de datos a las cuáles tiene acceso para poder trabajar en su proceso.

2.1 Tabla de Clientes

En esta tabla encontrará información de todos los clientes (2,000), de los cuáles, cuenta con 18 características enumeradas y descritas a continuación:

1. Person: Nombre del cliente.
2. Curren Age: Edad actual del cliente.
3. Retirement Age: Edad a la que se retiró o se retirará.
4. Birth Year: Año de Nacimiento del cliente.
5. Birth Month: Mes de Nacimiento del cliente.
6. Gender: Género del cliente.
7. Address: Dirección del cliente.
8. Apartment: Número de departamento del cliente (en caso de tener).
9. City: Ciudad de residencia del cliente.
10. State: Estado de residencia del cliente.
11. Zipcode: Código postal de la residencia del cliente.
12. Latitude: Latitud de la casa del cliente.
13. Longitud: Longitud de la casa del cliente.
14. Per Capita Income - Zipcode: Ingreso anual per cápita dado el código postal.
15. Yearly Income - Person: Ingreso Anual del cliente.
16. Total Debt: Deuda total del cliente.
17. FICO Score: Score crediticio del cliente.
18. Num Credit Cards: Número total de tarjetas asociadas al cliente (con la institución).

La llave primaria es el índice de la tabla, pues ese es el que se hereda como número de cliente.

2.2 Tabla Tarjetas

En esta tabla puede acceder a información a nivel producto, de forma que pueda identificar el comportamiento de los distintos tipos de productos. Se cuenta con un total de 6,146 tarjetas, y 13 características para cada una de ellas, las cuales son, las siguientes:

1. User: Número de cliente, es el cliente asociado a la tarjeta.
2. CARD INDEX: Número de la tarjeta del cliente.
3. Card Brand: Marca de la tarjeta.
4. Card Type: Tipo de la tarjeta.
5. Card Number: Número de la tarjeta.
6. Expires: Fecha de expiración de la tarjeta (formato(mes/año).
7. CVV: Código de seguridad de la tarjeta.
8. Has Chip: Variable indicadora para saber si la tarjeta tiene o no tiene chip.
9. Cards Issued: Número de tarjetas asignadas (o repuestas).

10. Credit Limit: Límite de crédito de la tarjeta.
11. Acct Open Date: Fecha de activación.
12. Year PIN las Changed: Año del último cambo de PIN.
13. Card on Dark Web: Indicadora, apoya a identificar si los datos de la tarjeta han sido filtrados a la Dark-Web

Dicha tabla tiene varias llaves, dependiendo de la tabla con la que la cruzarán será la que deberán utilizar.

3 Tabla Transacciones

En esta tabla es dónde se encuentra y vive la variable objetivo. Se cuenta con un total de 24,386,900 de transacciones y, con 15 características para cada una de ellas las cuáles son las siguientes:

1. User: Código de usuario, es decir número de cliente.
2. Card: Número de tarjeta del usuario.
3. Year: Año de la transacción.
4. Day: Día de la transacción.
5. Time: Hora de la transacción.
6. Amount: Total de la transacción.
7. Use Chip: Tipo de transacción.
8. Mechant Name: Nombre codificado del comercio, institución, particular, etc... en el que la transacción fue realizada.
9. Merchant City: Ciudad en la que se encuentra el comercio.
10. Mechant State: Estado en el que se encuentra el comercio.
11. Zip: Código postal del comercio.
12. MCC: Código de categoría del comercio (Merchant Category Code).
13. Errors?: Categoría indicando si hubo o no hubo error. En caso de haber error adicionalmente, especifica el error ocurrido.
14. Is Fraud?: Variable objetivo, bandera indicando si la transacción fue fraudulenta.

Como puede apreciar, tiene acceso a bastante información y que parece relevante, por lo que esperamos le sea posible plantear la solución adecuada para el presente problema de negocio. Recuérde utilizar todos sus conocimientos teóricos, para que pueda crear un modelo efectivo.

Como este es un trabajo para usted que es el experto en Ciencia de Datos, no hay más indicaciones, esperemos los datos le sean de utilidad, para poder trabajar en una solución para el problema que nos agobia, hoy en día.

4 Entrega:

La entrega se realizará vía classroom como de costumbre, dónde subirán uno por uno, es decir no zips, y tendrán los siguientes nombres. Archivo que no cumpla con el nombre, o que venga en zip, archivo que no será considerado.

- Presentación: OviedoQuezadaRolando_presentacion.pdf.
- Código: OviedoQuezadaRolando_Exam2code.ipynb
- Diagrama: OviedoQuezadaRolando_Diagrama.jpg ó OviedoQuezadaRolando_Diagrama.png
- Modelo(s): OviedoQuezadaRolando_NombreDelModelo.sav ó OviedoQuezadaRolando_NombreDelModelo.pkl ó OviedoQuezadaRolando_NombreDelModelo.joblib según sea el caso.
- Feedback: OviedoQuezadaRolando_feedback.doc ó OviedoQuezadaRolando_feedback.txt Documento de texto que contenga su feedback respecto al Módulo, con la intención de seguir mejorando. Recuérden que toda crítica puede ser tomada como algo constructivo, entonces pueden contarme lo que les gustó, y lo que no les gustó del módulo, así como los cambios sugeridos.

Nota: es importante respetar el nombre de los archivos y darle el orden como está establecido, ya que de esa forma será como su calificación será obtenida a partir de un proceso automático, es por esto, que resulta importante separar su apellido paterno, apellido materno y nombre con mayúscula. El nombre deberá llevar caracteres identificados por el idioma Inglés, esto porque durante la lectura algunos caracteres especiales como acentos, diéresis, ñ, le generan conflicto a python para la lectura de los archivos. Por lo que se les solicita de la forma más atenta omitirlos en el nombre de sus archivos.

5 Entrega

La entrega de las salidas indicadas en el apartado práctico se realizará en el classroom. Es importante subir cada archivo por separado.

Recuerden que tienen suficiente tiempo y la intención es que tengan esa experiencia como científico de datos, ya en un ambiente similar al real.

Les deseo mucho éxito, y espero se diviertan con el examen.