Prueba Técnica / Data Scientist (Risk & Antifraud Team)

El reto

Como sabes, Conekta es una plataforma de pagos electrónicos que procesa miles de transacciones al día de diversos tipos de negocios y servicios. Algunas de estas transacciones son fraudulentas y resultan en costos para los usuarios de Conekta o para nosotros mismos. Conekta busca maximizar la seguridad de los usuarios de su plataforma al reducir las transacciones fraudulentas por medio del trabajo conjunto de analistas humanos y modelos estadísticos y de machine learning.

Tu misión será desarrollar un modelo que clasifique cada una de las transacciones en el dataset que te compartimos como fraudulenta o no fraudulenta, y, de ser posible, proporcionar la probabilidad de que cada transacciones pertenezca a una u otra clase. Esto nos ayudara a automatizar nuestro sistema antifraude y optimizar el tiempo de nuestros analistas antifraude para que se enfoquen en las transacciones más fraudulentas.

El dataset y el stack

Te pedimos que utilices un lenguaje de programación como Python, Julia, Java o C. Siéntete libre de utilizar el lenguaje que prefieras, pero busca desarrollar un código legible y bien comentando para poder entender mejor la implementación de tu modelo.

El dataset que te compartimos corresponde a 2 meses de transacciones de tarjetas de crédito en la plataforma de Conekta en formato JSON. Toma en cuenta que algunos de los campos de este dataset, como la IP, el teléfono, correo electrónico, y tarjeta asociada a cada transacción están tokenizados, y no corresponden a los datos originales (aunque son tokens únicos). El campo BIN corresponde al Bank Identification Number, una clave que se utiliza para identificar el banco al que pertenece cada tarjeta bancaria. El campo *amount* corresponde mal monto por transacción en MXN en intervalos de 500 en 500 MXN.

Finalmente, el campo *status* se refiere a la clasificación que recibió cada transacción en nuestro sistema de antifraude. En este caso, te recomendamos que consideres las transacciones con status charged_back y antifraud_declined como transacciones fraudulentas, y aquellas con status paid como no fraudulentas. En el caso de las transacciones con status refunded puedes ignorarlas, pero también puedes proponer utilizarlas como transacciones fraudulentas o no fraudulentas para entrenar tu modelo, siempre y cuando justifiques tu decisión.

Lo que esperamos

No hay una única solución correcta para esto reto, así es que siéntete libre de proponer la solución que consideres mejor, pero toma en cuenta que utilizaremos los siguientes aspectos para evaluar tu desempeño:

- Capacidad para identificar de un modelo estadístico o de machine learning apropiado para el problema de clasificación al que te enfrentas, y capacidad de justificar tu elección.
- Capacidad para utilizar de forma útil y creativa las variables con las que cuentas, ya sean categóricos o continuos, y de explicar por qué los incluyes en el modelo.
- Implementación del modelo que escogiste en un lenguaje de programación como Python, Julia, Java o C.
- Capacidad para diseñar estrategias de evaluación del modelo (por ejemplo, cross validation o K-fold) y métricas para evaluar su desempeño (i.e., precision, recall, F1 score, etc.); identificar los tradeoffs en estas métricas entre distintos modelos de clasificación, y de justificar la elección de estas métricas

Te recordamos que el entregable de este ejercicio será: 1) el script del modelo que utilices y que resulte en la clasificación de las transacciones del dataset que te compartimos, 2) instrucciones para ejecutar este script, y 3) acceso al repositorio donde se encuentran tu código. De nuevo, toma en cuenta que tienes 1 semana a partir de que recibas este correo para mandar tu solución.