**Conjuntos de Datos Sintéticos para la Detección de Fraude**

**Acerca del Conjunto de Datos**

**Contexto**

Existe una falta de conjuntos de datos disponibles públicamente sobre servicios financieros y, especialmente, en el emergente dominio de las transacciones de dinero móvil. Los conjuntos de datos financieros son importantes para muchos investigadores y, en particular, para nosotros que realizamos investigaciones en el ámbito de la detección de fraude. Parte del problema es la naturaleza intrínsecamente privada de las transacciones financieras, lo que conduce a la ausencia de conjuntos de datos disponibles públicamente.

Presentamos un conjunto de datos sintético generado mediante el simulador llamado PaySim como un enfoque para tal problema. PaySim utiliza datos agregados del conjunto de datos privado para generar un conjunto de datos sintético que se asemeja a la operación normal de las transacciones e inyecta comportamiento malicioso para luego evaluar el rendimiento de los métodos de detección de fraude.

**Contenido**

PaySim simula transacciones de dinero móvil basadas en una muestra de transacciones reales extraídas de un mes de registros financieros de un servicio de dinero móvil implementado en un país africano. Los registros originales fueron proporcionados por una empresa multinacional, que es el proveedor del servicio financiero móvil que actualmente opera en más de 14 países en todo el mundo.

Este conjunto de datos sintético está reducido a 1/4 del conjunto de datos original y se crea solo para Kaggle. NOTA: Las transacciones que se detectan como fraude se cancelan, por lo que para la detección de fraude estas columnas (oldbalanceOrg, newbalanceOrig, oldbalanceDest, newbalanceDest) no deben usarse.

**Encabezados**

Esta es una muestra de 1 fila con la explicación de los encabezados:

1, PAYMENT, 1060.31, C429214117, 1089.0, 28.69, M1591654462, 0.0, 0.0, 0, 0

* **step**: Mapea una unidad de tiempo en el mundo real. En este caso, 1 paso es 1 hora de tiempo. Total de pasos 744 (simulación de 30 días).
* **type**: CASH-IN, CASH-OUT, DEBIT, PAYMENT y TRANSFER.
* **amount**: Cantidad de la transacción en moneda local.
* **nameOrig**: Cliente que inició la transacción.
* **oldbalanceOrg**: Saldo inicial antes de la transacción.
* **newbalanceOrig**: Nuevo saldo después de la transacción.
* **nameDest**: Cliente que es el destinatario de la transacción.
* **oldbalanceDest**: Saldo inicial del destinatario antes de la transacción. Nota: no hay información para clientes que comienzan con M (Comerciantes).
* **newbalanceDest**: Nuevo saldo del destinatario después de la transacción. Nota: no hay información para clientes que comienzan con M (Comerciantes).
* **isFraud**: Estas son las transacciones realizadas por los agentes fraudulentos dentro de la simulación. En este conjunto de datos específico, el comportamiento fraudulento de los agentes tiene como objetivo obtener ganancias tomando el control de las cuentas de los clientes y tratando de vaciar los fondos transfiriéndolos a otra cuenta y luego retirándolos del sistema.
* **isFlaggedFraud**: El modelo de negocio tiene como objetivo controlar transferencias masivas de una cuenta a otra y marca intentos ilegales. Un intento ilegal en este conjunto de datos es un intento de transferir más de 200,000 en una sola transacción.

**Investigación Pasada**

Hay 5 archivos similares que contienen la ejecución de 5 escenarios diferentes. Estos archivos se explican mejor en el capítulo 7 de mi tesis doctoral (Tesis Doctoral disponible aquí: [Enlace a la tesis](http://urn.kb.se/resolve?urn=urn:nbn:se:bth-12932).

Ejecutamos PaySim varias veces utilizando semillas aleatorias durante 744 pasos, representando cada hora de un mes de tiempo real, lo que coincide con los registros originales. Cada ejecución tomó alrededor de 45 minutos en un procesador Intel i7 con 16GB de RAM. El resultado final de una ejecución contiene aproximadamente 24 millones de registros financieros divididos en las 5 categorías de tipos: CASH-IN, CASH-OUT, DEBIT, PAYMENT y TRANSFER.

**Agradecimientos**

Este trabajo es parte del proyecto de investigación “Sistemas escalables y eficientes en recursos para el análisis de grandes datos” financiado por la Knowledge Foundation (subvención: 20140032) en Suecia.

Por favor, referirse a este conjunto de datos utilizando las siguientes citas:

Primer artículo de PaySim del simulador: E. A. Lopez-Rojas, A. Elmir, y S. Axelsson. "PaySim: Un simulador de dinero móvil financiero para la detección de fraude". En: El 28º Simposio Europeo de Modelado y Simulación-EMSS, Larnaca, Chipre. 2016

Con base en la información proporcionada y el objetivo de detectar fraudes en transacciones financieras, algunas de las columnas pueden ser menos relevantes o redundantes. Sin embargo, es importante considerar todas las posibles relaciones antes de descartar columnas. Aquí hay un análisis rápido de cada columna:

1. **step**: Representa el tiempo en horas. Puede ser útil para analizar patrones temporales en los fraudes.
2. **type**: Es crucial porque diferentes tipos de transacciones pueden tener diferentes probabilidades de ser fraudulentas.
3. **amount**: Es esencial, ya que el monto de la transacción puede influir en la probabilidad de fraude.
4. **nameOrig**: Identificador del cliente que inicia la transacción. Puede ser útil para identificar patrones de comportamiento fraudulento repetitivo. Sin embargo, podría ser codificado de manera diferente o anonimizado.
5. **oldbalanceOrg**: Saldo inicial del cliente antes de la transacción. Puede ser útil para analizar si los fraudes están relacionados con ciertos niveles de saldo.
6. **newbalanceOrig**: Saldo del cliente después de la transacción. Similar al anterior, puede ser útil.
7. **nameDest**: Identificador del cliente que recibe la transacción. Similar a nameOrig, puede ser útil pero podría ser anonimizado.
8. **oldbalanceDest**: Saldo inicial del destinatario antes de la transacción. Útil para analizar patrones de fraude en destinatarios.
9. **newbalanceDest**: Saldo del destinatario después de la transacción. Útil para analizar patrones de fraude en destinatarios.
10. **isFraud**: Variable objetivo. No se debe descartar.
11. **isFlaggedFraud**: Indica si la transacción ha sido señalada como potencialmente fraudulenta. Útil para ver el rendimiento de los sistemas de detección actuales.