Grupo 27

Entrega 1

Jhon Farley Adarve Diaz	j.adarve@uniandes.edu.co
Julian Alberto Gomez Gomez	j.gomez24@uniandes.edu.co
Giovvany Andres Cardona Rojo	g.cardonar@uniandes.edu.co

Contenido

Identificación de transacciones financieras fraudulentas mediante técnicas de machine learning

Problema

En nuestro proyecto, abordamos uno de los desafíos más críticos que enfrentan las instituciones financieras en la actualidad: la detección de fraude en transacciones bancarias. A medida que los sistemas financieros se digitalizan cada vez más, los métodos fraudulentos evolucionan en sofisticación, lo que hace necesario el desarrollo de sistemas de detección más robustos y adaptables.

Nos enfrentamos a un problema complejo caracterizado por:

- Datos altamente desbalanceados, donde las transacciones fraudulentas son eventos raros
- Necesidad de procesamiento en tiempo real o casi real
- Requerimiento de alta precisión para minimizar falsos positivos que podrían afectar a clientes legítimos
- Necesidad de adaptación continua a nuevos patrones de fraude

Pregunta de negocio y alcance del proyecto.

Nuestra pregunta de negocio es: ¿Cómo podemos desarrollar un sistema de detección de fraude que identifique eficazmente transacciones fraudulentas?

Nuestra pregunta de analítica será: ¿Cuáles son los modelos de machine learning supervisados más efectivos para predecir transacciones fraudulentas, y qué variables (monto de la transacción, ubicación, categoría, etc.) son los predictores más importantes en la detección de fraude?

El alcance de nuestro proyecto incluye:

- Desarrollo de modelos de machine learning para clasificación de transacciones
- Implementación de técnicas de análisis de datos para identificación de patrones
- Evaluación de rendimiento usando métricas específicas para datos desbalanceados
- Validación del modelo en un contexto simulado de transacciones bancarias
- Visualización mediante un tablero de control

Descripción de conjuntos de datos

Trabajamos con datos generados por BankSim, un simulador de pagos bancarios basado en datos transaccionales agregados de un banco español. Este conjunto de datos presenta características únicas que lo hacen especialmente valioso para nuestro proyecto:

- **Volumen de datos**: 594,643 registros totales
  - 587,443 transacciones normales
  - 7,200 transacciones fraudulentas (aproximadamente 1.2% del total)
- **Periodo de simulación**: 180 ejecuciones en BankSim (equivalente a aproximadamente 6 meses)
- **Características del fraude simulado**:
  - Simulación de ladrones que intentan robar un promedio de tres tarjetas por ejecución
  - Aproximadamente dos transacciones fraudulentas por día por tarjeta comprometida

La ventaja de utilizar estos datos simulados es que nos permiten trabajar con un conjunto que mantiene las características estadísticas de datos reales, pero sin comprometer información personal o confidencial, cumpliendo así con las regulaciones de privacidad mientras mantenemos la relevancia para el problema en cuestión.

Exploración de los datos.

Estructura del Conjunto de Datos

En nuestro análisis, trabajamos con un conjunto de datos que consta de 594,643 registros con 9 variables predictoras y 1 variable objetivo. Las características principales incluyen:

Variables Temporales y de Identificación

- **Step**: Variable numérica que representa el día desde el inicio de la simulación, abarcando un periodo de 180 días (aproximadamente 6 meses). Permite rastrear la evolución temporal de las transacciones.
- **Customer**: Identificador único para cada cliente que realiza transacciones en el sistema. Esta variable permite seguir el historial de transacciones por cliente.

Variables Geográficas

- **zipCodeOrigin**: Código postal asociado al origen de la transacción. Útil para identificar patrones geográficos en el comportamiento de los usuarios.
- **zipMerchant**: Código postal del comerciante donde se realiza la transacción. Permite analizar la distribución espacial de las actividades comerciales y posibles patrones de fraude por zona.

Variables Demográficas

- **Age**: Categorización de la edad del cliente en 8 grupos distintos:
  - 0: Menores o iguales a 18 años

  - 1: Entre 19 y 25 años
  - 2: Entre 26 y 35 años
  - 3: Entre 36 y 45 años
  - 4: Entre 46 y 55 años
  - 5: Entre 56 y 65 años
  - 6: Mayores de 65 años
  - U: Edad desconocida
- **Gender**: Género del cliente o tipo de cuenta, clasificado en 4 categorías:
  - E: Empresa (Enterprise)
  - F: Femenino
  - M: Masculino
  - U: Desconocido

Variables Transaccionales

- **Merchant**: Identificador único del comerciante que procesa la transacción. Permite analizar patrones de comportamiento por establecimiento.
- **Category**: Clasificación del tipo de compra o servicio adquirido. Esta variable categoriza las transacciones según el sector comercial.
- **Amount**: Monto de la transacción. Variable numérica que representa el valor monetario de la operación.

Variable Objetivo

- **Fraud**: Variable binaria que indica si la transacción es fraudulenta (1) o legítima (0). Esta es nuestra variable objetivo para el modelo de detección de fraude.

Hallazgos Clave

Calidad de los Datos

Un aspecto positivo destacable es la ausencia de valores faltantes en todas las variables, lo que eliminó la necesidad de realizar imputaciones y simplificó nuestro proceso de preparación de datos.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 594643 entries, 0 to 594642
Data columns (total 10 columns):
step           594643 non-null int64
customer       594643 non-null object
age            594643 non-null object
gender         594643 non-null object
zipcodeOri     594643 non-null object
merchant       594643 non-null object
zipMerchant    594643 non-null object
category       594643 non-null object
amount         594643 non-null float64
fraud          594643 non-null int64
dtypes: float64(1), int64(2), object(7)
memory usage: 45.4+ MB
```

Patrones de Fraude por Categoría

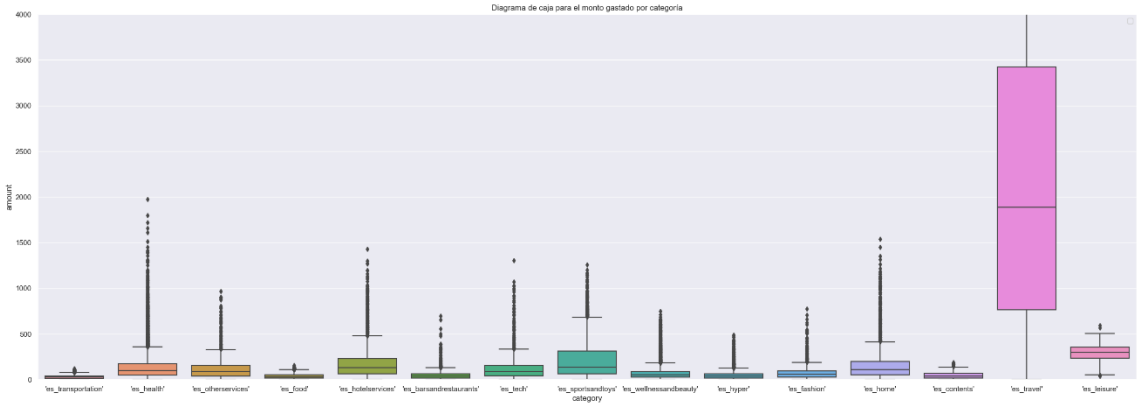
Identificamos patrones significativos en las transacciones fraudulentas:

1. **Categorías más vulnerables**:

Valores medios de las características por categoría	amount	fraud
category		
'es_barsandrestaurants'	43.461014	0.018829
'es_contents'	44.547571	0.000000
'es_fashion'	65.666642	0.017973
'es_food'	37.070405	0.000000
'es_health'	135.621367	0.105126
'es_home'	165.670846	0.152064
'es_hotelservices'	205.614249	0.314220
'es_hyper'	45.970421	0.045917
'es_leisure'	288.911303	0.949900
'es_otherservices'	135.881524	0.250000
'es_sportsandtoys'	215.715280	0.495252
'es_tech'	120.947937	0.066667
'es_transportation'	26.958187	0.000000
'es_travel'	2250.409190	0.793956
'es_wellnessandbeauty'	65.511221	0.047594

   a. Ocio/Leisure (94.99% de fraude)
   b. Viajes/Travel (79.39% de fraude)
   c. Deportes y juguetes (49.52% de fraude)
   d. Servicios hoteleros (31.42% de fraude)
2. **Montos de transacciones**:

   a. Las transacciones fraudulentas tienden a ser aproximadamente 4 veces mayores que las legítimas en la misma categoría
   b. La categoría "travel" destaca con montos significativamente más altos:



      i. Transacciones fraudulentas: ~2,660€
      ii. Transacciones legítimas: ~669€

Patrones Demográficos

En cuanto a la distribución por edad, encontramos un hallazgo interesante:

	Age	Fraud Percent
7	'U'	0.594228
6	'6'	0.974826
5	'5'	1.095112
1	'1'	1.185254
3	'3'	1.192815
2	'2'	1.251401
4	'4'	1.293281
0	'0'	1.957586

- La mayor incidencia de fraude se presenta en la categoría de edad 0 (≤18 años) con un 1.96%
- Existe un patrón ascendente en el porcentaje de fraude conforme la edad disminuye
- Las edades desconocidas ('U') muestran el menor porcentaje de fraude (0.59%)

Implicaciones para el Modelado

Estos hallazgos sugieren varios puntos importantes a considerar en nuestro modelo de detección:

1. La necesidad de prestar especial atención a transacciones en categorías de alto riesgo como viajes y ocio
2. La importancia de considerar la relación entre el monto de la transacción y el promedio histórico de la categoría
3. La relevancia de la edad como factor de riesgo, especialmente en transacciones asociadas a usuarios jóvenes

Estos insights serán fundamentales para el desarrollo de características (feature engineering) y la selección de nuestro modelo de detección de fraude.

Maqueta del prototipo

Descripción del tablero

Nuestro tablero de control integra visualizaciones dinámicas e interactivas diseñadas específicamente para el análisis y monitoreo de transacciones fraudulentas en tiempo real. El dashboard combina mapas de calor geográficos para identificar zonas de alto riesgo, gráficos de dispersión que relacionan montos y probabilidades de fraude, y series temporales que revelan patrones emergentes en diferentes escalas de tiempo. Estas visualizaciones se complementan con métricas clave de rendimiento del modelo (recall, precisión, F1-score) y análisis demográficos detallados a través de gráficos de distribución. La interfaz permite a los usuarios realizar análisis drill-down mediante filtros interactivos por fecha, categoría de transacción, ubicación y segmento demográfico, facilitando la identificación rápida de anomalías y patrones sospechosos.

El sistema de alertas automáticas, junto con los reportes periódicos, permitirá a los equipos de seguridad y análisis:

- Detectar proactivamente nuevos patrones de fraude
- Evaluar la efectividad de las estrategias de prevención
- Ajustar los parámetros del modelo en tiempo real
- Tomar decisiones informadas basadas en tendencias históricas y emergentes
- Optimizar la asignación de recursos para la investigación de casos sospechosos

Este enfoque integral no solo mejora la capacidad de detección de fraude, sino que también proporciona insights accionables para fortalecer las medidas preventivas y reducir las pérdidas financieras asociadas con actividades fraudulentas.

Tablero de Detección de Fraude en Transacciones Financieras

Métricas Clave

Total de transacciones: 594643

Total de fraudes detectados: 7200

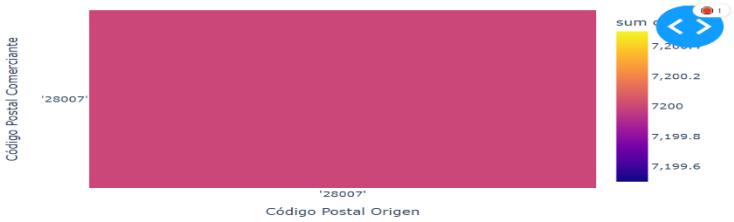
Tasa de detección de fraude (Recall): 1.21% (Meta: >90%)

Precisión en alertas: 88.09% (Meta: >85%)

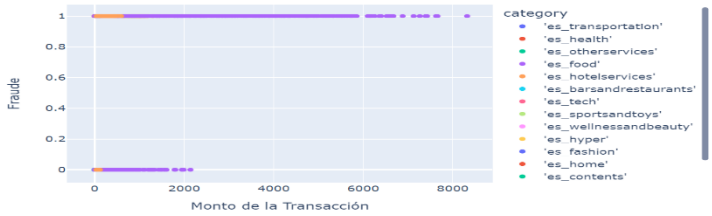
F1-Score: 89.26% (Meta: >87%)

Tiempo promedio de detección: 5.86 horas

Mapa de Calor de Fraudes por Código Postal

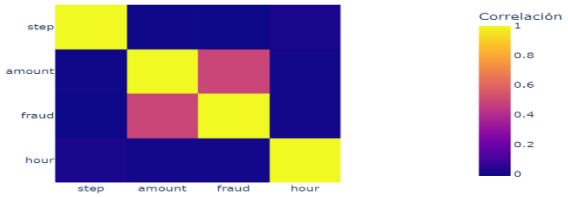


Monto vs. Probabilidad de Fraude

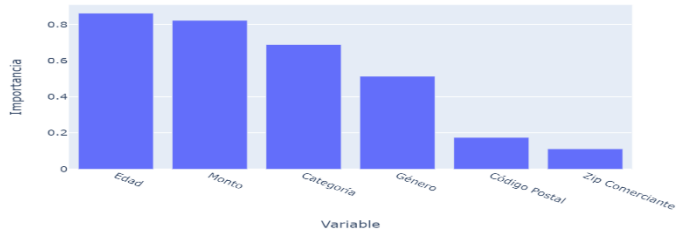


Tendencia de Fraude por Hora del Día

Matriz de Correlación entre Variables



Importancia de Variables Predictivas



Filtros Interactivos

Reportes Automáticos

Reportes automáticos diarios, semanales y mensuales estarán disponibles según los filtros aplicados.

Alertas y Umbrales

Visualizaciones del tablero

1. Métricas Clave (KPIs)

Indicadores de Rendimiento del Modelo

- **Tasa de Detección de Fraude** (Recall)
  - Porcentaje de fraudes reales detectados correctamente
  - Tendencia histórica por período
  - Meta: >90%
- **Precisión en Alertas**
  - Porcentaje de alertas que son realmente fraude
  - Desglose por categoría de transacción
  - Meta: >85%
- **F1-Score**
  - Equilibrio entre precisión y recall
  - Evolución temporal
  - Meta: >87%

Indicadores Operativos

- **Tiempo Promedio de Detección**
  - Desde la transacción hasta la alerta
  - Por tipo de fraude y categoría
- **Volumen de Alertas**
  - Total diario/semanal/mensual
  - Desglose por nivel de riesgo

2. Visualizaciones Principales

Análisis de Patrones de Fraude

1. **Mapa de Calor Geográfico**
   a. Distribución de fraudes por código postal
   b. Correlación entre ubicación origen-comerciante
2. **Gráfico de Dispersión Interactivo**
   a. Monto vs. Probabilidad de fraude
   b. Filtros por categoría y edad
3. **Series Temporales**
   a. Tendencias de fraude por hora/día/mes
   b. Comparativa con transacciones normales

Análisis de Variables Predictivas

1. **Gráfico de Importancia de Variables**
   a. Ranking de variables más predictivas
   b. Actualización periódica según reentrenamiento
2. **Matriz de Correlación**
   a. Relaciones entre variables

   b. Patrones emergentes

Segmentación y Perfiles

1. **Gráficos de Distribución**
   a. Por edad y género
   b. Por categoría de comercio
   c. Por monto de transacción
2. **Análisis de Comportamiento**
   a. Patrones de transacción por perfil
   b. Desviaciones del comportamiento normal

3. Filtros Interactivos

- Rango de fechas
- Categoría de transacción
- Rango de montos
- Segmento demográfico
- Ubicación geográfica

4. Alertas y Umbrales

- Configuración de umbrales por:
  - Monto de transacción
  - Frecuencia de transacciones
  - Score de riesgo
  - Desviación del comportamiento normal
- Sistema de notificaciones para:
  - Patrones anómalos emergentes
  - Degradación del rendimiento del modelo
  - Concentración inusual de alertas

Repositorios y anexos

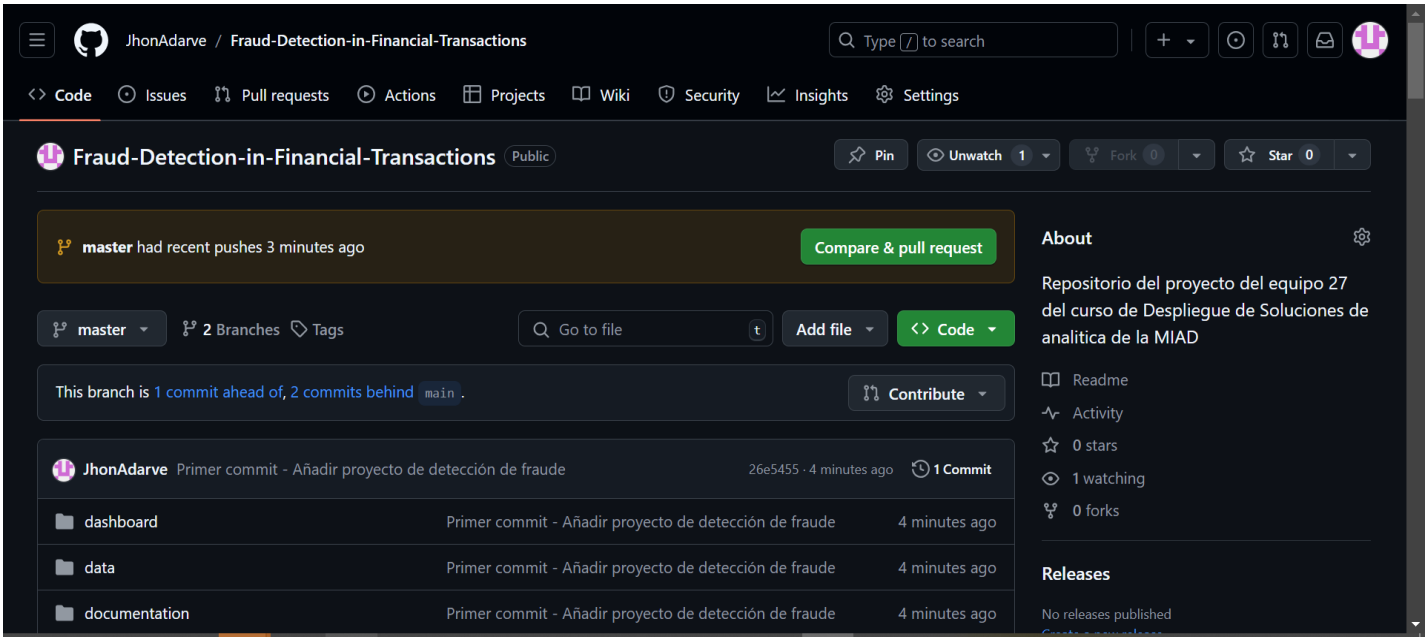
Enlace al boceto del tablero:

https://github.com/JhonAdarve/Fraud-Detection-in-Financial-Transactions/blob/master/dashboard/RadarFraud_Dashboard.py

Enlace al repositorio GitHub:

https://github.com/JhonAdarve/Fraud-Detection-in-Financial-Transactions/tree/master

```
MINGW64:/c/Users/jadarve/OneDrive - Grupo Bancolombia/Bancolombia/MIAD/Despliegue de soluciones de analytics/Proyecto/fraud-detection
Initialized empty Git repository in C:/Users/jadarve/OneDrive - Grupo Bancolombi
a/Bancolombia/MIAD/Despliegue de soluciones de analytics/Proyecto/fraud-detectio
n/.git/

jadarve@pb0b0923063 MINGW64 ~/OneDrive - Grupo Bancolombia/Bancolombia/MIAD/Desp
liegue de soluciones de analytics/Proyecto/fraud-detection (master)
$ git add .
warning: LF will be replaced by CRLF in notebooks/EDA.ipynb.
The file will have its original line endings in your working directory

jadarve@pb0b0923063 MINGW64 ~/OneDrive - Grupo Bancolombia/Bancolombia/MIAD/Desp
liegue de soluciones de analytics/Proyecto/fraud-detection (master)
$ git commit -m "Primer commit - Añadir proyecto de detección de fraude"
[master (root-commit) 26e5455] Primer commit - Añadir proyecto de detección de f
raude
 6 files changed, 595873 insertions(+)
 create mode 100644 README.md
 create mode 100644 dashboard/RadarFraud_Dashboard.py
 create mode 100644 data/banksim.csv
 create mode 100644 documentation/PlanManejoDatos_Equipo_27.pdf
 create mode 100644 notebooks/EDA.ipynb
 create mode 100644 requirements.txt

jadarve@pb0b0923063 MINGW64 ~/OneDrive - Grupo Bancolombia/Bancolombia/MIAD/Desp
liegue de soluciones de analytics/Proyecto/fraud-detection (master)
$ git remote add origin https://github.com/JhonAdarve/Fraud-Detection-in-Financi
al-Transactions.git

jadarve@pb0b0923063 MINGW64 ~/OneDrive - Grupo Bancolombia/Bancolombia/MIAD/Desp
liegue de soluciones de analytics/Proyecto/fraud-detection (master)
$ git push -u origin master
Enumerating objects: 12, done.
Counting objects: 100% (12/12), done.
Delta compression using up to 8 threads
Compressing objects: 100% (9/9), done.
Writing objects: 100% (12/12), 7.14 MiB | 2.24 MiB/s, done.
Total 12 (delta 0), reused 0 (delta 0), pack-reused 0
remote:
remote: Create a pull request for 'master' on GitHub by visiting:
remote:      https://github.com/JhonAdarve/Fraud-Detection-in-Financial-Transact
ions/pull/new/master
remote:
To https://github.com/JhonAdarve/Fraud-Detection-in-Financial-Transactions.git
 * [new branch]      master -> master
branch 'master' set up to track 'origin/master'.

jadarve@pb0b0923063 MINGW64 ~/OneDrive - Grupo Bancolombia/Bancolombia/MIAD/Desp
liegue de soluciones de analytics/Proyecto/fraud-detection (master)
$
```



Reporte de trabajo en equipo

1. **Jhon Adarve**: lideró la formulación de la pregunta de negocio, el alcance del proyecto y la definición del problema. Además, coordinó la búsqueda del conjunto de datos, también se enfocó en el desarrollo técnico del repositorio de Git, donde gestionó el código fuente y la versión del proyecto.
2. **Julian Gomez**: fue responsable de la creación de la maqueta del prototipo, asegurándose de que cumpliera con los requisitos funcionales y visuales.
3. **Giovanny Cardona**: se encargó de la exploración detallada de los datos, utilizando Python para descubrir patrones relevantes en los conjuntos de datos empleados. Además, colaboró en la redacción de los informes técnicos y la documentación sobre el análisis y las visualizaciones clave.

Cada miembro del equipo trabajó en conjunto para cubrir los diferentes aspectos del proyecto, contribuyendo de manera efectiva al cumplimiento de los entregables.