



PREVENCIÓN Y DETECCIÓN DE FRAUDE

NOS ENCARGAMOS DE CUIDAR TANTO DE USTED COMO DE SUS CLIENTES

José F. Ramos

CODERHOUSE



En el inframundo de las transacciones comerciales, donde el fraude acecha, Cerbero Seguridad se erige como el guardián definitivo. Al igual que el mítico perro de Hades, nuestras tres cabezas de protección –detección, prevención y respuesta– trabajan sin descanso para mantener tu negocio seguro.

Utilizando machine learning avanzado, analizamos cada transacción comercial para identificar y prevenir actividades fraudulentas. Nuestra tecnología aprende y se adapta continuamente, mejorando su eficacia con cada interacción.

Con Cerbero Seguridad, tu empresa está siempre bajo la protección de un guardián fiel y poderoso, combinando vigilancia implacable con tecnología de vanguardia para proteger la integridad de tus operaciones financieras.



1.

DETECCIÓN DE FRAUDE

¿SE PUEDE DETECTAR FRAUDE EN LAS
TRANSACCIONES DE LOS CLIENTES?



Abstract

A lo largo de la historia de la civilización, el fraude ha jugado un papel crucial en cualquier transacción comercial o intercambio de mercancías. Abordar este problema de manera efectiva es esencial debido a su impacto directo en el éxito de ambas partes involucradas. En este contexto, el análisis de datos emerge como una herramienta fundamental para comprender, prevenir y mitigar el riesgo de fraude, proporcionando un enfoque informado y proactivo en la gestión de transacciones comerciales. Este análisis no solo busca salvaguardar la integridad de las operaciones, sino también contribuir a un entorno empresarial más seguro y confiable.

Este proyecto investiga un conjunto de datos con numerosos factores presentes durante las transacciones, y aprende cuáles son los más importantes para detectar y prevenir fraudes.



Conjuntos de Datos

El corpus de datos empleado en este análisis se obtuvo de Kaggle, específicamente como parte de uno de los desafíos planteados por la plataforma en el ámbito de la detección de fraudes. Es relevante resaltar que Kaggle atribuye la provisión de estos datos a Vesta Corporation, una destacada pionera en soluciones de pago garantizadas para el comercio electrónico.

Organización de los Datos

El conjunto de datos consta de dos tablas fundamentales. En primer lugar, contamos con la "Tabla de Transacciones", que ofrece información detallada sobre cada transacción realizada. En segundo lugar, encontramos la "Tabla de Identidad", la cual proporciona datos cruciales relacionados con la conexión de red, incluyendo información como la dirección IP, proveedor de servicios de Internet (ISP), el uso de proxy, entre otros.

Esta estructura dual del conjunto de datos permite una comprensión integral de las transacciones al incorporar tanto los detalles transaccionales como la identidad digital asociada, enriqueciendo así el análisis y la interpretación de la información.

[Enlace al conjunto de datos] (<https://www.kaggle.com/competitions/ieee-fraud-detection/overview>)



Definición de Temática

En nuestro objetivo de desarrollar un sistema de detección de fraude temprana con el fin de mejorar la experiencia tanto del cliente como de las empresas involucradas en las transacciones, es esencial identificar las características más relevantes que permitan distinguir entre transacciones normales y fraudulentas.

Algunas de las características clave que podrían ser prioritarias en este análisis incluyen:

1. Patrones de Comportamiento:

- Frecuencia y horarios típicos de transacciones del usuario.
- Cambios inusuales en el comportamiento de transacciones, como volúmenes anómalos o ubicaciones atípicas.

2. Características de la Identidad Digital:

- Anomalías en direcciones IP, cambios frecuentes de dispositivos o ubicaciones.
- Inconsistencias en la información de firma digital, como versiones de navegador o sistemas operativos inesperados.

3. Historial de Transacciones:

- Comparación de patrones de compra habituales con la transacción actual.
- Evaluación de la coherencia en el historial de transacciones del usuario.

4. Datos de Autenticación:

- Fallos recurrentes o intentos repetidos de autenticación.
- Detección de actividad sospechosa durante el proceso de inicio de sesión.

5. Características Financieras:

- Transacciones de grandes cantidades o patrones de gastos inusuales.
- Cambios abruptos en el comportamiento financiero del usuario.

6. Análisis de Redes:

- Relaciones inusuales entre diferentes cuentas o usuarios.
- Actividad sospechosa en redes de conexión o uso de proxies.

7. Características Temporales:

- Detección de transacciones rápidas o simultáneas desde ubicaciones geográficas distantes.

8. Modelos de Machine Learning:

- Utilización de algoritmos de aprendizaje automático para identificar patrones no lineales y complejas relaciones entre variables.



En la fase inicial de nuestro análisis, nos enfrentamos a una abrumadora cantidad de columnas. Con el objetivo de simplificar y focalizarnos en esta etapa inicial, buscaremos reducir el conjunto de columnas al máximo posible.

Características Iniciales para el Análisis de Primera Etapa

- TransactionDT: Representa un delta de tiempo desde un punto de referencia.
- TransactionAMT: Indica la cantidad en USD para el pago de la transacción.
- ProductCD: Código del producto.
- addr: Dirección.
- dist: Distancia.
- P_ and (R_): Dominio de correo electrónico del comprador y del destinatario.
- deviceType
- DeviceInfo

Las siguientes características no se incluirán en esta fase inicial:

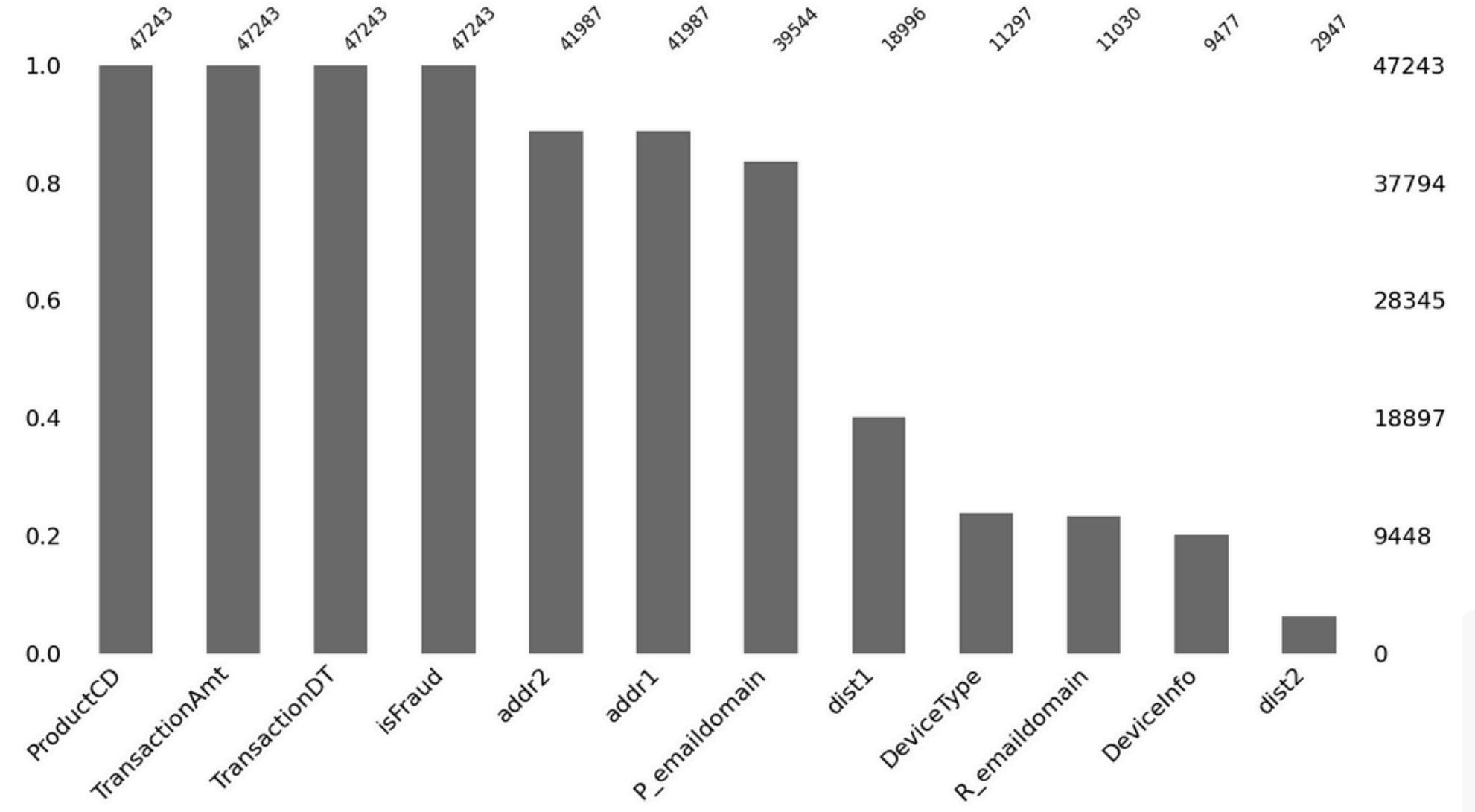
- C1-C14: Recuento, como la cantidad de direcciones asociadas a la tarjeta de pago, entre otros. El significado real está enmascarado.
- D1-D15: Timedelta, como los días entre la transacción anterior, etc.
- M1-M9: Coincidencias, como los nombres en la tarjeta y la dirección, etc.
- Vxxx: Características ricas desarrolladas por Vesta, que incluyen ranking, conteo y relaciones con otras entidades.
- id_12 - id_38



2. EDA ANÁLISIS EXPLORATORIO DE DATOS



Valores nulos y duplicados

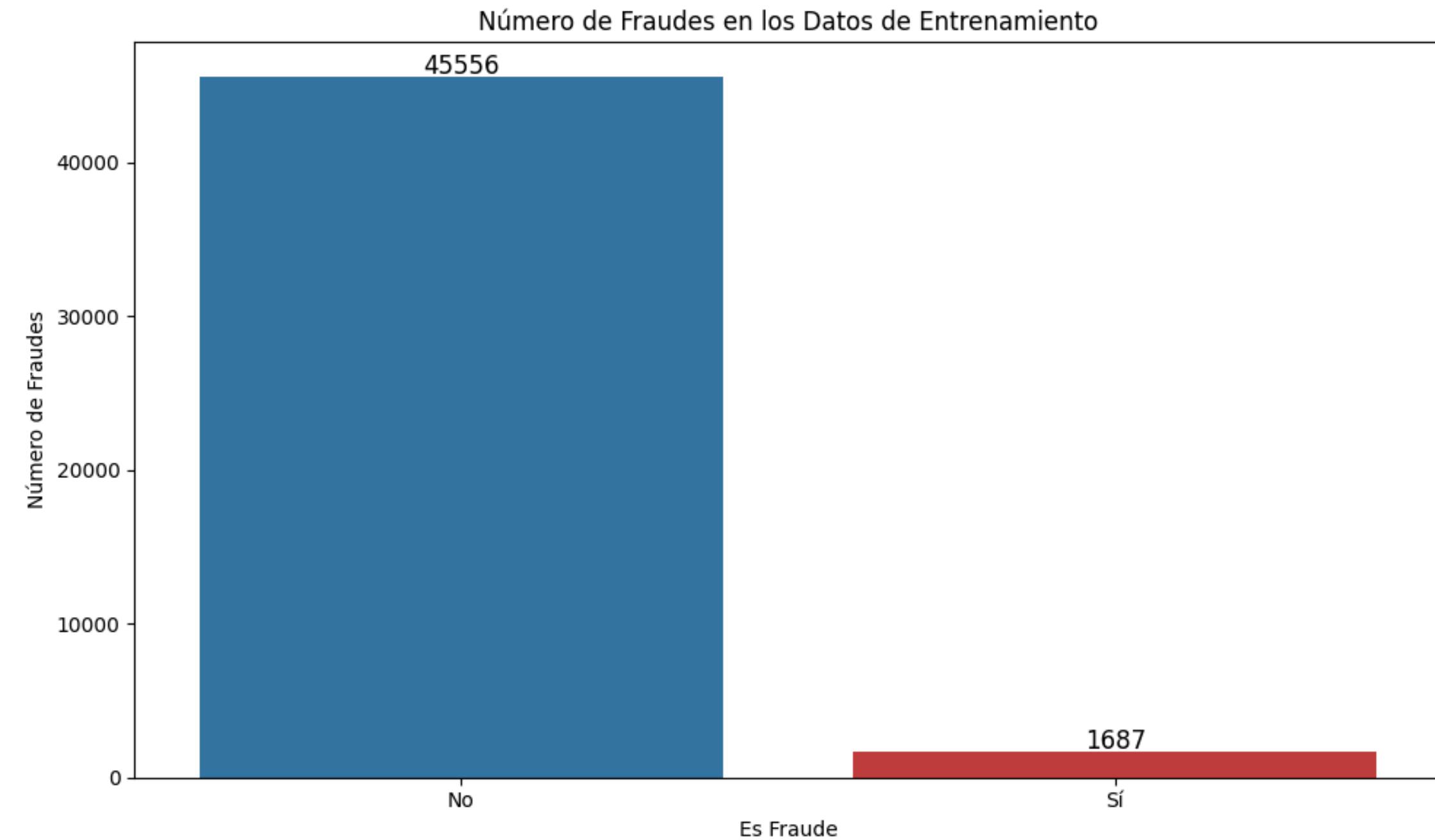


En la figura se presenta una visualización de la cantidad de valores nulos en forma descendente. Destacan especialmente cinco atributos con una significativa cantidad de valores nulos: 'dist1', 'DeviceType', 'R_emaildomain', 'DeviceInfo' y 'dist2'. Esta distribución nos ofrece una instantánea clara de las áreas de nuestros datos que pueden requerir mayor atención y manejo especial durante el proceso de limpieza y preparación de los datos.

nota: no hay datos duplicados



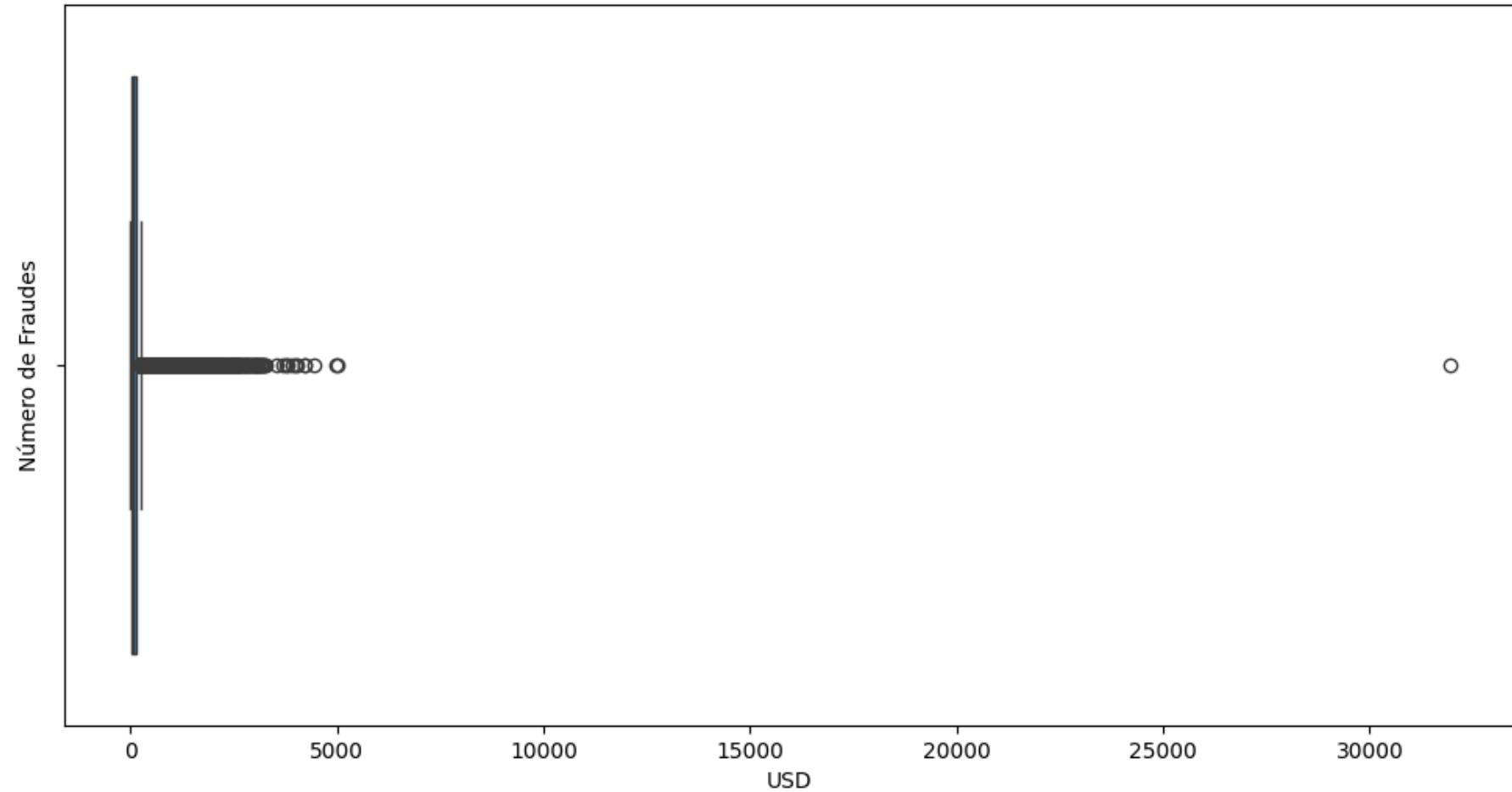
Análisis univariado – target



Es fraude	Porcentaje %
No	96.4
Si	3.6

Se destaca que la mayoría abrumadora de las transacciones se encuentran en la categoría no fraudulenta, mientras que solo una fracción mínima corresponde a casos de fraude. Este desbalance en los datos puede representar un desafío significativo para el proceso de aprendizaje automático, ya que los modelos pueden tener dificultades para identificar y aprender patrones en las clases minoritarias debido a su escasez relativa.

Análisis univariado – Feature: TransactionAmt



index	TransactionAmt
count	59054
mean	135.23
std	265.42
min	0.35
25%(Q1)	43.29
50%	68.5
75%(Q3)	125.0
max	31937.39
IQR	81.70
Límite Inferior	-79.27
Límite Superior	247.5

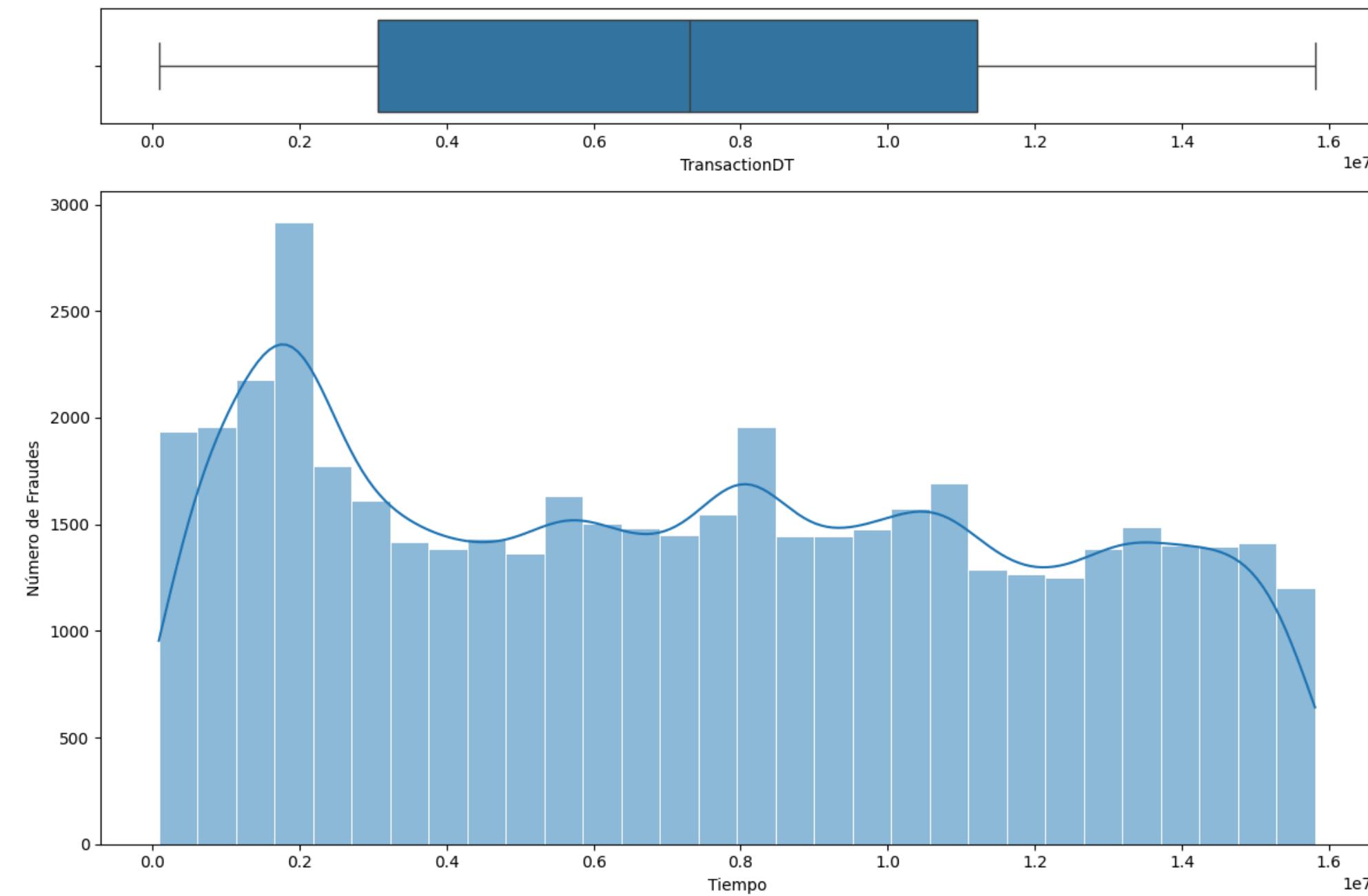
En la representación gráfica de la distribución y en la tabla de estadísticos asociada a la columna "TransactionAmt", se destacan las siguientes observaciones:

Se evidencia una notable disparidad del 50% entre la mediana y la media de los valores.

La comparación entre el límite superior del rango intercuartil y el valor máximo revela discrepancias significativas en órdenes de magnitud, indicando la presencia de datos atípicos.



Análisis univariado – Feature: 'TransactionDT',



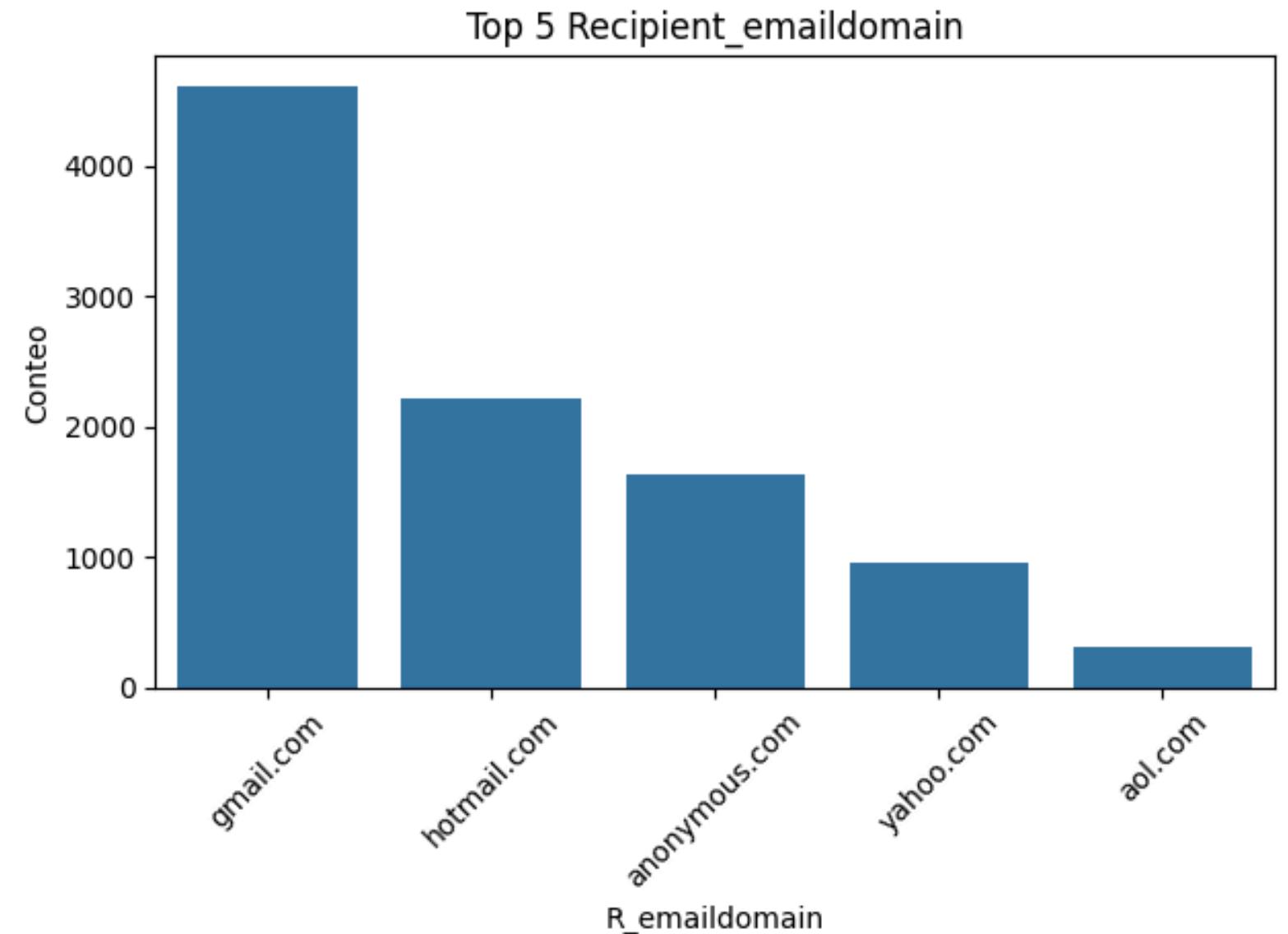
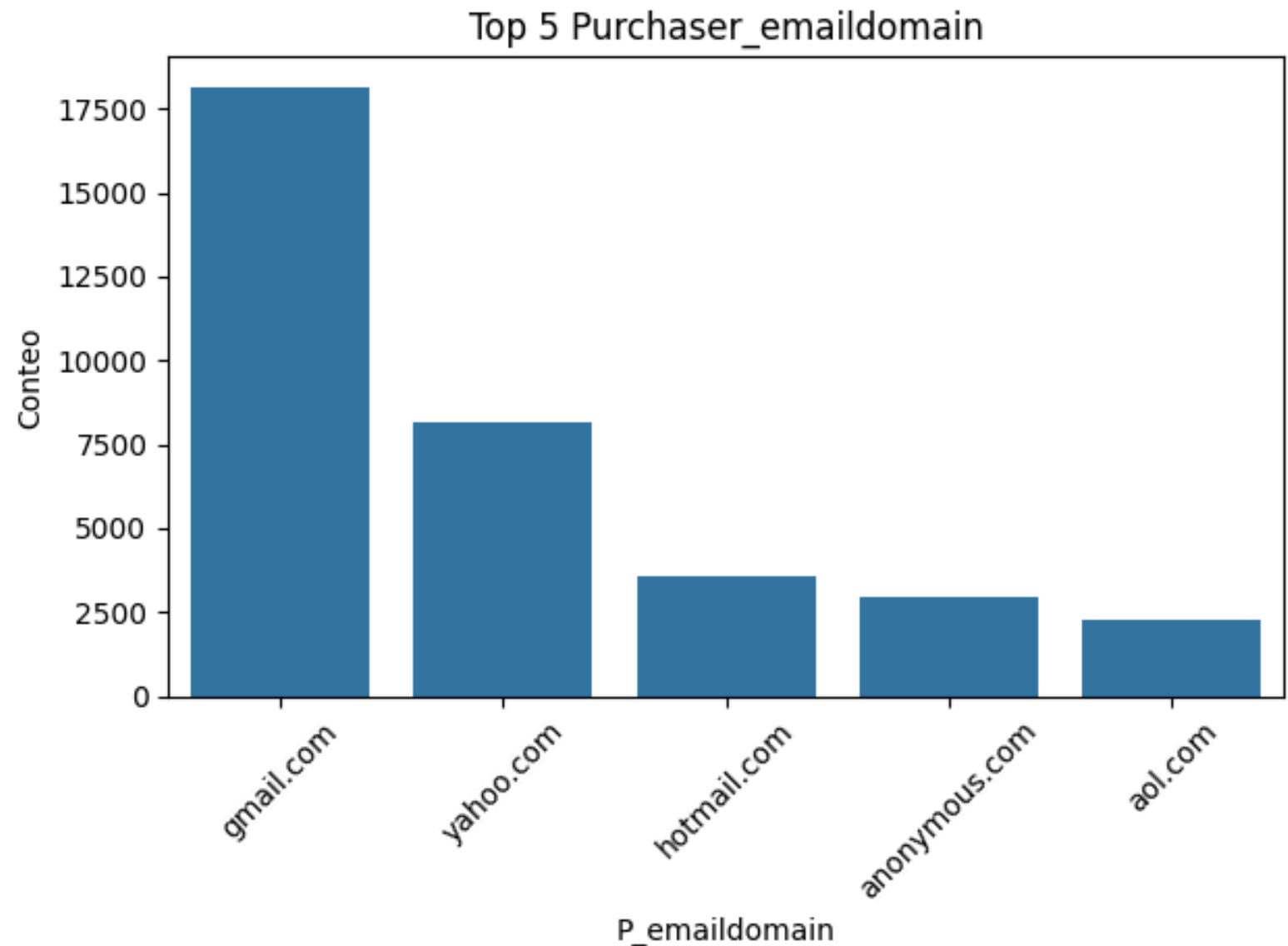
TransactionDT	
count	47243.00
mean	7371518.53
std	4607105.21
min	86469.00
25%	3065672.50
50%	7310923.00
75%	11216282.50
max	15810563.00

En la representación gráfica de la distribución y en la tabla de estadísticos asociada a la columna "TransactionAmt", se destacan las siguientes observaciones:

Se evidencia una notable disparidad del 50% entre la mediana y la media de los valores.

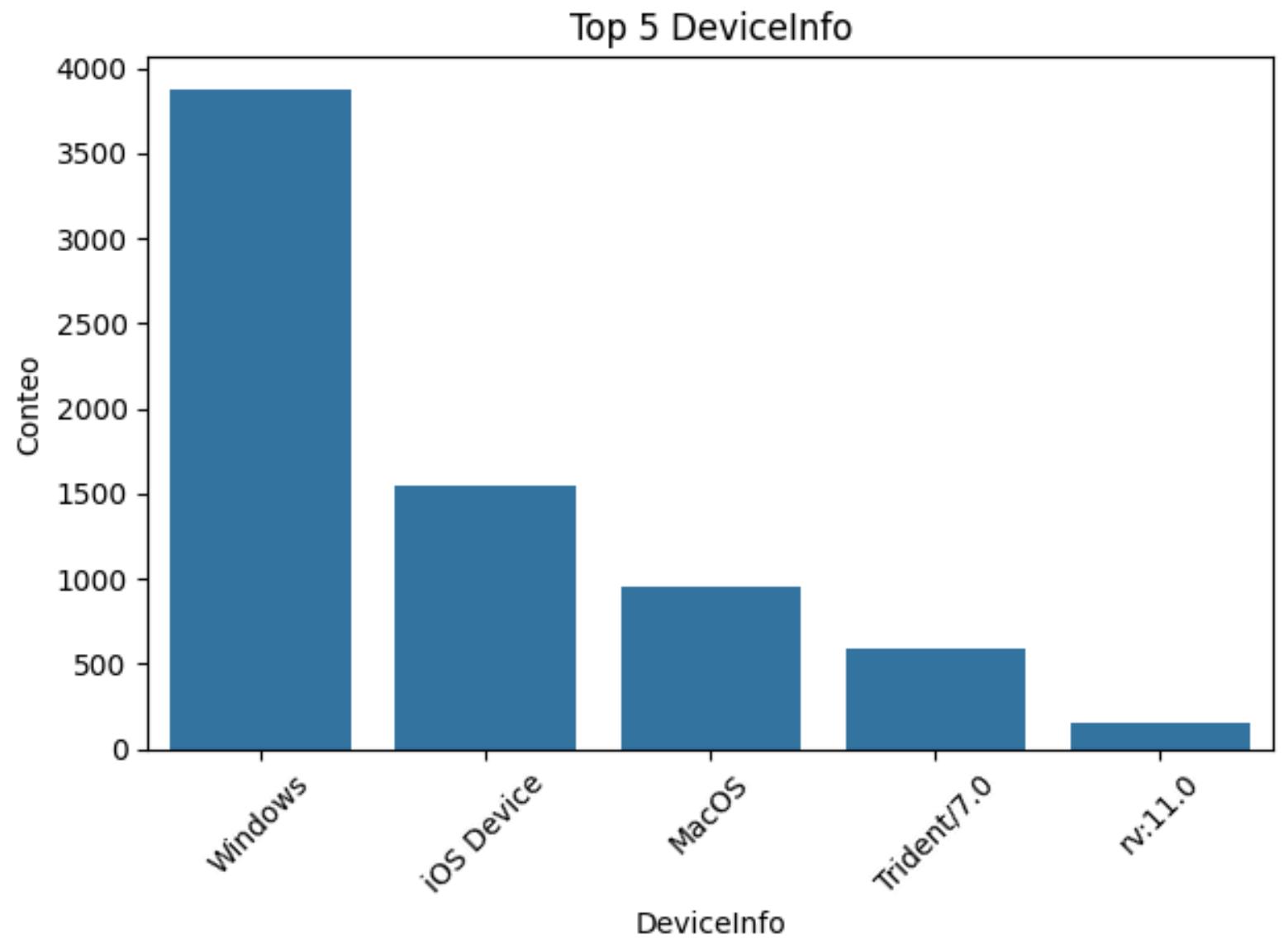
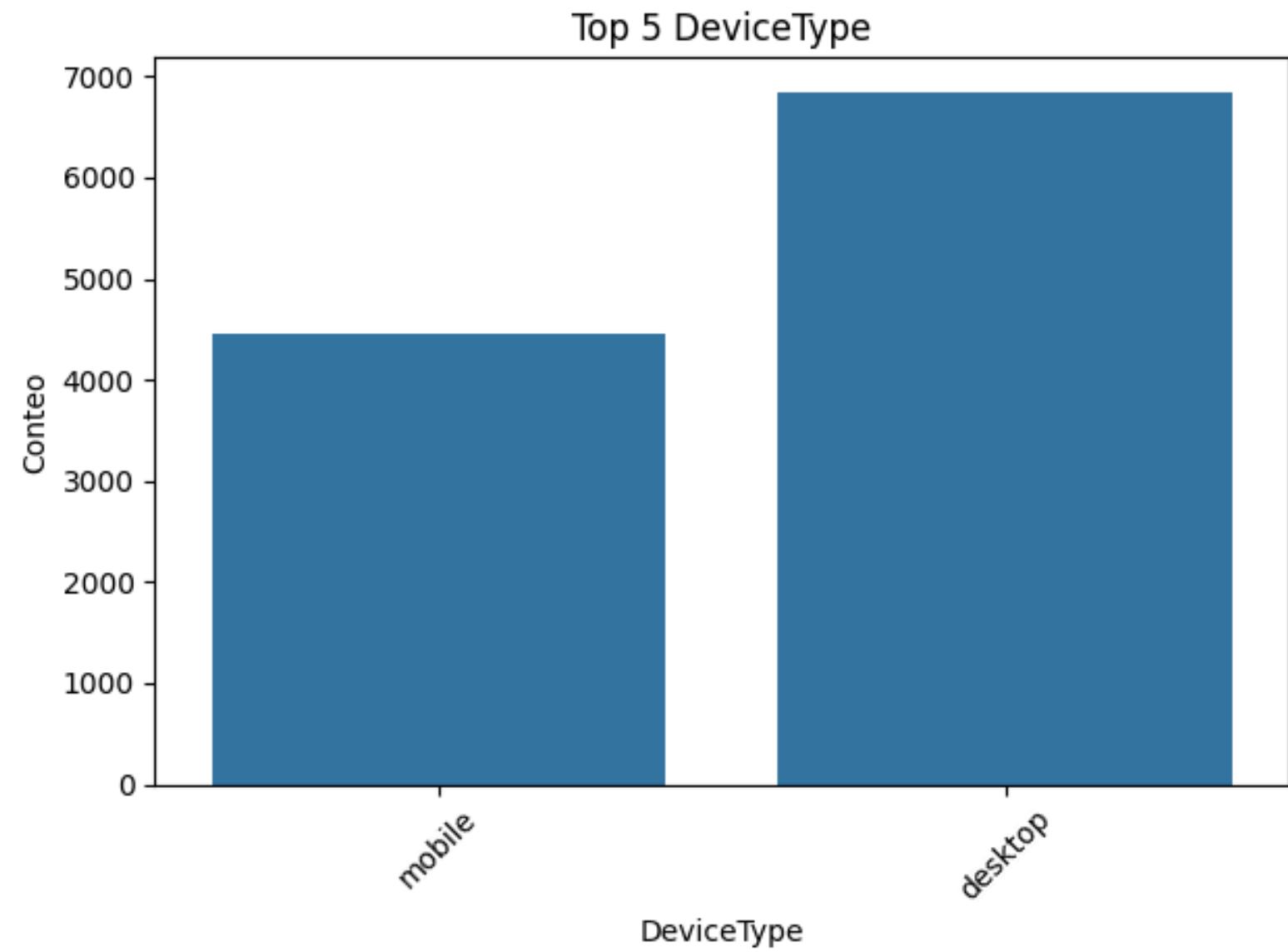
La comparación entre el límite superior del rango intercuartil y el valor máximo revela discrepancias significativas en órdenes de magnitud, indicando la presencia de datos atípicos.

Análisis univariado – Feature: mail – (purchaser) y del destinatario (recipient)



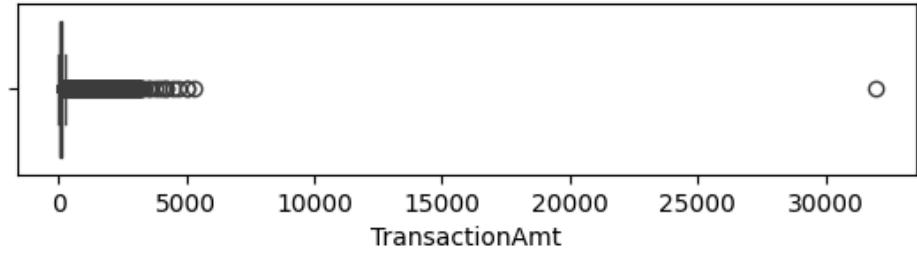
Dominio de correo electrónico del comprador (purchaser) y del destinatario (recipient) respectivamente la mayor parte de la transacción lo hace con "gmail"

Análisis univariado – Feature: Device type y Device Info



"Las transacciones realizadas desde dispositivos de escritorio lideran en cantidad, seguidas por las realizadas desde dispositivos móviles. Dentro de los dispositivos de escritorio, aquellos con sistema operativo Windows son los más frecuentes en las transacciones, seguidos por los dispositivos con iOS."

Análisis Bivariado- Feature: TransactionAmt – Target



TransactionAmt



TransactionAmt

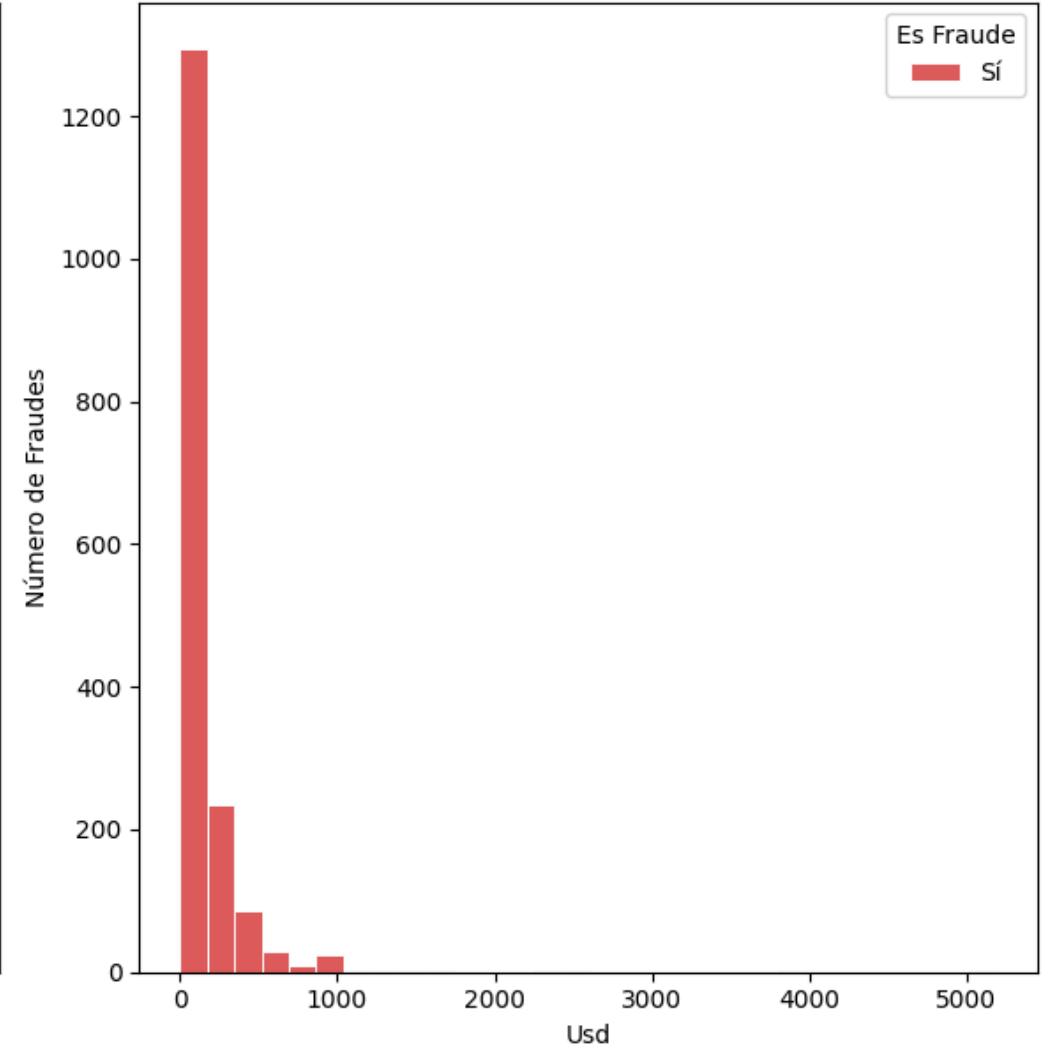
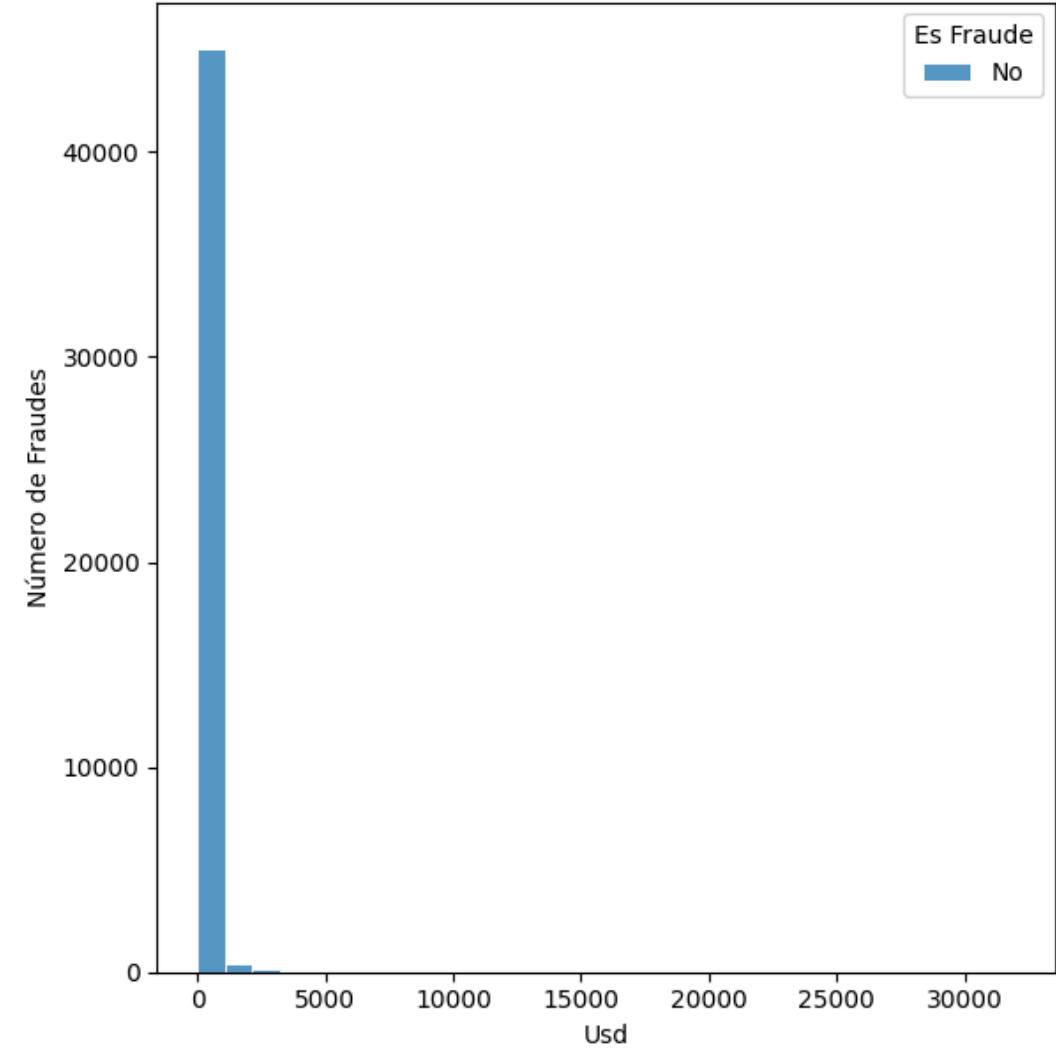


fig. Histograma de los importes de transacción de los datos de entrenamiento

	Fraude (Sí)	No Fraude	Total
Media	152.49	135.81	136.39
Mediana	77.00	68.02	68.50
Máximo	5191.00	31937.39	31937.39
Mínimo	0.29	0.88	0.29
75%	170.00	120.00	125.00

Análisis Bivariado- Feature: TransactionAmt – Target



Con la segmentación de fraude, se identificaron varias observaciones notables en las distribuciones de las transacciones con y sin fraude:

- Ambas distribuciones con y sin fraude muestran un sesgo hacia la izquierda, indicando una concentración de transacciones en los valores más bajos.
- El rango de los importes de las transacciones es muy amplio, abarcando desde 0.292 hasta 31937.391 dólares. Este amplio rango sugiere una variabilidad significativa en los datos.
- La diferencia entre el valor medio y la mediana es considerable, señalando una asimetría en la distribución.
- Se observaron valores atípicos (outliers) en ambas distribuciones, indicando transacciones inusuales o extremadamente altas.

Dada la naturaleza del rango extenso, el sesgo y la presencia de outliers, se propone segmentar el DataFrame en diferentes rangos de importe por lo que se propone dos formas cuales son las siguientes:

- 1. Separando los valores atípicos usando rango intercuartil
- 2. Separar con alguno criterio de segmentación de mercado.

Estas estrategias buscan mejorar la comprensión de los datos al agrupar las transacciones en categorías más manejables y reveladoras.

Elegiremos la opción número 1 y evaluaremos los resultados

Análisis Bivariado- Feature: Separación de los Valores Atípicos Utilizando el Rango Intercuartil

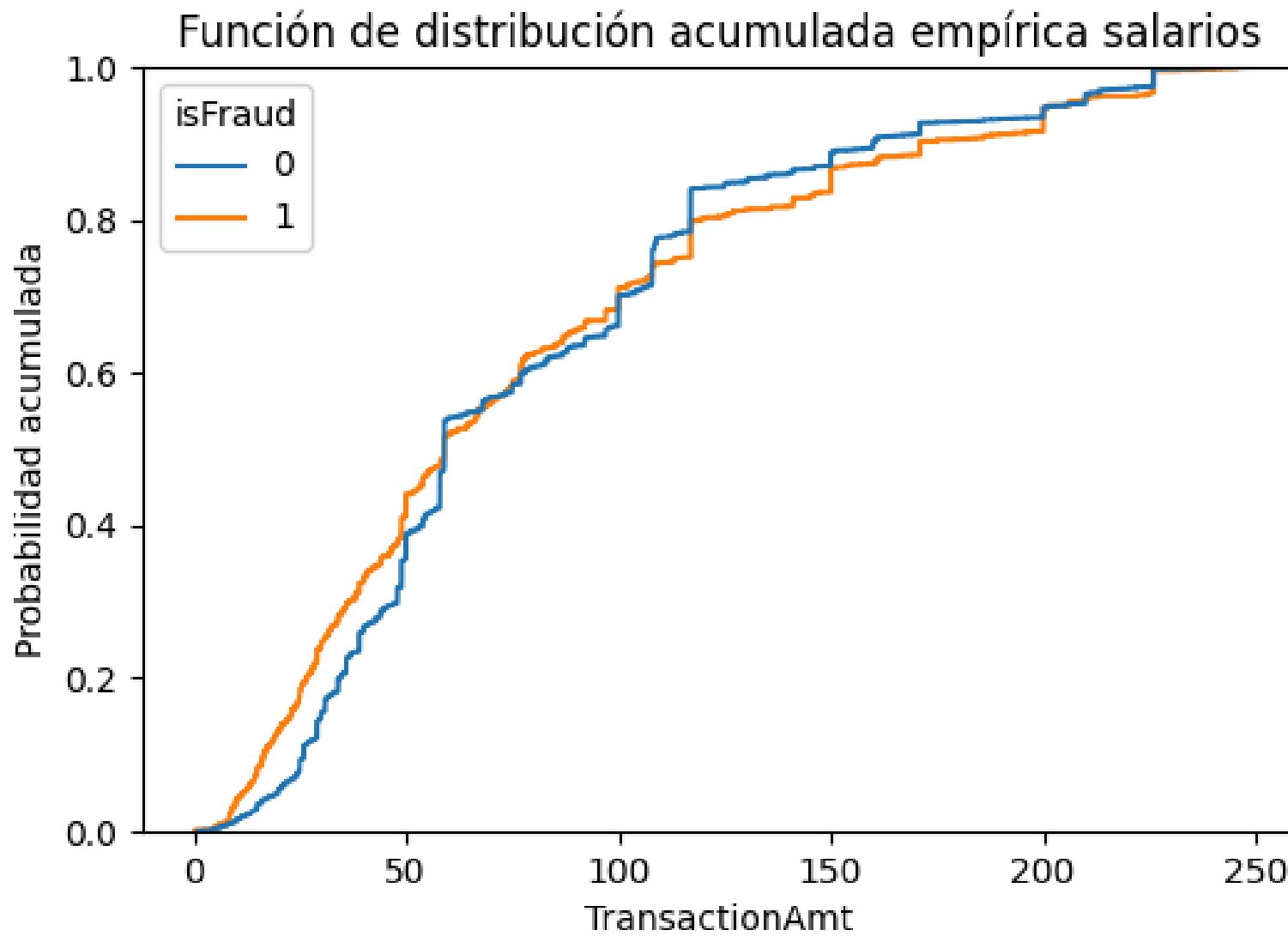
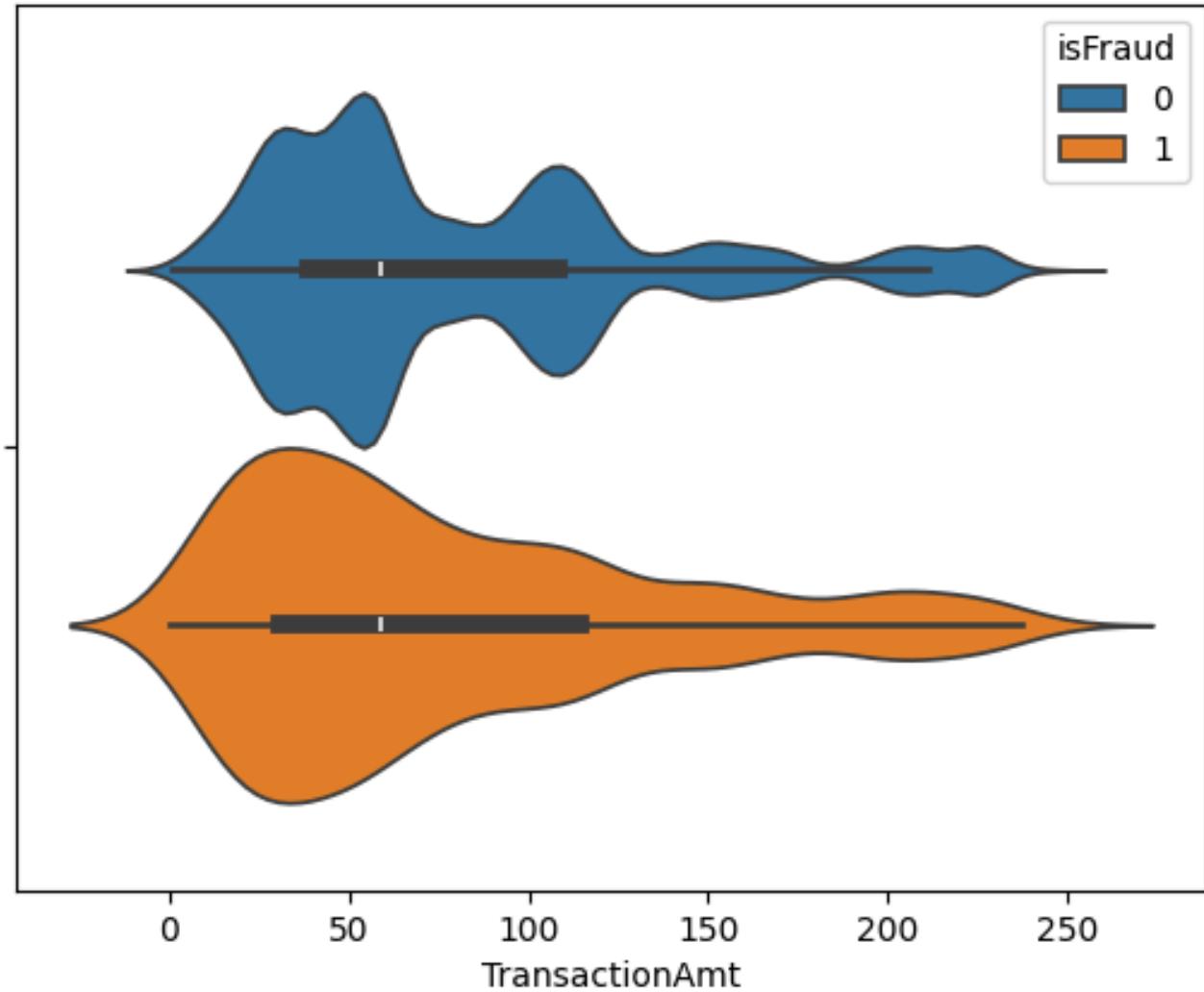


Rangos	Rangos (usd)	% Registro	%No Fraude	%Sí Fraude
R1	0 - 248	88,79	96.60	3.39
R2	248 - 999	9.93	94.59	5.40
R3	> 999	1.28	98.28	1.71

Teniendo en cuenta que la mayor cantidad de registros se encuentra en el rango 1, concentraremos nuestro análisis en esta categoría en la primera fase. Posteriormente, una vez completada esta etapa inicial, extenderemos el análisis para abarcar los demás rangos.



Prueba de Similitud entre las Distribuciones de importes con y sin Fraude en el Rango 1



realizamos una prueba estadística para determinar si las distribuciones de importes de transacciones con y sin fraude son similares. Esta evaluación nos proporcionará información crucial sobre la existencia de patrones distintivos en los importes de las transacciones fraudulentas en comparación con las no fraudulentas.

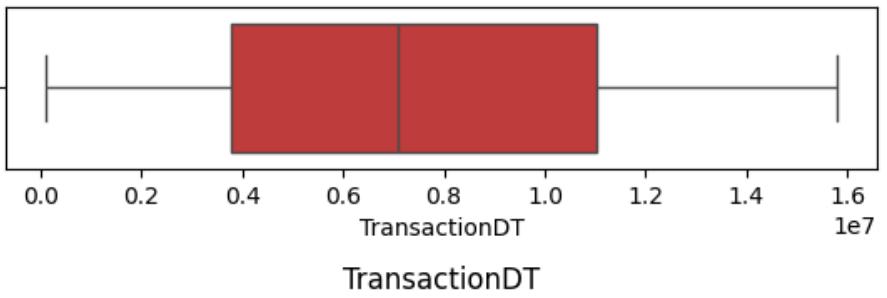
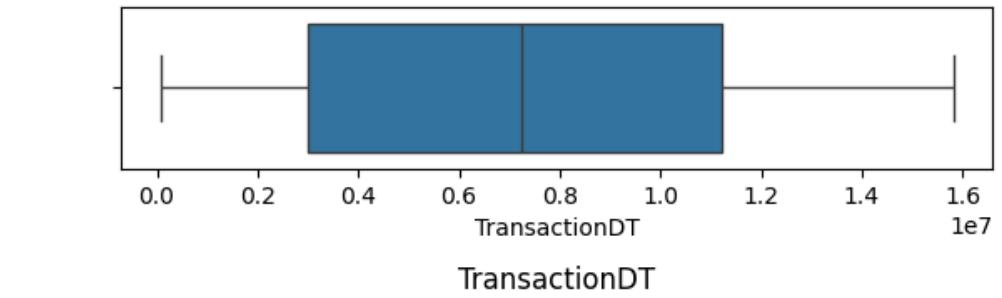
Prueba de Similitud entre las Distribuciones de importes con y sin Fraude en el Rango 1



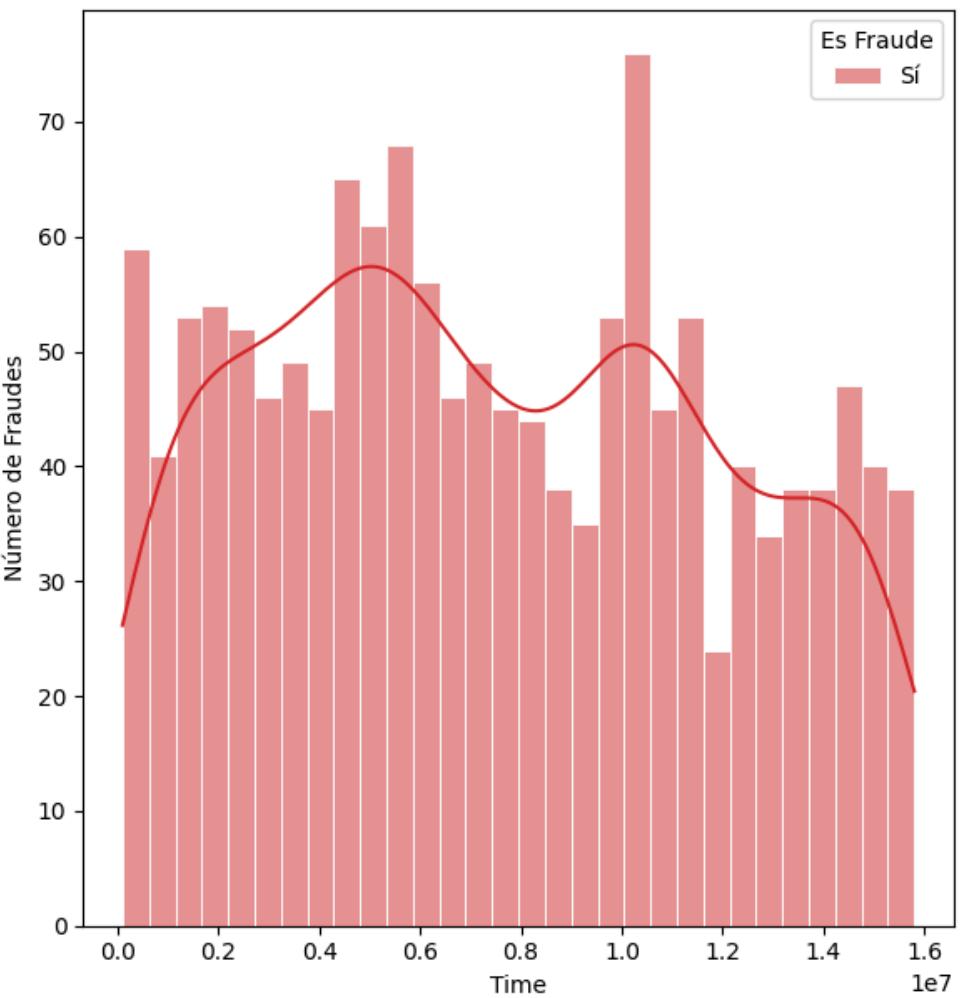
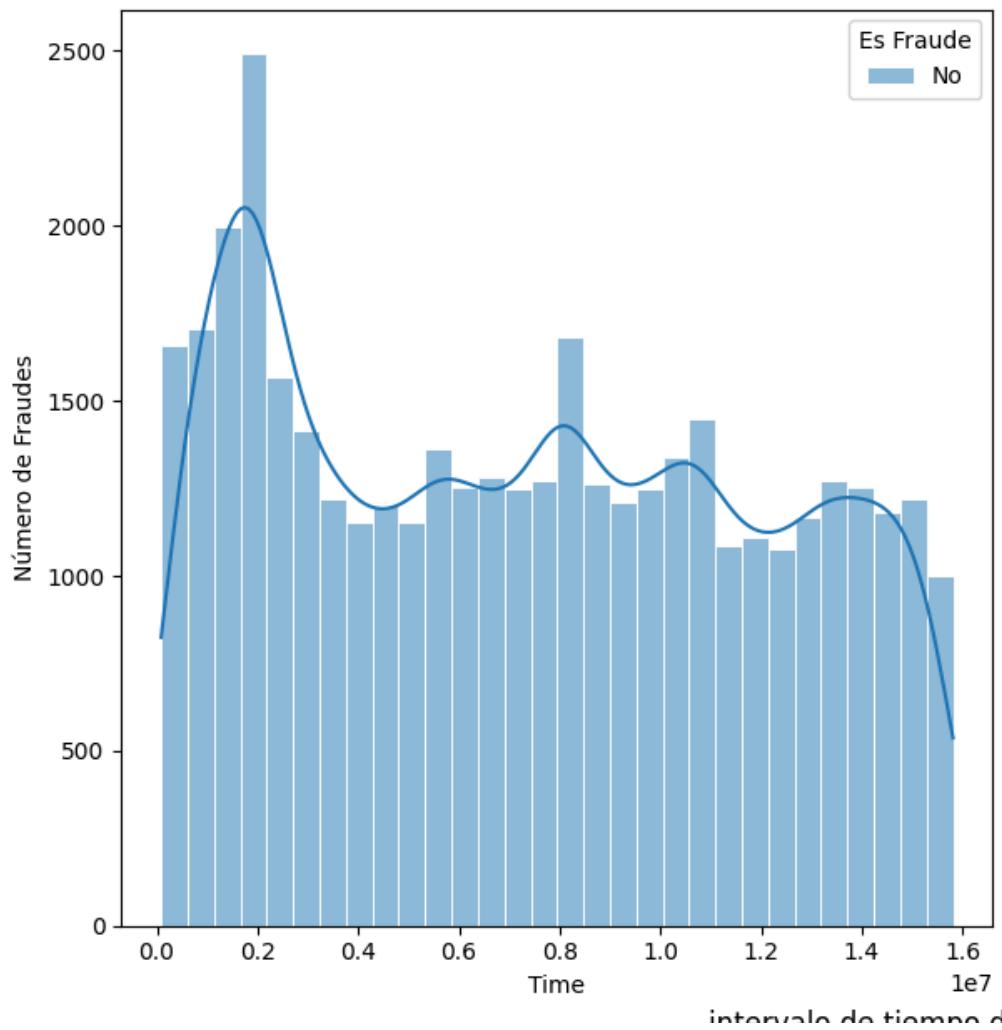
```
KstestResult(statistic=0.10152871255909689,  
            pvalue=7.283692335125864e-13,  
            statistic_location=28.917, statistic_sign=1)
```

El p-valor asociado a la prueba de Kolmogorov-Smirnov es notablemente bajo, específicamente 4.09e-14. Este valor de p es muy pequeño, lo que sugiere que la probabilidad de observar una distancia KS tan grande o mayor entre las distribuciones, suponiendo que provienen de la misma población, es extremadamente baja. Por lo tanto, se puede concluir con confianza que la hipótesis nula de que ambas muestras se originan en la misma distribución es rechazada. En otras palabras, existe **suficiente evidencia estadística para afirmar que las distribuciones son diferentes** y que estas diferencias no son simplemente el resultado de variaciones aleatorias. Este resultado proporciona una comprensión más profunda de las divergencias en el fenómeno analizado.

Análisis Bivariado- Feature:'TransactionDT' Vs isFraud en el Rango1



No Fraude



No Fraude

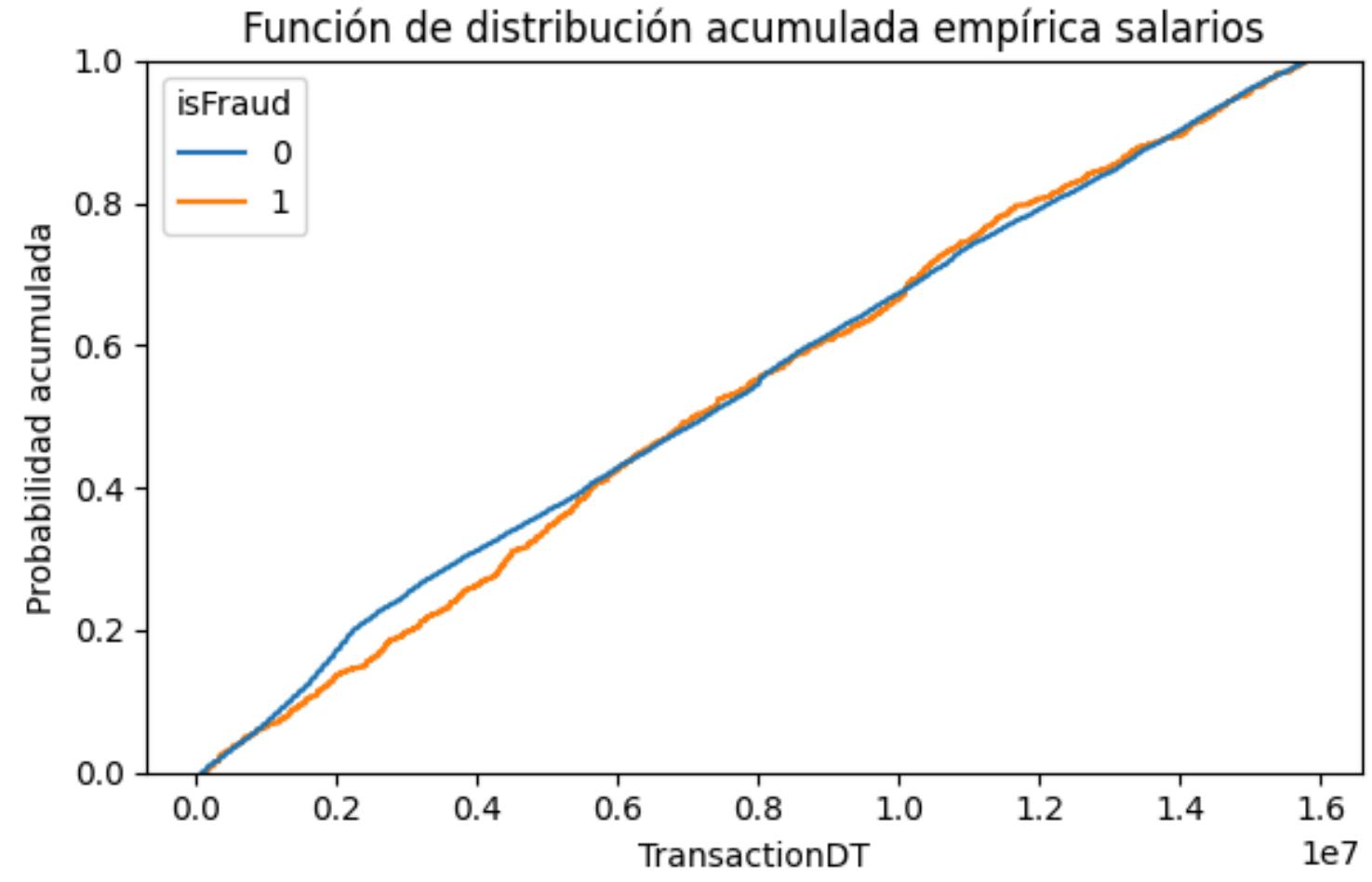
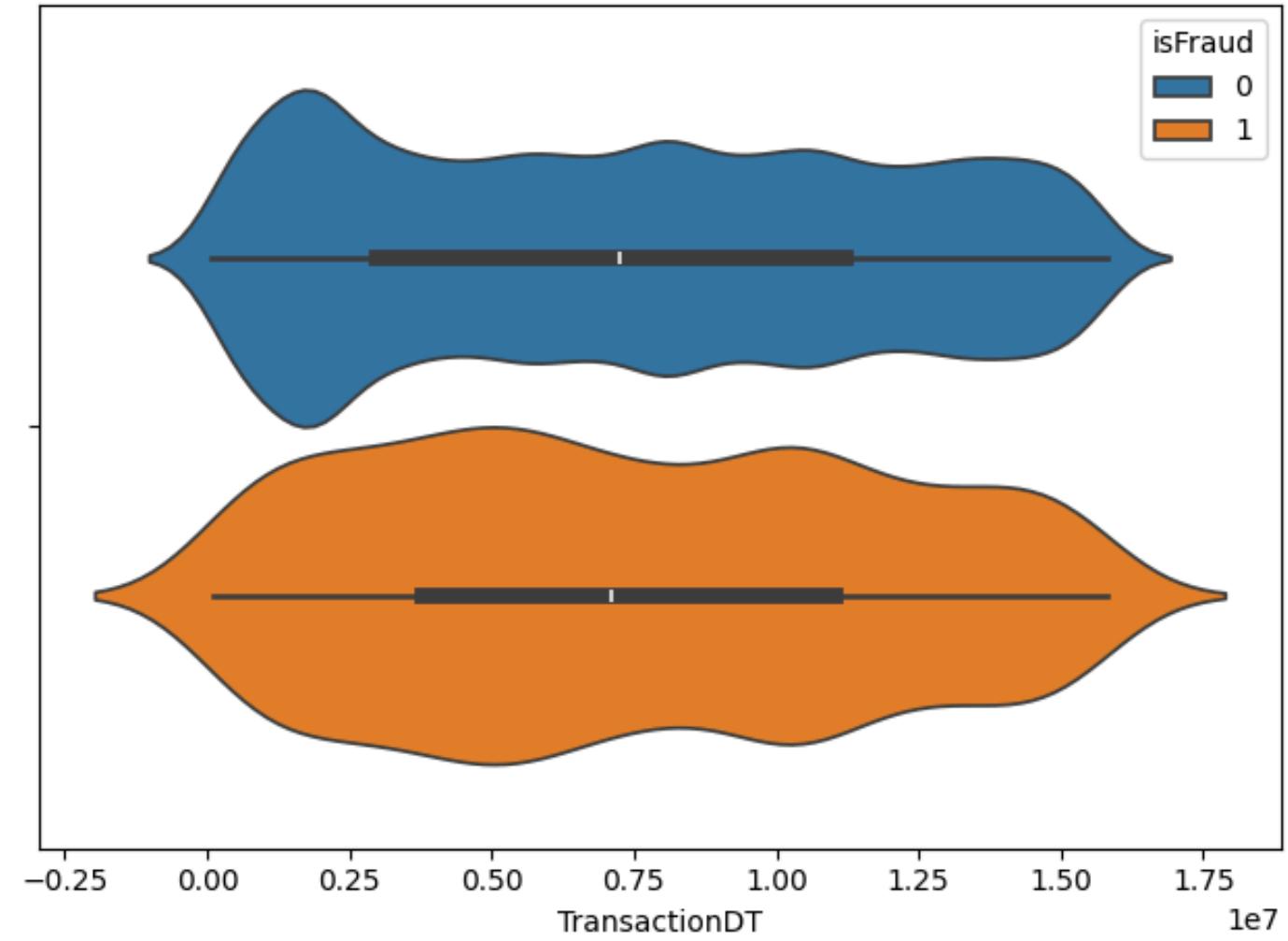
TransactionDT	
count	40538.00
mean	7342312.60
std	4620231.06
min	86469.00
25%	2999074.25
50%	7253138.00
75%	11219868.50
max	15810212.00

TransactionDT	
count	1432.00
mean	7485002.24
std	4437012.25
min	102188.00
25%	3778437.25
50%	7091826.50
75%	11034996.00
max	15801806.00

- Las media y la mediana son del mismo orden de magnitud.
- Los Histograma muestran distribuciones uniforme en ambos casos



Prueba de Similitud entre las Distribuciones de Tiempo (TransactionDT) con y sin Fraude en el Rango 1



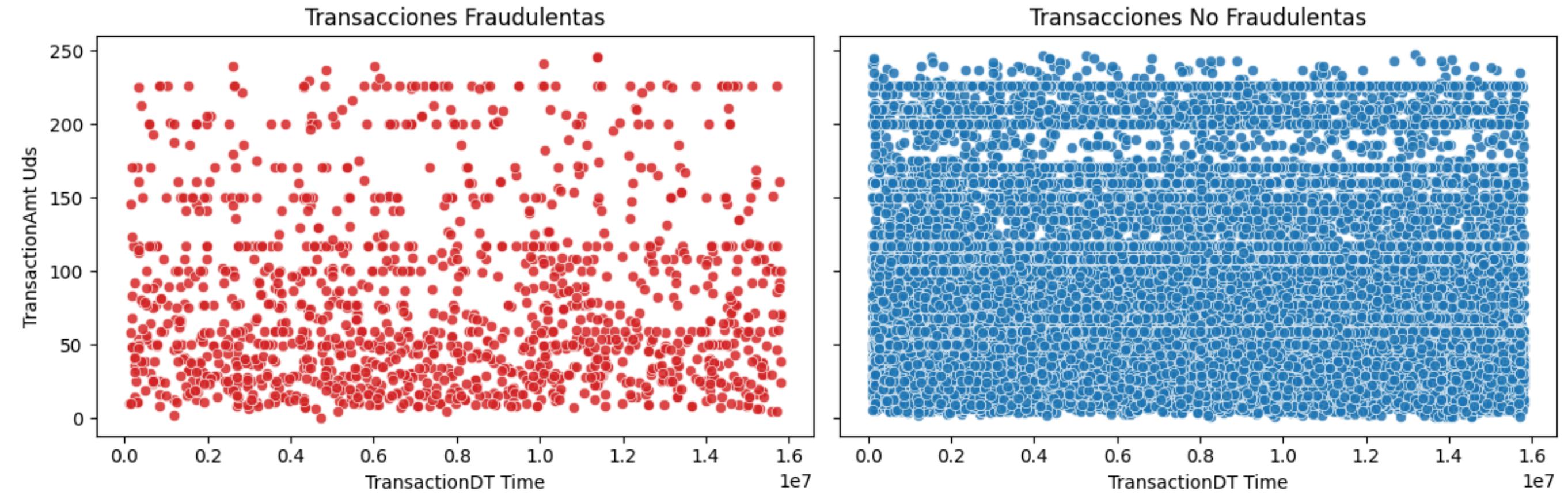
Prueba de Similitud entre las Distribuciones de Tiempo (TransactionDT) con y sin Fraude en el Rango 1



```
KstestResult(statistic=0.06200696305087633,  
            pvalue=4.583222564262201e-05,  
            statistic_location=2409964,  
            statistic_sign=-1)
```

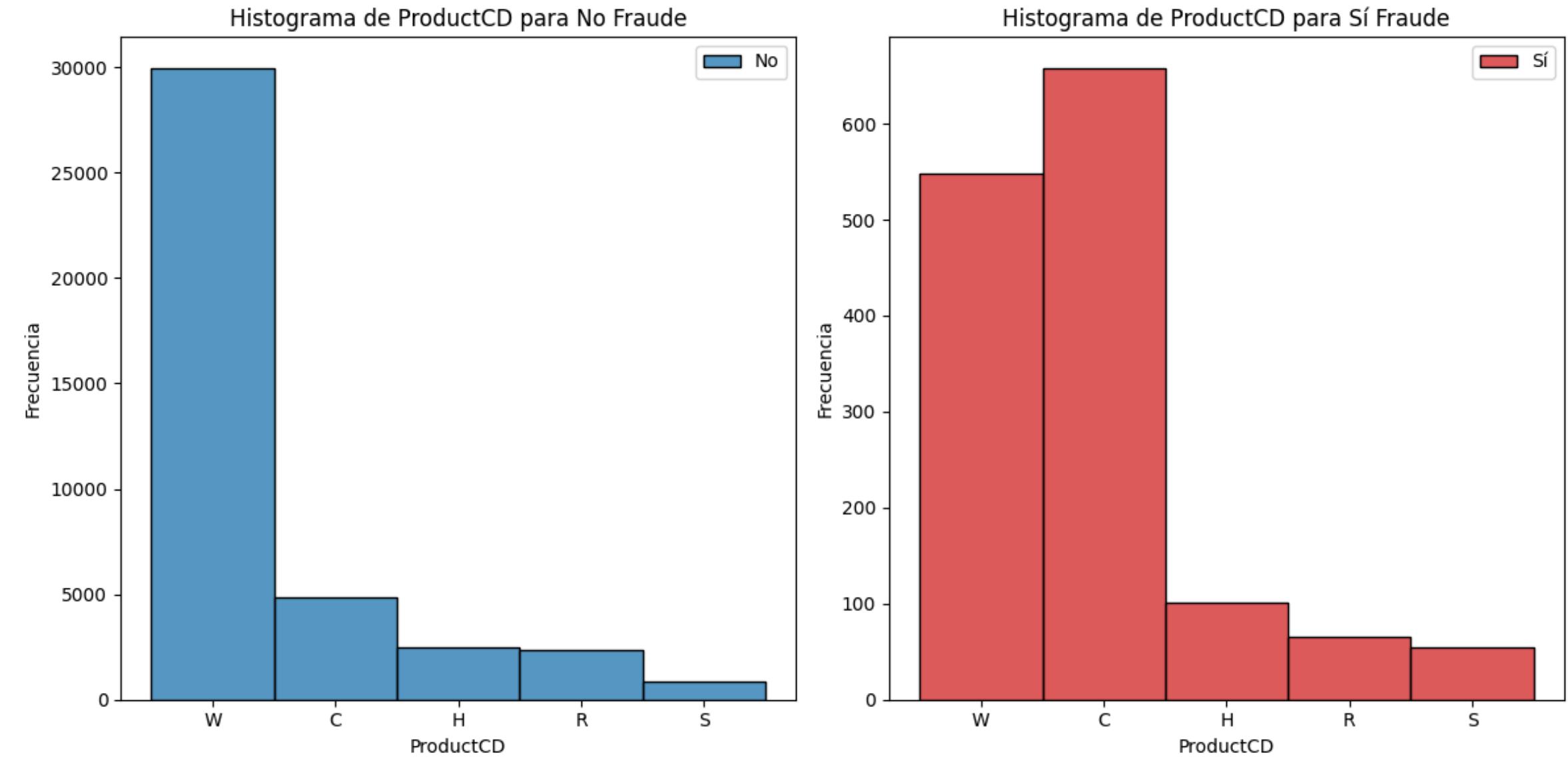
- Al considerar el p-valor, que es particularmente bajo ($3.35\text{e-}06$), llegamos a la conclusión de que las distribuciones son distintas y, por ende, podemos rechazar la hipótesis nula.
- Además, al observar las gráficas de las distribuciones acumuladas, notamos que estas siguen casi una línea recta. Este patrón sugiere que las distribuciones son casi uniformes, ya que la función de distribución acumulada (CDF) de una distribución uniforme debería incrementar de manera constante y lineal. La consistencia con esta forma lineal refuerza la evidencia de la uniformidad en las distribuciones analizadas.

Análisis Bivariado- Feature: Relación entre el Tiempo y los importes de las Transacciones



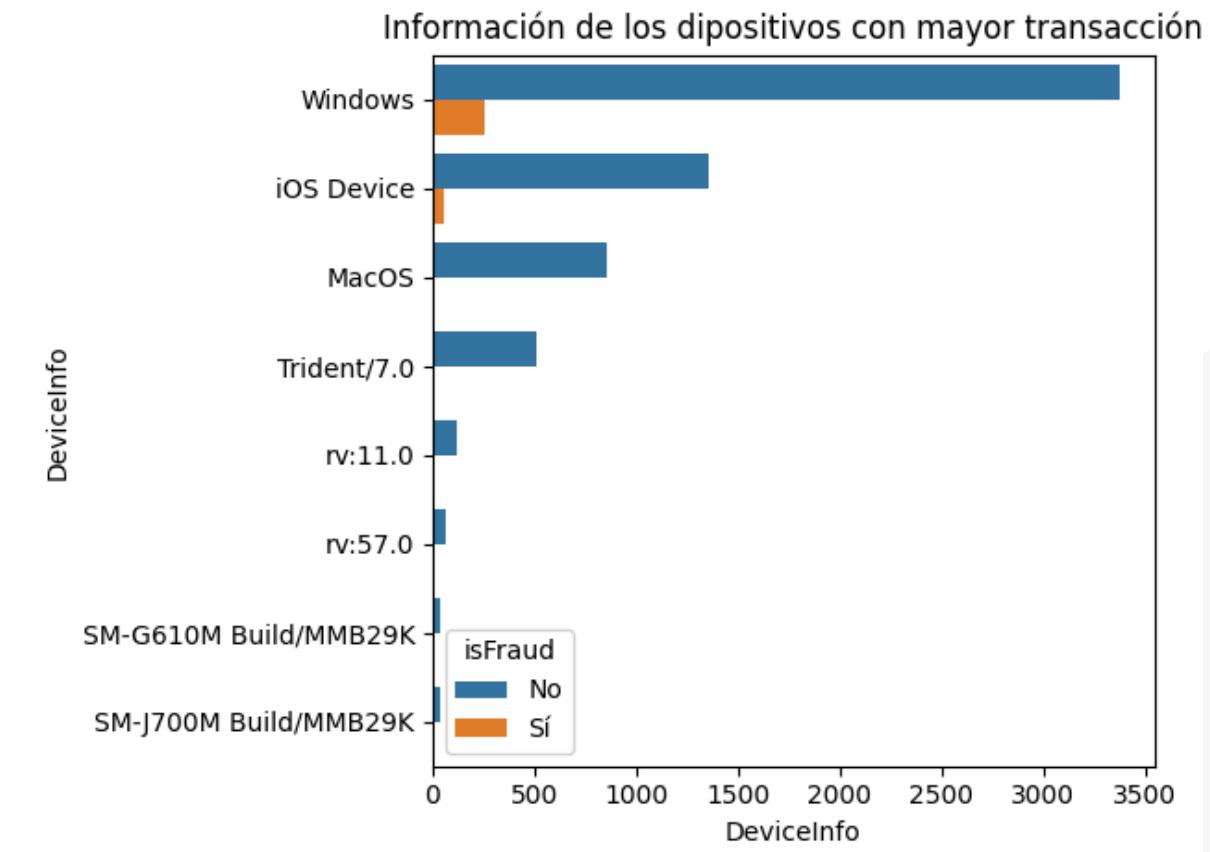
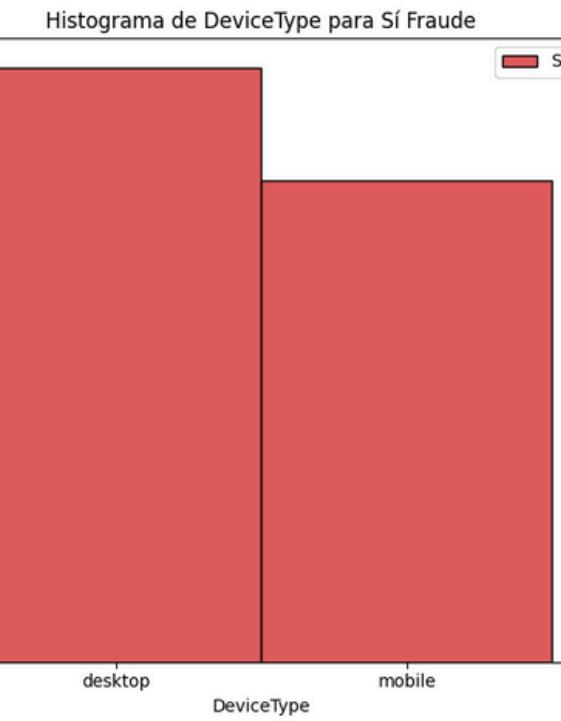
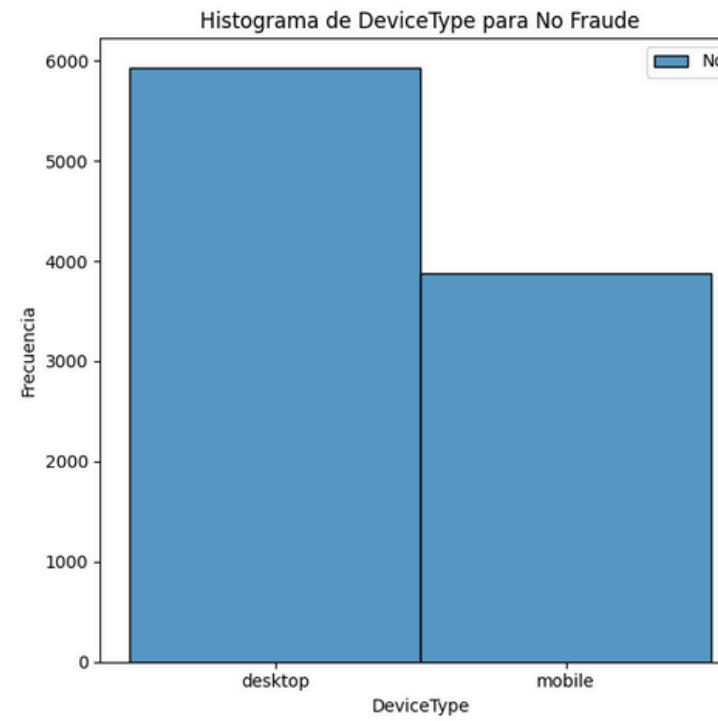
- Datos total: 0.006
- Con fraude -0.019
- sin fraude 0.07 Los valores de correlación son casi cero en todos los casos por lo tanto no hay correlación entre el importe y el tiempo

Análisis Bivariado- Feature: Columna ProductCD Vs isFraud



- los ProductCD con el código "C" y "W", tienen una mayor número de fraude en el Rango 1

Análisis Bivariado- Feature: Columna DeviceType Vs isFraud

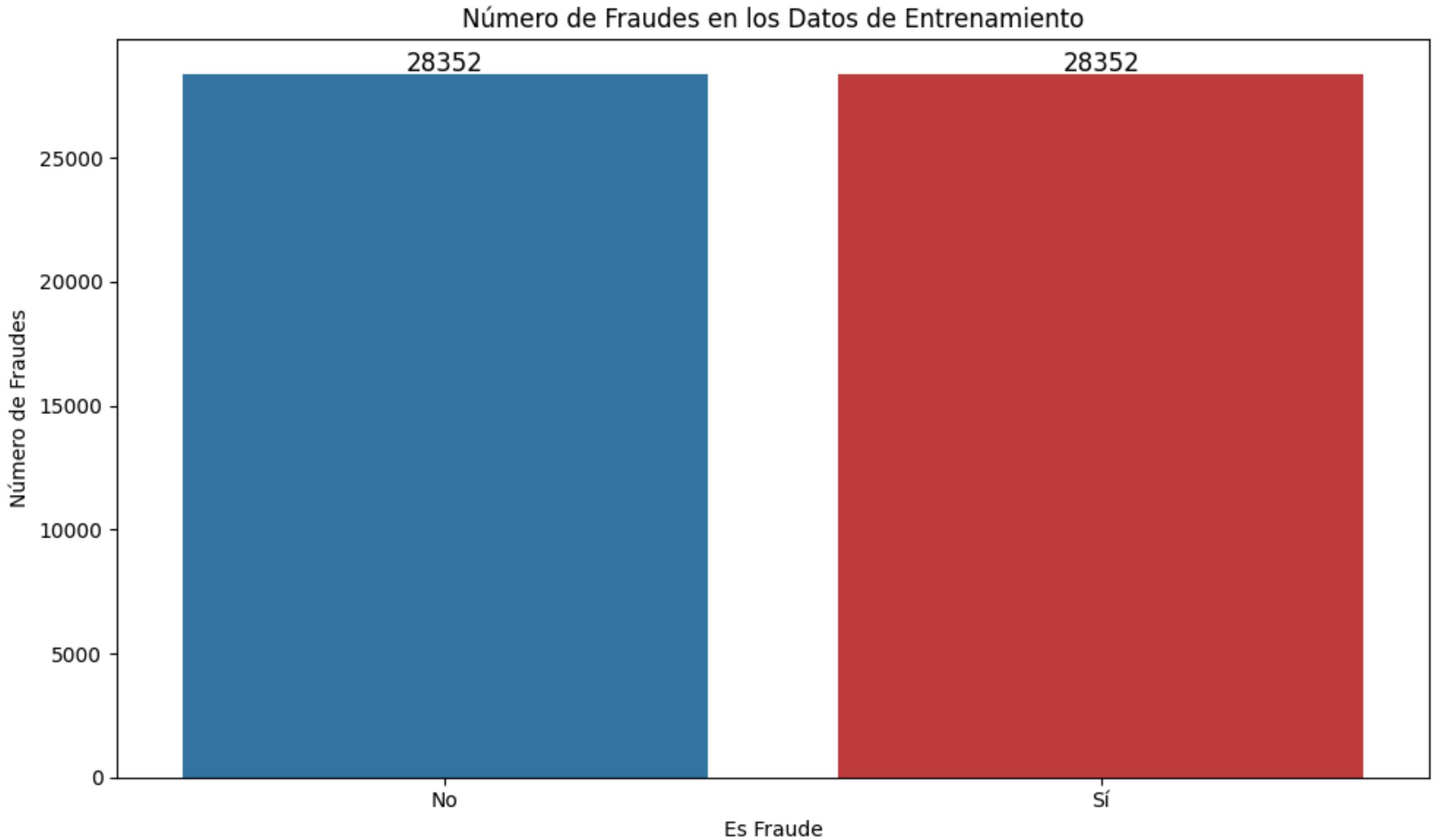


- los ProductCD con el código "C" y "W", tienen una mayor número de fraude en el Rango 1



3. ENTRENAMIENTO DEL MODELO

Balanceo de datos con respecto a la variable objetivo



La columna "objetivo" presentaba un desbalance significativo, por lo que se decidió utilizar la técnica SMOTE (Synthetic Minority Over-sampling Technique) para balancear los datos. SMOTE genera datos sintéticos para equilibrar las clases, mejorando así la capacidad del modelo para identificar correctamente las instancias minoritarias.

Resultado Modelo

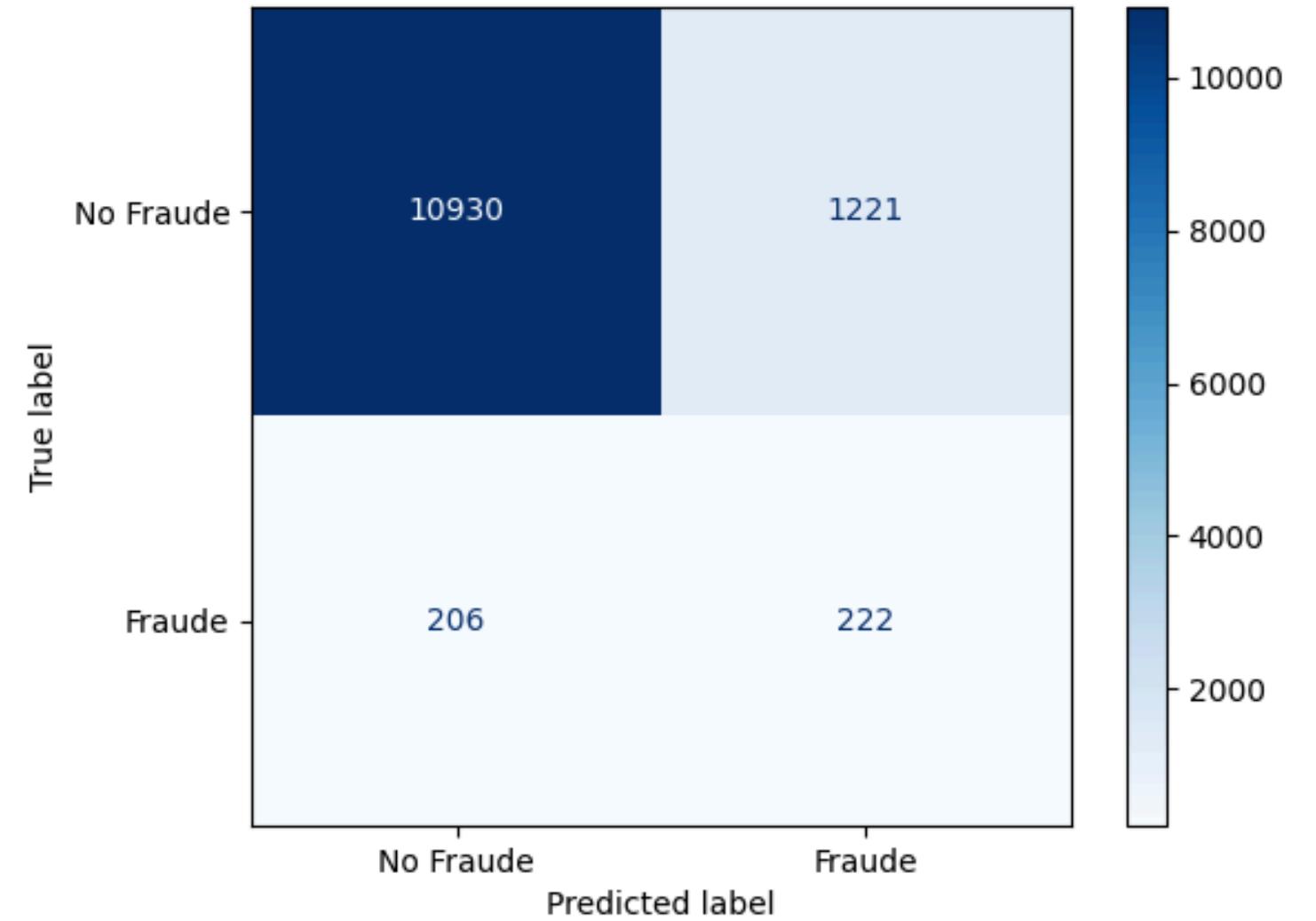


RandomForestClassifier

RandomForestClassifier(max_depth=40, random_state=42)

La precisión de nuestro modelo en los datos de entrenamiento es excepcionalmente alta, alcanzando un valor de 93%. En cuanto a los datos de evaluación, logramos una sólida precisión de 0.88. Estos resultados destacan la eficacia del modelo al clasificar correctamente las instancias tanto en el conjunto de entrenamiento como en el conjunto de evaluación, evidenciando su capacidad para aprender y generalizar patrones de manera robusta.

Matriz de Confusión



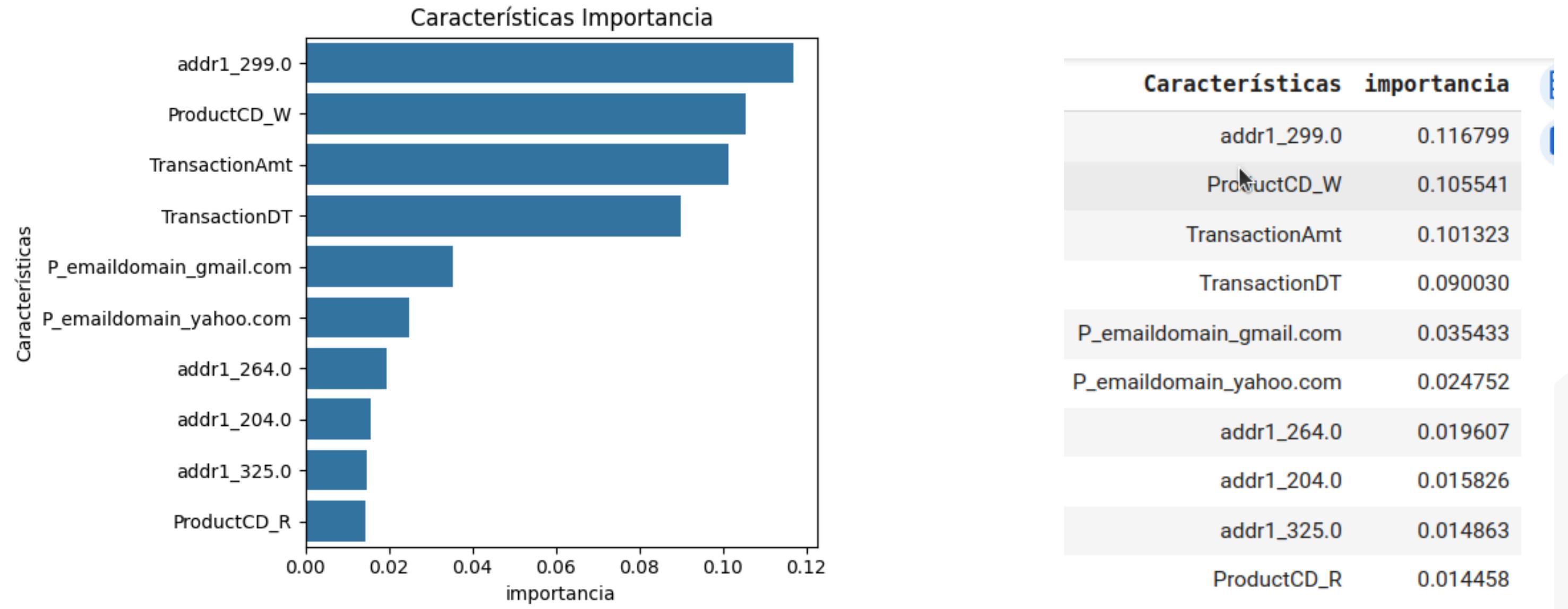
Reporte de Clasificación:

	precision	recall	f1-score	support
0	0.98	0.90	0.94	12151
1	0.15	0.52	0.24	428
accuracy			0.89	12579
macro avg	0.57	0.71	0.59	12579
weighted avg	0.95	0.89	0.91	12579

El modelo tiene una alta precisión del 98% para identificar transacciones legítimas, pero su precisión para detectar fraudes es del 15%. Esto sugiere la necesidad de mejorar la capacidad del modelo para identificar transacciones fraudulentas sin aumentar significativamente la tasa de falsos positivos.



Características Importantes



En la figura se observan los aportes de las 10 características que más influyen en el modelo. Se puede notar que las direcciones ("addr") están entre las primeras diez. El tipo de producto ("w") también es relevante, seguido por la cantidad por transacción y el tiempo de la transacción.

Conclusión



Durante el análisis, identificamos las variables más relevantes que influyen en la detección de fraude.

- La dirección de donde se realiza el fraude
- El tipo de producto
- los emails donde se hace los fraude
- los tiempo y el importe de la transacción

Comprender estas variables nos permite prevenir fraudes de manera más efectiva y mejorar la seguridad y satisfacción del cliente y el comerciante.

Recomendación



Para continuar mejorando, recomendamos:

- 1.Exploración Detallada de Características: Evaluar y optimizar las características utilizadas, crear nuevas derivadas y eliminar las no significativas.
- 2.Optimización de Hiperparámetros: Utilizar técnicas avanzadas como la búsqueda bayesiana para optimizar el modelo.
- 3.Exploración de Otros Modelos: Probar modelos alternativos de clasificación, como la regresión logística, redes neuronales o gradient boosting, para identificar el más eficaz.

Estas estrategias aumentarán la precisión del modelo, mejorando la toma de decisiones y la optimización de recursos.

