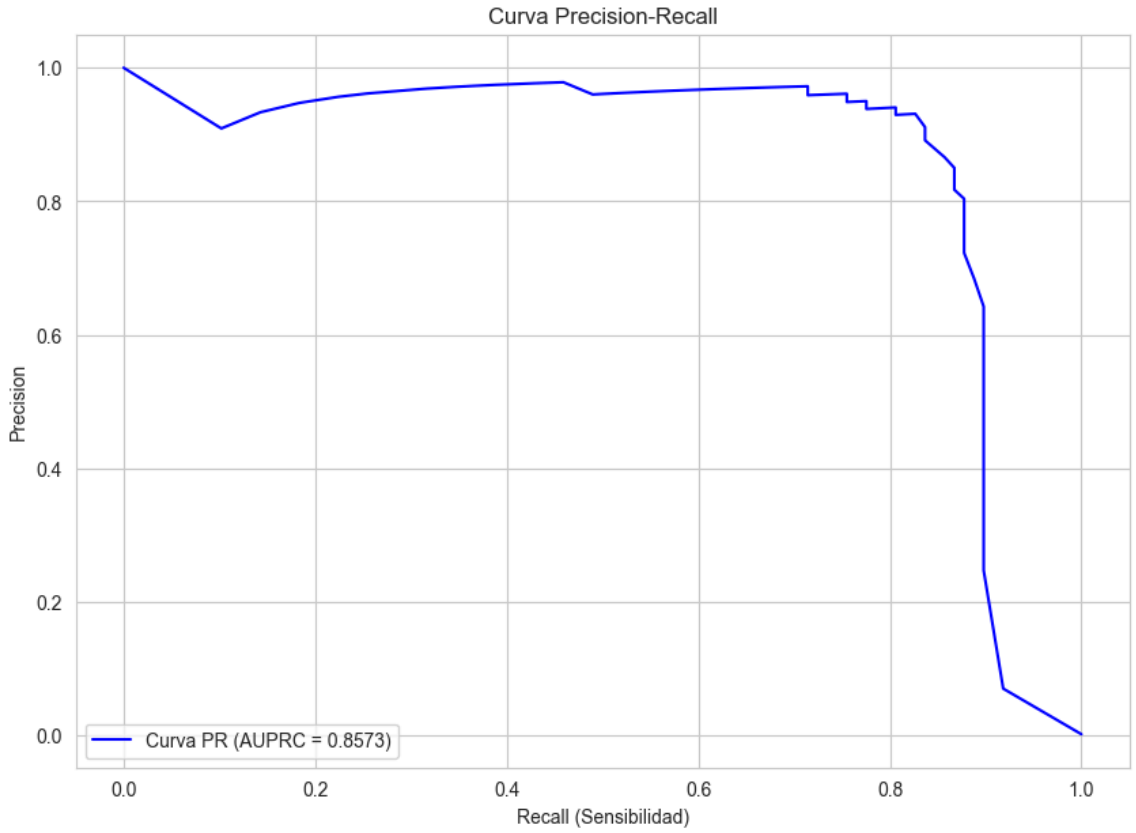
Reporte Final de Explicabilidad del Modelo de Fraude

Objetivo: Este reporte presenta un análisis de explicabilidad para un modelo de clasificación de fraude crediticio *usandoRandomForestClassifier*. Se aplican técnicas de explicabilidad global y local para entender el comportamiento del modelo, diagnosticar sus fortalezas y debilidades, y proponer recomendaciones accionables para su mejora.

1. Resumen del Desempeño del Modelo

Antes de analizar la explicabilidad, es importante recordar el rendimiento del modelo en el conjunto de prueba. El modelo fue entrenado considerando un severo desbalance de clases, priorizando la detección de fraudes (Recall) sin generar un número excesivo de falsas alarmas (Precision).

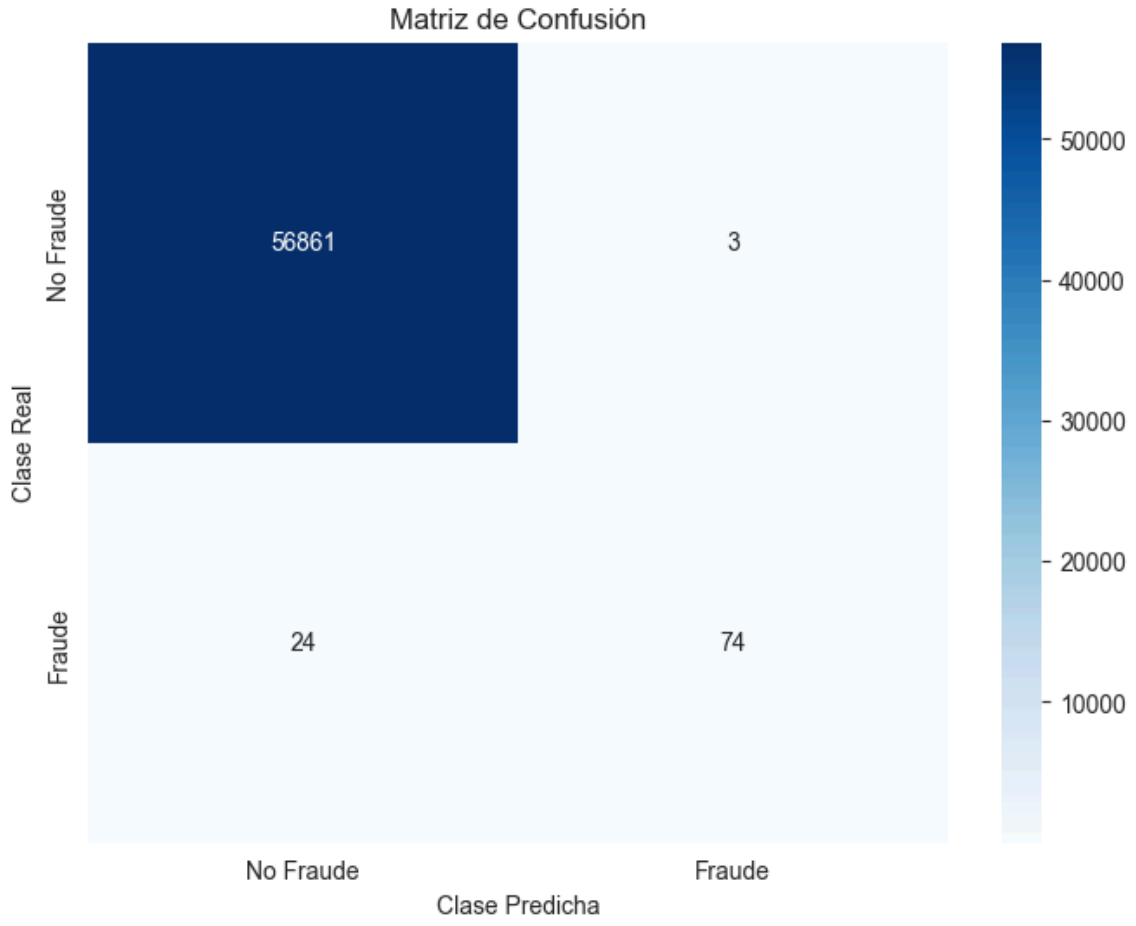
Métrica Principal: Área Bajo la Curva Precision-Recall (AUPRC).



AUPRC Obtenido: 0.8573 (calculado en el cuaderno 2)

Este es un resultado robusto, indicando que el modelo es significativamente mejor que un clasificador aleatorio. Podemos constatar dicha aseveración con la matriz de confusión, que

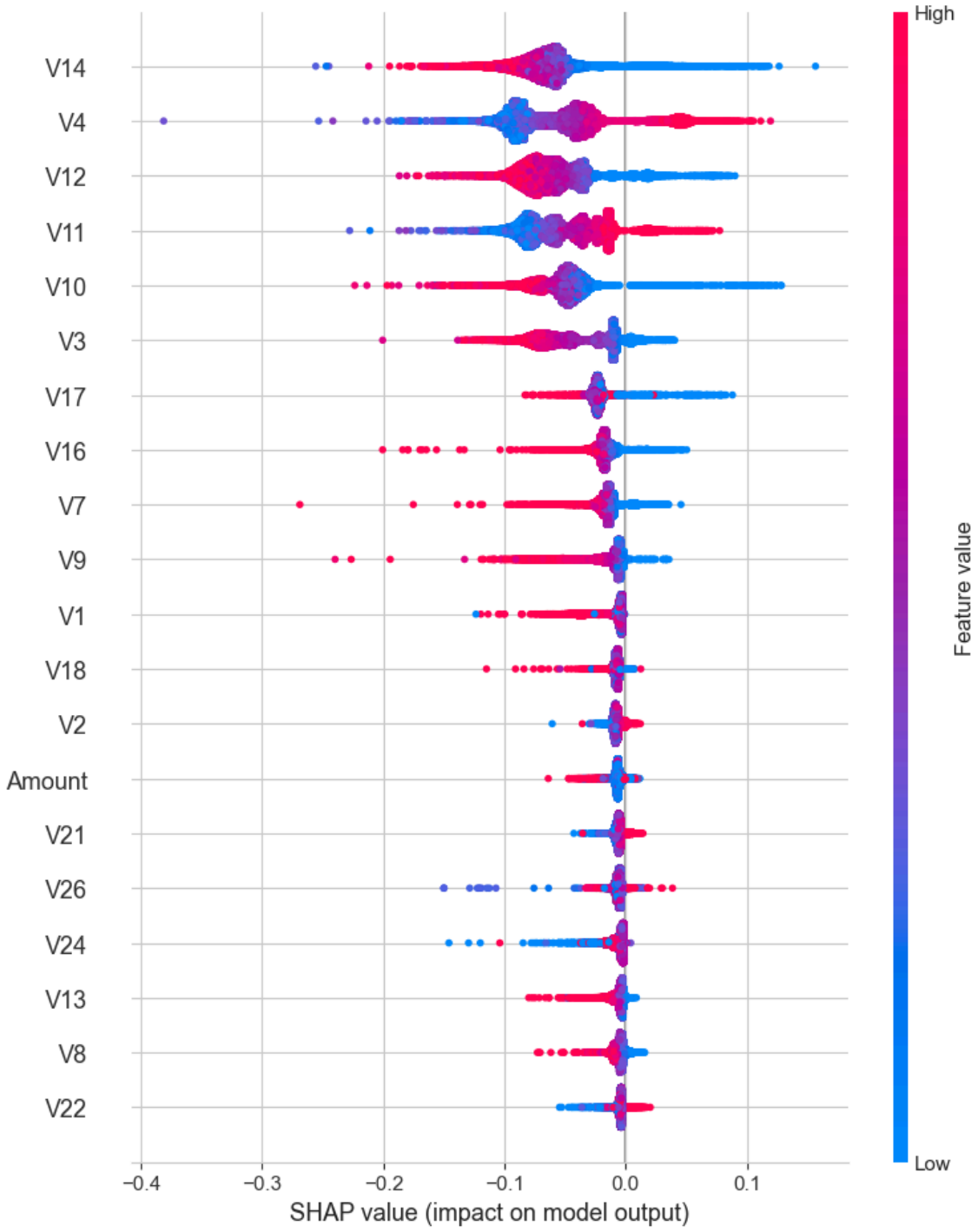
muestra una alta sensibilidad para detectar fraudes, aunque a costa de algunos Falsos Positivos.



---

2. Explicabilidad Global: ¿Cómo Piensa el Modelo en General?

Para entender los factores más importantes que el modelo utiliza para tomar decisiones, se utilizó un análisis SHAP.



Interpretación de los Hallazgos Globales:

El análisis global revela que el modelo ha aprendido patrones claros y consistentes.

- Características Dominantes: Las varaibles V14, V4, V12, V11 y V10 son, como se puede observar claramente, las más influyentes. El modelo depende fuertemente de ellas.
- Patrones de Transacción "Fraude": El gráfico muestra que valores bajos (puntos azules) en V14, V10, V12 y V17 están fuertemente correlacionados con una predicción de fraude (valor SHAP positivo).

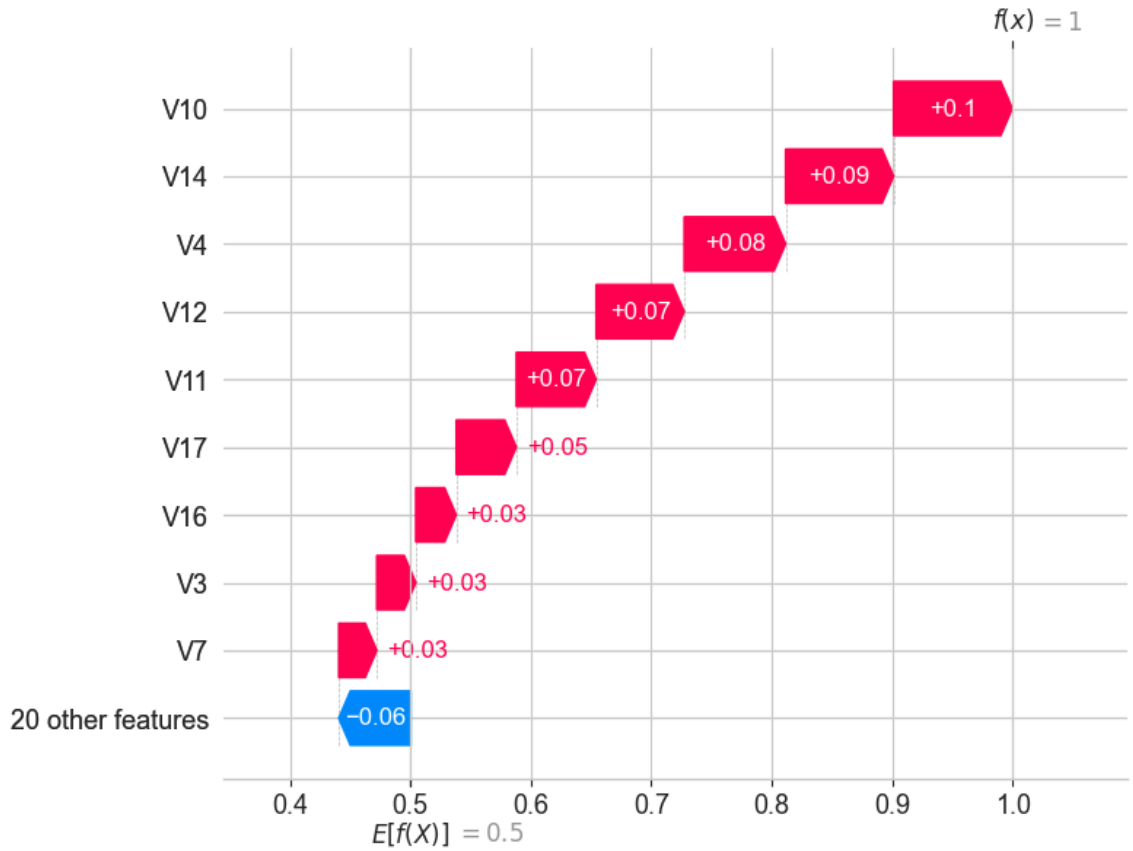
- Patrones de Transacción "Normalidad": Por el contrario, valores altos (puntos rojos) en variables como V4 y V11 son los que tienden a indicar una posible transacción fraudulenta.
- Conclusión Global: El modelo parece haber aprendido una lógica interna sólida. Sin embargo, su alta dependencia en un pequeño número de variables podría ser un riesgo si estas cambian su distribución en el futuro (data drift).

3. Explicabilidad Local: Análisis de 5 Casos de Estudio

Aquí demostramos el principio de que "One Explanation Does Not Fit All", analizando por qué el modelo tomó decisiones específicas en casos de éxito y de error.

Caso de Estudio 1: Falso Positivo (Alerta de Fraude Incorrecta)

Análisis del error más común del modelo: ¿Por qué se alertó sobre una transacción legítima?

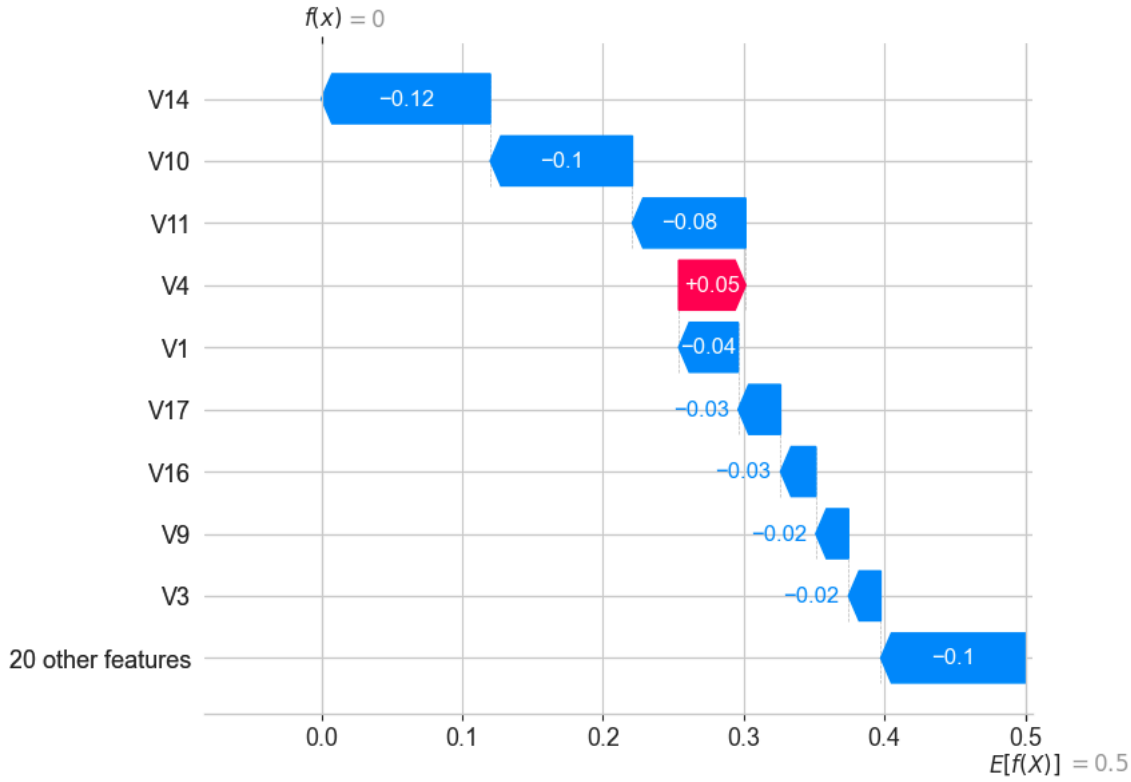


Análisis: Para esta transacción normal (no fraudulenta), el modelo generó una falsa alarma principalmente porque el valor de `V10`, `V14` y `V12` eran extremadamente bajo, un patrón que globalmente asocia con fraude.

Otro factor importante a observar es que el resto de características no intentan contrarrestar esta decisión o poco logran influir en la misma. Este caso muestra un posible punto de mejora: el modelo podría estar dando demasiado peso a `V10`, `V14` y `V12` y necesita aprender a considerar un contexto más amplio.

Caso de Estudio 2: Falso Negativo (Fraude No Detectado)

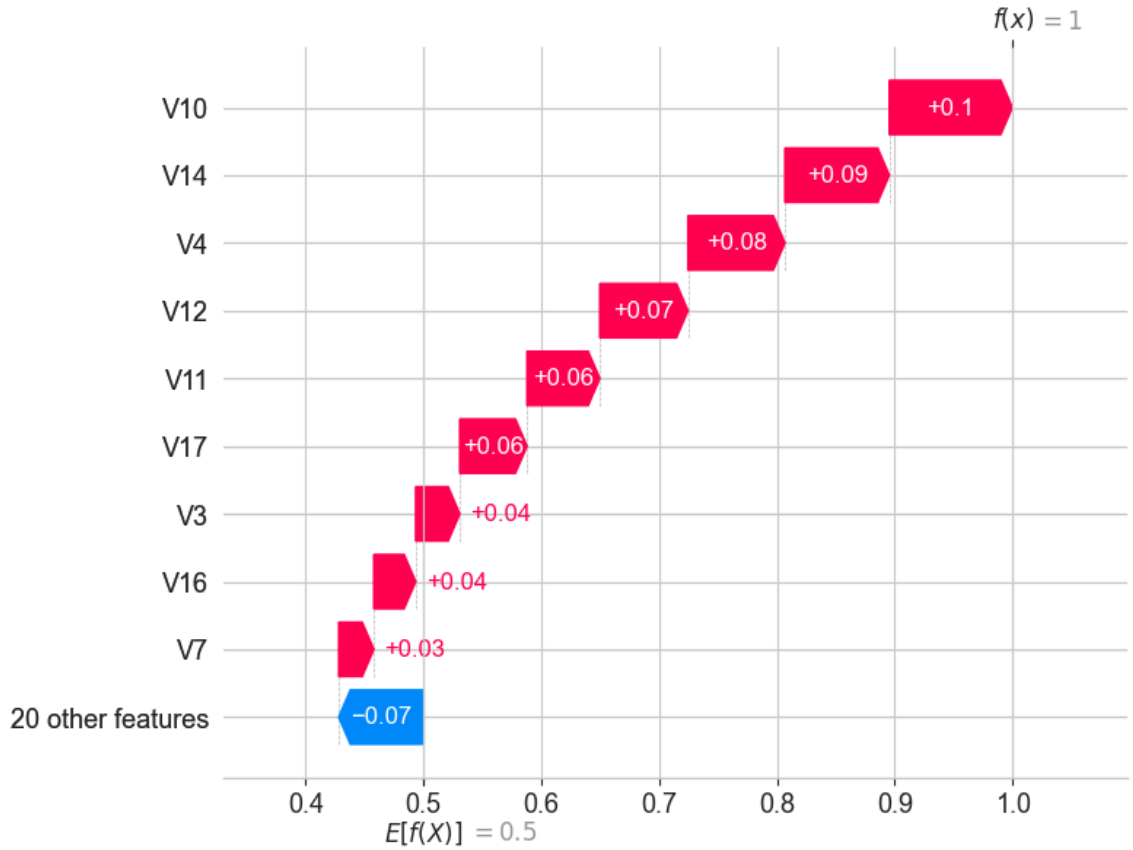
Análisis del error más costoso: ¿Por qué se nos escapó este fraude?



Análisis: Este es un caso de un fraude "silencioso". A pesar de ser una transacción fraudulenta, ninguna de sus features activó las alarmas principales del modelo. De hecho, `V14` y `V10` tenían valores que empujaban la predicción hacia la normalidad. Esto sugiere que este podría ser un **nuevo tipo de patrón de fraude** que el modelo no ha aprendido, o uno muy sofisticado. Estos casos son los más valiosos para el re-entrenamiento del modelo.

Caso de Estudio 3: Verdadero Positivo (Éxito del Modelo)

Confirmación del comportamiento esperado: ¿Por qué el modelo detectó correctamente este fraude?



Análisis: El éxito del modelo en este caso es claro. Múltiples features (`V14` , `V4` , `V10`) contribuyeron fuertemente a una alta probabilidad de fraude, alineándose con los patrones globales que el modelo ha aprendido. Esto confirma que el modelo está funcionando como se esperaba para los tipos de fraude que conoce.

---

4. Conclusiones y Recomendaciones Accionables

El análisis de explicabilidad nos ha permitido ir más allá de las métricas de rendimiento de entrenamiento del modelo para entender el "cómo" y el "porqué" de sus decisiones. Basado en los hallazgos, se proponen las siguientes acciones para mejorar el sistema:

1. **Investigación de Varaibles Clave:** La alta dependencia en `V14` , `V10` y `V4` es un riesgo. La primera acción debe ser colaborar con los expertos de negocio para **entender qué representan estas variables**. Este conocimiento permitirá validar si la lógica del modelo es sólida desde una perspectiva de negocio y guiará la creación de nuevas características o ajustarlas para hacer sentido de la mecánica del negocio.

2. **Análisis de Errores:**

    - **Falsos Positivos:** Se debe realizar un análisis profundo sobre una muestra representativa de los Falsos Positivos con mayor confianza. Si, como en nuestro caso de estudio, `V10` es consistentemente el culpable, se podría considerar la

creación de reglas de negocio que modulen las predicciones del modelo (ej. "si V10 es bajo pero V4 es normal, reducir la puntuación de fraude").

- **Falsos Negativos:** Estos sí conviene revisar cada caso a detalle, ya que cada Falso Negativo debe ser analizado por un experto en fraude. Los patrones identificados en estos casos deben ser utilizados para **crear nuevas variables** que capturen estas nuevas tácticas de fraude para el próximo re-entrenamiento del modelo.

3. **Implementar un Sistema de Monitoreo de Explicaciones:** Para un sistema en producción (MLOps), no es suficiente monitorear solo la precisión. Se debe implementar un sistema que monitoree la **distribución de los valores SHAP** a lo largo del tiempo. Un cambio repentino en la importancia de las features (ej. si `V5` de repente se vuelve muy importante) es un indicador clave de **concept drift** y una señal de que el modelo necesita ser re-evaluado.

Este enfoque basado en la explicabilidad nos permite crear un ciclo de mejora continua, haciendo que el modelo no solo sea más preciso, sino también más robusto, confiable y transparente.