Modelos Supervisados ML

Fraud Detection

Introducción: Detección de Anomalías en Transacciones Financieras

El presente estudio se enfoca en la identificación automatizada de transacciones fraudulentas mediante técnicas de aprendizaje supervisado. En el sector bancario, la detección temprana de fraude es un desafío crítico no solo por las pérdidas económicas directas, sino por la necesidad de mantener la confianza del usuario y la integridad operativa del sistema financiero.

Metodología y Datos

Para este análisis se utilizó el Credit Card Fraud Detection Dataset, el cual contiene transacciones realizadas por titulares de tarjetas europeas.

- Volumen de Datos: El dataset comprende un total de 284,807 transacciones.
- Naturaleza de las Variables: Debido a motivos de confidencialidad, las características originales han sido transformadas mediante un Análisis de Componentes Principales (PCA), resultando en 28 variables numéricas (V1 a V28).
- Desbalanceo Extremo: El dataset es altamente asimétrico; solo 492 transacciones (0.17%) están etiquetadas como fraude, frente a 284,315 transacciones normales. Esta disparidad define la complejidad técnica del proyecto.

Procesamiento y Modelado

El flujo de trabajo técnico se diseñó para mitigar el sesgo hacia la clase mayoritaria y garantizar la capacidad de detección:

1. Normalización: Se aplicó un escalado robusto a la variable Amount para alinear su magnitud con las componentes de PCA, eliminando la variable Time por su baja relevancia predictiva inicial.
2. Tratamiento del Desbalanceo (SMOTE): Se implementó la técnica SMOTE (Synthetic Minority Over-sampling Technique) para generar muestras sintéticas de fraude. Crítico: El balanceo se aplicó exclusivamente al conjunto de entrenamiento para evitar el Data Leakage.

3. División de Datos: Se utilizó una partición estratificada de 80% entrenamiento y 20% prueba, preservando la proporción de fraude en ambos conjuntos.
4. Evaluación Comparativa: Se entrenaron y contrastaron tres modelos con distintos enfoques algorítmicos:
    a. **Logistic Regression**: Evaluado por su eficiencia en problemas binarios.
    b. **Random Forest**: Utilizado por su robustez ante valores atípicos mediante bagging.
    c. **XGBoost**: Empleado por su alta capacidad de optimización mediante gradient boosting.

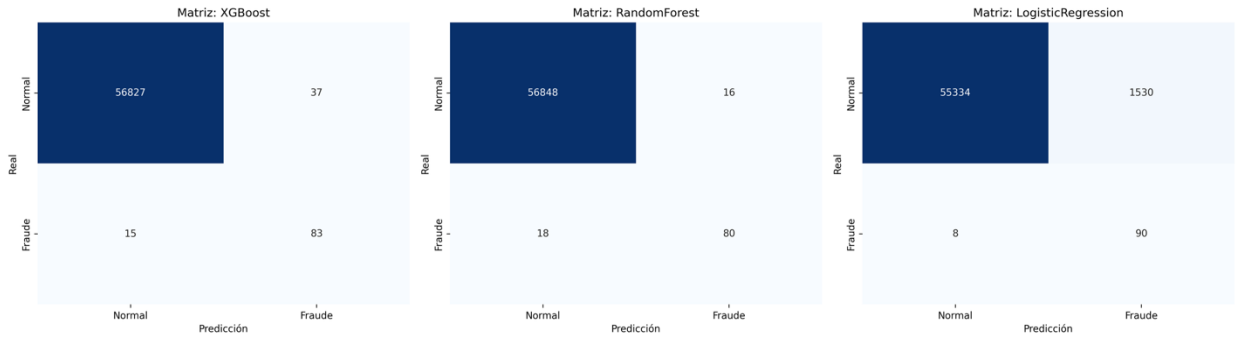
Objetivos del Análisis

Dado el desbalanceo extremo, el éxito del proyecto no se midió por la exactitud (Accuracy), sino por el Recall (capacidad de capturar el fraude) y la Precision (minimizar falsos bloqueos). La herramienta principal de decisión fue la Curva Precision-Recall (AUPRC), que proporciona una medida de rendimiento mucho más honesta que la curva ROC en este dominio específico.

Para la selección del mejor modelo, se evaluaron tres algoritmos bajo el mismo conjunto de prueba (85,443 transacciones originales). Los resultados se resumen en la siguiente tabla de la clase minoritaria (Fraude):

Modelo	Precisión	Recall (sensibilidad)	F1-score	Observación Principal
XGBoost	0.69	0.85	0.76	Mayor equilibrio para detección agresiva.
Random Forest	**0.83**	0.82	**0.82**	**Modelo más robusto y equilibrado.**
Logistic Regression	0.06	**0.92**	0.10	Inviable: Demasiadas falsas alarmas.

- **Modelos de Ensamble (Random Forest / XGBoost):** Ambos alcanzaron un AUPRC de 0.86, demostrando una superioridad técnica al manejar la complejidad no lineal del fraude.
- **Logistic Regression**: Mostró un rendimiento inferior (AUPRC = 0.76), evidenciando que una frontera de decisión lineal es insuficiente para separar los patrones sutiles de las transacciones fraudulentas sin generar un volumen inasumible de falsas alarmas.



El análisis de las matrices de confusión permite evaluar el desempeño de los modelos bajo la métrica de costo-beneficio operativa. En este escenario, el objetivo principal es maximizar la detección de fraude (Verdaderos Positivos) minimizando el impacto en clientes legítimos (Falsos Positivos).

Modelo	**Fraudes detectados (TP)**	**Fraudes omitidos (FN)**	**Falsas alarmas (FP)**	**Observación Principal**
XGBoost	83	15	37	**Recomendado para producción.**
Random Forest	80	18	**16**	Más preciso/menos fricción
Logistic Regression	**90**	**8**	1530	Inviable: Demasiadas falsas alarmas.

1. **XGBoost**: Equilibrio de Alto Rendimiento

El modelo XGBoost demuestra ser el más equilibrado para la operación bancaria:

- Detección de Fraude (Recall): Logró identificar 83 de los 98 fraudes presentes en el conjunto de prueba.
- Falsas Alarmas (Falsos Positivos): Solo generó 37 alertas erróneas en más de 56,000 transacciones normales.
- Impacto: Es el modelo con mejor precisión operativa, capturando el 85% del fraude con una fricción mínima para el cliente honesto.

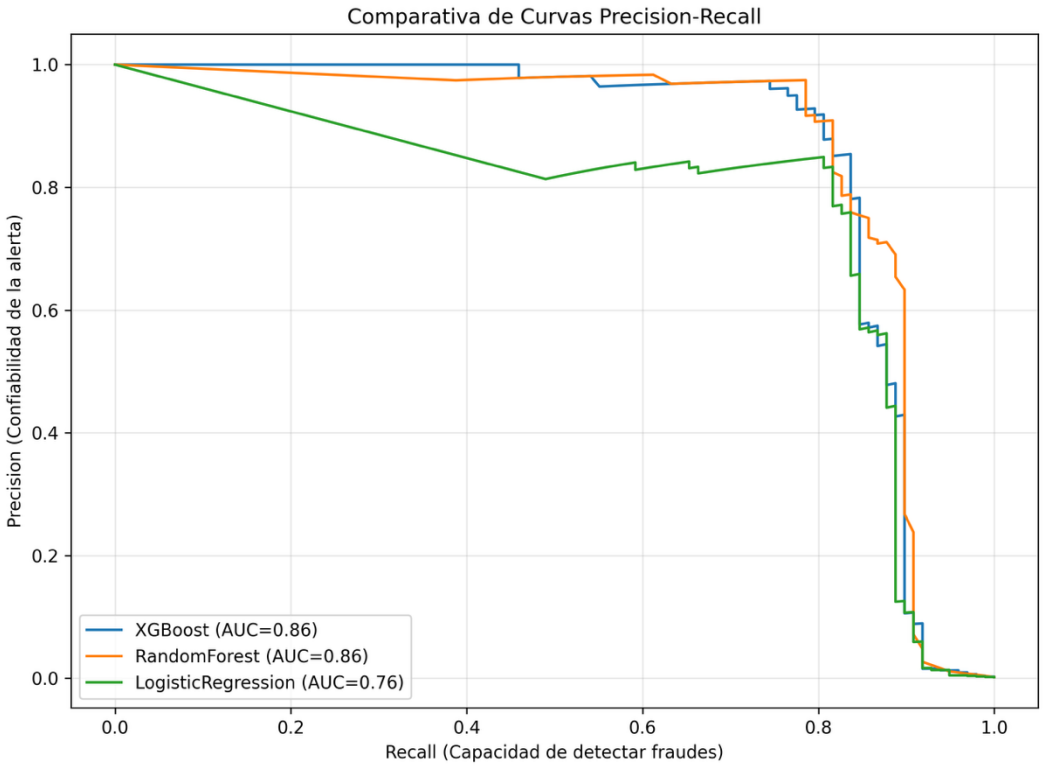
2. **Random Forest**: El Enfoque Conservador

- Muestra un comportamiento similar al XGBoost pero con una postura ligeramente más cautelosa:
- Falsas Alarmas: Es el modelo más preciso del grupo, con solo 16 falsos positivos.
- Detección de Fraude: Capturó 80 de los 98 fraudes.
- Impacto: Ideal si la prioridad absoluta del banco es evitar a toda costa el bloqueo injustificado de cuentas, aunque a cambio deje escapar 3 fraudes adicionales comparado con XGBoost.

3. **Logistic Regression**: El Modelo de Sensibilidad Extrema

La regresión logística ilustra perfectamente los peligros de un modelo con baja capacidad de discriminación en datos desbalanceados:

- Detección de Fraude: Alcanzó la cifra más alta detectando 90 de los 98 fraudes.
- Falsas Alarmas: Generó 1,530 falsos positivos.
- Impacto: Operativamente inviable. Aunque detecta más fraudes, el costo de investigar 1,530 alertas falsas colapsaría el departamento de seguridad y causaría una experiencia de usuario negativa masiva.



Al observar las curvas Precision-Recall para el ejercicio de detección de fraude revela una clara superioridad de los modelos basados en ensamble sobre el enfoque lineal. Tanto XGBoost como Random Forest alcanzaron un AUC de 0.86, manteniendo una precisión cercana al 100% incluso cuando el recall (capacidad de detección) supera el 80%, lo que es fundamental para evitar el bloqueo injustificado de tarjetas legítimas. En contraste, la Regresión Logística (AUC = 0.76) muestra una caída de precisión mucho más temprana y pronunciada; para lograr capturar la mayoría de los fraudes, este modelo sacrifica drásticamente la confiabilidad de sus alertas, generando un volumen de falsos positivos que resultaría operativamente inasumible en un entorno bancario real.

Sentiment Analysis

Introducción: Análisis de Sentimiento en Reseñas de Películas (IMDB)

El presente estudio aborda el problema de la clasificación binaria de texto mediante el análisis de sentimiento, una de las tareas fundamentales del Procesamiento de Lenguaje Natural (NLP). El objetivo principal es evaluar la capacidad de distintos algoritmos de aprendizaje automático para distinguir entre opiniones positivas y negativas en un entorno de lenguaje natural complejo y subjetivo.

Metodología y Datos

Para este experimento se utilizó el IMDB Dataset, una colección ampliamente reconocida en el ámbito académico que consta de 50,000 reseñas de películas altamente polarizadas. El conjunto de datos presenta una distribución equilibrada, con 25,000 muestras positivas y 25,000 negativas, lo que permite una evaluación directa de las métricas de rendimiento sin necesidad de técnicas de remuestreo.

Procesamiento y Modelado

El flujo de trabajo técnico se estructuró en tres fases críticas:

1. Vectorización: Se empleó la técnica TF-IDF (Term Frequency-Inverse Document Frequency), limitando el vocabulario a las 5,000 palabras más significativas, para transformar el texto no estructurado en una matriz numérica de alta dimensionalidad.
2. División de Datos: Se realizó una partición del dataset utilizando un 80% para entrenamiento (40,000 muestras) y un 20% para validación (10,000 muestras), asegurando la reproducibilidad mediante una semilla aleatoria.
3. Evaluación Comparativa: Se implementaron y compararon tres arquitecturas de modelos con fundamentos matemáticos distintos:
    a. Logistic Regression: Como modelo de referencia lineal.
    b. Random Forest: Como representante de métodos de ensamble basados en bagging.
    c. XGBoost: Como algoritmo de gradient boosting de alto rendimiento.

Objetivos del Análisis

La evaluación no solo se centró en la precisión global (accuracy), sino que se priorizó el análisis de la Curva Precision-Recall y el Área Bajo la Curva (AUPRC). Estas métricas son fundamentales para entender el compromiso entre la sensibilidad del modelo al detectar sentimientos y la confiabilidad de sus predicciones, proporcionando una visión integral de la robustez de cada arquitectura ante datos de texto.

Modelo	**Precisión**	**Recall (sensibilidad)**	**F1-score**	**Observación Principal**
Logistic Regression	**0.88**	**0.90**	**0.89**	**Óptimo y rápido.**
Random Forest	0.86	0.84	0.85	Sobreajusta un poco.
XGBoost	0.84	0.88	0.86	Menos efectivo con texto

En este ejercicio de clasificación de texto, se observa que la Regresión Logística ofrece el mejor rendimiento global. Esto demuestra que para tareas de Análisis de Sentimiento con representaciones TF-IDF, un modelo lineal robusto puede capturar mejor la semántica de las palabras clave que los modelos basados en árboles de decisión, los cuales requieren mayor ajuste de hiperparámetros para no caer en el sobreajuste (overfitting) ante vocabularios extensos.

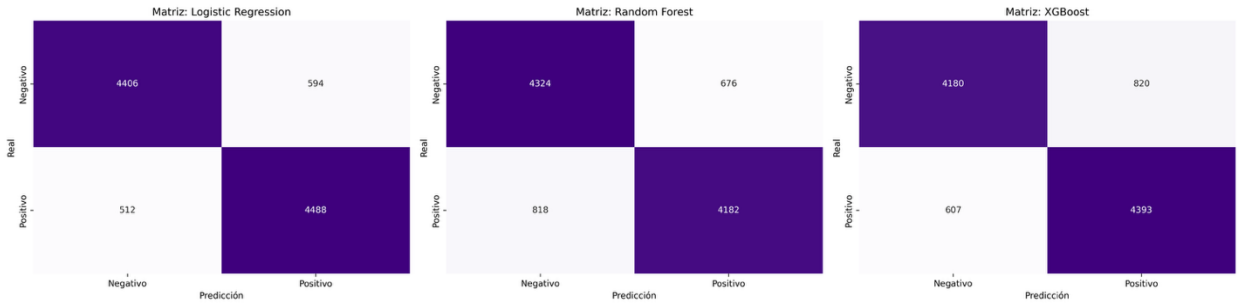


Tabla resumen de los datos mostrados en las matrices de confusión:

Modelo	**Falsos positivos (ruido)**	**Falsos negativos (omisión)**	**Error total**
Logistic Regression	**594**	**512**	**1106**
XGBoost	820	607	1427
Random Forest	676	818	1494

1. **Logistic Regression**: El equilibrio óptimo

Es el modelo con mejor desempeño visual y numérico.

- Falsos Positivos (594): Clasificó erronéamente 594 reseñas negativas como positivas.
- Falsos Negativos (512): Fue el modelo que menos reseñas positivas "perdió", detectando correctamente 4,488 de ellas.

2. **Random Forest**: Sesgo hacia la clase negativa

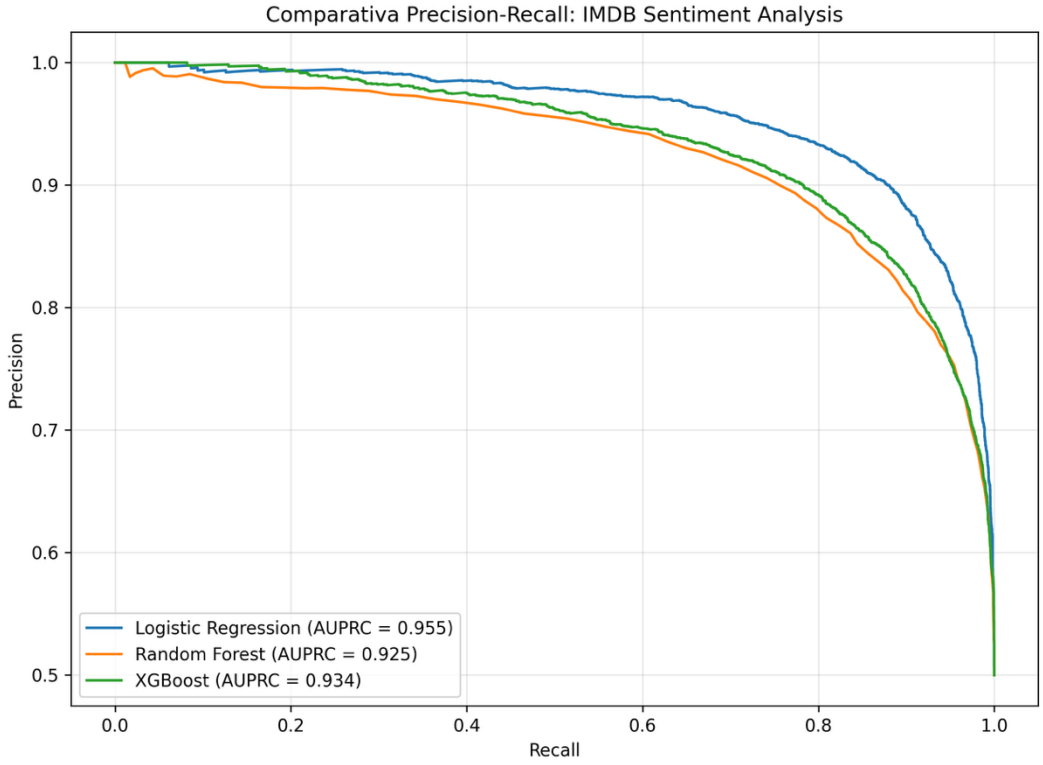
Este modelo muestra una dificultad mayor para identificar reseñas positivas.

- Falsos Negativos (818): Es la cifra más alta de las tres matrices. Esto indica que el Random Forest es el modelo más "pesimista"; tiende a calificar como negativas muchas reseñas que en realidad son positivas.
- Verdaderos Positivos (4,182): Tiene el acierto más bajo en la clase positiva.

3. **XGBoost**: El enfoque en la clase positiva

XGBoost presenta un comportamiento inverso al Random Forest.

- Falsos Positivos (820): Es el valor más alto en este cuadrante. El modelo tiende a ser más "optimista", clasificando más reseñas negativas como positivas de lo debido.
- Recall Positivo: Compensó sus errores detectando 4,393 reseñas positivas, superando al Random Forest en esta categoría, aunque sin alcanzar la eficiencia de la Regresión Logística.



Como se observa en la comparativa de curvas Precision-Recall, la Regresión Logística domina el espacio con un AUPRC de 0.955. En problemas de clasificación de texto con vectores de alta dimensionalidad (TF-IDF), los modelos lineales suelen generalizar mejor que los ensambles de árboles. La curva azul demuestra que podemos maximizar la detección de sentimientos sin inundar el sistema con falsos positivos, superando la complejidad estructural de XGBoost y Random Forest.