UNIVERSIDAD DEL VALLE DE GUATEMALA

Facultad de Ingeniería

Security Data Science

Entrenamiento Incremental en Modelos de Deep Learning y Machine Learning

MANUEL ALEJANDRO ARCHILA MORAN 161250

Primera Parte

Entrenamiento Incremental

El entrenamiento incremental en machine learning es una estrategia de entrenamiento en la que un modelo de aprendizaje automático se actualiza continuamente con nuevos datos sin necesidad de reentrenar el modelo desde cero. Esta técnica es especialmente útil en situaciones donde los datos fluyen continuamente o cuando es impracticable almacenar todo el conjunto de datos en memoria debido a su tamaño.

Algunos beneficios del entrenamiento incremental son:

- Flexibilidad: Permitiendo al modelo adaptarse a datos a medida que fluyen dentro del modelo
- Continuidad: Ofrece una manera de poder actualizarse sin tener que empezar desde cero
- Eficiencia de recursos: Al usar conjuntos de datos más pequeños se reduce el tiempo y la cantidad de poder computacional que requiere el aprendizaje.

El principal desafío del entrenamiento incremental es el problema del olvido catastrófico, donde un modelo puede olvidar información aprendida previamente al ser actualizado con nuevos datos. Soluciones como la regularización, el uso de técnicas de ensamblaje, o métodos específicos de aprendizaje incremental pueden ayudar a mitigar este problema. (Toolify.ai, 2023)

Investigación de Modelos

Redes Neuronales Artificiales (ANN)

Capacidades:

- **Adaptabilidad a nuevos datos**: Las ANN pueden adaptarse eficazmente a nuevos datos mediante el ajuste de pesos, lo cual es crucial en entornos dinámicos.
- **Capacidad de modelar relaciones no lineales complejas**: Pueden capturar relaciones intrincadas en los datos, lo que las hace adecuadas para tareas como el reconocimiento de imágenes y el procesamiento del lenguaje natural.
- **Soporte para diferentes estructuras de datos**: Son versátiles en términos de los tipos de datos de entrada que pueden manejar, incluyendo datos secuenciales y matriciales.

Limitaciones:

- **Requerimiento de grandes cantidades de datos para entrenamiento efectivo**: Esto puede ser un problema en entornos de entrenamiento incremental donde los datos llegan en lotes pequeños o uno a uno.
- **Riesgo de sobreajuste**: En escenarios de entrenamiento incremental, es posible que las ANN se ajusten demasiado a los últimos datos recibidos, perdiendo generalización.
- **Consumo computacional elevado**: El ajuste constante de los pesos y el procesamiento de nuevas entradas en tiempo real pueden requerir una gran cantidad de recursos computacionales. (FasterCapital, 2024)

Bosques Aleatorios (Random Forest)

Capacidades:

- **Robustez frente a ruido y datos atípicos**: Los bosques aleatorios son menos sensibles al ruido y a los datos atípicos, lo que los hace estables en entornos de entrenamiento incremental.
- **Facilidad para manejar características categóricas y continuas**: Pueden manejar de forma natural diferentes tipos de datos, lo que es útil en aplicaciones prácticas con entrenamiento incremental.
- **Buen desempeño en clasificación y regresión**: Ofrecen un buen equilibrio entre precisión y capacidad de generalización, lo cual es valioso al integrar nuevos datos progresivamente.

Limitaciones:

- **Complejidad en el tiempo de entrenamiento**: Aunque pueden manejar nuevos datos de manera incremental, el reentrenamiento de múltiples árboles puede ser computacionalmente costoso.
- **Dificultad en la actualización de modelos**: No están diseñados intrínsecamente para el entrenamiento incremental y pueden requerir métodos específicos para actualizar los modelos con nuevos datos.
- **Overfitting en escenarios de datos limitados**: Aunque generalmente son buenos evitando el sobreajuste, en situaciones con datos muy limitados o altamente correlacionados, pueden sobreajustarse a esos datos específicos. (INESDI, 2023)

Metodología para Reentrenamiento o Entrenamiento Incremental

1. **Evaluación del Volumen y la Variabilidad de los Nuevos Datos**
    - Volumen: Si el volumen de nuevos datos es relativamente pequeño comparado con el conjunto de datos original, el entrenamiento incremental podría ser adecuado. Esto permite actualizar el modelo sin el costo computacional de un reentrenamiento completo.

- Variabilidad: Si los nuevos datos representan cambios significativos o introducen características nuevas que podrían afectar la decisión del modelo, considera un reentrenamiento completo. Esto es particularmente importante si se sospecha que ha ocurrido un "drift" o desplazamiento en los datos.

2. **Evaluación del Rendimiento Actual del Modelo**
    - Realiza pruebas de rendimiento regulares con los datos actuales. Si el modelo sigue cumpliendo o superando los benchmarks de rendimiento, un entrenamiento incremental puede ser suficiente.
    - Si el rendimiento del modelo ha decaído significativamente, especialmente en los nuevos datos, esto podría indicar la necesidad de un reentrenamiento completo.
3. **Complejidad y Tipo de Modelo**
    - Modelos Simples: Modelos más simples como regresiones logísticas pueden ser más fáciles y rápidos de reentrenar completamente.
    - Modelos Complejos: Modelos como redes neuronales profundas o grandes ensambles pueden beneficiarse más del entrenamiento incremental debido a su alto costo computacional y tiempo de entrenamiento.
4. **Costo Computacional y Recursos Disponibles**
    - Considera los recursos computacionales disponibles. El reentrenamiento completo de modelos complejos puede ser prohibitivamente costoso en términos de tiempo y recursos computacionales.
    - El entrenamiento incremental puede ser una solución más eficiente si los recursos son limitados y el modelo no ha perdido significativamente en rendimiento.

Implementación Práctica

Para la parte práctica de este proyecto se probó con una iteración inicial de cada uno de los siguientes modelos:

- Artificial Neural Networks
- Random Forest
- LightGBM
- XG Boost
- Support Vector Machines

Sin embargo el rendimiento de algunos de estos modelos era deficiente en algunas métricas por lo que se seleccionó únicamente LightGBM y XGBoost para la continuación de este proyecto.

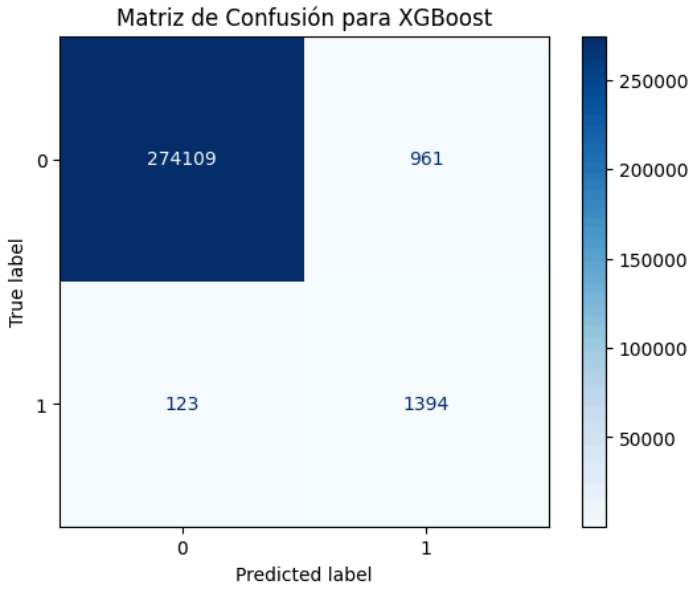
Evaluación de los Modelos

Resultados

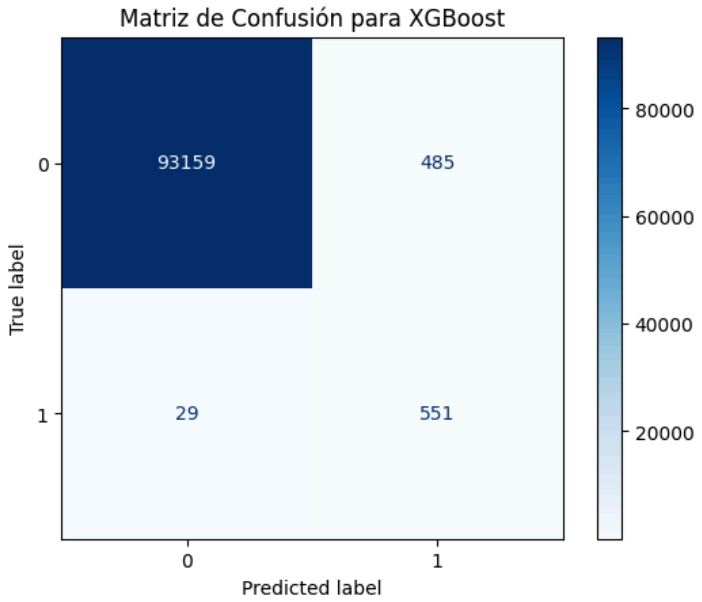
Modelo	AUC-ROC	F1-Score	Precisión	Recall	Accuracy
XG Boost	0.9946	0.7200	0.5919	0.9181	0.9961
XG boost reentrenado	0.9976	0.6819	0.5319	0.9500	0.9945
LightGBM	0.9933	0.8189	0.7500	0.9018	0.9978
LightGBM incremental	0.9558	0.1530	0.0832	0.9500	0.9353

Matrices de confusión

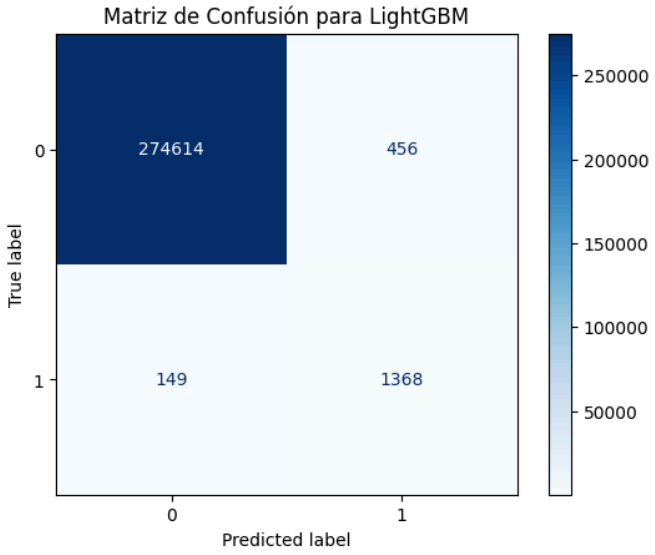
XG Boost



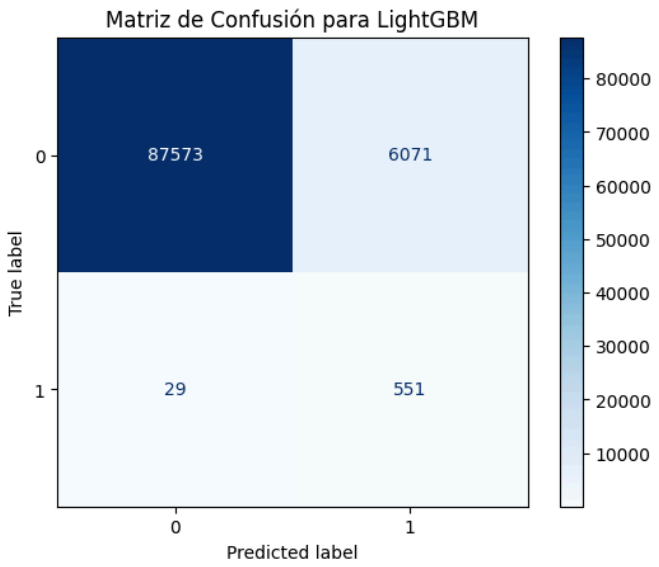
XG Boost reentrenado



LightGBM



LightGBM incremental



Al observar las métricas de rendimiento y las matrices de confusión, la primera iteración de los modelos devuelve resultados aceptables para XG Boost y LightGBM. Sin embargo, en ambos modelos se puede observar que la métrica más baja se trata de la precisión. Esta métrica se basa en la detección de verdaderos positivos. Esto se puede ver respaldado en las matrices de confusión, mostrando que varios resultados positivos son catalogados como falsos negativos. A pesar de esto, los modelos presentan métricas suficientemente altas para ser considerados como buenos.

Para la segunda iteración de XG Boost se decidió reentrenarlo completamente,ya que para entrenarlo de manera incremental tanto los datos como el modelo tendrían que ser transformados. Esta transformación presentó un reto, ya que afectaba de manera considerable las predicciones. De esta manera el modelo logró desempeñarse similar al modelo original. Es importante mencionar algunas métricas como el F1-Score y la precisión bajaron considerablemente, mientras que la curva ROC, Accuracy y Recall presentaron un aumento. Para finalizar con este modelo se observó la matriz de confusión, la cual presentaba pocos verdaderos positivos, ya que el modelo clasificaba muchos de los datos de prueba como transacciones no fraudulentas. Por último, este modelo fue el que mejores resultados presentó entre los dos seleccionados.

Finalmente para el modelo de LightGBM fue posible realizar el entrenamiento incremental sin necesidad de aplicar alguna transformación. A pesar de contar con él fue el modelo con las peores predicciones. Al tener una precisión de 0.08 es posible decir que el modelo es incapaz de clasificar transacciones fraudulentas y transacciones no fraudulentas. De igual manera al ver la matriz de confusión es más de 6000 transacciones fraudulentas fueron marcadas como no fraudulentas. Esto indica que el modelo es únicamente capaz de detectar transacciones no fraudulentas debido a la gran cantidad de transacciones no fraudulentas usadas en el entrenamiento.

Segunda parte

Metodología

1. Evaluación del Rendimiento del Modelo Actual
    - Monitoreo Continuo: Establecer un sistema de monitoreo continuo para evaluar el rendimiento del modelo en tiempo real o en intervalos regulares.
2. Análisis Temporal del Entrenamiento
    - Frecuencia de Entrenamiento: Registrar y analizar la frecuencia con la que se han llevado a cabo reentrenamientos totales e incrementales anteriores.
    - Decaimiento del Rendimiento en el Tiempo: Estudiar cómo decae el rendimiento del modelo con el tiempo, lo que puede ayudar a establecer un calendario óptimo para el reentrenamiento total basado en el tiempo desde el último entrenamiento total.
3. Identificación de Nuevas Tendencias en los Datos
    - Detección de Cambios: Implementar algoritmos de detección de cambios para identificar cambios significativos en la distribución de los datos de entrada que podrían hacer que el modelo actual quede obsoleto.
4. Evaluación de Costos y Beneficios
    - Costos de Entrenamiento: Comparar los costos computacionales y de tiempo entre el reentrenamiento total y el incremental.

Conclusiones y recomendaciones

Conclusiones

- XG Boost mostró un desempeño superior en general, especialmente en métricas como AUC-ROC y precisión después de un reentrenamiento completo, aunque hubo una disminución en F1-Score y precisión. El recall mejoró significativamente, lo que indica una mejor capacidad para detectar transacciones fraudulentas. LightGBM, por otro lado, tuvo un rendimiento considerablemente peor en la iteración de entrenamiento incremental. La precisión extremadamente baja sugiere que el modelo fue ineficaz para clasificar correctamente las transacciones fraudulentas.

- La implementación del entrenamiento incremental para LightGBM no resultó efectiva, lo que resultó en una precisión y F1-Score muy bajos. Esto sugiere que el modelo no se adaptó bien a los nuevos datos sin una reconfiguración significativa.

- El reentrenamiento completo de XG Boost demostró ser efectivo, logrando mantener un alto nivel de exactitud y mejorar el recall. Esto implica que, para ciertos modelos y

contextos, el reentrenamiento completo puede ser más adecuado para adaptarse a cambios significativos en los datos.

Recomendaciones

- Implementar un sistema robusto de monitoreo y evaluación continua para todos los modelos para identificar rápidamente la necesidad de reentrenamientos. Esto incluye el seguimiento de métricas de rendimiento críticas como precisión, recall, y AUC-ROC.
- Realizar análisis regulares sobre el volumen y la naturaleza de los nuevos datos para determinar la estrategia de entrenamiento más apropiada. Si los nuevos datos presentan variaciones significativas o introducen nuevas características, preferir el reentrenamiento completo para evitar el deterioro del rendimiento del modelo
- Para abordar el desequilibrio de clases y mejorar el rendimiento del modelo, se recomienda una estrategia combinada de oversampling y undersampling. Esta técnica equilibra las clases, aumentando la representación de las minoritarias y reduciendo ejemplos excesivos de las mayoritarias, lo que disminuye el riesgo de sobreajuste y mejora la eficiencia computacional. Asegúrate de validar estas mejoras con pruebas rigurosas como la validación cruzada para confirmar la efectividad de esta estrategia.

Referencias:

FasterCapital. (2024). Ventajas Y Limitaciones Del Uso De Redes Neuronales Artificiales Para El Modelado De Riesgo Crediticio - FasterCapital. Retrieved May 20, 2024, from FasterCapital website: https://fastercapital.com/es/tema/ventajas-y-limitaciones-del-uso-de-redes-neuronales-artificiales-para-el-modelado-de-riesgo-crediticio.html

INESDI. (2023). Random forest, la gran técnica de Machine Learning. Retrieved May 20, 2024, from Inesdi website: https://www.inesdi.com/blog/random-forest-que-es/

Toolify.ai. (2023, November 15). Aprende a realizar aprendizaje incremental en Machine Learning. Retrieved May 25, 2024, from Toolify.ai website: https://www.toolify.ai/es/ai-news-es/aprende-a-realizar-aprendizaje-incremental-en-machine-learning-466396