Javier Mombiela
Guatemala, 20 de mayo 2024

Entrenamiento Incremental en Modelos de Deep Learning y Machine Learning

Investigación Teórica

Para este proyecto, se seleccionaron XGBoost y LightGBM debido a su eficacia en problemas de clasificación y su capacidad para adaptarse a diferentes tipos de datos. Ambos algoritmos utilizan la metodología de *gradient boosting*, lo que los hace ideales para la detección de fraudes en transacciones financieras. (Saha, S. 2024).

El algoritmo *Extreme Gradient Boostin*g o XGBoost es potente y eficiente en términos de tiempo de entrenamiento y recursos computacionales. Utiliza un algoritmo de optimización de gradientes que permite actualizaciones incrementales con nuevos datos, siendo ideal para el entrenamiento incremental. Sin embargo, su complejidad aumenta con la adición de más árboles, lo que puede resultar en tiempos de entrenamiento más largos y requerimientos de memoria. (Jana, M. 2023).

Similar a XGBoost, el algoritmo *Light Gradient Boosting Machine* o LightGBM se destaca por su eficiencia y velocidad de entrenamiento. Su algoritmo basado en histogramas acelera el proceso de entrenamiento y reduce el uso de memoria, lo que lo hace adecuado para grandes conjuntos de datos. LightGBM también puede manejar el entrenamiento incremental de manera efectiva, pero puede tener un rendimiento peor que un reentrenamiento completo, dependiendo de los parámetros de entrenamiento y la validación de las predicciones (StackExchange, 2023).

Descripción de la implementación práctica

Para procesar los datos, se imputaron valores faltantes y se escalaron las características numéricas. Luego, los datos se dividieron en conjuntos de entrenamiento, desarrollo y prueba. Para abordar el desbalance de clases, se aplicó SMOTE al conjunto de entrenamiento, creando un conjunto de datos de entrenamiento resampleado con una proporción manejable de transacciones fraudulentas respecto a las legítimas (1:5), para mejorar la robustez de los modelos en la detección de fraudes.

La implementación del modelo XGBoost comienza con la conversión de los datos de entrenamiento resampleados y los datos de prueba al formato DMatrix de XGBoost. Luego, se definen los parámetros del modelo, configurando el objetivo como binary:logistic para clasificación binaria y la métrica de evaluación como logloss para medir la pérdida logística. Con estos parámetros y los datos preparados, se entrena el modelo utilizando 100 rondas de actualización.

La implementación del modelo LightGBM comienza con la conversión de los datos de entrenamiento resampleados al formato Dataset de LightGBM. A continuación, se definen los parámetros del modelo, configurando el objetivo como binary para clasificación binaria y la métrica de evaluación como binary_logloss para medir la pérdida logística. Con estos parámetros establecidos, se entrena el modelo utilizando 100 rondas de actualización.

Se exploraron estrategias de reentrenamiento incremental para mantener la efectividad de los modelos en entornos cambiantes. Se llevaron a cabo tres enfoques de reentrenamiento:

- Reentrenamiento Semestral: Se actualiza el modelo cada seis meses utilizando los datos más recientes disponibles del año 2020.
- Reentrenamiento Trimestral: Se actualiza el modelo cada tres meses utilizando los datos más recientes disponibles del año 2020.
- Reentrenamiento Total: Se entrena un nuevo modelo desde cero con todos los datos disponibles hasta el momento (2019 y 2020).

Análisis de los resultados

Se implementó un umbral basado en la curva ROC para maximizar el recall en la detección de transacciones fraudulentas. Aunque esto puede aumentar los falsos positivos, priorizamos la seguridad financiera al identificar con precisión las transacciones fraudulentas, incluso si significa clasificar algunas transacciones legítimas como sospechosas.

LGBM

Método	Recall Promedio	TN Promedio
LGBM Completo	93.61%	259,431
LGBM Semestral	87.80%	245,154
LGBM Trimestral	81.92%	229,201

Tabla 1: Rendimiento promedio de los 3 reentrenamientos LGBM

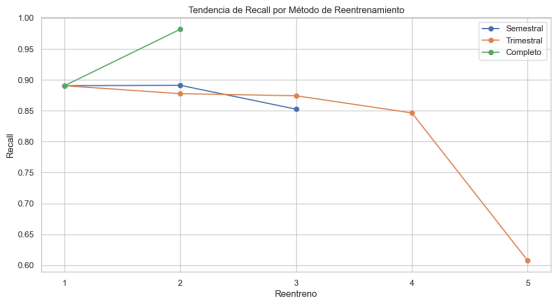

Gráfico 1: Tendencia del recall de cada reentrenamiento LGBM

El reentrenamiento total se destaca como el enfoque más efectivo, con un recall alcanzando el 98%, lo que sugiere una predicción casi perfecta de transacciones fraudulentas debido al acceso a todos los datos del conjunto.

En contraste, tanto el reentrenamiento semestral como trimestral muestran tendencias negativas en su rendimiento, con el semestral manteniéndose estable pero con una posible disminución futura.

El enfoque trimestral exhibe el peor rendimiento, con un recall disminuyendo hasta aproximadamente el 60%, atribuido a la menor disponibilidad de datos en cada reentrenamiento.

XGBoost

Método	Recall Promedio	TN Promedio
XGBoost Completo	84.71%	238,576
XGBoost Semestral	84.35%	237,505
XGBoost Trimestral	79.17%	240,865

Tabla 2: Rendimiento promedio de los 3 reentrenamientos XGBoost

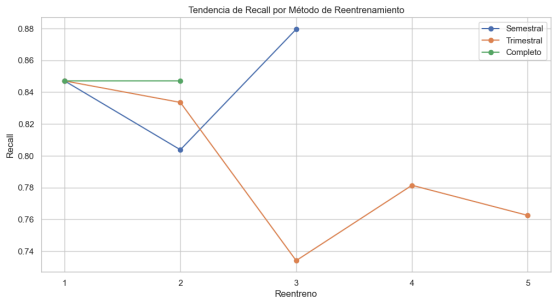

Gráfico 2: Tendencia del recall de cada reentrenamiento XGBoost

En el caso de XGBoost, tanto el reentrenamiento total como el semestral muestran tendencias positivas. Aunque el reentrenamiento total experimenta un aumento gradual, sugiriendo una viabilidad para este enfoque, se debe tener precaución debido a la posible sensibilidad de XGBoost a grandes cantidades de datos.

El enfoque semestral alcanza su punto máximo en el último reentrenamiento, pero el promedio entre ambos enfoques es similar, destacando la estabilidad del enfoque completo.

En contraste, el enfoque trimestral muestra un bajo rendimiento constante, con el promedio más bajo de todos los enfoques probados en ambos modelos.

Aunque XGBoost muestra mejoras en algunos reentrenamientos, su rendimiento general sigue siendo inferior al de LGBM, lo que sugiere que LGBM sigue siendo la mejor opción para predecir transacciones fraudulentas con precisión.

Metodología de reentrenamiento

En el proyecto se han explorado y comparado tres enfoques diferentes de reentrenamiento: reentrenamiento total, semestral y trimestral. Los resultados indican que, en el caso de LightGBM, el enfoque de reentrenamiento total supera consistentemente a los otros dos enfoques en términos de rendimiento, con una tendencia positiva en la capacidad de detección de transacciones fraudulentas a lo largo del tiempo. Esto respalda la recomendación de *StackExchange* de utilizar el reentrenamiento total como la metodología preferida para LightGBM.

Por otro lado, en el caso de XGBoost, si bien el enfoque de reentrenamiento total también muestra un rendimiento sólido, observamos que el enfoque semestral también puede proporcionar mejoras significativas en el rendimiento en comparación con el enfoque trimestral. Esto sugiere que el reentrenamiento semestral puede ser una opción eficaz si el reentrenamiento total no es factible debido a la cantidad excesiva de datos o recursos computacionales requeridos. Sin embargo, si se dispone de suficientes datos y recursos, se recomienda el reentrenamiento total para garantizar la máxima efectividad del modelo.

En general, la recomendación es utilizar un enfoque de reentrenamiento que tenga suficientes datos disponibles para mantener y mejorar la efectividad del modelo en la detección de transacciones fraudulentas. Esto significa que, en la mayoría de los casos, el reentrenamiento total será la mejor opción, ya que siempre tendrá datos suficientes para un buen rendimiento.

Sin embargo, si el reentrenamiento total resulta prohibitivamente costoso en términos de recursos, el reentrenamiento semestral puede ser una alternativa efectiva. Esto es debido a que, un semestre también tiene una gran cantidad de datos, lo cual podría resultar en un modelo eficiente. Adicionalmente, también es recomendable usar un reentrenamiento que cuente con más datos, como vendría siendo uno anual. Habiendo dicho esto, es importante mencionar que se tiene que evitar el enfoque trimestral, ya que los resultados indican que puede ser menos efectivo debido a la menor cantidad de datos disponibles en cada reentrenamiento.

Conclusiones

- El reentrenamiento total se muestra como la opción más efectiva, especialmente para modelos basados en LightGBM, al ofrecer una mejora continua en el rendimiento y una mayor efectividad en la detección de transacciones fraudulentas.
- El reentrenamiento semestral emerge como una alternativa viable en escenarios donde el reentrenamiento total no es factible, proporcionando mejoras significativas en el rendimiento en comparación con el enfoque trimestral, especialmente para modelos basados en XGBoost.
- El reentrenamiento trimestral no es opción en ninguno de los casos, ya que para ambos modelos, cada reentrenamiento resulto en una tendencia negativa. Esto es debido a la pequeña cantidad de datos que tiene cada reentrenamiento.

Recomendaciones

- Utilizar umbral basado en la curva ROC para mejorar la detección de fraudes, ajustando las métricas de los modelos para maximizar el recall y optimizar la detección de verdaderos negativos.
- Aplicar SMOTE con una proporción de 5 a 1 para abordar el desbalance de clases, generando datos sintéticos de transacciones fraudulentas y conservando los datos originales de transacciones no fraudulentas. Esta estrategia puede mejorar significativamente la capacidad del modelo para detectar transacciones fraudulentas mientras mantiene un bajo índice de falsos positivos.
- Se recomienda evitar el enfoque trimestral debido a la menor cantidad de datos disponibles en cada reentrenamiento, lo que puede afectar negativamente la efectividad del modelo en la detección de fraudes.

Bibliografía

GeeksForGeeks. (29 de abril de 2024). LightGBM (Light Gradient Boosting Machine). Recuperado de: https://www.geeksforgeeks.org/lightgbm-light-gradient-boosting-machine/

Jana, M. (11 de septiembre de 2023). Exploring Machine Learning Models: A Comprehensive Comparison of Logistic Regression, Decision Trees, SVM, Random Forest, and XGBoost. Recuperado de: https://medium.com/@malli.learnings/exploring-machine-learning-models-a-comprehensive-comparison-of-logistic-regression-decision-38cc12287055

Saha, S. (16 de abril de 2024). XGBoost vs LightGBM: How Are They Different. Recuperado de: https://neptune.ai/blog/xgboost-vs-lightgbm

StackExchange. (24 de septiembre de 2023). Recuperado de: https://stats.stackexchange.com/questions/453540/how-does-lightgbm-deals-with-incremental-learning-and-concept-drift