Universidad del Valle de Guatemala

Facultad de Ciencias e Ingeniería

Security Data Science

Proyecto 2 - Entrenamiento Incremental en Modelos de Deep Learning y Machine Learning

Javier Valle
Carnet 20159
Guatemala, 7 de abril de 2024

Parte 1 - Investigación teórica

Redes neuronales artificiales

Las redes neuronales son uno de los exponentes de la inteligencia artificial, que pertenecen propiamente al aprendizaje automático o el machine learning. Lo anterior permite que las máquinas puedan realizar tareas y/o funciones que antes no éramos capaces de automatizar con los sistemas que ya se tenían. Uno de los aspectos más importantes de este modelo es que trata de simular el comportamiento del cerebro humano.

Una red neuronal artificial está formada por neuronas artificiales, que son unidades que reciben información del exterior o de otras neuronas. Lo anterior simula hasta cierto punto los impulsos nerviosos que reciben las neuronas del cerebro humano, las procesan y generan un valor de salida que alimenta a otras neuronas de la red o son la salida hacia el exterior de la red. (Unir, 2023)

Capacidades:

- Las redes neuronales artificiales son capaces de poder procesar secuencias.
- Son capaces de detectar correlaciones entre los sets de datos que se les proporciona a las redes neuronales.(*4.13 - Ventajas Y Limitaciones de las Redes Neuronales Recurrentes Y LSTM | Codificando Bits*, s. f.)

Limitaciones:

- Los gradientes pueden crecer o disminuir sin ninguna limitación.
- La presencia de los gradientes que explotan y se desvanecen limitan la memoria de largo plazo.

Random forest

Random forest es un algoritmo de aprendizaje automático supervisado que generalmente se usa para solucionar problemas de clasificación y regresión. Este algoritmo construye árboles de decisión a partir de diferentes muestras y toma su voto mayoritario para decidir la clasificación y el promedio en caso de regresión. Algo bastante importante de este algoritmo es que puede manejar conjuntos de datos que contengan variables continuas, como es el caso de regresión; o variables categóricas como es el caso de la clasificación.

Capacidades:

- Se puede usar en problemas de clasificación y regresión.
- Funciona bien incluso si los datos contienen valores nulos/ausentes.
- Cada árbol de decisión creado es independiente del otro, por lo que este algoritmo muestra la propiedad de paralelización.

Limitaciones:

- Este algoritmo es bastante complejo en comparación con los árboles de decisión en los que se pueden tomar decisiones siguiendo la ruta del árbol.
- El tiempo de entrenamiento de este algoritmo es mayor en comparación con otros modelos, dado que es bastante complejo.
- (*Random Forest, la Gran Técnica de Machine Learning*, 2023)

Parte 2 - Criterios para Re-entrenamiento

Desarrollo de Metodología

Para poder realizar un reentrenamiento, se debe pensar en que los modelos que se tienen aún sigan haciendo predicciones acorde a lo esperado sin afectar las métricas de negocio que se tienen desde un principio. Es importante mencionar que el reentrenamiento funciona excelente cuando el modelo que se entrenó en la fase de research quedó obsoleto debido al concept drift y, por lo tanto, hay que re-entrenarlo. Asimismo, es importante mencionar que el re-entrenamiento es de alta eficiencia cuando se quiere evitar el trabajo manual de investigación de optimización de modelos que realizará un científico de datos en caso de que no exista un proceso de re-entrenamiento automático. (Neira, 2023)

Por otro lado, es importante mencionar que en algunas ocasiones es útil reentrenar el modelo en base al tiempo, por ejemplo se podría volver a entrenar cada 3 meses o cada 10,000 etiquetas nuevas. Lo anterior depende también del volumen de datos que se tienen en el dataset o de la situación que esté pasando, por ejemplo podría ser una pandemia (como el COVID) o alguna recesión a nivel nacional/mundial. (Neira, 2023)

Finalmente, es importante mencionar que los algoritmos de aprendizaje incremental se pueden utilizar cuando el origen de datos (o el dataset) es demasiado grande o los datos están sobre-muestreados. Los algoritmos de aprendizaje incremental se pueden utilizar para continuar el entrenamiento utilizando los datos restantes en un origen sobre-muestreado, dividiendo los datos restantes en un origen sub-muestreado dividiendo así los datos restantes en lotes, de ser necesario. Cada lote de datos de entrenamiento se puntúa de forma independiente utilizando la métrica optimizada, para que se pueda revisar el rendimiento de cada lote cuando se están explorando los resultados.(IBM Corporation, s. f.)

Validación Empírica

En el caso del dataset que se nos proveyó, se aplicó un entrenamiento incremental, dado que en el dataset origen existe un sobre-muestreo con respecto a los datos de fraude de tarjetas de crédito. Asimismo, es importante mencionar que a raíz de que se tienen dos años en el dataset (2019 y 2020) en los cuales el 2020 se tuvo una pandemia mundial, se optó por tomar dos

años completos para poder realizar el entrenamiento incremental de los modelos de ANN y de Random Forest.

Análisis de resultados

Redes neuronales 2019

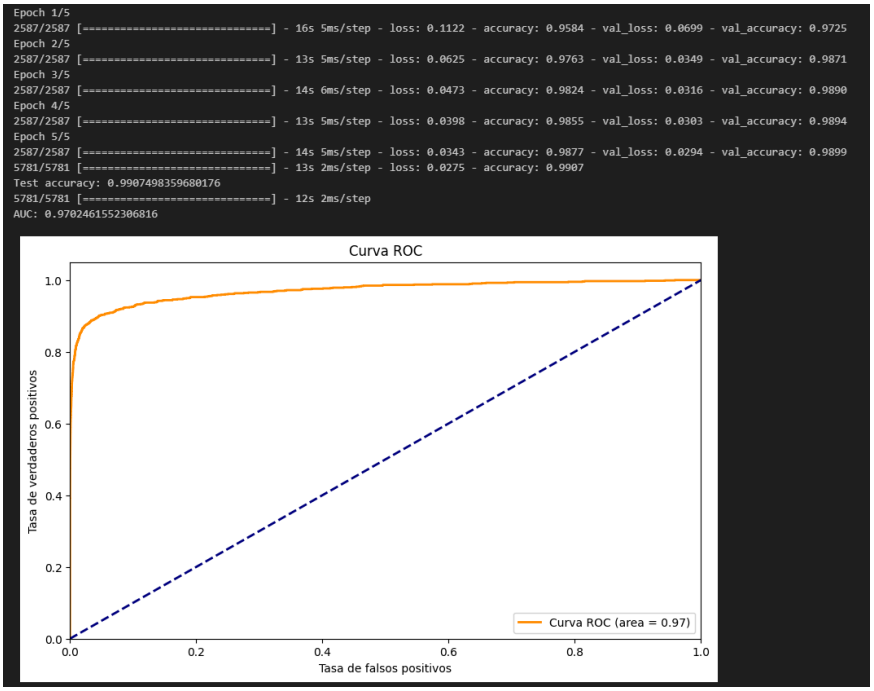


Figura 1: "Curva ROC" de las redes neuronales con los datos de 2019

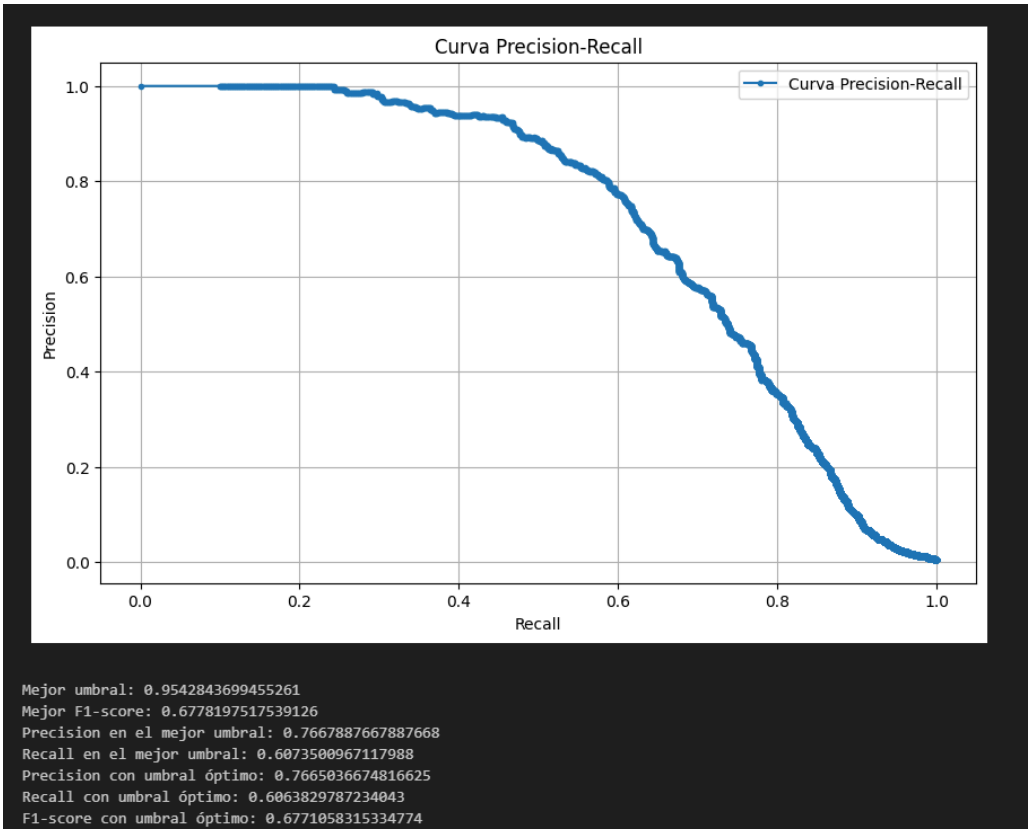
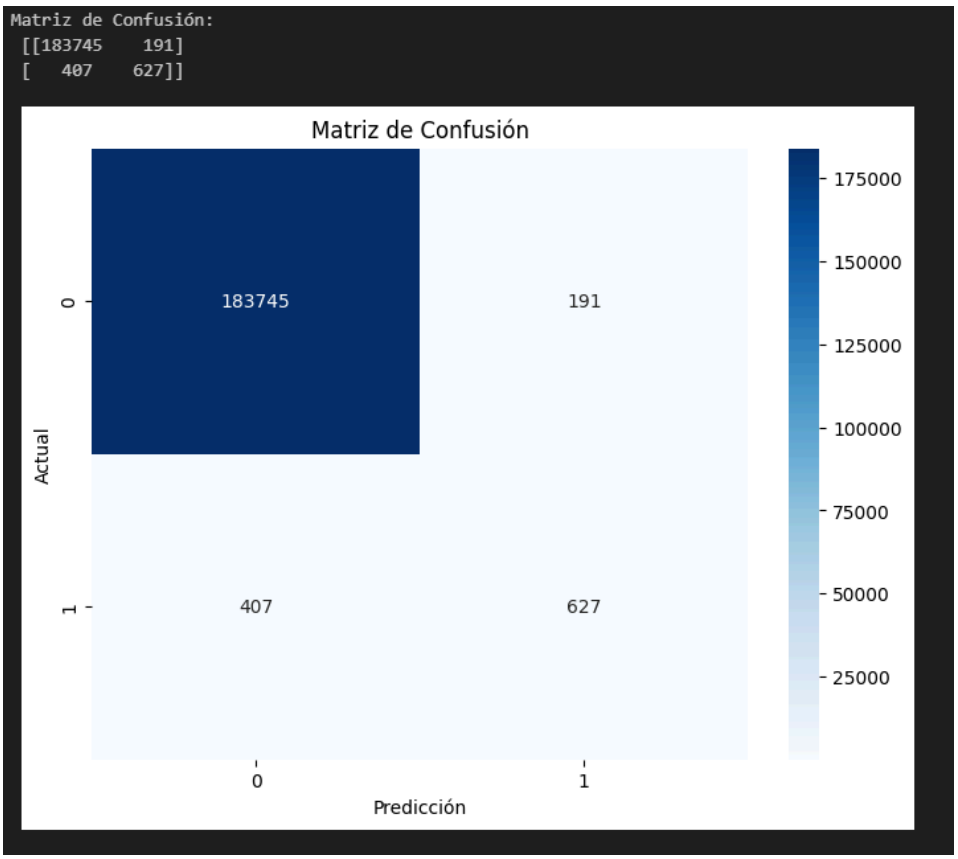


Figura 2: Curva "Precision Recall" y resultados de las redes neuronales con los datos de 2019

Figura 3: Matriz de confusión de las redes neuronales con los datos de 2019



En los resultados del modelo de ANN para el 2019, se logra observar que el modelo obtuvo resultados aceptables para el análisis de fraude en las transacciones realizadas en el año 2019. Esto quiere decir que logró detectar una cantidad promedio de fraudes cometidos durante el 2019 en los comercios a nivel mundial.

Random Forest 2019

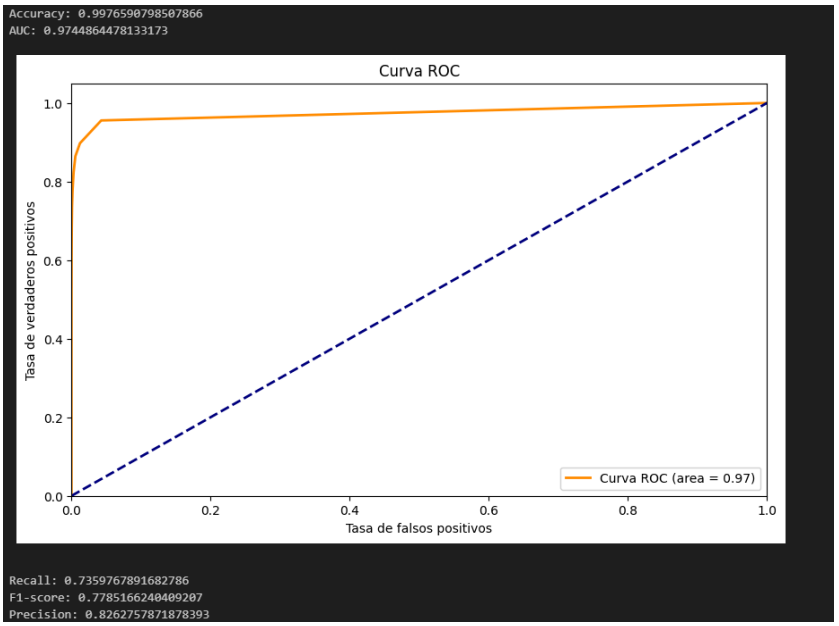


Figura 4: Curva ROC y resultados obtenidos con la data de 2019 usando Random Forest

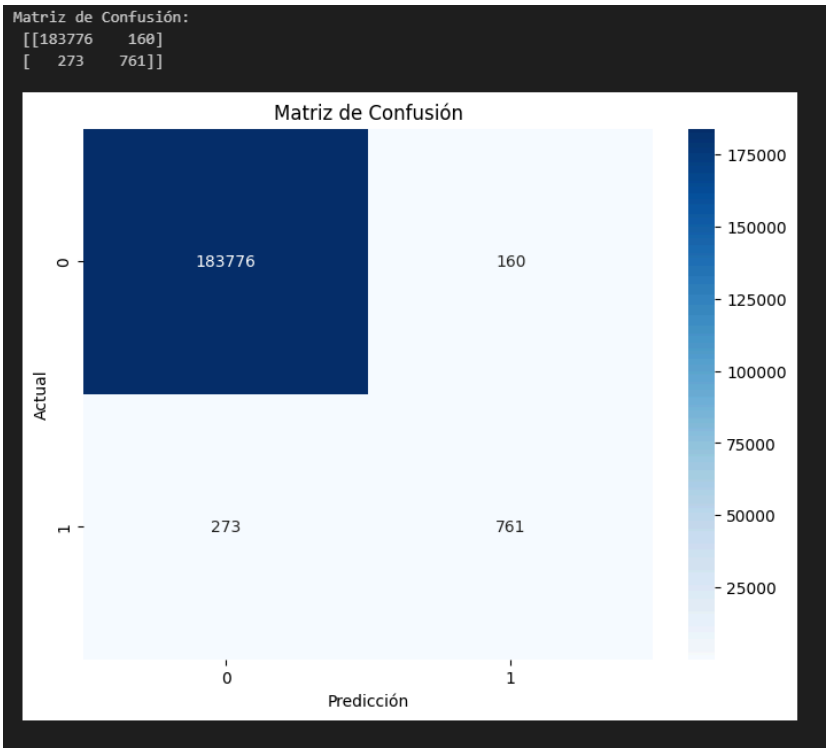


Figura 5: Matriz de confusión del Random Forest con los datos de 2019

Se puede observar que los resultados del modelo de Random Forest en el año 2019 también fueron bastante aceptables al igual que los de las redes neuronales. Lo anterior quiere decir que el modelo pudo detectar una buena parte de los fraudes que se realizaron en los comercios a nivel mundial en el año 2019.

Redes Neuronales luego del entrenamiento incremental de 2020

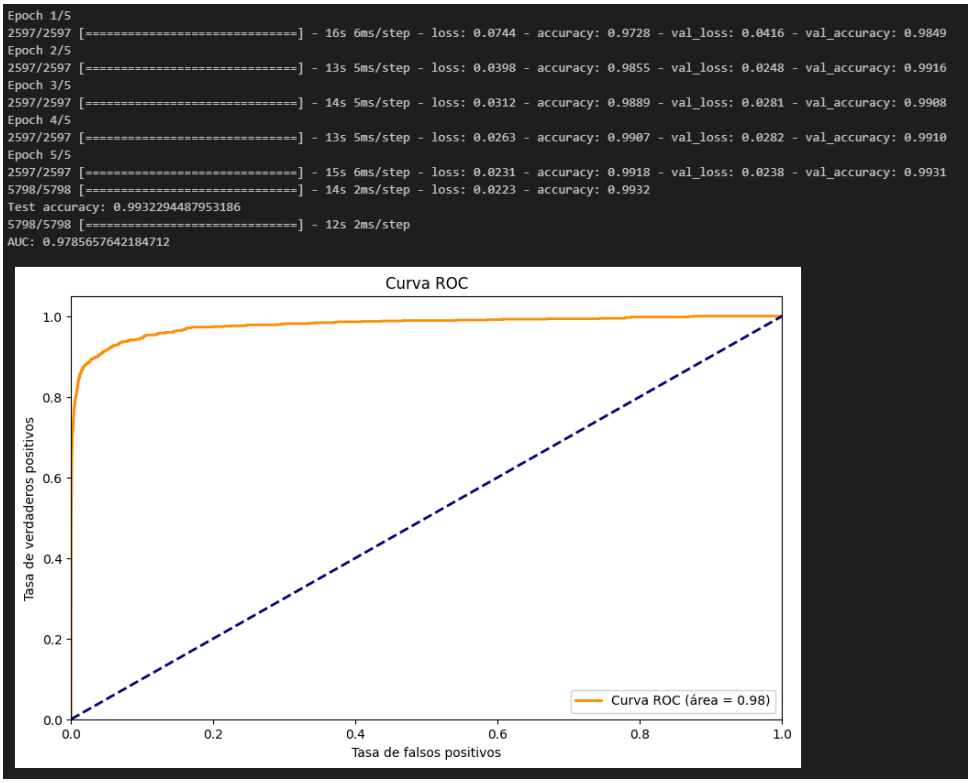


Figura 6: Curva ROC de las redes neuronales luego del entrenamiento incremental de 2020

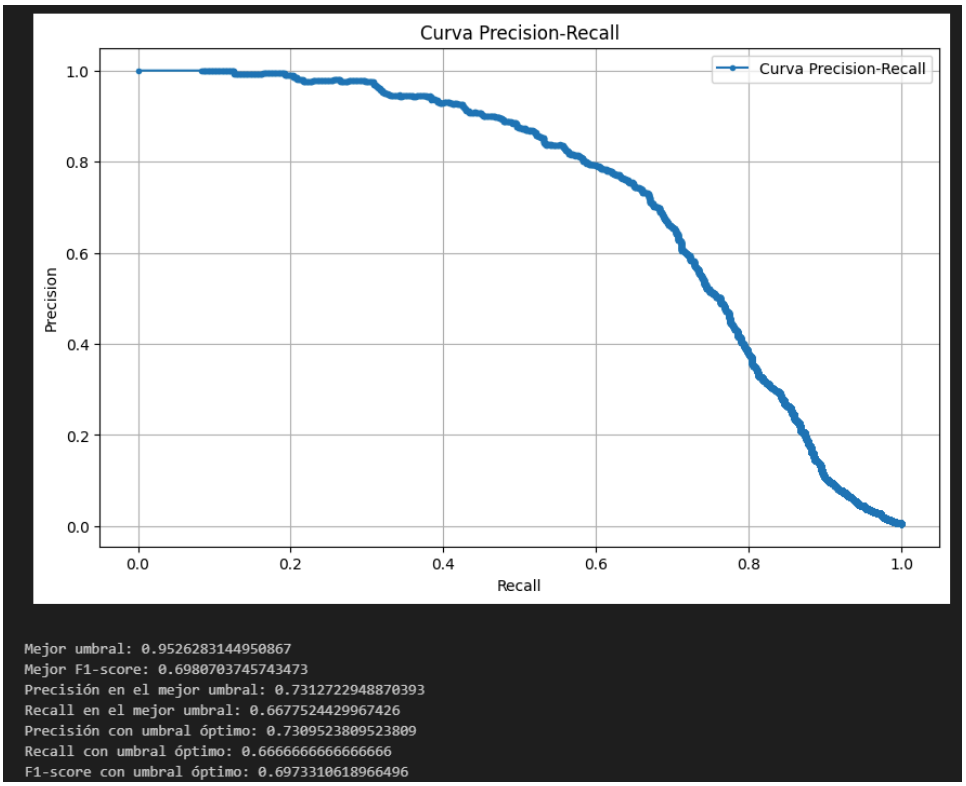


Figura 7: Curva "Precision Recall" y resultados de las redes neuronales luego del re-entrenamiento con la data de 2020

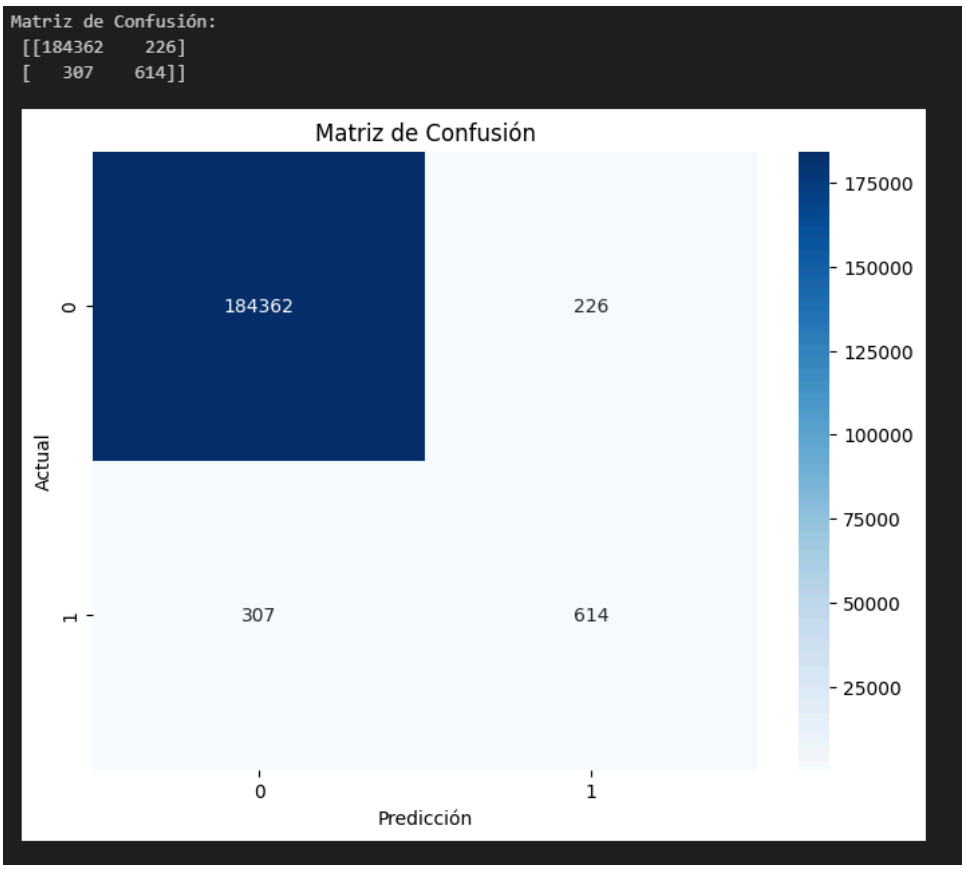


Figura 8: Matriz de confusión de las redes neuronales luego del re-entrenamiento con la data de 2020

Dado que se realizó un entrenamiento incremental con los datos de 2020 con las redes neuronales, es importante mencionar que las métricas mejoraron bastante y que el modelo ya pudo detectar más casos de fraude para el año 2019 y pudo detectar bastantes casos de fraude en el año 2020.

Random Forest luego del entrenamiento incremental de 2020

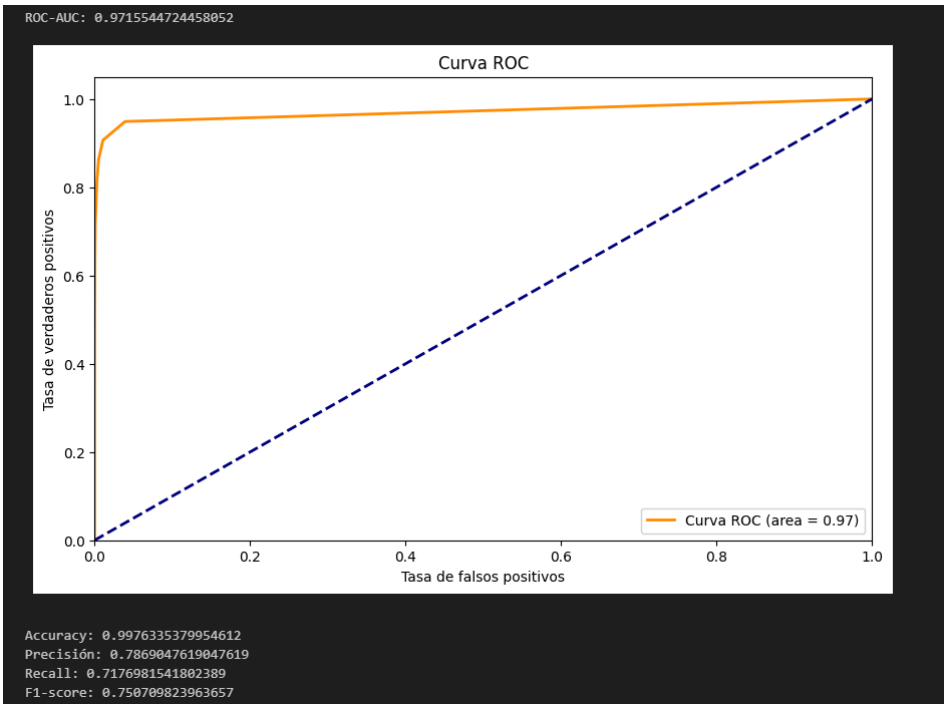


Figura 9: Curva ROC y resultados usando la data de 2020 luego del re-entrenamiento del Random Forest

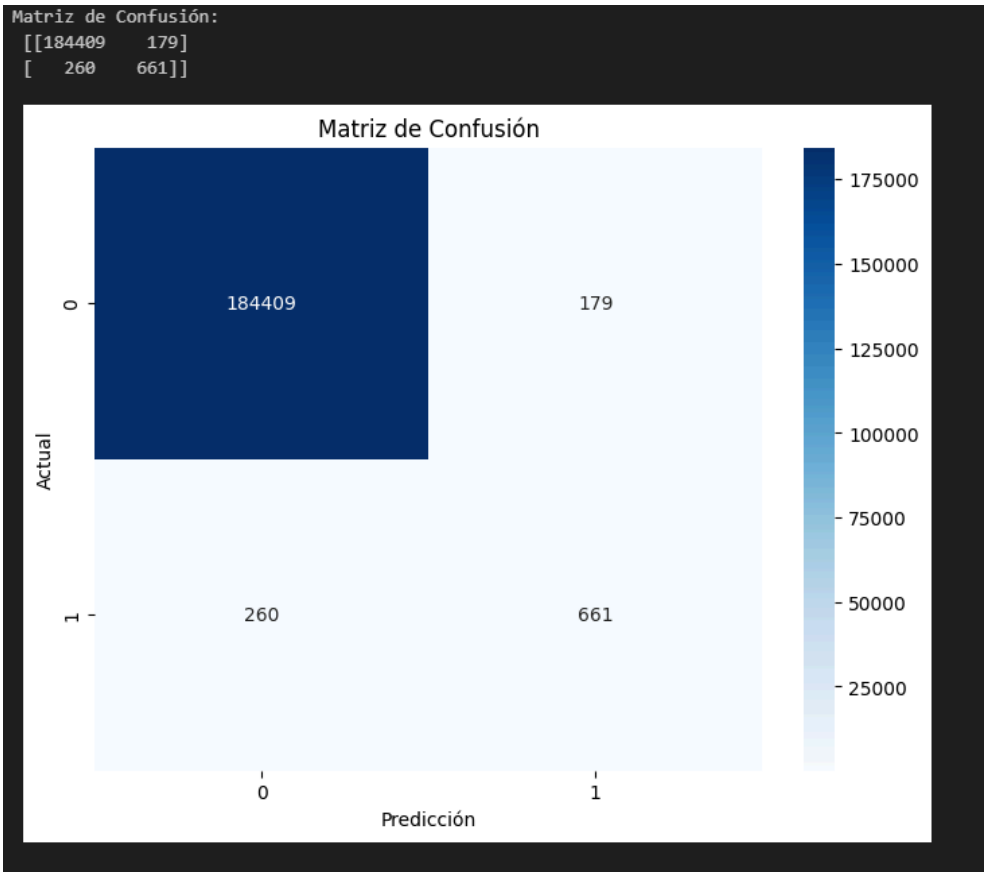


Figura 10: Matriz de confusión del Random Forest luego del re-entrenamiento con la data de 2020

Luego de que se realizó el entrenamiento incremental con los datos de 2020 con el random forest, es importante mencionar que el modelo pudo mejorar la detección de fraude de las tarjetas de crédito en los negocios a lo largo de los años 2019 y 2020. Asimismo, es importante mencionar que el modelo ya pudo tener una mayor amplitud de detección de fraude con respecto al dataset proveído.

Conclusiones

- El modelo presentó un mejor rendimiento al aplicar el entrenamiento incremental usando dos años de datos de fraude.
- La técnica de balance de clases que mejor funcionó para el dataset fue la de SMOTE y balanceando clases de 1 a 5 para evitar el sobreajuste.
- El modelo que mejores resultados presentó a lo largo de los entrenamientos fue el de Random Forest.

Recomendaciones

- Para poder entrenar modelos detectores de fraude con datasets similares a la presente práctica, se recomienda usar un entrenamiento incremental, dado que se puede controlar mejor el rendimiento de los modelos y se puede verificar mejor los resultados.
- Es importante detectar si el dataset presenta algún desbalance de clases, dado que este factor podría afectar el rendimiento del modelo y podría sobre ajustar el modelo.
- Los modelos que mejores resultados presentan son todos los relacionados a Random Forest.
- Se recomienda buscar diferentes técnicas de balance de clases, dado que algunas técnicas pueden no ser eficientes y podrían sobre ajustar el modelo a entrenar.

Citas

- Unir, V. (2023, 29 noviembre). ¿Qué son las redes neuronales? Concepto y usos principales. *UNIR*. https://www.unir.net/ingenieria/revista/redes-neuronales-artificiales/
- *4.13 - Ventajas y limitaciones de las Redes Neuronales Recurrentes y LSTM | Codificando Bits*. (s. f.). Codificando Bits. https://www.codificandobits.com/curso/fundamentos-deep-learning-python/redes-recurrentes-13-ventajas-limitaciones/

- *Random forest, la gran técnica de Machine Learning*. (2023, 27 enero). Inesdi. https://www.inesdi.com/blog/random-forest-que-es/
- Neira, M. S. (2023, 10 abril). ¿Cómo es una estrategia exitosa de continuous training en ML? *Medium*. https://medium.com/latinxinai/c%C3%B3mo-es-una-estrategia-exitosa-de-continuous-training-en-ml-e00198e0eae5
- IBM Corporation. (s. f.). *Utilización del aprendizaje incremental para entrenar con un conjunto de datos grande | IBM Cloud Pak for Data as a Service*. © Copyright IBM Corporation 2021. https://dataplatform.cloud.ibm.com/docs/content/wsj/analyze-data/autoai-incr-learn.html?pos=2&locale=es&context=cpdaas