Entregable 2

Integrantes:

Cristian Andres Henao Londoño

Jhon Janer Torres Restrepo

Asignatura:

Introducción a la Inteligencia Artificial

Docente:

Raul Ramos Pollan

Informe de Progreso: Detección de Fraudes en Solicitudes de Apertura de Cuentas

Introducción

Este informe documenta el avance y los resultados del proyecto orientado a detectar fraudes en las solicitudes de apertura de cuentas bancarias. El objetivo principal de este proyecto es desarrollar un modelo de aprendizaje automático que sea capaz de distinguir entre solicitudes legítimas y fraudulentas, contribuyendo así a reforzar la seguridad y eficiencia en los procesos bancarios.

El dataset utilizado para abordar este problema, se encuentra alojado en el siguiente enlace de Kaggle https://www.kaggle.com/datasets/sgpjesus/bank-account-fraud-dataset-neurips-2022

El proyecto ha transitado por diversas etapas, y se destacan a continuación los aspectos más sobresalientes de su evolución.

Descripción del Progreso

Recopilación de Datos

La primera fase del proyecto consistió en la recopilación y exploración de datos relevantes para la detección de fraudes en solicitudes de apertura de cuentas. Estos datos incluyen una variedad de características, como información demográfica del cliente, detalles de la solicitud, datos financieros y comportamientos del cliente. Además, se identificaron dos características categóricas, "payment_type" y "housing_status," que requerían una codificación adecuada para su inclusión en el proceso de modelado.

Preprocesamiento de Datos

Para preparar los datos para su utilización en el modelado, se llevaron a cabo varias acciones:

- **Codificación One-Hot**: Se aplicó una codificación one-hot a las características categóricas para asegurar que el modelo pudiera comprender y utilizar estas variables. Sin embargo, se enfrentó a un desafío de datos insuficientemente codificados, que se resolvió posteriormente.

- **División de Datos:** El conjunto de datos se separó en conjuntos de entrenamiento y prueba, permitiendo así una evaluación adecuada del rendimiento del modelo.

Modelado Inicial

En una fase temprana, se implementó un modelo inicial basado en Random Forest para evaluar su rendimiento con los datos preprocesados. Sin embargo, este modelo arrojó errores debido a la presencia de datos inadecuadamente codificados, lo que se manifestó en un mensaje de error: "ValueError: could not convert string to float."

Solución de Problemas

Se investigó y resolvió el problema relacionado con la codificación insuficiente de los datos categóricos. La solución consistió en aplicar la codificación one-hot a todas las columnas categóricas, lo que finalmente permitió que el conjunto de datos fuera apto para su utilización en el modelado.

Modelado con Random Forest

Se optó por la implementación de un modelo Random Forest como punto de partida:

Puesta en Marcha de Random Forest

- Se procedió a dividir los datos en conjuntos de entrenamiento y prueba.
- Se inicializó un modelo Random Forest con 100 árboles y se entrenó con el conjunto de entrenamiento.
- Las predicciones se efectuaron en el conjunto de prueba y se evaluó su desempeño.

Rendimiento del Modelo Random Forest

El modelo Random Forest exhibió una precisión elevada, alcanzando 0.9889983070407784. Sin embargo, el recall resultó extremadamente bajo, con un valor de 0.0013863216266173752. Esto indicó que el modelo experimentó dificultades en la detección de solicitudes fraudulentas.

Métricas de Desempeño

Se calcularon métricas adicionales, como el F1-Score y la matriz de confusión:

- Precisión (Accuracy): 0.9889983070407784
- Recall (Sensibilidad): 0.0013863216266173752
- F1-Score: 0.0027649769585253456

- Matriz de Confusión:

```
[[194530      3]
[  2161      3]]
```

Conclusiones

Basado en el progreso alcanzado hasta la fecha, se pueden extraer las siguientes conclusiones:

- El preprocesamiento de datos fue esencial para abordar problemas iniciales de codificación insuficiente de variables categóricas y para garantizar la adecuada preparación del conjunto de datos para el modelado.

- El modelo Random Forest logró una alta precisión, pero presentó un recall extremadamente bajo. Esto sugiere que el modelo enfrenta dificultades en la identificación de solicitudes de apertura de cuentas fraudulentas.

- Es evidente la necesidad de abordar el desequilibrio en el conjunto de datos para mejorar la capacidad del modelo para detectar fraudes.

- Se requieren ajustes adicionales en el modelado, la selección de algoritmos y, posiblemente, la ingeniería de características para mejorar el poder predictivo del modelo.

Próximos Pasos

Los siguientes pasos para el proyecto abarcan:

- La exploración de técnicas de remuestreo para atender el desequilibrio de clases en el conjunto de datos.

- La investigación en ingeniería de características y la consideración de algoritmos más apropiados.

- La evaluación de otros modelos de aprendizaje automático, como XGBoost, con el propósito de mejorar el rendimiento en términos de recall.

Modelado con XGBoost

Tras enfrentar desafíos con el modelo Random Forest, se optó por investigar una alternativa: el modelo XGBoost. Se acondicionó el conjunto de datos preprocesado para este nuevo modelo y se siguieron los pasos siguientes:

Implementación de XGBoost

Se puso en funcionamiento y entrenó un modelo XGBoost con el conjunto de datos. Este modelo arrojó un desempeño destacable en términos de precisión, registrando un valor de 0.9887847806524757. Esto implica que el modelo puede categorizar la mayoría de las solicitudes de apertura de cuentas de manera acertada.

Desempeño del Modelo XGBoost

Precisión del Modelo XGBoost

La precisión del modelo XGBoost es una métrica importante que indica la proporción de solicitudes de apertura de cuentas que se han clasificado correctamente como fraudulentas o legítimas. En este caso, el modelo XGBoost alcanzó una alta precisión de 0.9888, lo que significa que la mayoría de las solicitudes se clasifican correctamente.

Métricas de Desempeño

Recall (Sensibilidad)

El recall, o sensibilidad, es crucial para determinar cuántas de las solicitudes de apertura de cuentas fraudulentas se detectaron correctamente. En el caso del modelo XGBoost, el valor de recall es 0.0365, lo que indica que solo se logra capturar un pequeño porcentaje de las solicitudes fraudulentas.

F1-Score

El F1-Score es una métrica que combina tanto la precisión como el recall y es especialmente útil cuando se enfrenta a desequilibrios en las clases del conjunto de datos. El modelo XGBoost obtiene un F1-Score de 0.0668, lo que sugiere un equilibrio aceptable entre la precisión y la capacidad de detectar fraudes.

Matriz de Confusión

La matriz de confusión proporciona una visión detallada de cómo se están clasificando las solicitudes de apertura de cuentas, incluyendo los falsos positivos y los falsos negativos. En el caso del modelo XGBoost, la matriz muestra 121 falsos positivos y 2085 falsos negativos, lo que indica que existen áreas de mejora en la detección de fraudes.

[[194412 121]

[2085 79]]

Próximos Pasos

Los siguientes pasos para el proyecto permanecen sin cambios, tal como se mencionaron en el informe previo. Asimismo, se examinarán posibles mejoras en la ingeniería de características y en la elección de algoritmos, con el propósito de lograr un modelo más eficaz en la detección de fraudes en solicitudes de apertura de cuentas bancarias.

En conjunto, estos hallazgos reflejan un progreso positivo en el proyecto, a pesar de la presencia de desafíos que deben superarse para alcanzar un modelo altamente eficaz y equilibrado en términos de precisión y recall.