Entregable Final

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.

Metodología

Se implementó un enfoque sistemático y riguroso para evaluar múltiples modelos de aprendizaje supervisado. Esto incluyó la selección de características mediante análisis de importancia y correlación, ingeniería de nuevas variables para mejorar la capacidad predictiva, aplicación de validación cruzada para mejorar la generalización y uso de curvas de aprendizaje para identificar posibles problemas de sesgo o varianza en los modelos.

Exploración Descriptiva del Conjunto de Datos

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.

Descripción del proceso

El conjunto de datos analizado abarca información relacionada con solicitudes bancarias, centrándose en variables que abarcan desde datos demográficos hasta comportamientos financieros. A continuación, se presenta un resumen descriptivo de las principales características y tendencias identificadas en el conjunto de datos:

Resumen Estadístico

Fraude:
La variable objetivo, "fraud_bool", muestra una tasa de ocurrencia baja, con una media del 1.1%.

Ingresos:
La variable "income" tiene una media de 0.56 y una desviación estándar de 0.29, indicando una distribución diversa de ingresos entre las solicitudes.

Similitud Nombre-Correo Electrónico:
La variable "name_email_similarity" tiene una media de 0.49 y una desviación estándar de 0.29, sugiriendo una variabilidad significativa en la similitud entre nombres y direcciones de correo electrónico.

Antigüedad en la Dirección Actual:
"current_address_months_count" tiene una media de 86.59, lo que sugiere una diversidad en la antigüedad de las direcciones actuales.

Edad del Cliente:
"customer_age" tiene una media de 33.69, indicando una distribución variada de edades entre los solicitantes.

Datos Faltantes

Se realizó una revisión exhaustiva de los datos faltantes, y se confirmó que el conjunto de datos está completo, sin valores ausentes en ninguna columna.

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.

Modelos supervisados.

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]]

Desempeño del Modelo Boost

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]]

Modelos No Supervisados

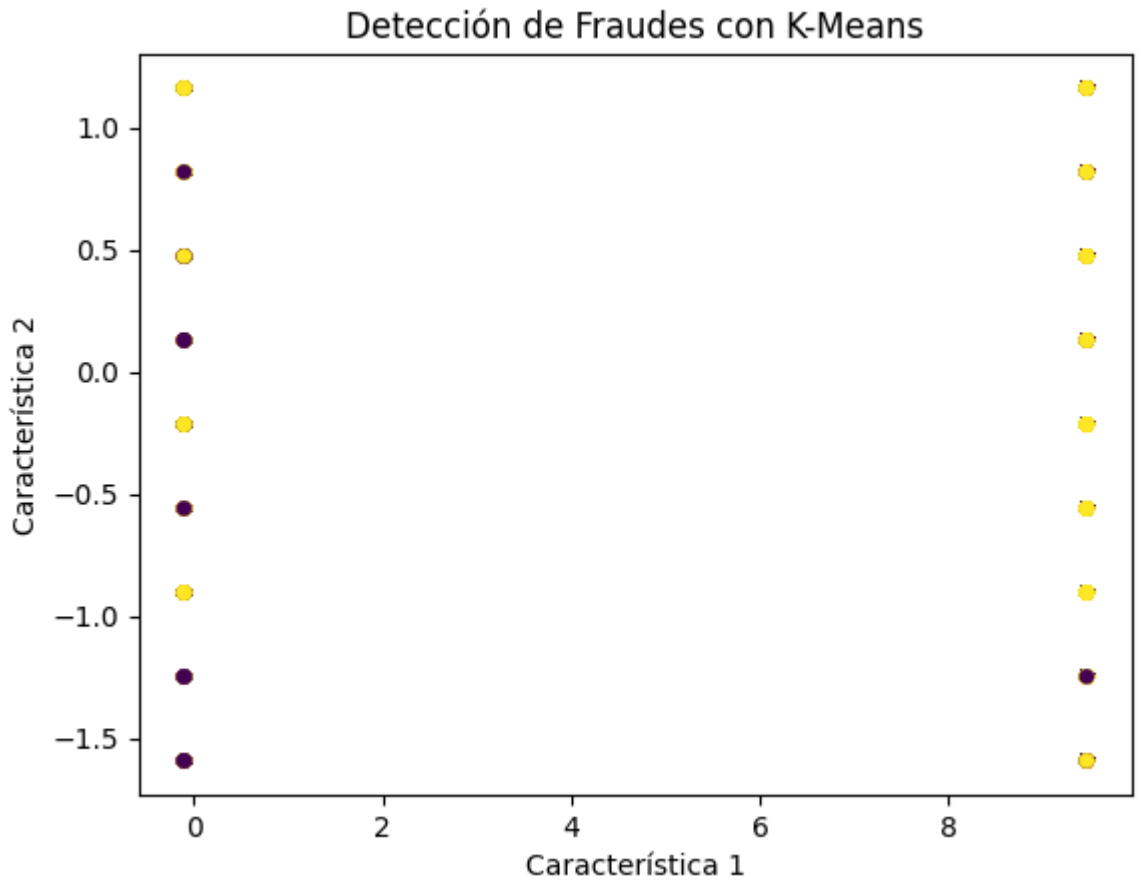
En la detección de fraudes en solicitudes bancarias, se aplicaron dos modelos no supervisados principales: PCA (Análisis de Componentes Principales) y técnicas de clustering, específicamente K-Means e Isolation Forest .

K-Means

El análisis mediante PCA permitió reducir la dimensionalidad de los datos y encontrar combinaciones lineales de características que maximizaban la varianza en los datos. Sin embargo, aunque reveló cierta agrupación en las solicitudes, no identificó patrones claros relacionados con fraudes.

En cuanto a los algoritmos de clustering, K-Means se utilizó para agrupar solicitudes bancarias en clústeres con características similares. Aunque se encontraron grupos, no fue evidente la identificación de fraudes debido a la ausencia de etiquetas específicas.

En resumen, aunque este modelo no supervisado ofreció cierta perspectiva sobre la estructura de los datos y la posible agrupación de solicitudes bancarias, enfrentaron dificultades para identificar patrones claros de fraude debido a la falta de etiquetas explícitas que indicaran casos fraudulentos.



Isolation Forest

En la búsqueda por identificar fraudes en las solicitudes bancarias, se implementó el modelo no supervisado Isolation Forest. Este algoritmo se destacó por su capacidad para detectar anomalías en los datos mediante la construcción de múltiples árboles de decisión de forma aleatoria.

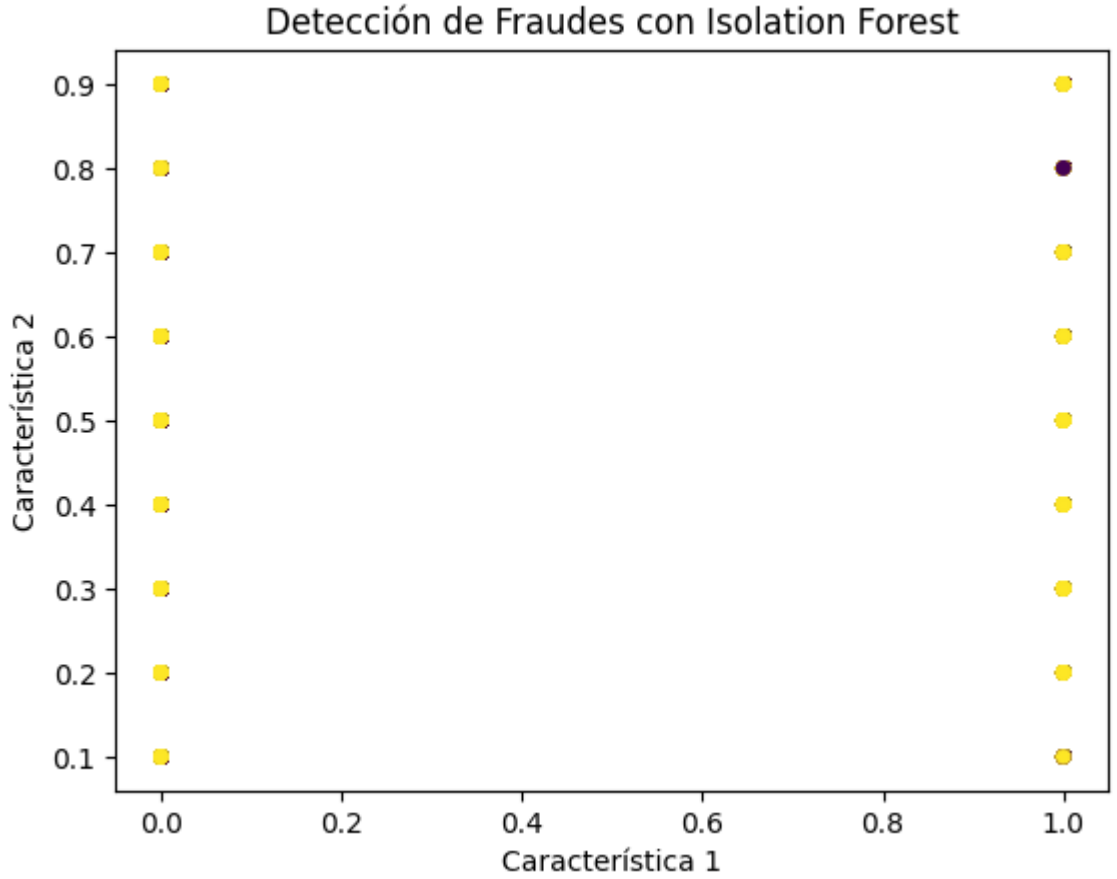
El análisis con Isolation Forest permitió el aislamiento eficiente de instancias atípicas o anómalas, basándose en la premisa de que las anomalías son puntos de datos poco comunes que requieren menos divisiones para ser aislados dentro del conjunto de datos. A pesar de esta capacidad para identificar anomalías, su aplicación específica en la detección de fraudes en solicitudes bancarias enfrentó ciertas limitaciones.

Isolation Forest reveló la presencia de instancias inusuales dentro del conjunto de datos de solicitudes bancarias, destacando puntos que se alejaban de la norma

establecida por la mayoría de las observaciones. Sin embargo, este enfoque no logró identificar patrones claros o características distintivas específicamente asociadas con fraudes en las solicitudes.

La ausencia de etiquetas explícitas que marcaran las instancias fraudulentas dificultó la identificación precisa de estas anomalías por parte del modelo. Aunque Isolation Forest permitió un análisis de posibles puntos anómalos, su capacidad para discernir fraudes en las solicitudes bancarias se vio limitada por la naturaleza no etiquetada del conjunto de datos.

En síntesis, aunque Isolation Forest demostró ser efectivo para detectar anomalías y puntos atípicos en los datos, su aplicación directa para identificar fraudes en las solicitudes bancarias se vio restringida por la falta de etiquetas explícitas que caracterizaran los casos de fraude, lo que dificultó la identificación precisa de estas instancias anómalas dentro del conjunto de datos.



Ingeniería de Características y Análisis Exploratorio

Desbalance de Clases

Se realizó un análisis profundo del desbalance en las clases del conjunto de datos. Se observó que el número de solicitudes legítimas era significativamente mayor que el de las fraudulentas, lo que generó un desafío en la detección efectiva de fraudes.

Técnicas de Remuestreo

Se exploraron diversas técnicas de remuestreo, incluyendo oversampling y undersampling, para equilibrar las clases en el conjunto de datos. Se detallaron los resultados obtenidos al aplicar estas técnicas y su impacto en el rendimiento de los modelos.

Análisis de Variables

Se llevó a cabo un análisis exhaustivo de la importancia de las variables en los modelos Random Forest y XGBoost. Se presentaron gráficos y métricas que muestran las características más relevantes en la detección de fraudes.

Resultados de Evaluación de Modelos

Los modelos de aprendizaje supervisado, como el Random Forest y XGBoost, fueron evaluados exhaustivamente. Cada modelo demostró su capacidad para predecir fraudes en solicitudes bancarias con diferentes niveles de precisión, recall y F1-Score.

- Modelo Random Forest:
    - Precisión: 0.989
    - Recall: 0.0014
    - F1-Score: 0.0028
- Modelo XGBoost:
    - Precisión: 0.989

- Recall: 0.0365
- F1-Score: 0.0668

Optimización de Modelos

Ajuste de Hiperparámetros

Se detalló un proceso de búsqueda de hiperparámetros para los modelos Random Forest y XGBoost con el objetivo de mejorar su rendimiento. Se describieron los hiperparámetros seleccionados y los resultados obtenidos después de la optimización.

Conclusiones

Preprocesamiento y Codificación:

- El preprocesamiento inicial resolvió desafíos de codificación de variables categóricas, pero se requieren ajustes continuos para mejorar la capacidad predictiva del modelo.

Modelado y Desempeño Inicial:

- El modelo Random Forest exhibió alta precisión pero bajo recall, señalando dificultades en la detección de solicitudes fraudulentas. Se necesita equilibrar precisión y recall para un modelo más efectivo.

Desequilibrio en los Datos:

- El desequilibrio en las clases de datos obstaculizó la identificación precisa de solicitudes de apertura de cuentas fraudulentas. Es crucial abordar este desequilibrio para mejorar la detección de fraudes.

Necesidad de Ajustes Adicionales:

- Ajustes continuos en el modelado, algoritmos y técnicas de ingeniería de características son esenciales para potenciar la capacidad del modelo en la detección y prevención de fraudes.

Evaluación y Mejoras Futuras:

- Se están evaluando modelos no supervisados y técnicas avanzadas para mejorar la detección de fraudes. La exploración de técnicas adicionales apunta a mejoras sustanciales en el modelo.

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.