Estimación de transacciones sospechosas para clientes de una aseguradora de Colombia aplicando Isolation Forest y Local Outlier Factor para el control de riesgo Sarlaft

Estimation of suspicious transactions for customers of an insurance company in Colombia applying Isolation Forest and Local Outlier Factor for Sarlaft risk control

Silvia Alexandra Guevara González*

* Magíster en Finanzas, Universidad Externado de Colombia. Líder de operaciones Seguros del Estado, Bogotá (Colombia). [salexandraguevarag@hotmail.com].

Artículo recibido: 12 de julio de 2024.
Aceptado: 10 de octubre de 2024.

Para citar este artículo:
Guevara González, S. A. (2024). Estimación de transacciones sospechosas para los clientes de una aseguradora de Colombia aplicando Isolation Forest y Local Outlier Factor para el control de riesgo Sarlaft. *Odeon*, *27*, 171-207.
DOI: https://doi.org/10.18601/17941113.n27.06

Resumen

El artículo aborda la detección de transacciones sospechosas en clientes de una aseguradora en Colombia, enfocándose en la necesidad de fortalecer el control de riesgo bajo la normativa del Sistema de Administración del Riesgo de Lavado de Activos y de la Financiación del Terrorismo (Sarlaft), mediante dos métodos de detección de anomalías: Isolation Forest y Local Outlier Factor (LOF).

El estudio se enfoca en la importancia de identificar patrones inusuales en las transacciones para prevenir el lavado de activos y la financiación del terrorismo en Colombia, un país afectado por conflictos sociales y económicos, lo cual es crítico para la integridad del sistema financiero. A través de la aplicación de estos algoritmos de aprendizaje automático, se busca mejorar la precisión en la identificación de comportamientos anómalos que podrían indicar actividades delictivas.

El método Isolation Forest se basa en la creación de árboles aleatorios para aislar observaciones, mientras que LOF evalúa la densidad local de los datos para identificar puntos atípicos. El artículo incluye un análisis de la efectividad de estas técnicas y su aplicabilidad en el contexto de la aseguradora, así como recomendaciones para su implementación.

En conclusión, la investigación busca fortalecer los controles a partir del uso de estos dos enfoques, para optimizar la detección de transacciones sospechosas, contribuyendo a un mejor cumplimiento de las normativas de riesgo y robusteciendo las prácticas de prevención de lavado de activos y la financiación del terrorismo (LAFT) en el sector asegurador colombiano.

Palabras clave: transacciones; clientes; Isolation Forest; control; riesgo; Local Outlier Factor.

Clasificación JEL: G22, C45

Abstract

The article addresses the detection of suspicious transactions in clients of an insurance company in Colombia, focusing on the need to strengthen risk control under the SARLAFT regulations within the framework of the (Money Laundering and Terrorist Financing Risk Management System). By means of two anomaly detection methods: Isolation Forest and Local Outlier Factor (LOF).

The study focuses on the importance of identifying unusual patterns in transactions to prevent money laundering and terrorist financing in Colombia, a country affected by social and economic conflicts, which is critical for the

integrity of the financial system. Through the application of these machine learning algorithms, the aim is to improve accuracy in identifying anomalous behavior that could indicate criminal activity.

The Isolation Forest method is based on the creation of random trees to isolate observations, while LOF evaluates the local density of the data to identify outliers. The article includes an analysis of the effectiveness of these techniques and their applicability in the insurance context, as well as recommendations for their implementation.

In conclusion, the research seeks to strengthen controls based on the use of these two approaches to optimize the detection of suspicious transactions, contributing to better compliance with risk regulations and strengthening LAFT prevention practices in the Colombian insurance sector.

Key words: Transactions; clients; Isolation Forest; control; risk; Local Outlier Factor.

JEL classification: G22, C45

Introducción

En Colombia, un país con una creciente y difícil situación que involucra temas como conflictos sociales, económicos, narcotráfico, corrupción, tráfico de drogas, trata de personas, contrabando y enriquecimiento ilícito, la prevención y detección del lavado de activos y la financiación del terrorismo son desafíos cruciales para los sistemas financieros y las entidades reguladoras. En este contexto, el Sistema de Administración del Riesgo de Lavado de Activos y Financiación del Terrorismo (Sarlaft) se convierte en una herramienta fundamental para garantizar la integridad y transparencia en las operaciones financieras.

Nuestro objeto de estudio se centra en las entidades financieras, en particular las aseguradoras, que buscan promover la transparencia, responsabilidad y confianza en todos los productos que ofrecen. Esta responsabilidad conlleva la obligación de promover una cultura de cumplimiento para evitar delitos, y, de esta manera, dejar atrás el fuerte impacto que comenzó en la década de los setenta, cuando el lavado de activos emergió como uno de los principales delitos, vinculado al surgimiento de diversos carteles impulsados por la venta de estupefacientes en los Estados Unidos (Rocha, 2014). La aparición de bandas criminales generó un aumento en los indicadores económicos y provocó distorsiones macroeconómicas como la inflación, además de agudizar problemas sociales.

Desde el surgimiento de los carteles, Colombia ha ocupado un lugar de suma importancia en el delito de lavado de activos, junto con la percepción de corrupción e inseguridad a nivel internacional debido al tráfico de drogas. Esto permitió que entrara más dinero al país, proveniente de actividades ilícitas, lo que generó distorsiones en la imagen de Colombia frente a otros países. En respuesta, el país ha venido fortaleciendo su marco normativo en materia de riesgo de LAFT e implementando acciones para robustecer los controles y las estrategias que permitan mitigar estos delitos.

Según diferentes estudios, Colombia ocupa el tercer lugar en la lista de los países más vulnerables al lavado de dinero en América Latina, de acuerdo con la publicación de Infobae del 26 de noviembre de 2017. A nivel mundial, según una publicación del 24 de enero de 2023, Colombia se encuentra en la posición 83 con un índice de 4,74 en el listado de países elaborado por el Instituto de Gobernanza de Basilea (BGI), una organización internacional de investigación y desarrollo sin fines de lucro que se dedica a mejorar la calidad de la gobernanza en ámbitos corporativos, estatales, internacionales y sociales. Este índice otorga a cada país un puntaje del 1 al 10, donde 1 significa riesgo bajo y 10 riesgo máximo. En 2022, Colombia mostró una ligera mejoría en el nivel de riesgo con respecto a 2021.

Dada la diversidad de conflictos internos que enfrenta Colombia, es necesario intensificar y promover la adopción de medidas para la lucha contra el lavado de activos, la financiación del terrorismo y la proliferación de armas de destrucción masiva, cumpliendo con toda la normativa que exige la ley. Por lo tanto, es imprescindible que las organizaciones refuercen sus herramientas de prevención para evitar estas prácticas ilícitas y de riesgo, protegiendo la integridad del sistema financiero y contribuyendo a la estabilidad y seguridad nacional.

Este artículo busca realizar un experimento dentro del marco normativo vigente, implementando herramientas para evaluar la efectividad en la lucha contra el lavado de activos y fortalecer los desafíos que enfrentan las entidades reguladoras. El objetivo es identificar soluciones que promuevan la integridad y transparencia en las operaciones financieras, así como la seguridad del sistema en su conjunto.

La presente investigación tiene como objeto de estudio la estimación de transacciones sospechosas para los clientes de una aseguradora en Colombia, aplicando los algoritmos Isolation Forest y Local Outlier Factor para el control del riesgo Sarlaft. El objetivo es realizar un análisis detallado para las

entidades vigiladas por la Superintendencia Financiera y demás entes de control, específicamente para una aseguradora, dada su importancia para la economía del país y su contribución socioeconómica en Colombia. Es fundamental que estas entidades cuenten con todas las herramientas necesarias para apoyar la adopción de medidas de autocontrol, prevención, detección y gestión de riesgos contra el lavado de activos, la financiación del terrorismo y la financiación de la proliferación de armas de destrucción masiva.

En resumen, en este artículo se presenta un estado del arte, seguido de una exposición de la metodología; se exploran los controles establecidos por el Estado colombiano para combatir el lavado de activos y la financiación del terrorismo; se identifican los departamentos del país más propensos al riesgo de estas actividades ilícitas. Seguidamente, se explica la aplicación de los métodos Isolation Forest y Local Outlier Factor dentro del Sarlaft en el sector asegurador; luego, se aplican estos métodos para estimar transacciones sospechosas en una aseguradora específica. La investigación concluye con un análisis general de los resultados obtenidos y su impacto en la lucha contra el lavado de activos en Colombia.

1. Estado del arte

Numerosos estudios han abordado la problemática del lavado de activos, la financiación del terrorismo, el Sarlaft, así como los métodos de detección de anomalías como Isolation Forest y Local Outlier Factor, y los riesgos que enfrentan las compañías financieras. A continuación, se presenta una revisión de los principales trabajos relevantes para esta investigación, organizada de manera que siga una progresión lógica desde el contexto institucional hasta las metodologías específicas aplicadas.

Mariño *et al*. (2014), en su estudio titulado "Determinantes en la prevención del riesgo para el lavado de activos y la financiación del terrorismo", exploran las instituciones y los mecanismos implementados por el Estado colombiano para enfrentar estos flagelos, con especial énfasis en la Unidad de Investigación y Análisis Financiero (UIAF). Este organismo, creado por la Ley 526 de 1999 y posteriormente modificado por la Ley 1762 de 2015, ha desempeñado un papel fundamental en la inteligencia financiera del país, recibiendo y analizando reportes de actividades sospechosas de entidades financieras y sociedades comerciales. Este estudio es relevante para nuestra investigación ya que proporciona un contexto sobre los controles estatales existentes contra el lavado de

activos y la financiación del terrorismo, lo cual es esencial para el desarrollo de nuevas estrategias preventivas en el sector asegurador.

Zabala (2023), en "La eficacia del Sarlaft en Colombia", evalúa la eficacia de este sistema en la lucha contra la instrumentalización de las instituciones financieras en actividades ilícitas. El estudio subraya que el Sarlaft no es un sistema homogéneo, sino que debe adaptarse a las particularidades de cada organización, desarrollándose según las necesidades y especificidades de cada entidad. Este enfoque es crucial para nuestro trabajo, que busca aplicar los métodos Isolation Forest y Local Outlier Factor en una aseguradora específica, adaptando el Sarlaft a su contexto particular.

Un ejemplo destacado de la aplicación de metodologías avanzadas en la detección de actividades sospechosas es la investigación de Guevara y Granados (2021), titulada "Machine learning methodologies against money laundering in non-banking correspondents". Este estudio se centra en la detección de anomalías en corresponsales no bancarios en Colombia, utilizando varios algoritmos de *machine learning* para identificar transacciones sospechosas. Estos hallazgos son valiosos para nuestra investigación, ya que destacan la importancia de aplicar enfoques tecnológicos avanzados en la vigilancia de transacciones sospechosas, ofreciendo un puente hacia la discusión sobre los métodos específicos que utilizaremos.

Hyder John y Sameena Naaz, en su artículo "Credit card fraud detection using Local Outlier Factor and Isolation Forest", abordan la problemática del fraude en transacciones con tarjetas de crédito, un desafío creciente debido al aumento del comercio electrónico. Los autores comparan los algoritmos Local Outlier Factor e Isolation Forest utilizando datos públicos, y concluyen que el primero alcanza una precisión del 97%, mientras que el segundo logra un 76%. Estos resultados son significativos, ya que demuestran la efectividad de estos algoritmos en la detección de fraudes, lo que respalda su aplicación en nuestro estudio sobre transacciones sospechosas en el sector asegurador.

Liu y Ting (2008), en su artículo "Isolation Forest", proponen un método innovador para la detección de anomalías basado en el aislamiento de puntos atípicos en lugar de la identificación de perfiles normales. Este algoritmo, diseñado para manejar grandes volúmenes de datos con alta eficiencia, ha demostrado obtener mejores resultados que métodos como ORCA, LOF y Random Forest en términos de AUC (Area Under Curve) y tiempo de procesamiento. La capacidad de Isolation Forest para detectar anomalías en conjuntos de datos de alta dimensión es particularmente relevante para nuestra investigación, ya

que permitirá identificar con mayor precisión transacciones sospechosas en un entorno financiero complejo.

El enfoque metodológico cuantitativo de esta investigación, basado en la comprensión de fenómenos desde sus particularidades naturales, es defendido por autores como Hernández *et al*. (2010), quien subraya la importancia de analizar los múltiples factores explicativos que envuelven un objeto de estudio dentro de su contexto dinámico y subjetivo. Denzin y Lincoln (2020) complementan esta visión al destacar que la investigación cualitativa implica un enfoque multimétodo que estudia los fenómenos en su entorno natural, buscando interpretar los significados que las personas les otorgan. Este marco metodológico es instrumental para nuestra investigación, ya que permite entender el problema del lavado de activos y la financiación del terrorismo dentro del contexto específico del sector asegurador colombiano.

2. Metodología

i. Identificar las variables en la base de datos es crucial para aplicar el tratamiento adecuado y desarrollar el modelo (qué tipo de variables existen, categóricas, numéricas, etc.), esto implica que para el tratamiento de las categóricas se asigna un valor que las identifique. Las variables categóricas identificadas son: el producto, el tipo de documento, el segmento, la ciudad, si es nacional o extranjero, el género, la profesión, el estado civil y el código CIIU.
ii. Realizar el análisis descriptivo de la base de datos, con el fin de conocer de primera mano las distribuciones de los datos, la identificación de patrones o los comportamientos estables dentro del comportamiento de clientes de la entidad en todas las variables recolectadas.
iii. Rediseñar la estructura de la base de datos con el fin de facilitar la inclusión de nuevos elementos y el testeo de estos sin importar su fecha de aportes. La base de datos contiene 43 columnas correspondientes al flujo de aportes de cada cliente de manera mensual.
iv. Dado que el propósito del estudio es la estimación de operaciones inusuales de clientes existentes y nuevos en cuanto al flujo de aportes, se debe condensar la información de estas 43 columnas en variables temporales que resuman el comportamiento transaccional de cada cliente.
v. Para este propósito, se desarrolló una función que permite calcular variables tales como valor aporte promedio, aporte mínimo, aporte máximo, transacciones distintas de 0, pendiente, cuantiles. Para efectos de esta variable, se

calculan de acuerdo con las fechas establecidas para la data de entrenamiento, validación y prueba, es decir, estas variables se calculan con una fecha de corte específica. La data también presenta procesos de normalización debido a la naturaleza de estos modelos (*clustering*).

vi. Para los modelos de *machine learning* la data se divide en tres partes:

- Data de entrenamiento: es la data con la que se realiza el entrenamiento del modelo entre enero de 2020 a mayo de 2021 (fechas de aporte), es toda la información comprendida entre ese rango de fechas.
  Por consiguiente, en la data de entrenamiento se obtienen y analizan datos que revelan similitudes y diferencias desconocidas encontrando patrones ocultos. Para la investigación de la data que se usa para estimar el modelo es común utilizar un *Split* de datos con enfoque 70-20-10 %. Sin embargo, para el caso de las transacciones sospechosas, tema de la investigación, se recomienda dividir o hacer un *Split* por fechas en lugar de realizarlo por porcentaje (70% para entrenamiento, 20% para validación y 10% para prueba) y, dado este *Split*, la data de entrenamiento es la que se utiliza para estimar el modelo.
- Data de validación: set de data entre junio de 2021 a junio de 2022.
  Una vez realizada la etapa de partición de la data en tres partes, se procede a la ejecución del modelo y a identificar cuál metodología es la adecuada para el desarrollo del modelo entre los diferentes enfoques disponibles y recomendados; según la exploración, se encuentran Isolation Forest y Local Outlier Factor, ambas metodologías utilizadas en diferentes investigaciones para comprobar su enfoque y cálculos a fin de demostrar cuál de las dos metodologías es mejor en la aplicación.
- Data de prueba: es la data de testeo entre julio de 2022 a julio de 2023.

vii. A medida que se tienen conformadas las bases de datos para cada uno de los pasos, se procede a la definición de hiperparámetros y semilla con la que se va a evaluar el modelo con respecto a la data de validación y prueba, es decir, una vez definidos los hiperparámetros se calcularán los modelos y, posteriormente, las métricas de Silhouette Score y Calinski Harabasz Score en validación y prueba.
  Para el caso de Isolation Forest se definieron los siguientes:
  N_Estimators: número de árboles por utilizar dentro del modelo. Los definidos en este caso por criterio investigativo son: 50, 100, 200, 500, 1000.

Estos se definen de esta manera entendiendo que cuando existen más de 1000 árboles por lo general el modelo se sobreestima, y si son muy pocos, el modelo será subestimado.

Contamination Parameter: el parámetro de contaminación corresponde al porcentaje de la data que se considerará *outlier* dependiendo de qué tan cerca o alejados se encuentren los mismos. Para este caso se considera como número mínimo de outliers 1% y máximo de 30%, entendiendo que es posible de acuerdo con el comportamiento temporal de las transacciones con picos en diciembre y enero.

Para el caso de Local Outlier Factor se definen los mismos parámetros en cuanto a contaminación con diferencia de Número de vecinos o N_Neighbours, que especifica el número de vecinos que se deben considerar al calcular la densidad local alrededor de cada punto de datos.

viii. Una vez evaluados los modelos en cada una de sus combinaciones de parámetros a partir de la data de validación, se toma el mejor valor de cada uno de los criterios escogidos para este propósito (Silhouette y Calinski) entendiendo estos como:

El Silhouette Score, introducido por Peter J. Rousseeuw en 1987, es una herramienta gráfica para interpretar y validar el análisis de clústeres. Este método evalúa la cohesión y separación de los grupos, indicando qué objetos pertenecen a cada grupo y cuáles se encuentran en una posición intermedia. El ancho promedio de la Silhouette permite evaluar la validez de la agrupación y ayuda a determinar el número adecuado de clústeres. Los valores del Silhouette Score oscilan entre -1 y 1, donde 1 indica una excelente separación y -1 una muy mala.

El coeficiente de Silhouette se calcula mediante la distancia media dentro del grupo (a) y la distancia media al grupo más cercano (b) para cada muestra. Es importante señalar que este coeficiente solo se define si existe un número adecuado de etiquetas para los grupos:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Donde:

- $a(i)$ es la distancia promedio desde la muestra 'i' a otros puntos en el mismo grupo.

- $b(i)$ es la distancia promedio desde la muestra 'i' hasta los puntos en el grupo más cercano del que 'i' no forma parte.

El Calinski-Harabasz Score, propuesto por Tadeusz Calinski y Jerzy Harabasz en 1974, es una métrica interna para evaluar la calidad de algoritmos de agrupamiento. También conocido como criterio de varianza (VRC), se centra en los datos y los resultados de la agrupación sin depender de etiquetas externas:

$$CH = \frac{B}{W} \times \frac{N-K}{K-1}$$

Donde:

- CH es el Calinski Harabasz score.
- B es la dispersión entre grupos.
- W es la dispersión dentro del grupo.
- N es el número total de puntos de datos.

Para el caso de este indicador, cuanto más alto sea su valor, mejor es la calidad del clúster calculado.

A partir de los mejores modelos, se calcula el Feature Importance o Importancia de Variables para cada uno con el fin de conocer el comportamiento de las variables y las de mayor importancia que logren identificar el comportamiento atípico de cada uno de los clientes de la compañía.

Una vez con los resultados, se identifican en la data de prueba los posibles datos atípicos y sus características más populares –en este caso la descripción–, y se realiza un análisis descriptivo para cada una de las variables con el fin de generalizar el comportamiento. Esto también se hace para la data de entrenamiento y prueba.

3. Análisis descriptivo

Se deben explorar las variables que componen el conjunto de datos con el objetivo de comprender su distribución, características y patrones más relevantes. Esta exploración proporciona una visión general de cómo está conformado dicho conjunto de datos.

Variables:

a) Tipo producto
Es una variable categórica, compuesta por cinco clases, su distribución se describe mediante un gráfico de barras.

Preferencias de producto

- Los productos OMPEV y CREA son los más populares entre los asegurados, representando más del 90 % del total.
- OMPEV, que es el seguro de vida con ahorro SVA, es el producto preferido, con casi 17.500 asegurados (58,7%).
- CREA, seguro de vida SV, ocupa el segundo lugar, con alrededor de 10.000 asegurados (34,6%).

Productos menos comunes

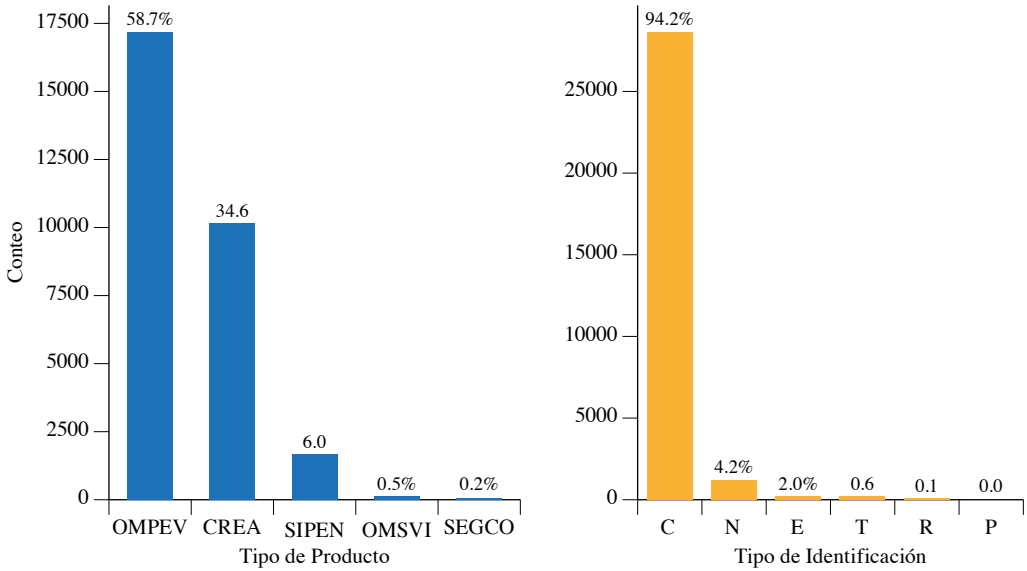
- SIPEN, el seguro para ahorrar para la pensión, es elegido por un 6% de los asegurados, lo que equivale a menos de 2.500 personas.
- OMSVI es el seguro de vida a cinco años, y SEGCO es el seguro para proteger la pensión SP, estos tienen una participación aún menor, con menos del 1% de los datos.

En resumen, la mayoría de los asegurados están en OMPEV y CREA, mientras que los otros productos tienen una presencia mucho más pequeña en el conjunto de datos (figura 1).

b) Tipo de identificación
Esta variable categórica está distribuida en seis clases: clase C identifica a los asegurados con tipo de documento cédula ciudadanía, donde se concentra la mayoría de los datos con un 94,2%. Clase N (Nit) representa a los asegurados con representación jurídica o empresas, con un 4,2% de los datos. Clase E son los asegurados con identificación cédula de extranjería con un 1%. Las tres clases restantes: T (tarjeta de identidad), R (registro civil), p (pasaporte) solo representan el 0,8%.

El set de datos está conformado en su mayoría por personas naturales de nacionalidad colombiana y una porción minoritaria de empresas, extranjeros y otros (figura 1).

Figura 1. Resultado producto y tipo de identificación



Fuente: elaboración propia.

c) Edad

Variable continua, con un promedio de 47 años y una desviación de 3,5; su distribución se representa por medio de un histograma de 8 intervalos, donde se visualiza una asimetría positiva y una porción minoritaria en las edades altas. El intervalo más representativo se sitúa entre los 45 a 55 años, que es consistente con el promedio de los datos (figura 2).

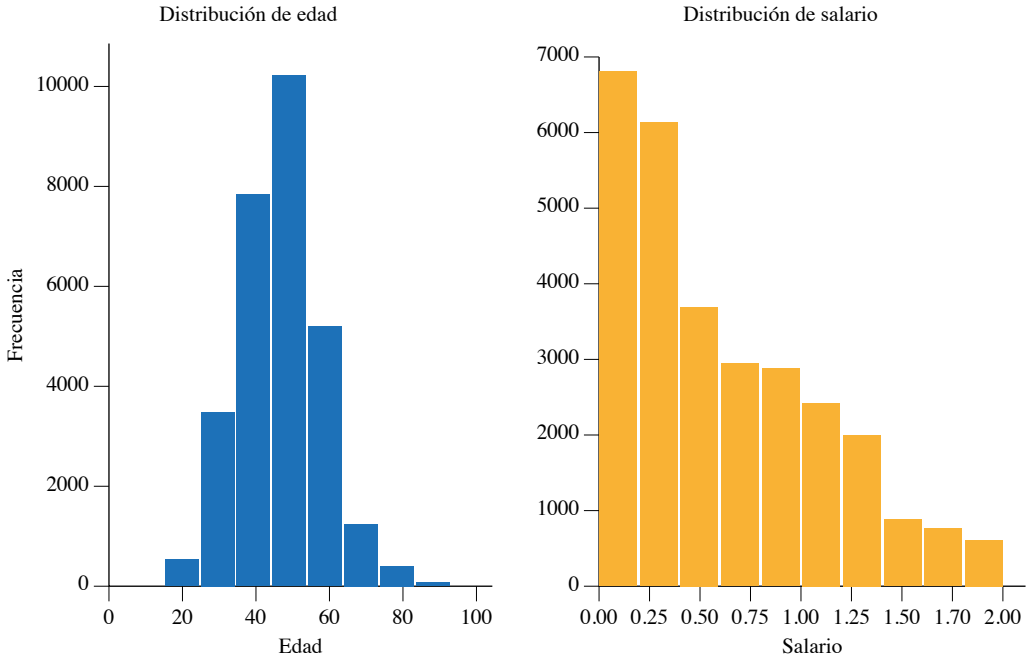
d) Salario

Variable continua con un promedio de $3.500.000, en su histograma se evidencia una asimetría positiva, lo que permite concluir que la mayoría de los asegurados se concentran en rangos salariales bajos. La concentración más alta se sitúa en el primer intervalo con salarios menores a $2.500.000, de ese valor en adelante todos los intervalos disminuyen. Lo que lo hace inversamente proporcional, a mayor salario menor cantidad de asegurados (figura 2).

e) Segmento comercial

Variable discreta ordinal en la que se identifican 3 segmentos: finanzas personales, élite y corporativo. Finanzas personales tiene la tasa representativa más alta con 59,7% y la más baja es para corporativo con un porcentaje de 1,7%,

Figura 2. Resultado distribución de edad y distribución de salario



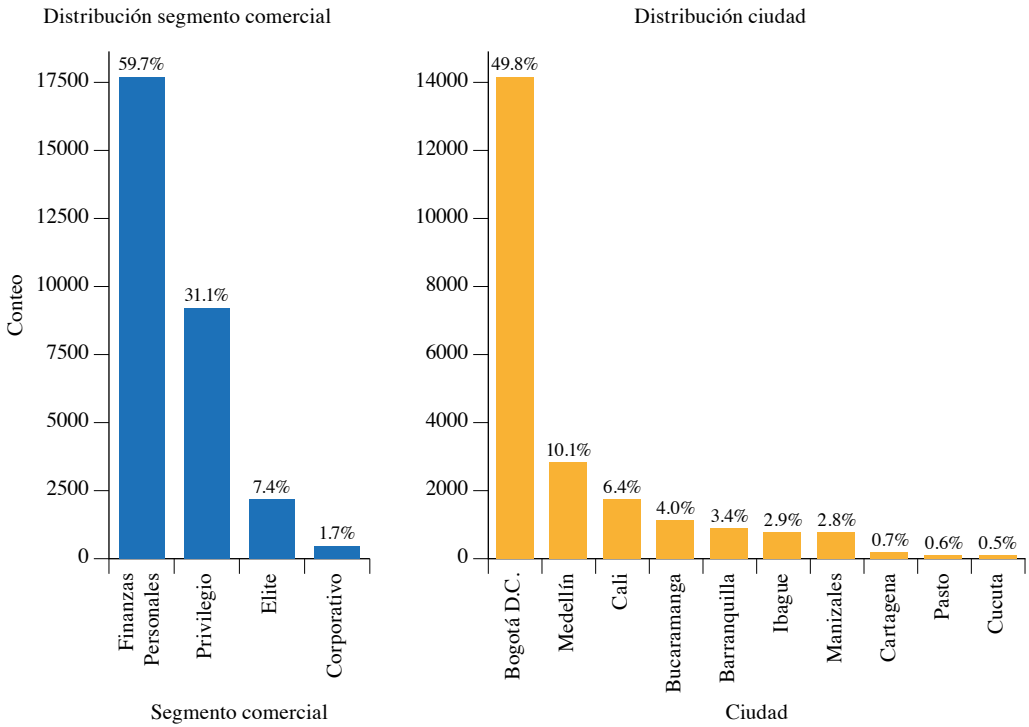
Fuente: elaboración propia.

lo que refleja también el comportamiento histórico que ha tenido la compañía debido al tipo de clientes atraído, siendo finanzas personales los clientes con más bajos activos y corporativo los clientes con mayores activos (figura 3).

f) Ciudad

Variable discreta que ubica a la ciudad de Bogotá en el primer puesto con mayor participación (49,8%), liderándola. Esto evidencia que la mayor parte del mercado está concentrada en la capital del país debido a que la operación nació en esta ciudad. Las siguientes dos ciudades son Medellín y Cali. Según el artículo Infolaft, "De acuerdo con los datos, Cundinamarca tiene 15.960 registros de personas y empresas con alguna mención negativa en medios, seguida de Antioquia con 6044 registros y de Valle del Cauca con 5424", es decir, la compañía presenta la mayor participación en las ciudades de mayor riesgo Sarlaft (figura 3).

Figura 3. Resultado de la distribución segmento comercial y ciudad



Fuente: elaboración propia.

g) CIIU

La variable discreta muestra un promedio superior de 67,2% de la participación con la actividad de elaboración de productos alimenticios (figura 4).

h) Distribución nacional y extranjera

El 96% de los clientes son nacionales y el riesgo por admitir ciudadanos extranjeros es muy bajo (figura 5).

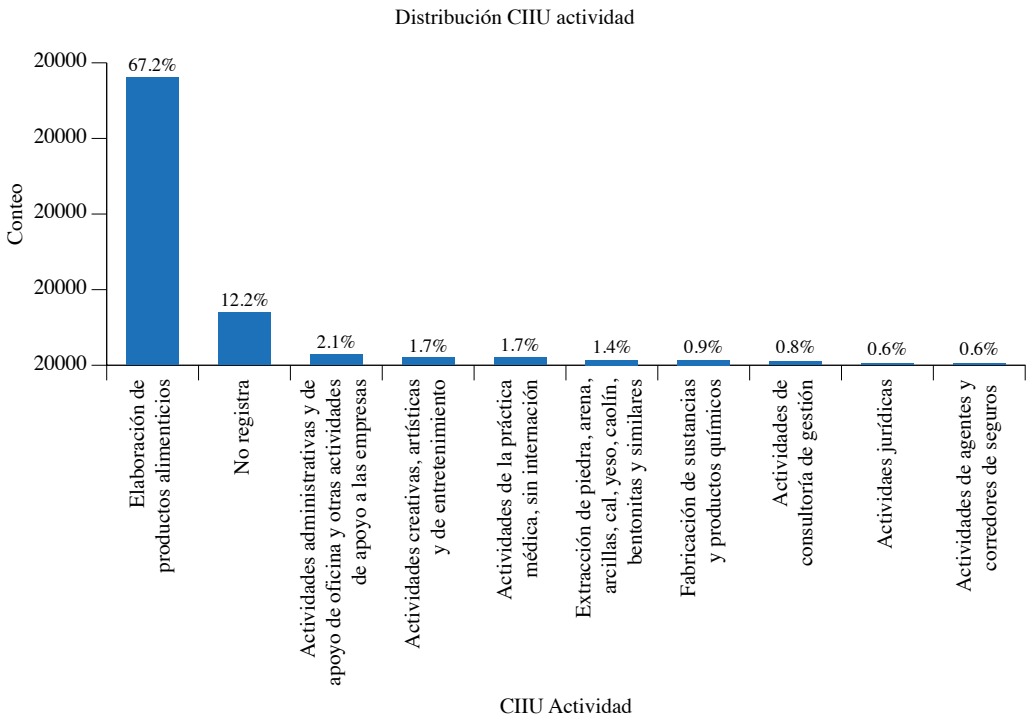
i) Género

Mismo nivel de hombres y mujeres y poca porción de empresas, es decir, la mayoría son personas naturales (figura 5).

j) Profesión

En la profesión, un 84,2% de la base de clientes se presenta como no definida, lo que significa que la variable se encuentra despoblada. Bajo estos datos se

Figura 4. Resultado distribución CIIU por actividad



Fuente: elaboración propia.

espera que esta variable no tenga mucho peso en el modelo debido a su baja calidad (figura 5).

k) Estado civil

La variable estado civil casado representa 39,1%, y en segundo lugar se ubica estado civil soltero con 29,4% (figura 5).

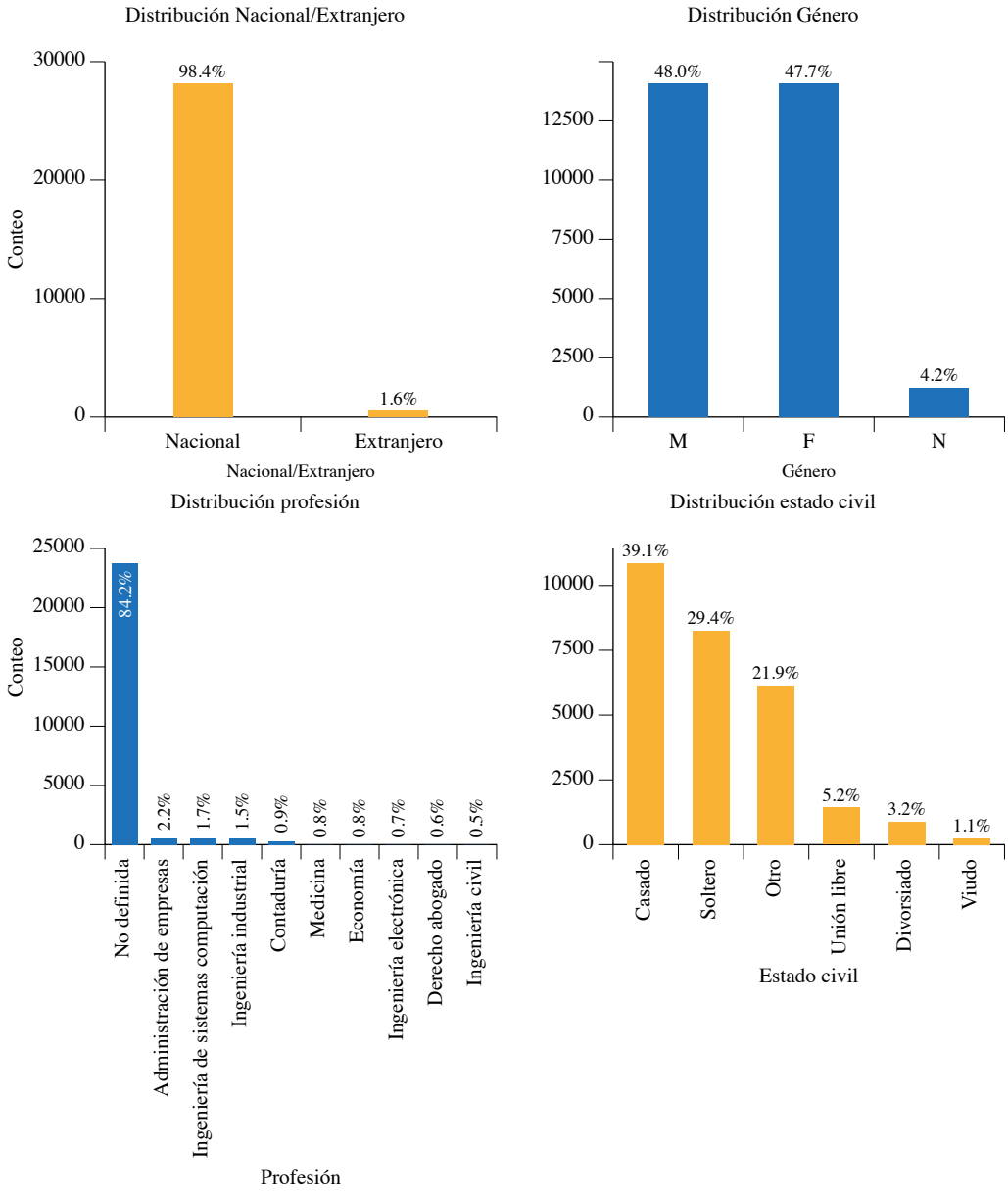
l) Años actualización

La proporción más grande está en los clientes que han tenido actualización en los primeros dos años, lo que implica que son clientes nuevos o se han podido actualizar rápidamente (figura 6).

m) Años antigüedad

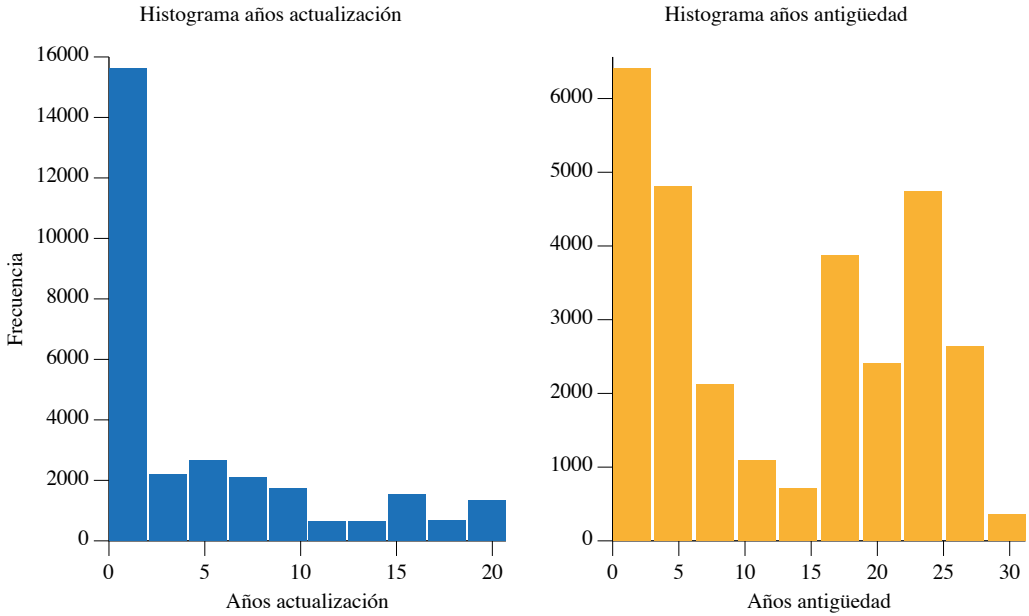
Los clientes de 0 a 5 años son aquellos que llevan poco tiempo en la compañía y normalmente permanecen hasta los 10 o 15 años, luego de este intervalo

Figura 5. Resultado distribución nacionalidad, género, profesión y estado civil



Fuente: elaboración propia.

Figura 6. Resultado histogramas de años de actualización y años de antigüedad



Fuente: elaboración propia.

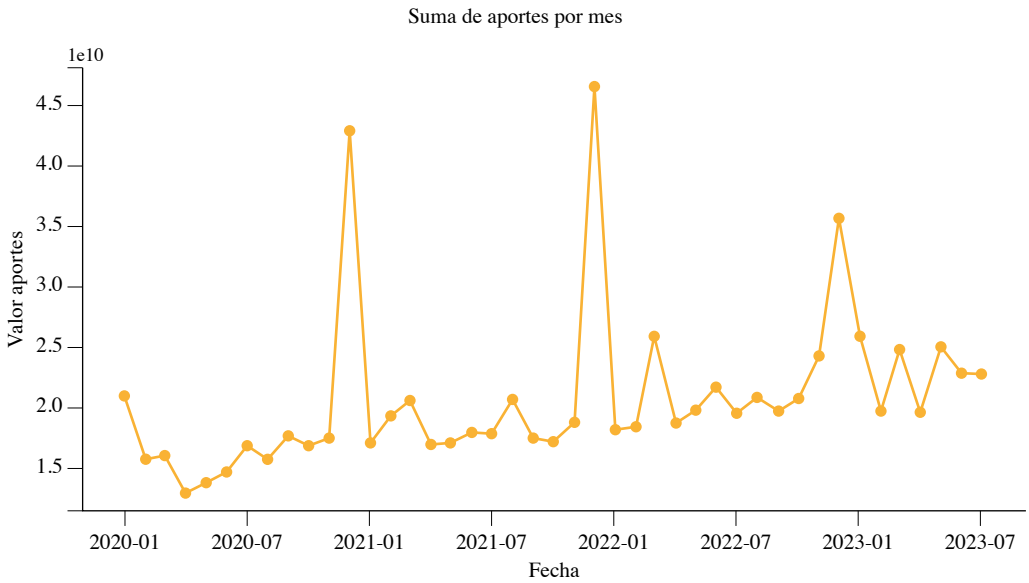
existen los clientes muy antiguos que presentan mayor fidelidad a la compañía (figura 6).

n) Suma de aportes por mes
El aumento de los aportes por mes normalmente presenta un crecimiento más elevado en diciembre, lo que implica para Sarlaft una gestión más exhaustiva en este mes debido al alto número de operaciones en las cuales puede aumentar la probabilidad de una operación sospechosa (figura 7).

4. Modelo Isolation Forest

Los resultados de la aplicación del modelo con ajuste de *tunning* de hiperparámetros obtenidos con las combinaciones dadas en la metodología al aplicar el modelo en el conjunto de datos de la aseguradora desde julio de 2020 hasta 2023, de manera mensual para un total de número de clientes de 29.405. Ubicación de la escala de Silhouette en la data de validación según la partición previamente realizada (figura 8).

Figura 7. Resultado suma de aportes por mes



Fuente: elaboración propia.

Figura 8. Escala de calificación de *Silhouette Score*

Silhouette Coefficient	Interpretación
0,7 < SC ≤ 1.0	Strong Structure
0,5 < SC ≤ 0,7	Medium Stucture
0,25 < SC ≤ 0,5	Weak Structure
SC ≤ 0,25	No Structure

Fuente: tomado de Benaya *et al*. (2023).

Al aplicar el modelo y evaluar los resultados mediante la métrica Silhouette score, que mide la cohesión dentro de los grupos y la separación entre grupos, el resultado obtenido es 0,82, que proporciona una estructura fuerte y sólida, lo que indica un modelo fuerte y óptimo, ya que los valores ubicados en la escala más cercanos a 1 indican una buena separación.

Sin embargo, para la aplicación de la métrica de Calinski Score, al buscar maximizar la relación dentro de los grupos, la métrica no cuenta con una escala que proporcione un rango para clasificar los resultados obtenidos, por lo tanto, la metodología y la aplicación indican que el valor más alto en este indicador corresponde al mejor modelo, lo que da como resultado un valor de 557.

En la tabla 1 se evidencian 25 modelos con distintos hiperparámetros, de donde se escoge el mejor por los criterios mencionados anteriormente. Según estas métricas, el mejor modelo es el que tiene 200 árboles y un parámetro de contaminación de 1%. En algunos modelos, variando el número de árboles, su resultado en las métricas es el mismo, por ende, en esta situación se define el mejor modelo bajo el criterio de parsimonia, lo que indica que un menor número de parámetros es más recomendable.

Tabla 1. Resultado *Silhouette Score* a partir del *tunning* de hiperparámetros

N_Estimators	Contam_ Param	Val_Silhouette_ Score	Val_Calinski_ Score	Val_Silhouette_ Score_Order	Val_Calinski_ Score_Order
200	0,01	0,82311	568,89152	2	2
500	0,01	0,82311	568,89152	2	2
1000	0,01	0,82311	568,89152	2	2
50	0,01	0,8219	557,6904	4,5	4,5
100	0,01	0,8219	557,6904	4,5	4,5
1000	0,05	0,55467	155,52031	6	6
100	0,05	0,5526	154,99498	7	8
500	0,05	0,55135	154,7881	8	9
200	0,05	0,55099	155,01287	9	7
50	0,05	0,54941	153,66406	10	10
1000	0,1	0,41826	102,99249	11	11
500	0,1	0,41439	102,31612	12	12
100	0,1	0,41241	101,77972	13	14
200	0,1	0,40888	102,18267	14	13
50	0,1	0,40858	97,08648	15	15
1000	0,2	0,29873	88,26348	16	16
500	0,2	0,29638	85,29538	17	17
200	0,2	0,29524	83,85754	18	18
50	0,2	0,29161	77,6397	19	23
100	0,2	0,29152	81,00094	20	20

N_Estimators	Contam_ Param	Val_Silhouette_ Score	Val_Calinski_ Score	Val_Silhouette_ Score_Order	Val_Calinski_ Score_Order
1000	0,3	0,2208	83,1742	21	19
500	0,3	0,21873	80,57637	22	21
200	0,3	0,21594	77,9551	23	22
100	0,3	0,21406	76,03873	24	24
50	0,3	0,20356	65,55971	25	25

Fuente: elaboración propia.

5. Importancia de las variables Isolation Forest

De acuerdo con los resultados, el análisis y procesamiento de las variables y la clasificación de la importancia según el modelo de Isolation Forest, la que más se destaca es la variable que tiene el porcentaje de la siguiente manera (figura 9):

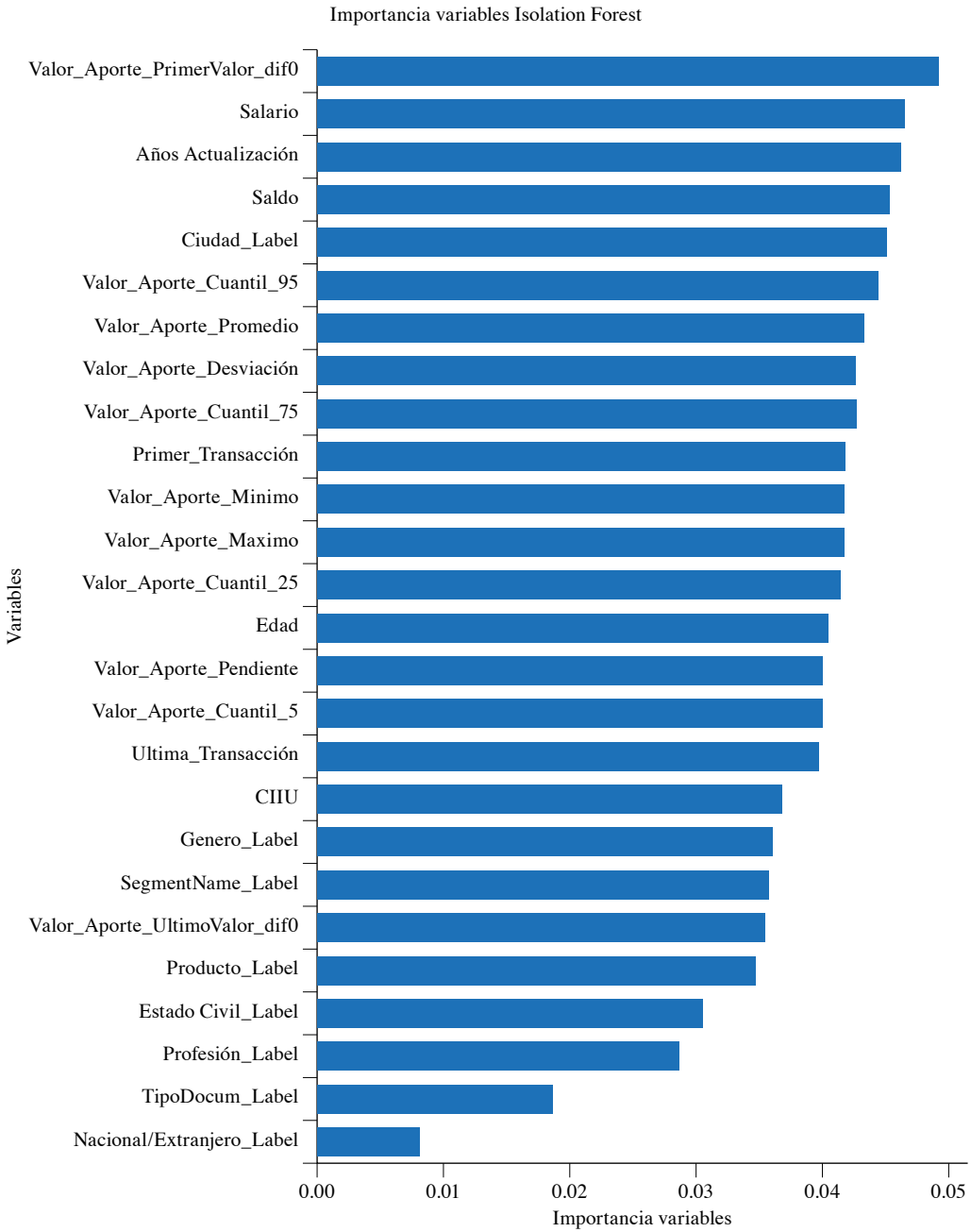
En primer lugar de importancia de las variables de Isolation Forest se ubica la variable de saldo de los clientes con 5%, variable importante dado que el objetivo principal en la detección de transacciones sospechosas en la aseguradora es el movimiento transaccional de los clientes, en el que se identifican cambios atípicos como consignaciones o retiros que establecen y perfilan el comportamiento de estos como posibles transacciones sospechosas que refuerzan la gestión de riesgo Sarlaft.

En segundo lugar, el modelo arroja como resultado de la importancia de las variables el valor aporte promedio con el 4,7%; esta variable tiene un gran sentido según el análisis y procesamiento del modelo, dado que las contribuciones significativamente diferentes ubican e identifican las transacciones sospechosas y aporta al perfilamiento de los clientes con esta anomalía.

En tercer lugar, con un resultado de 4,5%, se ubica la variable Años de actualización, que para la aseguradora es una variable determinante de acuerdo con el monitoreo y los resultados, ya que cambios drásticos en el aporte se consideran sospechosos, lo que genera la alerta para la aseguradora.

La última variable de nacional o extranjero, con resultado de 1%, se ubica en el último lugar dado que es una variable que presenta menos relevancia debido a que todos son nacionales (figura 9).

Figura 9. Resultado de la importancia de las variables *Isolation Forest*



Fuente: elaboración propia.

Dada la naturaleza del modelo (no supervisado), no existe una forma confiable de calcular la matriz de confusión dado que no hay etiquetas, no obstante, se da paso a la creación de una pseudovariable en la que se identifican como *outlier* los datos superiores al percentil 99 y no *outlier* los que estén por debajo. Con esta métrica se calcula la matriz de confusión y se compara con el resultado de Isolation Forest (figura 10).

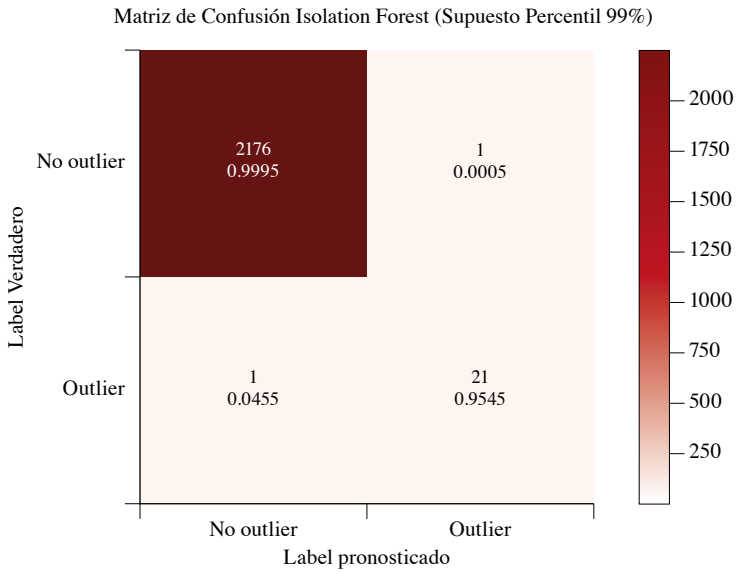
- Verdaderos negativos (no *outlier* identificados): 2176 (99,95%).
- Falso positivos (no *outlier* etiquetados como *outlier*): 1 (0,05%).
- Falsos negativos (*outlier* no detectados): 1 (4,55%).
- Verdaderos positivos (*outlier* correctamente detectados): 21 (95,45%).

Pasos para la construcción de la pseudovariable

i. Se toma la data de validación y se normaliza utilizando z-score.
ii. Una vez normalizada la data, se suman los valores de cada columna por fila, obteniendo como resultado una sola columna con la suma de *n* filas, es decir, *n* clientes.
iii. A partir del dato anterior, de la suma obtenida se calcula el percentil 99 asumiendo que los valores superiores a este percentil son considerados como atípicos.
iv. Se selecciona el mejor modelo de LOF y de Isolation Forest, y se extraen los resultados de cada cliente para cada modelo.
v. Con los resultados obtenidos, se calcula el percentil 99 para LOF y para Isolation Forest, e igualmente se consideran atípicos los superiores a este percentil.
vi. Finalmente, se calcula la matriz de confusión teniendo en cuenta el criterio de atípicos del Z-score vs. el criterio de atípicos de ambos modelos.

De acuerdo con la matriz de confusión, se identifica que solamente dos de los datos evaluados se encuentran en diferente categoría, denotando que con esta pseudovariable la precisión es relativamente alta, lo que demuestra que Isolation Forest es capaz de identificar los *outliers*.

Figura 10. Resultado de matriz de confusión *Isolation Forest* (supuesto percentil 99%)



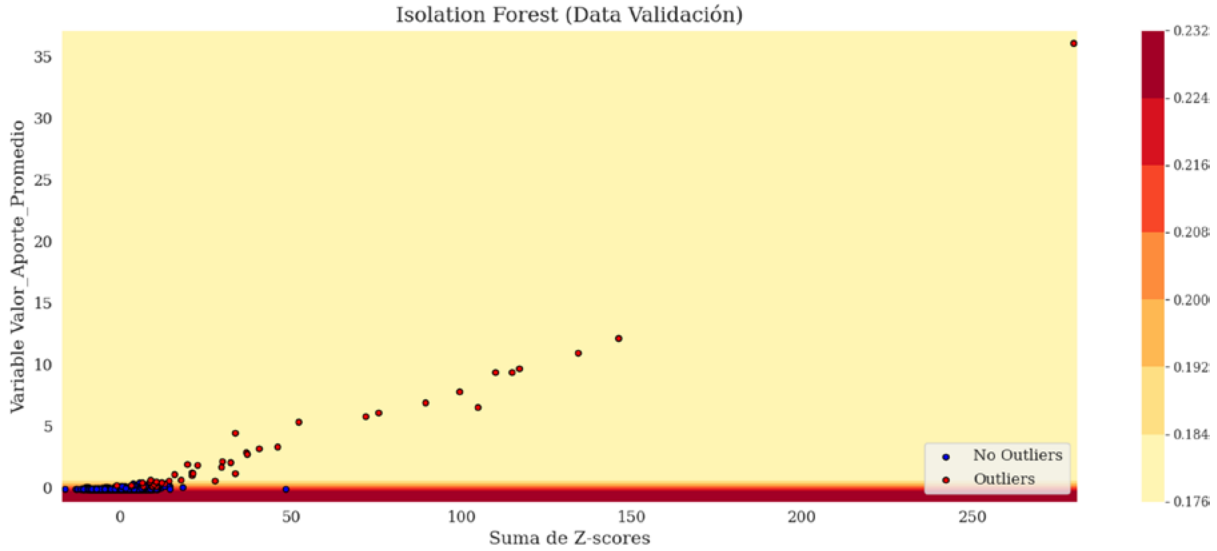
Fuente: elaboración propia.

Este modelo es altamente eficaz tanto en los datos no *outlier* como en los *outlier*, con una precisión significativa, lo que indica que el modelo está bien ajustado. Comparado con el LOF, Isolation Forest parece ser más equilibrado y preciso en la detección de ambas clases (figuras 11 y 12).

Las figuras muestran los dos grupos (*outlier* y no *outlier*) por escala para la data de validación y de prueba identificando los *outliers* por *grid* (malla). El *grid* más rojo identifica a los clientes que no son *outliers* y colores más tenues los que son *outliers*, comprobando así la lógica del modelo que divide distintos grupos hasta encontrar las regiones en las que se presentan los *outliers* y los no *outliers*.

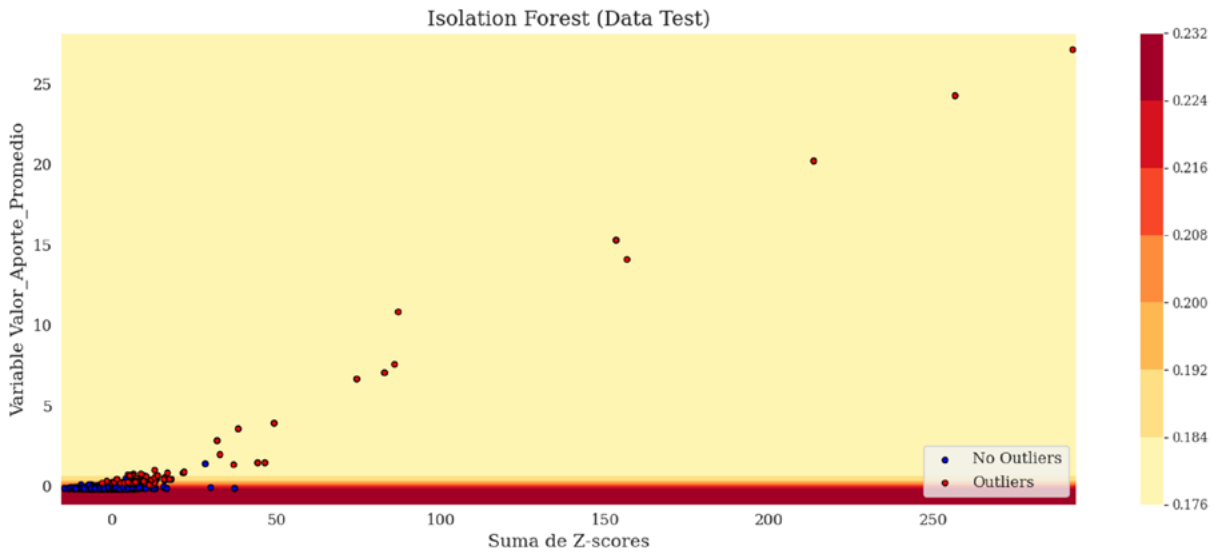
La diferencia es que los datos que se encuentran como *outlier* en la data de prueba (figura 12) tienen valores más pequeños que la data de validación (figura 11), lo que indica una clasificación un poco más robusta que los *outliers* comparada con la data de validación.

Figura 11. Resultado *Isolation Forest* (Data de validación) (Data Test)



Fuente: elaboración propia.

Figura 12. Resultado *Isolation Forest*



Fuente: elaboración propia.

6. Modelo Local Outlier Factor (LOF)

Al aplicar el modelo y evaluar los resultados obtenidos, la aplicación de la métrica Silhouette score es 0,827, la cual proporciona una estructura fuerte y sólida que indica como resultado un modelo fuerte y óptimo.

Sin embargo, para la aplicación de la métrica de Calinski Score, al maximizar la relación dentro de los grupos, la metodología y la aplicación da como resultado para el mejor modelo un valor de 344 evaluando la calidad de los grupos. El resultado es menor que el de Isolation Forest anteriormente comparado con los otros modelos reflejando diferencias atípicas.

Según la tabla 2, en la cual los experimentos se realizaron con diferentes parámetros y la evaluación dada utilizando métricas de calidad de agrupación LOF, que es un algoritmo eficaz en la detección de anomalías en conjunto de datos, los parámetros con mayores puntajes son las transacciones anómalas que requieren mayor atención y análisis en la prevención y el monitoreo.

Tabla 2. Resultado Local Outlier Factor

N_Neighbors	Contam_ Param	Val_Silhouette_ Score	Val_Calinski_ Score	Val_Silhouette_ Score_Order	Val_Calinski_ Score_Order
30	0,01	0,82742	344,41749	1,000	1,000
20	0,01	0,81226	267,80359	2,000	2,000
10	0,01	0,67905	70,54068	3,000	6,000
5	0,01	0,65507	50,17341	4,000	9,000
30	0,05	0,55417	124,25954	5,000	3,000
20	0,05	0,54965	117,80346	6,000	4,000
10	0,05	0,47976	72,84733	7,000	5,000
30	0,1	0,38799	66,6605	8,000	7,000
20	0,1	0,37189	63,20371	9,000	8,000
10	0,1	0,32521	49,17226	10,000	10,000
5	0,05	0,31226	18,86096	11,000	17,000
30	0,2	0,22001	37,42568	12,000	11,000
20	0,2	0,19584	29,3654	13,000	12,000
5	0,1	0,19062	11,39712	14,000	18,000

N_Neighbors	Contam_ Param	Val_Silhouette_ Score	Val_Calinski_ Score	Val_Silhouette_ Score_Order	Val_Calinski_ Score_Order
10	0,2	0,17279	28,04149	15,000	13,000
30	0,3	0,12704	27,24816	16,000	14,000
20	0,3	0,11188	21,15233	17,000	15,000
10	0,3	0,10297	20,49295	18,000	16,000
5	0,2	0,09293	7,59477	19,000	19,000
5	0,3	0,04885	5,8983	20,000	20,000

Fuente: elaboración propia.

En la figura 13, para el análisis y procesamiento de las variables, la clasificación de la importancia según el modelo de Local Outlier Factor que más se destaca es la variable que tiene el porcentaje de la siguiente manera:

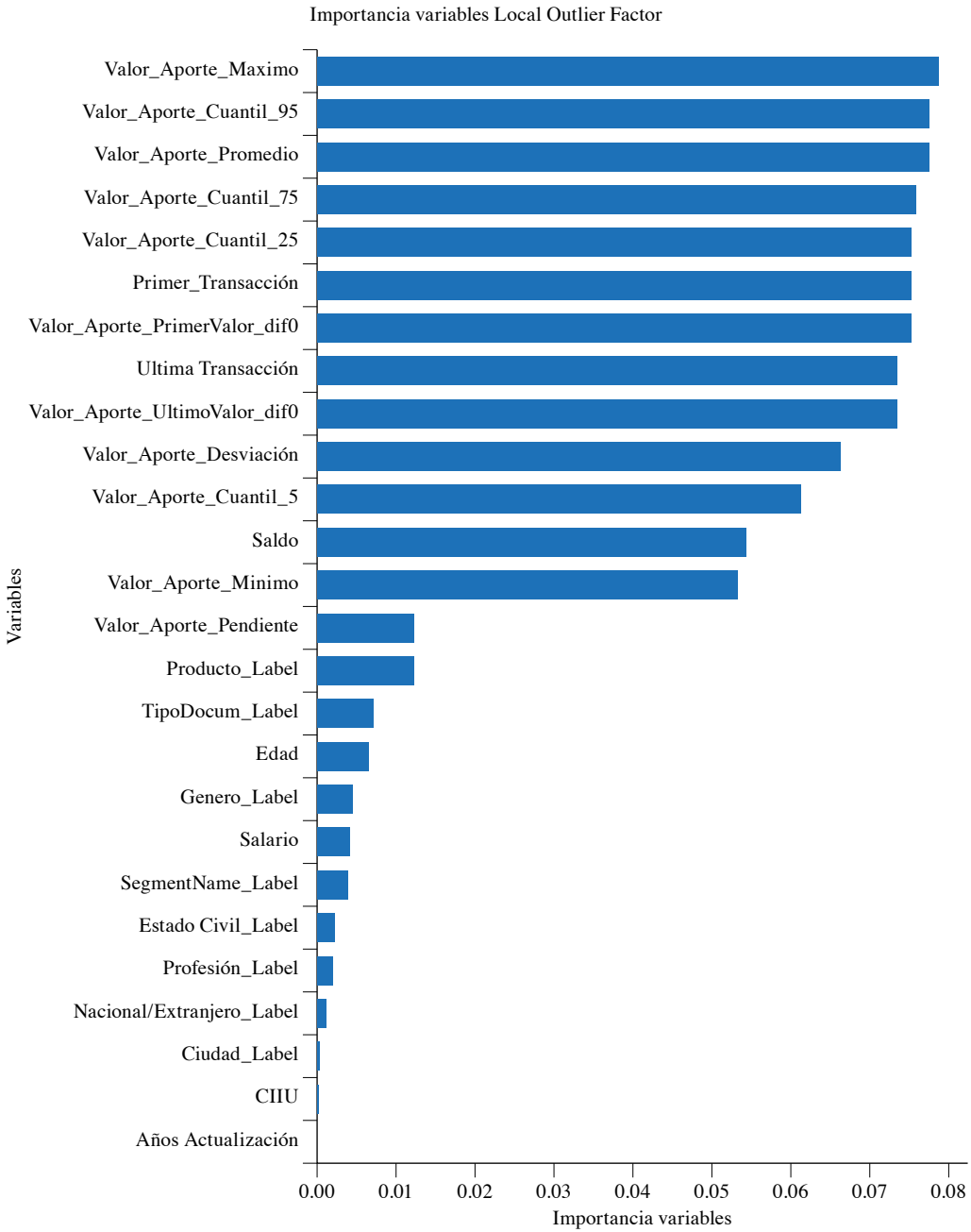
En primer lugar de importancia de las variables de LOF se ubica la de valor aporte máximo con 8%, variable importante dado que el objetivo principal en la detección de transacciones sospechosas en la aseguradora es el movimiento transaccional y los montos de los clientes, en los que se identifican cambios atípicos en las consignaciones o movimientos que permiten identificar y perfilar el comportamiento de estos como posibles transacciones sospechosas que refuerzan la gestión de riesgo Sarlaft.

En segundo lugar, el modelo arroja como resultado de la importancia de las variables el valor aporte cuantil_95 con el 7%.

En tercer lugar, con un resultado de 7%, se ubica el valor del aporte promedio, que para la aseguradora es una variable determinante de acuerdo con el monitoreo y los resultados, ya que cambios en los aportes promedio, contribuciones financieras regulares y que tengan un cambio atípico se consideran sospechosos y generan la alerta para la aseguradora.

La última variable, años actualización, con un resultado de 0%, se ubica en el último lugar dado que presenta menos relevancia debido a que, según el análisis del modelo, la actualización de los clientes no es determinante en la detección (figura 13).

Figura 13. Resultado importancia de las variables Local Outlier Factor



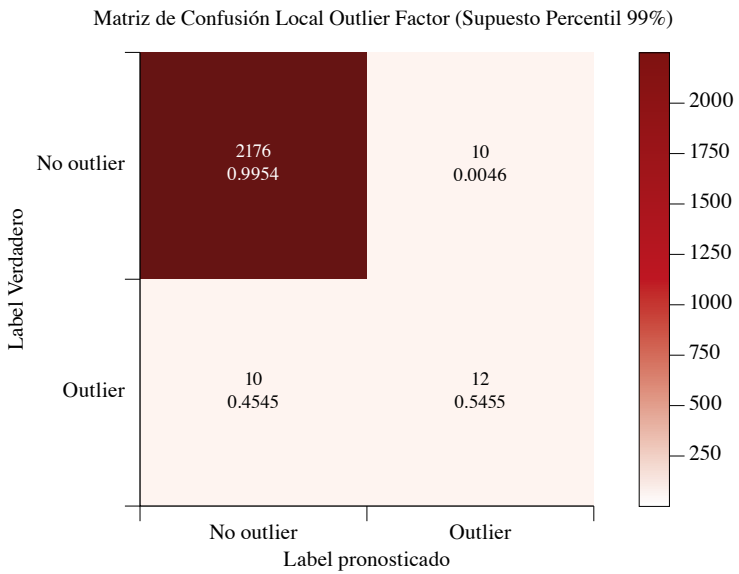
Fuente: elaboración propia.

La matriz de confusión según la aplicación del modelo Local Outlier Factor clasifica los datos de la siguiente forma (figura 14).

- Verdaderos negativos: 2167 instancias identificadas como no atípicas (0,9954).
- Falsos positivos: 10 instancias etiquetadas como atípicas (0,0046).
- Falsos negativos: 10 instancias etiquetadas como no atípicas (0,4545).
- Verdaderos positivos: 12 instancias identificadas como atípicas (0,5455).

El LOF es bastante eficaz para identificar los datos no *outlier*, pero tiene una efectividad moderada en la detección de *outlier* ya que casi la mitad de los casos que deberían ser detectados como *outlier* no lo son, lo cual concluye que el modelo es conservador y podría necesitar ajustes adicionales.

Figura 14. Resultado Matriz de confusión Local Outlier Factor (supuesto percentil 99%)



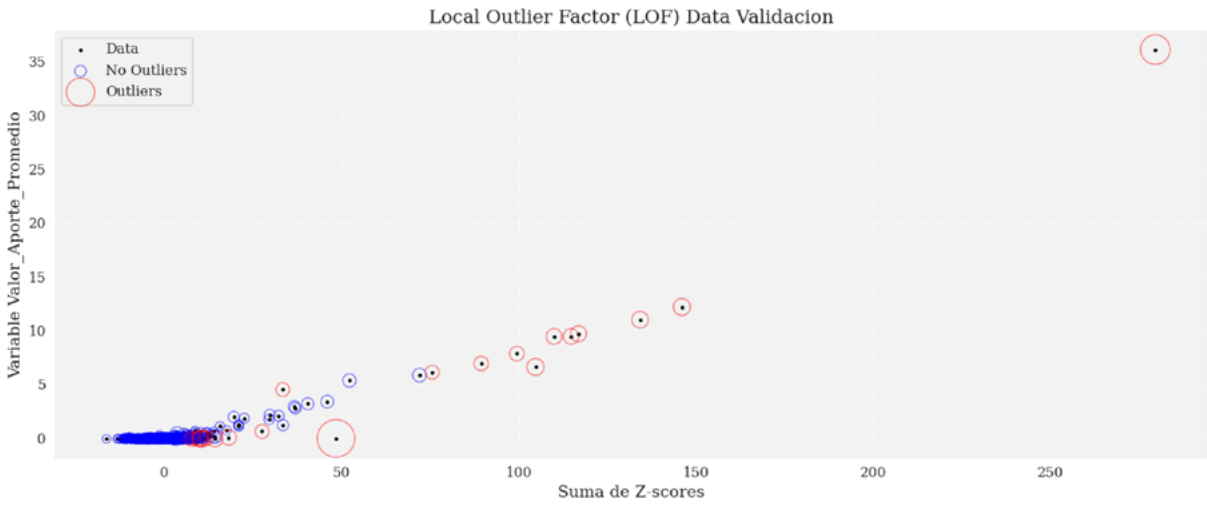
Fuente: elaboración propia.

En las figuras 15 y 16, en la aplicación del modelo Local Outlier Factor para la data de validación y prueba, se muestra el resultado de la identificación de los valores atípicos en el conjunto de datos, dada la densidad de sus vecinos más cercanos, entendiendo como Outlier la magnitud de cada punto, es decir, entre más grande su magnitud más cumple la característica de Outlier. A continuación se explica brevemente en qué consiste la gráfica:

- Eje *X*: representa la suma de las puntuaciones Z, indicando las desviaciones estándar valor por encima o por debajo de la media.
- Eje *Y*: representa la variable del conjunto de datos.
- Puntos negros (Data): representan los datos sin categorizar.
- Círculos azules (no *outliers*): puntos que no se consideran *outliers* en el modelo LOF.
- Círculos rojos (*outliers*): representan los puntos que son considerados *outliers* en el modelo LOF.
- El tamaño de los círculos rojos de los *outliers* indica su grado de atipicidad.

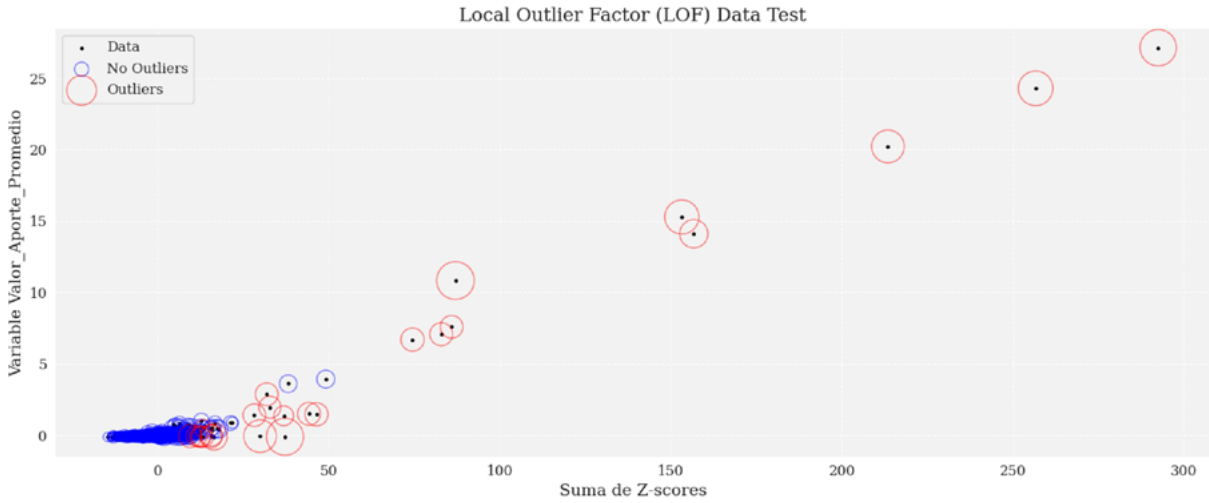
La mayoría de los datos se agrupan en la esquina inferior, lo que resulta en valores normales y los *outliers* están dispersos lejos del grupo principal, con características diferentes comparadas con la mayoría, esto quiere decir que cada dato se realiza el análisis de forma independiente y sus características de cada valor comparado con otros valores. De esta manera, de acuerdo con el tamaño y el círculo que se utiliza en el método Local Outlier Factor, la anomalía de los datos se identifica a partir del grado de atipicidad dado por el tamaño del círculo. Esto último ayuda a identificar de forma más precisa comportamientos atípicos de los clientes para que se realice la gestión correspondiente en la aseguradora de manera más efectiva para la prevención del lavado de activos y financiación del terrorismo (figura 15).

Figura 15. Resultado lof (data de validación)



Fuente: elaboración propia.

Figura 16. Resultado lof (data Test)



Fuente: elaboración propia.

7. Comparación Isolation Forest

De la comparación de los resultados en la data de prueba de los datos identificados como *outliers* vs, no *outliers* se extraen indicadores descriptivos como la media, el mínimo, el máximo, el número y cuantil 25, cuantil 75 (tabla 3).

Como se evidencia en los resultados, la variable del salario no está segmentando el cambio ni evidenciando cuáles son *outlier* y cuáles no lo son; para la variable de la edad de las personas, la que se identifica con mayores cambios a nivel de transaccionalidad en los datos atípicos es la de personas mayores. Para la variable de saldo se observa una importante diferencia al identificar como *outliers*, lo cual refleja que los clientes con estos datos atípicos tienen saldos mayores.

En la tabla 4 se encuentra el resumen detallado de los resultados obtenidos de las métricas de Isolation Forest en el conjunto de datos analizado.

Tabla 3. Resultados data de prueba de los datos identificados como outliers vs. No outliers isolation forest

Variable	Isolation Forest					
	Mean_outlier	Min_outlier	Max_outlier	Mean_no_outlier	Min_no_outlier	Max_no_outlier
Salario	$ 6.930.372.308,00	$ 1.161.200.000,00	$ 19.106.300.000,00	$ 6.811.447.922,00	$ 1.160.000.000,00	$ 19.891.000.000,00
Edad	52.231	19.000	87.000	43.086	19.000	95.000
Saldo	$ 52.920.771.701,00	$ 120.350.240,00	$ 580.535.583.210,00	$ 16.434.335.144,00	$ 287.710,00	$ 6.494.445.513.320,00
Valor_Aporte_Promedio	$ 34.217.468.355,00	$ 3.090.112.909,00	$ 334.900.000.000,00	$ 880.350.244,00	$ 36.000.000,00	$ 18.968.881.500,00
Valor_Aporte_Minimo	$ 25.245.607.552,00	$ 1.000.000.000,00	$ 334.900.000.000,00	$ 630.281.256,00	$ 3.737.000,00	$ 5.304.500.000,00
Valor_Aporte_Maximo	$ 56.282.647.737,00	$ 4.917.273.000,00	$ 467.500.000.000,00	$ 1.371.309.222,00	$ 38.000.000,00	$ 140.200.000.000,00
Valor_Aporte_Desviacion	$ 14.180.654.717,00	0.000	$ 220.325.498.664,00	$ 309.788.780,00	0.000	$ 48.992.859.899,00
Valor_Aporte_PrimerValor_dif0	$ 44.900.239.938,00	$ 1.000.000.000,00	$ 334.900.000.000,00	$ 794.266.728,00	$ 34.000.000,00	$ 20.000.000.000,00
Valor_Aporte_UltimoValor_dif0	$ 28.257.052.428,00	$ 1.000.000.000,00	$ 334.900.000.000,00	$ 863.564.297,00	$ 38.000.000,00	$ 10.025.149.480,00
Primer_Transaccion	$ 44.900.239.938,00	$ 1.000.000.000,00	$ 334.900.000.000,00	$ 794.266.728,00	$ 34.000.000,00	$ 20.000.000.000,00
Ultima_Transaccion	$ 28.257.052.428,00	$ 1.000.000.000,00	$ 334.900.000.000,00	$ 863.564.297,00	$ 38.000.000,00	$ 10.025.149.480,00
Valor_Aporte_Pendiente	-$ 5.525.073.614,00	-$ 264.398.527.110,00	$ 23.221.000.000,00	$ 19.572.720,00	-$ 8.133.865.952,00	$ 3.211.047.000,00
Valor_Aporte_Cuantil_5	$ 25.577.392.175,00	$ 1.000.000.000,00	$ 334.900.000.000,00	$ 656.867.106,00	$ 34.000.000,00	$ 5.304.500.000,00
Valor_Aporte_Cuantil_25	$ 26.848.990.249,00	$ 1.000.000.000,00	$ 334.900.000.000,00	$ 738.338.161,00	$ 34.000.000,00	$ 5.304.500.000,00
Valor_Aporte_Cuantil_75	$ 37.316.921.364,00	$ 2.000.000.000,00	$ 379.952.540.250,00	$ 937.341.352,00	$ 38.000.000,00	$ 11.588.665.000,00
Valor_Aporte_Cuantil_95	$ 51.112.175.392,00	$ 4.097.727.500,00	$ 457.372.500.000,00	$ 1.248.402.230,00	$ 38.000.000,00	$ 92.372.868.200,00

Fuente: elaboración propia.

Tabla 4. Resultados de las métricas Isolation Forest

Resultado Metricas Isolatión Forest	
Accuracy	0.9990904956798545
F1-Score	0.9545454545454546
AUC	0.9770430534096128

Fuente: elaboración propia.

ODEON, ISSN: 1794-1113, E-ISSN: 2346-2140, N° 27, julio-diciembre de 2024, pp. 171-207

8. Local Outlier Factor

En la tabla 5 se presenta el resultado de *outlier* vs. no *outlier*.

Tabla 5. Resultado tabla descriptiva de outlier vs. no outlier Local Outlier Factor

Variable	Local Outlier Factor					
	Mean_outlier	Min_outlier	Max_outlier	Mean_no_outlier	Min_no_outlier	Max_no_outlier
Salario	$ 6.941.461.538,00	$ 1.161.200.000,00	$ 15.954.100.000,00	$ 6.813.192.700,00	$ 1.160.000.000,00	$ 19.891.000.000,00
Edad	52.577	37.000	87.000	43.225	19.000	95.000
Saldo	$ 676.260.663.182,00	$ 37.843.850,00	$ 6.494.445.513.320,00	$ 10.504.184.261,00	$ 287.710,00	$ 1.345.958.923.100,00
Valor_Aporte_Promedio	$ 69.489.906.467,00	$ 419.879.000,00	$ 334.900.000.000,00	$ 1.034.006.166,00	$ 36.000.000,00	$ 50.000.000.000,00
Valor_Aporte_Minimo	$ 52.155.388.265,00	$ 300.000.000,00	$ 334.900.000.000,00	$ 734.711.030,00	$ 3.737.000,00	$ 46.000.000.000,00
Valor_Aporte_Maximo	$ 112.963.176.423,00	$ 500.000.000,00	$ 467.500.000.000,00	$ 1.639.198.695,00	$ 38.000.000,00	$ 90.000.000.000,00
Valor_Aporte_Desviacion	$ 27.960.420.646,00	$ -	$ 220.325.498.664,00	$ 383.069.910,00	$ -	$ 27.080.128.015,00
Valor_Aporte_PrimerValor_dif	$ 86.459.223.308,00	$ 300.000.000,00	$ 334.900.000.000,00	$ 1.050.824.846,00	$ 34.000.000,00	$ 90.000.000.000,00
Valor_Aporte_UltimoValor_dif	$ 52.316.227.765,00	$ 336.000.000,00	$ 334.900.000.000,00	$ 1.041.184.626,00	$ 38.000.000,00	$ 46.000.000.000,00
Primer_Transaccion	$ 86.459.223.308,00	$ 300.000.000,00	$ 334.900.000.000,00	$ 1.050.824.846,00	$ 34.000.000,00	$ 90.000.000.000,00
Ultima_Transaccion	$ 52.316.227.765,00	$ 336.000.000,00	$ 334.900.000.000,00	$ 1.041.184.626,00	$ 38.000.000,00	$ 46.000.000.000,00
Valor_Aporte_Pendiente	-$ 14.531.733.551,00	-$ 264.398.527.110,00	$ 14.429.585.700,00	$ 26.765.641,00	-$ 14.000.000.000,00	$ 23.221.000.000,00
Valor_Aporte_Cuantil_5	$ 52.687.000.817,00	$ 300.000.000,00	$ 334.900.000.000,00	$ 763.987.304,00	$ 34.000.000,00	$ 46.000.000.000,00
Valor_Aporte_Cuantil_25	$ 54.772.585.641,00	$ 300.000.000,00	$ 334.900.000.000,00	$ 855.587.239,00	$ 34.000.000,00	$ 46.000.000.000,00
Valor_Aporte_Cuantil_75	$ 75.939.940.278,00	$ 419.879.000,00	$ 379.952.540.250,00	$ 1.103.649.065,00	$ 38.000.000,00	$ 52.500.000.000,00
Valor_Aporte_Cuantil_95	$ 102.715.584.131,00	$ 500.000.000,00	$ 457.372.500.000,00	$ 1.490.278.895,00	$ 38.000.000,00	$ 82.500.000.000,00

Fuente: elaboración propia.

En la tabla 6 se encuentra el resumen detallado de los resultados obtenidos de las métricas de Local Outlier Factor en el conjunto de datos analizado.

Tabla 6. Resultados de las métricas Local Outlier Fator

Resultado Metricas Local Outlier Factor	
Accuracy	0.9909049567985448
F1-Score	0.5454545454545454
AUC	0.7704305340961289

Fuente: elaboración propia.

Dados los resultados de los mejores modelos, el Silhouette score más alto fue de 0,827 con el Local Outlier Factor, y para Isolation Forest fue de 0,82311, una diferencia mínima, escogiendo el mejor modelo por el valor más alto en este criterio (tabla 7). Para el Calinski-Harabasz score, el Isolation Forest obtuvo el mejor resultado con 568,89152, mientras que el Local Outlier Factor alcanzó 344,417, mostrando una diferencia significativa. Estos resultados sugieren que es necesario utilizar otras métricas para una evaluación más completa. Por lo tanto, se recurre a una pseudovariable para determinar de manera más precisa cuál modelo es mejor en la detección de *outliers*.

Tabla 7. Resultados mejor modelo

Resultado mejor modelo		
	Val_Silhouette_Score	Val_Calinski_Score
Isolatión Forest	0.82311	568.89152
Local Outlier Factor	0.827	344.417

Fuente: elaboración propia.

Una vez evaluada la matriz de confusión con las pseudovariables, se encontraron los siguientes resultados con menor tasa de error: para no *outlier* 1 y *outlier* 1 en Isolation Forest y en Local Outlier Factor fue de 10 tanto para no *outlier* como para *outlier*.

Respecto a los resultados de las métricas (tablas 3 y 5) para Isolation Forest en comparación con Local Outlier Factor, para Accuracy los dos modelos tienen alta exactitud, sin embargo, Isolation Forest presenta un valor superior. F1- Score en Isolation Forest es más alto, lo que indica que tiene un mejor equilibrio entre precisión y sensibilidad. El Isolation Forest tiene un AUC más alto, lo que

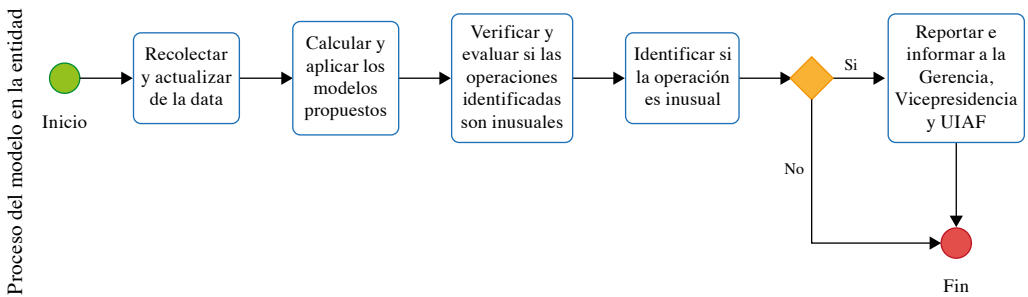
sugiere que es mejor para distinguir entre clases positivas y negativas. Por lo tanto, Isolation Forest es más efectivo en general, en términos de precisión y sensibilidad, basado en la pseudovariable.

9. Implementación

Para una implementación adecuada del modelo dentro de la entidad se propone un flujograma (figura 17) que permite resumir de manera breve cómo debería funcionar el modelo calculado para la correcta gestión de Sarlaft.

i. Recoger y actualizar la data permanentemente para asegurar que los modelos trabajen con la información más reciente.
ii. Calcular y aplicar los modelos propuestos para identificar los datos atípicos (operaciones inusuales).
iii. Verificar y evaluar si las operaciones identificadas como inusuales son inusuales o son sospechosas.
iv. Reportar e informar a la Gerencia, Vicepresidencia y UIAF sobre las operaciones detectadas.

Figura 17. Proceso del modelo en la entidad



Fuente: elaboración propia.

Conclusiones

Los resultados obtenidos resaltan diferencias en cuanto al desempeño de cada algoritmo con las metodologías utilizadas.

Al utilizar los modelos Isolation Forest y Local Outlier Factor, en los cuales se ajustaron los mejores hiperparámetros (Isolation Forest: N_Estimators. Contamination Parameter / Local Outlier Factor: N_Neighbours y Contamination

Parameter), según las métricas utilizadas (Calinski_Score, Silhouette_Score, Accuracy, F1-Score, AUC), se puede concluir que Isolation Forest tiene mejor desempeño en términos de precisión y Local Outlier Factor tiene una mejor capacidad en la identificación de casos específicos de acuerdo con su grado de atipicidad.

Según la comparación y los resultados, ambos métodos en general tienen sus particularidades, dadas las tablas 3 y 5 calculadas para la data de prueba. Ambos métodos detectan *outliers* con medias y rangos muy similares para los aportes en general. Local Outlier Factor es más sensible con valores extremos en algunas variables, mientras que Isolation Forest ofrece una detección más conservadora, es decir, de forma grupal. Los *outliers* calculados con LOF presentan una desviación estándar más alta que Isolation debido a la sensibilidad y especificidad de transacciones con mayor potencial de ser *outlier* dado el comportamiento histórico de los clientes.

Los resultados anteriores solamente sugieren las transacciones atípicas dada la historia de estas. Sin embargo, para la gestión de Sarlaft es importante tener en cuenta la capacidad tanto humana como tecnológica que permita evaluar las operaciones detectadas como inusuales para posteriormente catalogarlas como sospechosas o no sospechosas.

Los resultados obtenidos son propuestas que deben implementarse en tiempo real dentro de la entidad debido a la naturaleza de las transacciones. Una vez aplicadas en la compañía, se valida la efectividad y robustez dentro de los procesos y procedimientos de riesgo Sarlaft.

Adicionalmente, se sugiere realizar y desarrollar un modelo más complejo, aumentando el grado de precisión en la detección de datos atípicos como redes neuronales convolucionales o de aprendizaje no supervisado robusto como: Recurrent Neural Networks (RNN), Bayesian Networks, Gaussian Mixture, Modelos generativos, entre otros.

Referencias

Amat Rodrigo, J. (s. f.). *Detección de anomalías: Isolation Forest*. https://www.cienciadedatos.net/documentos/66_deteccion_anomalias_isolationforest.html

Arriagada, M. (2020). Ciencia de datos: hacia la automatización de las decisiones. *Ingeniare. Revista chilena de ingeniería*, *28*(4), 556-557. https://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-33052020000400556.

Artasanchez, A. y Prateek, J. (2020). *Artificial Intelligence with Python* (pp. 100-101). Packt Publishing.

Aung, M., Drachen, A. y Bonomett, V. (2018). *Predicting skill learning in a large, longitudina*. MOBA dataset.

Bagley, B. M. (2000). Narcotráfico, violencia política y política exterior de Estados Unidos hacia Colombia en los noventa. *Colombia Internacional*, (49-50), 5-38. https://doi.org/10.7440/colombiaint49-50.2000.01.

Benaya, R., Sibaroni, Y. y Ihsan, A. F. (2023). Clustering content types and user roles based on tweet text using k-medoids partitioning based. *Journal of Computer System and Informatics (JoSYC)*,*4*(4), 749-756.

Blanco, J. I. (2023). *¿Por qué la normalización es clave e importante en Machine Learning y Ciencia de Datos?* https://jorgeiblanco.medium.com/por-qu%C3%A9-la-normalizaci%C3%B3n-es-clave-e-importante-en-machine-learning-y-ciencia-de-datos-4595f15d5be0

Brownlee, J. (2020, agosto). *Métricas para evaluar algoritmos de aprendizaje automático en Python*. *Machine Learning Mastery*. https://machinelearningmastery.com/metrics-to-evaluate-machine-learning-algorithms-in-python/

Buczak, A. L. y Guven, E. (2016). *A survey of data mining and machine learning methods for cybersecurity intrusion detection*. IEEE.

Caliński, T., Harabasz, J. (1974). Un método de dendritas para el análisis de conglomerados. *Comunicaciones en Estadística*, *3*(1), 1-27.

Cuevas, E., Avalos O., Díaz P., Valdivia A. y Pérez M. (2021). Introducción *al Machine learning con MATLAB*. AUM.

Denzin, N. K., Lincoln, Y. S. (1994). Manual de investigación cualitativa 1. (Artículo Ingresando al campo de la investigación cualitativa).

Fernández, R. y Martínez, A. (2020). Gestión de riesgos y análisis de transacciones sospechosas en la Sagrilaft: un estudio de caso en el sector financiero. *Revista de Investigación en Contabilidad, Finanzas y Administración*, *15*(1), 78-95.

Guevara, J. I. y Granados, O. (2021). Machine learning methodologies against money laundering in non-banking correspondents (Tesis de grado). U. Tadeo Lozano.

Hariri, S., Kind, M. C. y Brunner, R. J. (2021). Extended Isolation Forest. *IEEE Transactions on Knowledge and Data Engineering.*

Hernández, R., Fernández, C., Baptista, M. P. (2010). *Metodología de la investigación* (6ta ed.). McGraw Hill.

Liu, F. T. y Ting, K. M. (2008). *Isolation Forest*. Gippsland School of Information Technology Monash University. https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf.

Mariño, G., Chaparro, F. y Medina, I. (2014). Determinantes en la prevención del riesgo para el lavado de activos y la financiación del terrorismo (LA/FT) en el sector real. *AD-minister*, (25), 7-35.

Soria Olivas, E., Sánchez, M. A., Montañés Isla, R., Gamero Cruz R., Castillo Caballero, B. y Cano Michelena, P. (2023, mayo). *Sistema de aprendizaje automático*. Edicion Ra-Ma.

Zabala, T. La eficacia del Sarlaft en Colombia. *Revista de la Ciencia y la Investigación*, *1*(16). http://167.249.40.87/index.php/DERROTERO/article/view/244. (2023).

Zou, K. H., O'Malley, A. J. y Mauri, L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*, *115*(5), 654-657.