Análisis Exploratorio para la prevención de fraude en Positiva Compañía de Seguros

Martínez H. Andrés F
Facultad de Inegniería y Ciencias Básicas
Universidad Central
Maestría en Analítica de Datos
Curso de Automatización e integración de datos
Bogotá, Colombia
amartinezh3@ucentral.edu.co

April 27, 2024

Contents

1 Introducción

Dentro de la economía colombiana, el sector asegurador es uno de los más fuertes. Durante el 2023, esta industria emitió un total de $60.6 billones pesos, representando un crecimiento del 7% con respecto al 2022 (La República, 2024). Esto equivale a un crecimiento según el indicador de penetración (calculado por la diferencia entre el valor consolidado de las primas y el PIB calculado del 1% para el 2023), ya que pasó del 3.24% al 3.42%. (Portafolio, 2024).

Por otra parte, los eventos que fueron atendidos por las aseguradoras también aumentaron en un 3%, lo cual representó $21.89 billones de pesos. (Portafolio, 2024). La siniestralidad más alta, se presenta en los seguros de vida (Gráfico 1) del cuál se pagaron cerca de $2.471 mil millones de pesos.

Dentro del análisis de la ocurrencia de estos siniestros, es importante tener en cuenta el factor fraude en la industria, que permanece muy latente dentro de los valores pagados. El sector asegurador ha implementado diferentes acciones durante el 2022 para prevenir estos fraudes y así se pudieron controlar cerca de 41.000 posibles eventos, cuyo pago hubiese significado entregar cerca de $202.000 millones en reclamaciones ilegales (Fasecolda, 2024).

Sin embargo, el riesgo de fraude puede haberse materializado en algunos casos, aún con los controles definidos para evitarlo. Los seguros más vulnerables en este sentido resultan siendo el SOAT, riesgos laborales, autos, salud y sustracción. La mayor parte de estos sucesos corresponden a eventos que no sucedieron. Para estos casos las estrategias más utilizadas fueron la simulación del siniestro, el uso de información falsa o adulterada, los cobros dobles o el uso de pólizas que están a nombre de terceros (Fasecolda, 2024).

2 Características del proyecto de investigación que hace uso de Integración y Automatización de Datos para IA

2.1 Titulo del proyecto de investigación

Análisis Exploratorio para la prevención de fraude en Positiva Compañía de Seguros.

2.2 Objetivo general

Diseñar y desarrollar un tablero de control automatizado para la identificación de los factores de riesgo de fraude predominantes en las pólizas de Positiva Compañía de Seguros.

2.2.1 Objetivos especificos

- Recolectar, y análizar datos históricos de reclamaciones y transacciones de pólizas de seguros de Positiva Compañía de Seguros durante 2023, para

identificar patrones y tendencias relacionadas con el fraude.

- Identificar y seleccionar las variables relevantes que podrían inidicar la presencia de actividades fraudulentas en las pólizas de seguros, como anomalías en los reclamos, inconsistencia en la información proporcionada por los asegurados, entre otros.
- Diseñar e implementar algoritmos y modelos estadísticos supervisados para el análisis de datos, y de machine learning, con el fin de detectar los principales factores de riesgo de fraude en las pólizas de seguro.
- Desarrollar un tablero de control automátizado que visualice de manera clara y concisa los resultados del análisis de riesgo de fraude, destacando los factores de riesgo identificados, proporcionando herramientas interactivas para explorar los datos en detalle.

2.3 Alcance

Positiva Compañía de Seguros S.A. es una empresa colombiana dedicada a ofrecer seguros a personas naturales y jurídicas, que cuenta con más de 60 años de experiencia en el mercado, cuando se constituyó Seguros Tequendama de Vida, la cual fue adquirida en 1995 por La Previsora S.A. Tras la integración de operaciones con ARP Seguro Social en 2008, ambas empresas formaron la actual compañía. Actualmente, Positiva cuenta con una base de 6 millones de personas aseguradas (equivalente al 8% del mercado) en toda Colombia, a través de productos que protegen riesgos laborales, salud, vida, rentas, etc.

El presente estudio pretende, tomando como referencia una base de datos ofrecida por la empresa que contiene los registros de las pólizas vendidas por la compañía y que fueron cobradas por siniestros (riesgos materializados) durante el año 2023 (actualizada trimestralmente) en Colombia, identificar los diferentes patrones que se presentan en los eventos de fraude en el cobro de estos seguros. De esta manera, se espera generar una herramienta que ayude a la compañía a identificar y generar acciones para mitigar los riesgos asociados con los cobros ilegales de sus pólizas.

2.4 Pregunta de investigación

¿Cuáles son los principales agentes identificados en los casos de fraude externo reportados por Positiva Compañía de Seguros durante el año 2023, y cuáles son los factores de riesgo asociados que deben ser considerados en el diseño e implementación de medidas preventivas y de detección de fraudes en la compañía?

2.5 Hipótesis

A partir del uso de los datos obtenidos por Positiva Compañía de Seguros derivados de eventos fraudulentos ocurridos con el pago de sus seguros, se pueden generar herramientas que permitan determinar las correlaciones existentes entre

los factores de ocurrencia y muestren estos resultados en un dashboard en tiempo real, junto con algunos indicadores generales de control. Con esto, se busca entregar un instrumento que apoye análisis más profundos, los cuales permitirán tomar decisiones que mitiguen la materialización de estos riesgos.

3 Reflexiones sobre el origen de datos e información

Con base en el enfoque del presente estudio, y para dar claridad en la obtención y manejo de los datos a utilizar, es necesario establecer:

3.1 ¿Cuál es el origen de los datos e información?

El presente estudio se realiza tomando como base un dataset aportado por Positiva Compañía de Seguros S.A., con fines educativos y con compromiso de manejarla conforme a los requisitos exigidos por la ley para tal fin. Esta base de datos contiene los registros de las pólizas activas que presentaron alguna reclamación durante el año 2023, y para su manejo, se han identificado 42 variables, de las cuales 28 son cualitativas y 14 son cuantitativas.

En la primera fase del proceso, en la cual se estandarizará la base de datos y se eliminarán los datos que no se utilizarán, quedarán 39 variables.

3.2 ¿Cuáles son las consideraciones legales o éticas del uso de la información?

Es importante establecer el marco legal bajo el cual se realizará el presente estudio, pues este estará bajo los lineamientos establecidos para la manipulación y uso de los datos recibidos por parte de la compañía. En principio, todo el proceso estará en función al cumplimiento de las siguientes normas:

Ley 1581 de 2012: Esta ley establece el marco general para la protección de datos en Colombia. Regula la recolección, almacenamiento, uso, circulación y supresión de datos personales, y garantiza los derechos de las personas sobre su información personal.

Decreto 1377 de 2013: Este decreto establece disposiciones adicionales a la Ley 1581 de 2012. Define aspectos como el registro de las bases de datos, los procedimientos para el ejercicio de los derechos de los titulares de la información, y las responsabilidades de los encargados y responsables del tratamiento de los datos.

Al mismo tiempo, es importante aclarar que, para garantizar el cumplimiento de estas leyes y el acceso a la documentación de sus procesos, los resultados obtenidos serán compartidos exclusivamente con la compañía y todos los datos fueran anonimizados, esto a fin de proteger la información de sus afiliados.

3.3 ¿Cuáles son los retos de la información y los datos que utilizara en Integración y Automatización de Datos para IA?

La base de datos a utilizar contiene una mezcla de 43 variables, entre variables cuantitativas y cualitativas, que describen las pólizas reclamadas por siniestros ocurridos durante un periodo de tiempo establecido. Esta base de datos es alimentada de forma continua cada 3 meses por la misma compañía, esto con el

fin de asegurar una trazabilidad de los riesgos materializados y de los productos que más se afectaron por ello.

El reto en este proyecto es diseñar un tablero de control que se actualice de forma constante e integre los datos que se vayan generando en tiempo real. De esta manera, se asegura la disponibilidad de los principales indicadores de riesgo actualizados en todo momento y servirá como herramienta en su diseño de estrategias de prevención de estos riesgos.

3.4 ¿Qué espera de la utilización de Integración y Automatización de Datos para IA para su proyecto?

Con la realización del presente estudio, se espera generar una herramienta dinámica que la empresa Positiva Compañía de Seguros pueda utilizar para la estructuración de sus procesos de prevención de fraude en el cobro de pólizas de seguro de sus afiliados en un futuro. De esta manera, se aplicarán conocimientos y técnicas de tratamiento de bases de datos para identificar los principales ecosistemas de fraude que se presentan actualmente en esta industria y los diferentes agentes que participan en él, fungiendo como un punto de partida en el uso de análisis posteriores más profundos.

4 Diseño de integración y Automatización de Datos para IA

Como primer paso en el desarrollo del presente estudio, y con el fin de entender las variables que presenta la base de datos y como estas se relacionan entre sí, se genera un diagrama entidad-relación. Con este, se pretende entregar al lector un panorama global del objeto de estudio, que le permitirá identificar de manera orgánica la integración de los diferentes registros dentro del proyecto.

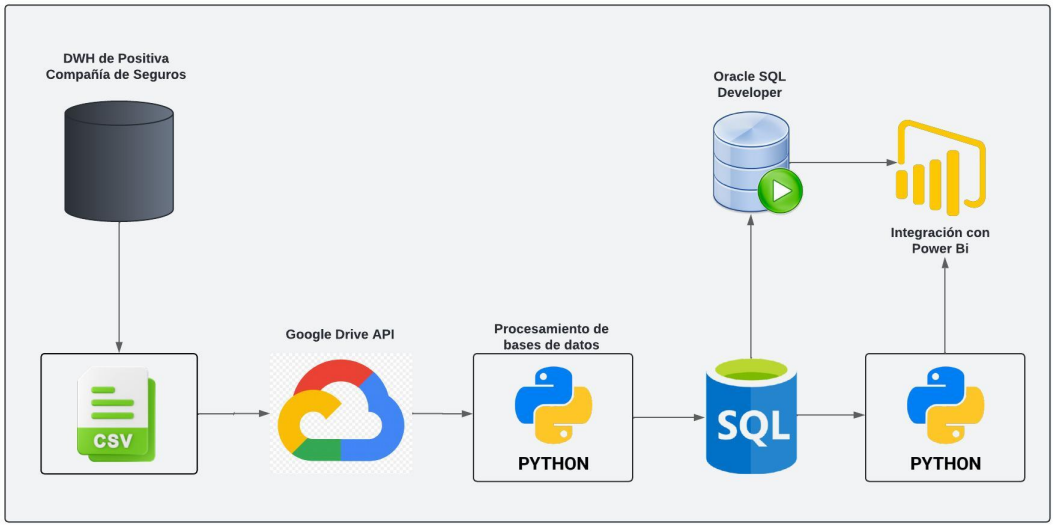

Figure 1: Diseño de integración y automatización de datos para IA.
(Elaboración propia)

4.1 Fuente de datos

- Datos entregados por Positiva Compañía de Seguros como fuente primaria y custodiados en su DWH (Bodega de datos).
- Actualización de datos almacenados a través de archivos .csv o .xlsx.

4.2 Procesamiento de datos

- Con el fin de asegurar la actualización periodica de los datos de forma automática, se usará Google API y un cliente de está última que permite su integración con Python.
- Para garantizar la calidad de los datos y la imputación de registros nulos en la base de datos, se usará Python, principalmente sus librerías Pandas,

Numpy, Matplotlib, Scikit-learn, cx Oracle, entre otras.

- Posterior a la estandarización y preparación de las bases de datos, se usará Oracle SQL developer como herramienta de almacenamiento y manipulación de los mismos.

4.3 Difusión de resultados

- los resultados entregados por los modelos predictivos serán visualizados a través de un dashboard de Power Bi. Esto se hará a través de su función "Scripts de Python".

- Los indicadores descriptivos también se visualizarán en Power Bi utilizando el conector nativo de bases de datos de Oracle.

5 Bibliografía

- Revista Fasecolda, 2024. *Mercado mundial de seguros 2021 - 2023.* Extraído el 10 de marzo de 2024 de https://revista.fasecolda.com/index.php/revfasecolda/article/download/830/787/144
- Swiss Re Institute, 2024. *Sigma 3/2023 - World insurance: Stirred, and not shaken.* Extraído el 10 de marzo de 2024 de https://www.swissre.com/institute/research/sigmaresearch/sigma-2023-03.html
- Fasecolda, 2024. *Más colombianos protegidos, la meta de 2024.* Extraído el 10 de marzo de 2024 de https://www.fasecolda.com/cms/wp-content/uploads/2024/02/ComunicadoCifras-2024-F.pdf
- La República, 2024. *Sura, Bolívar y Alfa, aseguradoras líderes en primas de seguros de personas en 2023.* Extraído el 10 de marzo de https://www.larepublica.co/finanzas/surabolivar-y-alfa-las-aseguradoras-lideres-en-seguros-de-personas-el-ano-pasado-3810082
- Portafolio, 2024. *Un colombiano gasta en promedio $970.000 en seguros al año.* Extraído el 10 de marzo de 2024 de https://www.portafolio.co/mis-finanzas/ahorro/balance-del-sector-asegurador-en-2023-59825
- Senado de Colombia, 2012. *Ley estatutaria 1581 de 2012.* Extraido el 01 de abril de 2024 de http://www.secretariasenado.gov.co/senado/basedoc/ley_1581_2012.html
- Gobierno de Colombia, 2013. *Decreto 1377 de 2013.* Extraido el 01 de abril de 2024 de https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=53646.