ESCUELA DE NEGOCIOS

MAESTRÍA EN INTELIGENCIA DE NEGOCIOS Y CIENCIA DE DATOS

APLICACIÓN DE MODELO DE DETECCIÓN DE ANOMALIAS ISOLATION FOREST PARA LA DETECCIÓN DE RECLAMOS FRAUDULENTOS EN UNA COMPAÑÍA DEL SECTOR DE MEDICINA PREPAGADA

Profesor

Mario Salvador González

Autor

Mateo Bernabé Rodríguez Salguero

2023

RESUMEN

El sector de medicina prepagada es uno de los que más riesgo de fraude presenta debido al giro mismo del negocio y a los riesgos presentes en la actualidad. Para las áreas de auditoría es complejo monitorear los reclamos que presentan los clientes, debido a esto, el uso de herramientas tecnológicas acompañadas de un correcto análisis de datos es imprescindible para tener un monitoreo sobre posibles casos de fraude. El uso de técnicas de Machine Learning (ML) para la detección de casos anómalos, que podrían considerarse fraude, es un enfoque que ha tenido crecimiento en diferentes sectores. Estás técnicas permiten la detección de patrones y pueden ser implementadas sin tener necesariamente etiquetas de casos fraudulentos. Al no tener etiquetas el enfoque de estudio puede ser tratado como un problema de aprendizaje no supervisado. Existen algoritmos especializados en detección de anomalías, como es Isolation Forest, que trabaja aislando los registros anómalos en función de variables relevantes. Para este caso de estudio se analizó los reclamos presentados por clientes durante los años 2021 al 2023 a una compañía de medicina prepagada. El objetivo de estudio es identificar posibles casos de fraude aplicando el algoritmo de Isolation Forest que analizará variables cuantitativas y cualitativas. Para la variable cualitativa del diagnóstico médico se empleó el algoritmo word2vec que permite vectorizar esta variable y aportar más información al modelo. Al implementar este tipo de algoritmos de ML las áreas de las empresas de seguros dedicadas al control mejorarán su gestión en prevención de fraudes en reclamaciones presentadas.

Keywords: Medicina prepagada, machine learning, anomalía, fraude, isolation forest, word2vec

ABSTRACT

The medical insurance sector is one of the areas that presents a higher risk of fraud due to the nature of the business itself and the current risks. For audit departments, monitoring customer claims is complex. Therefore, the use of technological tools along with proper data analysis is essential for monitoring potential fraud cases. The utilization of Machine Learning (ML) techniques for detecting anomalous cases, which could be considered fraud, has gained traction across various sectors. These techniques enable the identification of patterns and can be implemented without necessarily having labels for fraudulent cases. Without labels, the study approach can be treated as an unsupervised learning problem. There are specialized algorithms for anomaly detection, such as the Isolation Forest, which works by isolating anomalous records based on relevant variables. In this case study, customer claims submitted to a medical insurance company during the years 2021–2023 were analyzed. The study's objective is to identify potential fraud cases by applying the Isolation Forest algorithm, which will analyze both quantitative and qualitative variables. For the qualitative variable of medical diagnosis, the word2vec algorithm was employed to vectorize this variable and provide more information to the model. By implementing such ML algorithms, the insurance company departments focused on control will enhance their fraud prevention management for submitted claims. This will contribute to an overall improvement in fraud prevention within the insurance sector.

Keywords: Medical insurance, machine learning, anomaly, fraud, isolation forest, word2vec

Índice

Índice de tablas

Índice de figuras

1. Introducción

Según la NHCAA (National Health Care Anti-Fraud Association), las pérdidas por fraudes cometidos en Estados Unidos relacionados a temas sanitarios o de salud se estimaron en un 3% del gasto total del sector en el año analizado (2019). En el caso del Ecuador no existe alguna investigación o datos que nos indiquen cual fue la pérdida estimada por casos de fraude, pero se conoce a nivel mundial que, debido a la pandemia, se incrementó los riesgos de fraude en el sector de seguros médicos al igual que en otros sectores.

Dentro del mercado asegurador existen diferentes esquemas de fraude, donde los principales actores son los pacientes o asegurados, prestadores médicos y farmacias. Entre los esquemas de fraude más populares están las irregularidades en la facturación y prestación médica, donde, según Ron Cresswell (2018) en un estudio realizado para la ACFE, muestra que los más comunes son: upcoding, desagregación, doble facturación, atenciones ficticias y facturación por servicios sin cobertura. En la tabla a continuación se detallan los casos.

Tabla 1. Tipos de fraude en el sector de medicina prepagada

Tipo de fraude	Esquema
Upcoding	Presentar un código de alguna prestación médica el cual tenga una mayor cobertura del seguro médico.
Desagregación	Facturar por separado procedimientos que se realizaron como una única intervención.
Doble facturación	El prestador médico factura varias veces por un servicio prestado. Por lo general se alteran fechas de facturación.
Atenciones ficticias	Es un fraude poco común y consiste en facturar por servicio médicos que no fueron brindados e incluso puede aparecer la figura de pacientes ficticios.

Servicios sin cobertura	Cubrir procedimientos sin cobertura (por ejemplo, procedimientos estéticos: cirugías bariátricas, balones gástricos, operaciones estéticas de nariz, entro otros) debido a una incorrecta categorización del diagnóstico o procedimiento.

El enfoque que se tiene dentro de las empresas de medicina prepagada para mitigar el riesgo de fraude está centrado en tres ejes que son: personal capacitado, políticas internas y actualmente se acompaña con el uso de tecnologías de la información. Para las empresas de medicina prepagada analizar todos los reclamos entrantes por parte de los clientes es insostenible y poco eficiente, por lo que el uso de tecnologías es necesario con el fin de mejorar la experiencia del cliente y mantener una cultura de control basada en riesgos.

2. Revisión de literatura

Por varios años los estudios actuariales han servido para los análisis estadísticos dentro del sector asegurador para medir riesgo, primas entre otros estudios (Rawat et. al, 2021). En las últimas décadas y gracias a los avances tecnológicos se ha popularizado el uso de inteligencia artificial dentro de los análisis de riesgo en empresas aseguradoras. Uno de los principales objetivos al emplear estas nuevas tecnologías se enfoca en la prevención y detección de fraude. Al igual que en el sector financiero, los procesos transaccionales son el núcleo de los servicios de medicina prepagada. Estos procesos transaccionales involcuran el procesamiento de gran cantidad de reclamaciones por parte de los asegurados para la cobertura de sus atenciones médicas. Para las áreas de control o auditoría representa un reto la detección fraude dentro del proceso reembolsos. El uso de herramientas tecnológicas busca identificar anomalías dentro de este proceso, el cual se basa en su mayoría en comparar datos actuales contra los reclamos históricos (Mehbodniya et. al, 2021). Al comparar los los datos historicos contra los actuales el objetivo del área de auditoría es

encontrar irregularidades en reclamaciones que permitan considerarlas cómo anómalas. La correcta detección de anomalías o *outliers* es de gran utilidad para identificar posibles patrones de fraude y esquemas utilizados por los actores involucrados.

Al utilizar técnicas de Machine Learning por lo general se emplean dos métodos para la detección de fraude y anomalías. El primero es el aprendizaje no supervisado, donde para el caso de estudio no existan etiquetas de casos fraudulentos. Por otro lado, el aprendizaje supervisado es empleado en casos donde si se cuentan con etiquetas de casos de fraude y una ventaja de contar con etiquetas es que es posible evaluar que tan preciso es el modelo detectando estos casos. Entre algunos de los retos que existen para la detección de fraude están el desbalance de los datos (transacciones verdaderas y transacciones fraudulentas), gran cantidad de datos que demanden procesamiento excesivo y el comportamiento dinámico de las personas al cometer fraude (Zareapoor y Shamsolmoali, 2015).

En las investigaciones de los úlitmos años para detección de anomalías se ha venido implementando modelos más complejos, como son redes neuronales y la combinación de varios algoritmos de machine learning. Sin embargo, las técnicas habituales de detección de outliers son útiles al momento de identificar anomalías, como son: medición por cuartiles, Local Outlier Factor (LOF) y Isolation Forest (Carletti et. al, 2023). Para los algoritmos de ML de aprendizaje supervisado se busca resolver un problema de clasificación, el cual se enfocará en identificar clases binarias (fraude o no fraude), basándose en su mayoría en funciones de riesgo por distancias entre las características (Kose et. al, 2015). Según Rukhsar (2022), entre los algoritmos de clasificación más utilizados para detección de anomalías y fraudes se pueden destacar:

- Support Vector Machine (SVM): Usado para problemas de clasificación no linear.
- Naive Bayes: Probabilidad de las diferentes instancias.
- Adaboost: enfocado en mejorar la capacidad predictiva.
- Random forest: Empleados en grandes conjuntos de datos (usualmente en fraude trabajamos con gran cantidad de datos).

Para el caso del aprendizaje no supervisado existen algunas complicaciones debido a que no se puede comparar contra etiquetas de fraude. En el sector asegurador de salud muchas veces es complicado tener etiquetas de reclamos fraudulentos y en caso de que las transacciones estuvieran etiquetadas, dificilmente estos casos serían representativos en comparación a la gran cantidad de datos que se procesa. Es por esta razón que muchos de los estudios para la detección de anomalías utilizan las técnicas tradicionales mencionadas (LOF, isolation forest, etc), y algoritmos de aprendizaje no supervisado. Entre los métodos tradicionales más usados está el Isolation Forest que permite identificar anomalías de una manera eficiente y también se adapta a conjunto de datos de gran tamaño (Domingues, et. al, 2018). Por otro lado, en técnicas no supervisadas, el estudio propuesto por Settipalli y Gangadharan (2023), realiza un análisis de los patrones en las atenciones brindadas por los prestadores, donde los algoritmos no supervisados permiten identificar relaciones entre el médico y el tipo de prestación, junto con evaluar la información compartida que pueda existir entre prestadores. Junto con los estudios mencionados anteriormente, en la detección de fraude ha sido de gran utilidad el llamado aprendizaje semi supervisado, donde se combinan técnicas de aprendizaje supervisado y no supervisado. Al utilizar este nuevo tipo de aprendizaje se pueden utilizar pocas etiquetas y ayudar a los modelos no supervisados a detectar casos atípicos con mayor precisión. Esta combinación de técnicas ha permitido mejorar los resultados de los modelos de detección, donde un claro ejemplo sería la evaluación de transacciones de tarjetas de crédito,

en el cual se mejoró la identificación de casos atípicos (posibles fraudes) y la precisión (Mehbodniya et. al, 2021).

3. Identificación objeto de estudio

Durante el 2019 en Estados Unidos se detectó que los casos por pagos indebidos en el sector de medicina prepagada ascendieron a 46.2 billones de dólares (U.S. Government Accountability Office, 2020). El sector de seguros médicos es un blanco para las personas que buscan beneficiarse de su cobertura empleando técnicas fraudulentas. Para las empresas de seguros es un reto poder identificar si los reclamos presentados por los clientes son fraudulentos o presentan irregularidades, es por eso que el uso de tecnologías es indispensable. Técnicas de inteligencia artificial, machine learning y algoritmos para la detección de anomalías, han demostrado ser eficientes para la detección de anomalías y casos fraudulentos. Esto significa reducir tiempos y costos en la revisión de casos por parte de los auditores y focalizar las revisiones en casos verdaderamente atípicos.

Utilizar algoritmos para la detección de outliers aun cuando no se tenga etiquetas en los datos, permitirá identificar posibles patrones o irregularidades en los reclamos presentados por clientes, que deberán ser corroboradas con investigaciones puntuales de los casos.

4. Planteamiento del problema

Existen varios factores de riesgo para las empresas de medicina prepagada en cuanto al proceso de reembolso o cobertura médica. Con la modalidad de telemedicina y el uso constante de la tecnología para hacer uso de la cobertura médica, es necesario que las áreas de control

cuenten con herramientas de control, detección de anomalías y puedan llevar un monitoreo continuo.

La falta de etiqueta en los datos ya sea porque no están documentados, no han exisitido casos relevantes o los casos no han sido identificados como fraudes, llevan a manejar este problema con agloritmos de aprendizaje no supervisado para la detección de outliers. Esto permitirá identificar patrones irregulares en el conjunto de datos, junto con casos irregulares. El fin de utilizar estos algoritmos permitirá generar conocimiento acerca de los datos históricos de la organización.

5. Objetivos

5.1. Obejtivo general

El objetivo de este estudio es identificar casos atípicos en las atenciones médicas cubiertas por una empresa de medicina prepagada, utilizando algoritmos de machine learning para la detección de anomalías que permitan identificar posibles casos de fraude.

5.2. Objetivos específicos

Cómo objetivos dentro del estudio estarán:

- Identificar posibles grupos de acuerdo con sus características como puede ser el diagnóstico médico, lugar de atención y grupo de diagnóstico.
- Utilizar variables cualitativas dentro del modelo de detección de anomalías utilizando el método que mayor valor genere al modelo y aporte información.
- Emplear el algoritmo word2vec para la codificación de la variable de diagnóstico con el fin de generar vectores de las palabras y relacionar diagnósticos para aportar información al modelo.

6. Justificación y aplicación de la metodología

Una vez conocidos los objetivos que busca este estudio se trabajará con una base de datos privada de una empresa dedicada a brindar servicios de medicina prepagada. Debido al tipo de datos que se trata no se expondrán nombres ni ningún tipo de dato que permita la identificación de personas ni de prestadores médicos.

6.1. Recopilación de los datos

Para la recopilación de datos se empleó la herramienta tecnológica Qlik Sense la cual permite la conexión a las bases de datos de la organización y la extracción de estos. La ventaja de esta herramienta es que por medio de ciertas sentencias permite un preprocesamiento de los datos bajos los criterios seleccionados y la depuración de algunos campos. Para el análisis realizado se decidió tomar los reclamos pagados a los asegurados en el periodo de tiempo de enero del 2021 hasta julio del 2023. El concepto de estos reclamos llamados pagos de contado consiste en que el asegurado recibió una atención médica por su cuenta y pago por los servicios médicos. Posterior a esa atención el cliente tiene que presentar los soportes para que la aseguradora pueda cubrir esta prestación bajo los términos contractuales acordados. En la siguiente tabla se detallan los criterios adicionales de la base de datos seleccionada para la preparación de datos y el modelado:

Tabla 2. Criterio de filtrado por variables

Criterio de variables	Consideración
Tipos de contratos	Individuales, experience, corporativos y pooles (Se excluye contratos de cobertura oncológica).

Estado de los reclamos	Únicamente reclamos que hayan generado costo a la organización.
Diagnósticos	Después de una revisión preliminar de los datos se decidió excluir los diagnósticos relacionados a COVID y donde el diagnóstico principal sea algún tipo de cáncer. Esta decisión se consideró debido al costo elevado que se generaron por estas atenciones.
Tipo de procedimientos realizados	Se excluyen del análisis procedimientos en los cuáles no existe cobertura por parte de la aseguradora por términos contractuales.

Después de las consideraciones mencionadas en la tabla anterior se detallan las variables a considerar para el modelo. Debido a que el problema que se maneja en este estudio tiene un enfoque no supervisado se realizó un análisis descriptivo de las variables numéricas, así como un análisis explicativo de las variables determinantes en la detección de fraude según el departamento de auditoría.

Tabla 3. Variables escogidas para el análisis y modelamiento

Nombre variable	**Descripción**	**Tipo de variable**	**Tipo de dato (Python)**
numero-reclamo	Identificador único del reclamo presentado por el cliente	Cualitativa	object
persona-numero	Identificador único del cliente	Cualitativa	int64
codigo-diagnostico	Código CIE-10 del diagnóstico por el cual se dio la prestación médica	Cualitativa	object
nombre-prestador	Nombre del prestador médico el cuál realizó la atención médica	Cualitativa	object
tipo-prestador	Tipo de prestador: puede ser médico, clínica, farmacia, laboratorio, especial.	Cualitativa	object
codigo-producto	Se identifica si el producto es individual, experience, corporativo o pool	Cualitativa	object
region	La región de donde es el cliente: Sierra, Costa o Austro	Cualitativa	object

lugar-atencion	Categoría del lugar donde ocurrió la atención médica: Consulta médica, Hospital, Hospital del Día, Emergencia, Domicilio, Punto Médico	Cualitativa	object
periodo-incurrencia	Tiempo que transcurrió desde la inclusión del beneficiario al plan médico hasta que se dio la atención	Cuantitativa	int64
periodo-presentacion	Tiempo que transcurrió desde la atención médica hasta la presentación del reclamo a la aseguradora	Cuantitativa	int64
edad-beneficiario	Edad que tenía el beneficiario al momento de la atención médica	Cuantitativa	int64
nivel-prestador-desde	Niveles mínimos de planes de medicina prepagada que tienen cobertura en ese prestador	Cualitativa	float64
nivel-prestador-hasta	Niveles máximos de planes de medicina prepagada que tienen cobertura en ese prestador	Cualitativa	float64
grupo-diagnostico	Grupos según diagnósticos médicos	Cualitativa	int64
nombre-cabecera	Nombre en español del diagnóstico principal por el que se brindó la atención	Cualitativa	object
cantidad-presentada	Cantidad de servicios médicos que se brindó	Cuantitativa	int64
valor-presentado	Valor que el cliente presenta a la aseguradora por la atención recibida	Cuantitativa	float64
valor-bonificado	Valor que la aseguradora bonifica al cliente según sus condiciones contractuales	Cuantitativa	float64
nombre-beneficio	Es el nombre en español del beneficio cubierto por la aseguradora: Medicinas, honorarios médicos, suministros, etc.	Cualitativa	object
monto-cobertura	Es el monto de cobertura máximo del plan de medicina prepagada contratado por el cliente	Cuantitativa	int64
fechas	fecha-incurrencia: fecha en la cual se dio la atención médica fecha-presentación: fecha donde el cliente presentó el reclamo a la aseguradora fecha-liquidación: fecha en la cual la aseguradora bonifico al cliente el valor cubierto	N/A	object

Estas son las variables que se extrajeron de la base de datos, donde en su mayoría son de tipo cualitativas. Para el estudio de detección de fraude que se propone, estás variables tienen relevancia para logar aislar de una mejor manera las anomalías. A continuación, se presentan visualizaciones de interés para conocer el costo generado

por atenciones médicas brindadas en el periodo enero 2021 a julio 2023.

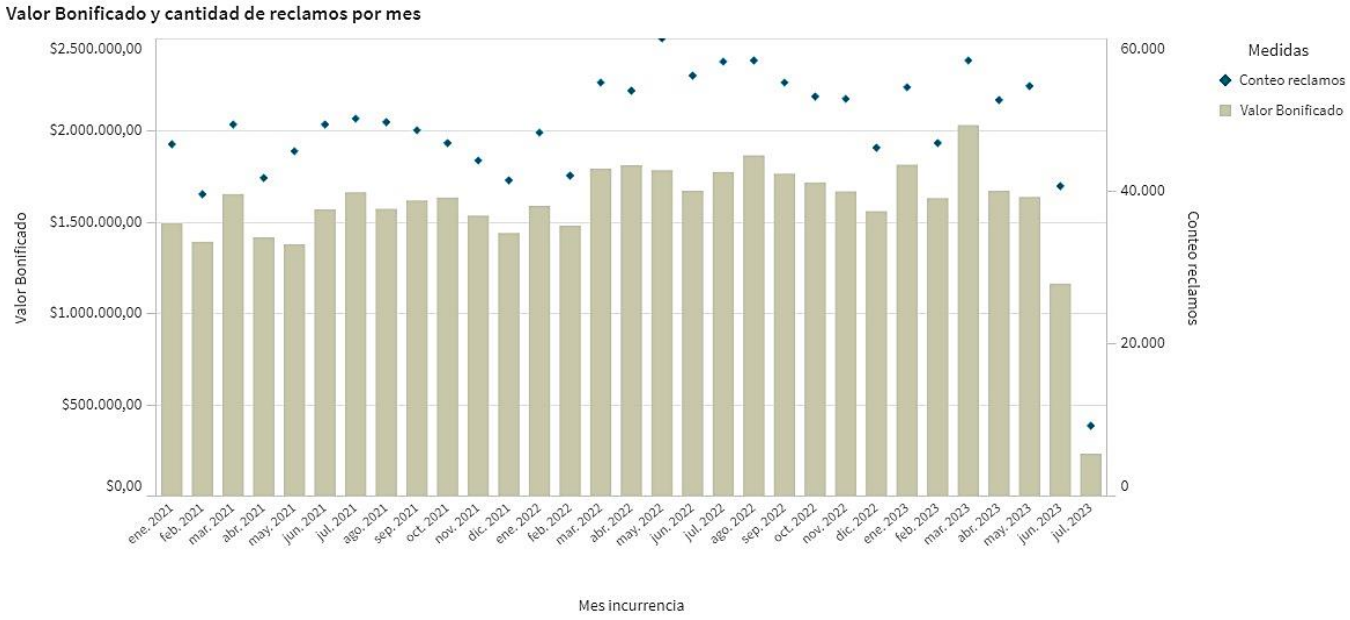

Ilustración 1. Valor bonificado por meses vs cantidad de reclamos presentados por clientes

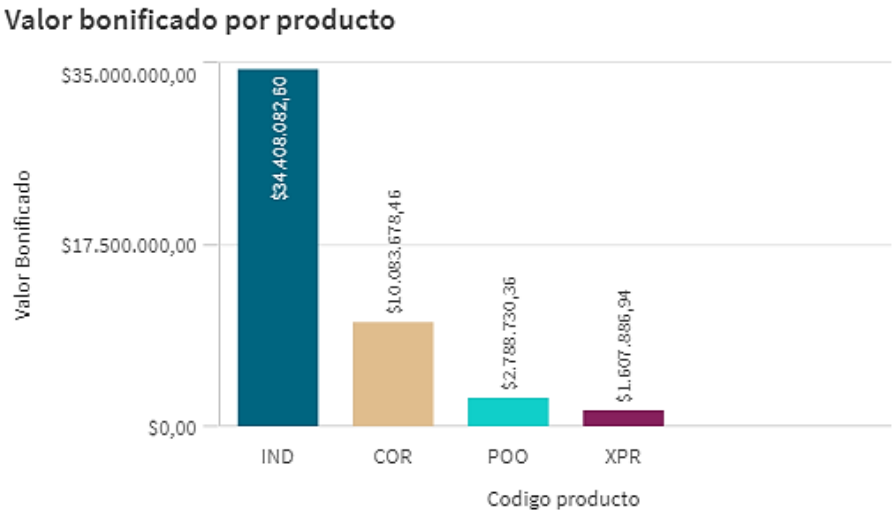

Ilustración 2. Valor bonificado por producto

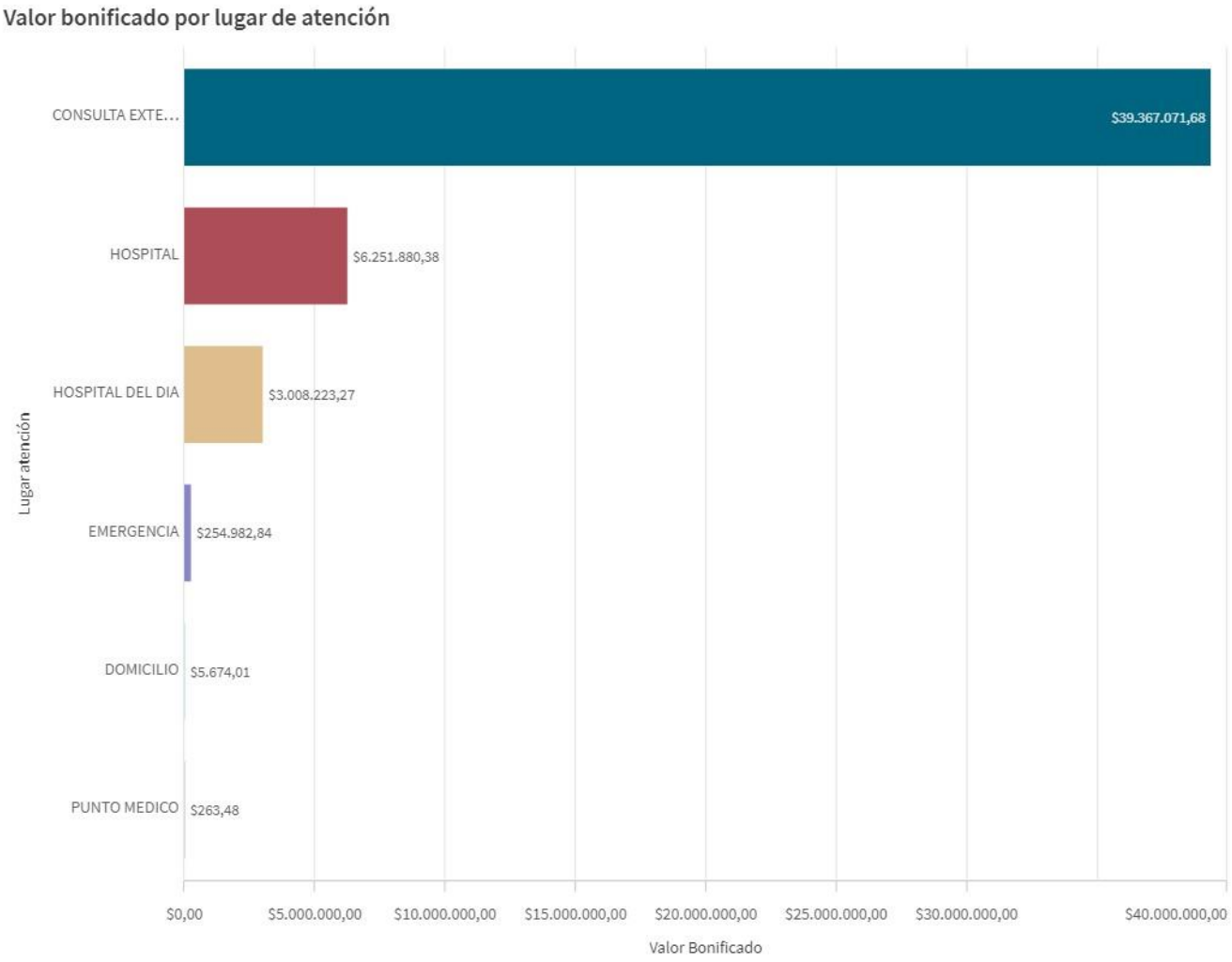

Ilustración 3. Valor bonificado por lugar de atención médica

En los gráficos anteriores se puede observar el costo generado por los reclamos de los clientes presentados a la aseguradora para su devolución. Debido a que el periodo de análisis es de tan solo dos años es difícil interpretar alguna tendencia a lo largo del tiempo, pero en la ilustración número uno se puede evidenciar la cantidad de reclamos incurridos según el mes, junto con el valor que generaron esas atenciones médicas. La cantidad de reclamos que se procesan cada mes oscila entre 40.000 a 60.000 reclamos. Esto nos indica la gran carga de trabajo que se genera para los liquidadores auditar estos casos y pagarlos.

El objetivo de este estudio se centra en identificar casos atípicos utilizando técnicas de machine learning. Antes de emplear estás técnicas se presenta el diagrama de caja, que tiene como objeto analizar la distribución de los datos e identificar *outliers*. Al utilizar únicamente datos cuantitativos (cómo puede ser el valor bonificado o presentado), se pueden utilizar técnicas como LOF o la fórmula $Q3 + 1.5 * (IQR)$ para identificar valores. En el siguiente gráfico utilizando el valor presentado por la atención médica no se puede observar correctamente la distribución de los datos. Debido a la dispersión de los valores se podría dar una conclusión errónea al considerar ciertos valores de reclamos como *outliers*.

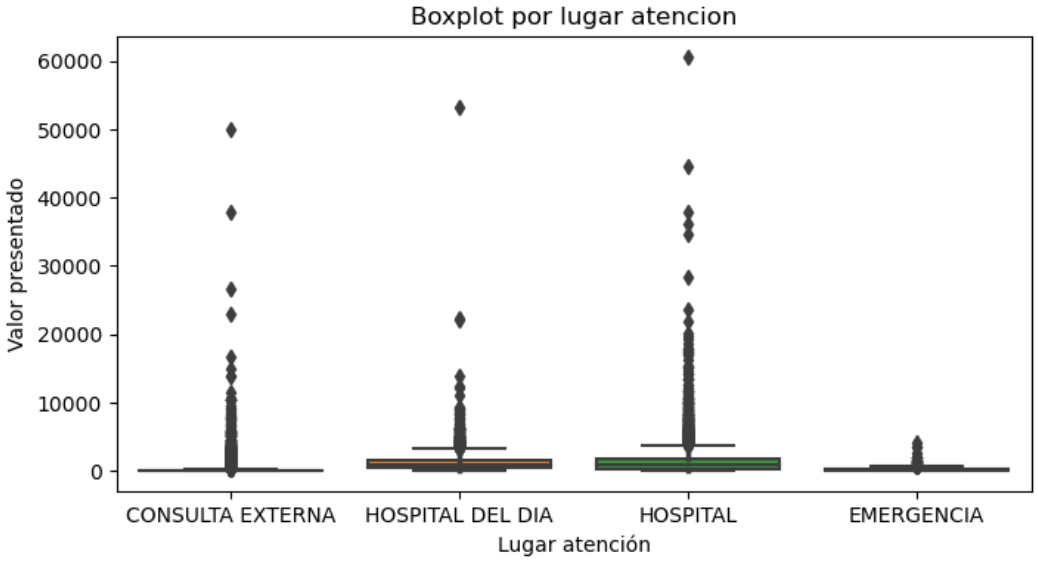

Ilustración 4. Boxplot del valor presentado por lugar de atención

Es por eso la importancia de agregar variables cualitativas al modelo para la detección efectiva de anomalías. En el gráfico a continuación podemos ver que al segmentar por diagnóstico (para este ejemplo apendicitis), se puede analizar mejor la distribución, sobre todo para los casos hospitalarios.

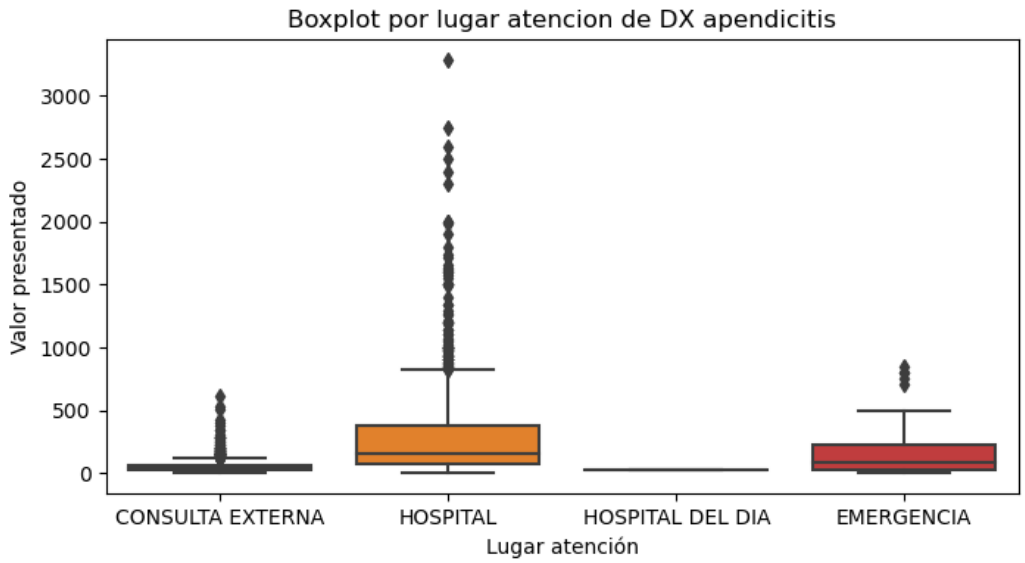

Ilustración 5. Boxplot de valor presentado por lugar de atención de diagnóstico apendicitis

Para este estudio se busca implementar estas variables cualitativas en el modelo de detección de anomalías asegurarse que la transformación que se realice a las variables cualitativas sea significativa para el modelo y genere valor.

6.2. Preparación de los datos

La preparación de los datos se enfoca en la corrección de algunos registros vacíos, eliminación de registros con datos incoherentes o que no aporten al modelo y estandarización de los campos que contengan texto. En los siguientes puntos se detalla el proceso de preparación de los datos y las técnicas utilizadas para poder modelar los datos.

a. Corrección de registros:
    i. Completar el nombre de beneficio que corresponde al 40000 por Laboratorio Imagen.
    ii. Unificar el lugar de atención Punto Médico y Domicilio dentro de Consulta Externa debido a que existen muy pocos registros.

b. Eliminación de registros
    i. Eliminar valores bonificados en 0
    ii. Eliminar periodos de incurrencia y periodos de presentación menores a 0.
    iii. Eliminar registros con fechas incoherentes (Años: 3020-8020-5020-9020)
    iv. Eliminar registros donde no exista un nombre de beneficio.
    v. Eliminar edades negativas y mayores a 100 años.
c. Estandarización de texto
    i. Aplicar la función unicodedata. normalize para los campos nombre-beneficio y nombre-cabecera.
    ii. Eliminar símbolos de los campos de texto como paréntesis, corchetes, números, etc.
    iii. Aplicar la función stop. words para eliminar los artículos de las palabras y únicamente utilizar los diagnósticos médicos.

Una vez realizada la limpieza de los datos y su preprocesamiento se aplicará una técnica de vectorización de palabras para el campo de nombre-cabecera (diagnóstico principal). El modelo empleado para convertir las palabras en vectores será Word2vec de la librería gensim. Este modelo de vectorización de palabras consiste en transformar las palabras en vectores dimensionales utilizando el contexto de Skip-Gram y CBOW (Naili et.al, 2017). Esto quiere decir que al utilizar este modelo es posible analizar el contexto en que una palabra aparece dentro de un texto, lo que permite generar relaciones entre palabras. Según la documentación de la librería Gensim así es como funciona word2vec:

- Skip-gram: Junta las palabras por pares y va deslizando por todo el texto, después se entrena una red neuronal con una

capa oculta para poder asignar una probabilidad a las palabras cercanas de una palabra de entrada. En otras palabras, skip-gram trata de conocer el contexto de la palabra de entrada.

- Continous-bag-of-words (CBOW): Aquí el algoritmo utiliza la media de las palabras dadas para predecir la palabra central. Esta red neuronal hace lo contrario de skip-gram, conociendo el contexto trata de predecir cuál sería la palabra en la mitad del texto analizado.

En el conjunto de datos analizado la variable del diagnóstico médico comprende alrededor de 1450 categorías diferentes. Al ser una variable con gran cardinalidad no sería efectivo utilizar codificaciones típicas como son *one-hot encoder* o *label-encoder*, por lo que emplear técnicas de *word embeddings* (word2vec), ayudaría a enriquecer el modelo junto con optimizar el procesamiento de datos. En el estudio realizado por Churgin y Bansal (2022), el cual está enfocado en predecir posibles atenciones médicas según su historia clínica, al emplear word2vec los resultados mejoran en 4 puntos la métrica de AUC para una regresión logística. Debido a que el problema tratado en este estudio es no supervisado no se contará con métricas claras de evaluación, sin embargo, se analizará el modelo word2vec midiendo las relaciones entre diagnósticos médicos.

6.3. Modelado de datos

6.3.1. Modelo word2vec

Para el entrenamiento del modelo word2vec se extrajo una base adicional de diagnósticos médicos, en la cual se unificó los diferentes diagnósticos de una persona por atención (por ejemplo, tos y rinofaringitis aguda). Al realizar este proceso la intención es poder relacionar los diferentes diagnósticos médicos cuando sean transformados en vectores. Después de vectorizar el campo de nombre-cabecera se obtiene un vector de tamaño 15, este vector es

el promedio de la suma de los vectores de las palabras que componen ese diagnóstico. En el siguiente cuadro se detalla un ejemplo de esto:

Tabla 4. Ejemplo de codificación de variable diagnóstico aplicando word2vec

Diagnóstico	**Vector**
Rinofaringitis	[-1.9758183 -1.0550183 -0.44513655 1.566888 …]
Aguda	[-1.6970391 1.2559682 -0.30808872 0.35042754 …]
Rinofaringitis aguda	[-1.8364286 0.10047495 -0.37661263 0.95865774 …]

Después de vectorizar los diagnósticos se empleó el algoritmo de reducción de dimensionalidad UMAP (Uniform Manifold Approximation and Projection) para analizar las relaciones de los vectores. En el gráfico a continuación se observa los vectores de diagnósticos reducidos a dos dimensiones y coloreados con el grupo diagnóstico, esto con el objetivo de ver las asociaciones entre diferentes enfermedades.

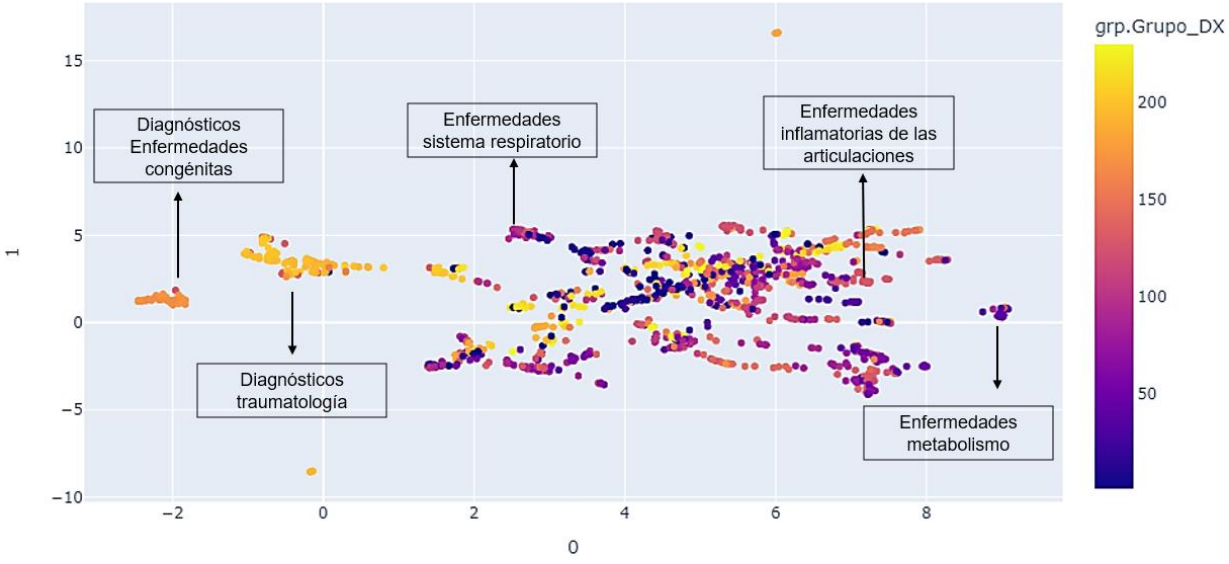

Ilustración 6. Agrupación de diagnósticos similares aplicando técnica de embedding word2vec y UMAP

En el gráfico anterior se evidencia que los diagnósticos se agrupan por enfermedades relacionadas, por ejemplo: enfermedades traumatológicas, congénitas, metabólicas, entre otras. Esto indicaría que el modelo word2vec está cumpliendo con el objetivo de relacionar diagnósticos similares para aportarán información al modelo de detección de anomalías. Debido a que la naturaleza de word2vec es de tipo aprendizaje no supervisado es necesario realizar pruebas de similitud entre palabras, esto es posible empleado la función coseno a dos vectores.

6.3.2. Modelo Isolation Forest

Una vez vectorizado la variable del diagnóstico médico que fue considerada relevante para el estudio planteado, es posible utilizar esta variable dentro del modelo de detección de anomalías. Como se mencionó anteriormente existen varios algoritmos no supervisados y supervisados que son comunes para la detección de fraude (casos anómalos). En este caso tendremos un enfoque de aprendizaje no supervisado debido a que no existen los registros necesarios para trabajar con etiquetas de fraude. El algoritmo escogido para tratar este problema es Isolation Forest, usado justamente para la identificación y tratamiento de *outliers*. Este algoritmo, debido a su estructura basada en árboles, es un modelo que tiene buen desempeño con conjuntos de gran cantidad de datos, donde su enfoque principal es el aislamiento de las características anómalas, (Vijayakumar et.al, 2020). Este modelo trabaja bajo el concepto que es mucho más fácil aislar los registros que no son comunes dentro del conjunto de datos estudiado por medio de particiones aleatorias. El modelo Isolation Forest de la librería sklearn. ensemble de Python será el que permita la identificación de anomalías del conjunto de datos escogidos. Los hiperparámetros empleados para el modelo son los siguientes:

- N_estimators=1000 (Es la cantidad de árboles del modelo)

- Max_samples= ‘auto’ (Número de muestras utilizadas ajustada automáticamente según el tamaño del conjunto utilizado)
- N_jobs: -1 (Emplear todos los núcleos del CPU disponibles)
- Contamination: 0.01 (Se establece que al menos el 1% de los registros sean considerados anómalos)
- Random_state: 123 (Establecer una semilla para replicación de resultados)
- Bootstrap: True (Permite que una muestra aparezca varias veces en la construcción de árboles para mejorar la detección de anomalías)

7. Resultados

Para los modelos de aprendizaje no supervisado no existen métricas que permitan evaluar al modelo, a diferencia del aprendizaje supervisado que tiene métricas como F1 score, precisión y especificidad cuando se tratan problemas de clasificación. Para este caso se decidió evaluar el modelo comparando dos instancias:

- Primera: Entrenar y correr el modelo sin utilizar la variable del diagnóstico vectorizado.
- Segunda: Entrenar y correr el modelo utilizando la variable diagnóstico-vectorizada después de aplicar word2vec.

La eficiencia y eficacia de los modelos se evaluarán con casos puntuales de fraude que ocurrieron en la compañía. Esto con el fin de comparar si al agregar la variable de diagnóstico vectorizada se puede mejorar el rendimiento del modelo para la detección de anomalías. A continuación, se presenta la distribución de los valores de anomalía, que como bien se indicó anteriormente, serán el 1% del conjunto de datos.

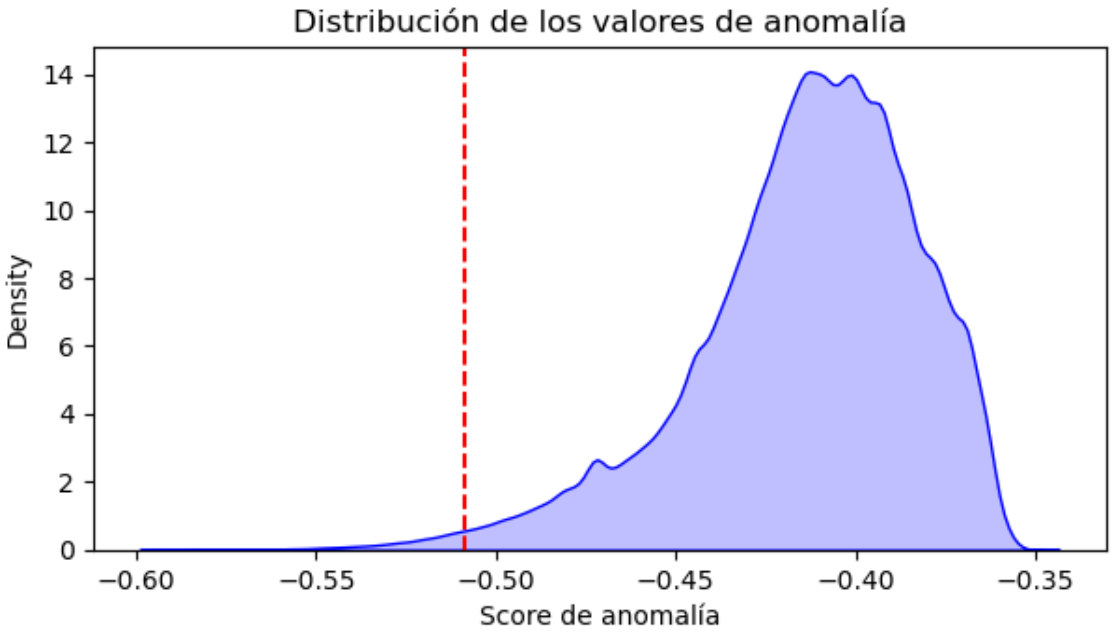

Ilustración 7. Distribución de los valores de anomalía del modelo Isolation Forest

El gráfico representa todos los valores de anomalía menores al primer cuartil, lo que sería el 1% del total del conjunto de datos. Estos valores se obtienen aplicando el método score_samples, lo que nos entrega la puntuación de anomalía normalizada, por lo que los valores de anomalía menor a -0.5089 serán categorizados como anomalías.

7.1. Discusión de los resultados

Para lograr entender cómo se evaluarán los resultados se expondrán los dos sistemas de fraude identificados en la compañía que se buscan detectar con el modelo. El primer caso es por atenciones ficticias, donde se facturó por servicios médicos no prestados. El segundo caso es por una denuncia de posible facturación de servicios médicos los cuáles no tienen cobertura médica (procedimientos estéticos). Ambos resultados resumidos se pueden evidenciar en los anexos del presente trabajo. En el segundo caso analizado no fue posible comprobar el

fraude, sin embargo, se evaluará en función de cuantos casos atípicos de este prestador médico detecta el modelo.

Tabla 5. Comparación de resultados obtenidos de modelo IFO

Tipo de casos de fraude	Modelo IFO sin diagnóstico vectorizado	Modelo IFO diagnóstico vectorizado word2vec
Atención ficticia (Fraude comprobado)	4 casos identificados	8 casos identificados
Posibles procedimientos sin cobertura o con exceso en facturación	46 atenciones médicas con posibles procedimientos sin cobertura o con montos mayores.	66 atenciones médicas con posibles procedimientos sin cobertura o con montos mayores.

Como se muestra en la tabla es posible evidenciar que al agregar la variable del diagnóstico vectorizada se mejoran los resultados de detección de anomalías. Esto posterior a evaluar que el modelo de vectorización esté capturando la información de esta variable de manera significativa. Además, cabe señalar que al agregar esta variable disminuyeron los casos que podrían ser considerados falsos positivos, debido a que existían casos considerados como atípicos que, debido a la naturaleza de la enfermedad y el procedimiento médico, no suelen ser comunes. Entre estos casos mencionados están atenciones en el extranjero, atenciones a clientes experience con mayores beneficios y procedimientos médicos especiales.

8. Conclusiones y recomendaciones

Emplear técnicas de machine learning para la detección de fraude en una empresa de medicina prepagada puede ayudar a mejorar la eficiencia y generar ahorro (costo evitado por fraude). En la introducción del presente trabajo se mencionó que alrededor del 3% del gasto médico en los Estados Unidos se estima que fue por

atenciones fraudulentas. Para el caso de la empresa analizada podemos estimar un conservador 1% por pérdidas de fraude o irregularidades correspondientes a las atenciones generas en cada año. Para el año 2022 los costos generados por atenciones médicas de contado ascienden a 20.3 millones de dólares, lo que podría indicar que aproximadamente USD 200.000 equivaldrían a reclamos con irregularidades o incluso fraudes. Como no existen cifras ciertas en el Ecuador utilizaremos este indicador cómo un posible ahorro al lograr detectar este tipo de reclamos empleando técnicas de ML, principalmente detección de anomalías. A continuación, se presentan conclusiones puntuales acerca del modelo y su desempeño en el caso de estudio.

8.1. Conclusiones

- El modelo Isolation Forest es un algoritmo eficiente para la detección de anomalías que se adapta bien a gran cantidad de datos. Para el sector de medicina prepagada este algoritmo lo que busca es identificar posibles reclamos de clientes que estén categorizados dentro de los esquemas de fraude mencionados al principio de este trabajo. En los resultados obtenidos se puede evidenciar que este modelo realiza correctamente su trabajo de aislamiento para detectar casos que no sean comunes en las prestaciones médicas, lo que podrían conducir a un posible fraude.
- Las variables cualitativas muchas veces son relevantes al momento de realizar un modelo. Utilizar estas variables de una manera adecuada ayudará a mejorar los resultados y las métricas analizadas. En el caso puntual, para la detección de anomalías dentro del sector de seguros médicos, la variable de diagnóstico es relevante y aporta información de la atención médica. En los resultados obtenidos vemos un mejor rendimiento del modelo al agregar la variable de diagnóstico, lo

que permite generar un mayor aprovechamiento de los datos y relacionarlos entre ellos, haciendo uso del análisis de lenguaje natural.

8.2. Recomendaciones

- Para enriquecer al modelo de detección de anomalías se podrían agregar variables adicionales que se consideran relevantes como son: especialidad del médico que realiza la atención junto con la especialidad de la clínica, en caso de que exista. Es con el fin de poder aportar más información al modelo y que exista un mejor aislamiento de registros con irregularidades. Se han presentado casos donde un médico sin la especialidad requerida realiza procedimientos que no le compete a su área por lo que agregar este tipo de variables ayudará a identificar estas inconsistencias.
- El departamento de auditoría debe trabajar en la creación de una base de fraudes para poder implementar un algoritmo supervisado o semi-supervisado y tener una mayor comprensión de estos casos y poder evaluar el desempeño del modelo con métricas.
- Junto con la implementación de variables cualitativas adicionales se puede evaluar el uso de variables conocidas como banderas rojas, para identificar a prestadores o clientes con características riesgosas, como, por ejemplo: contratos morosos, inclusiones en varios contratos, prestadores con denuncias, clientes con quejas frecuentes, entre otras.

Bibliografía

Carletti, M., Terzi, M., & Susto, G. A. (2023). Interpretable Anomaly Detection with DIFFI: Depth-based feature importance of Isolation Forest. Engineering Applications of Artificial Intelligence, 119(105730), 105730. https://doi.org/10.1016/j.engappai.2022.105730

Cresswell R. (2018). *Health care fraud: 5 common billing schemes*. ACFE Insights. https://www.acfeinsights.com/acfe-insights/2018/12/12/health-care-fraud-5-common-billing-schemes

Churgin, M & Bansal, J. (2022, julio 19). Embedding medical journeys with machine learning to improve member health at CVS Health. CVS Health Tech Blog. https://medium.com/cvs-health-tech-blog/embedding-medical-journeys-with-machine-learning-to-improve-member-health-at-cvs-health-957148339cd6

Domingues, R., Filippone, M., Michiardi, P., & Zouaoui, J. (2018). A comparative evaluation of outlier detection algorithms: Experiments and analyses. Pattern Recognition, 74, 406–421. https://doi.org/10.1016/j.patcog.2017.09.037

R. Řehůřek, Gensim: topic modelling for humans. (s/f). Radimrehurek.com. Recuperado el 7 de agosto de 2023, de https://radimrehurek.com/gensim/auto_examples/tutorials/run_word2vec.html

Kose, I., Gokturk, M., & Kilic, K. (2015). An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance. Applied soft computing, 36, 283–299. https://doi.org/10.1016/j.asoc.2015.07.018

Mehbodniya, A., Alam, I., Pande, S., Neware, R., Rane, K. P., Shabaz, M., & Madhavan, M. V. (2021). Financial fraud detection in healthcare using machine learning and deep learning techniques. *Security and Communication Networks*, *2021*, 1–8. https://doi.org/10.1155/2021/9293877

Naili, M., Chaibi, A. H., & Ben Ghezala, H. H. (2017). Comparative study of word embedding methods in topic segmentation. Procedia Computer Science, 112, 340–349. https://doi.org/10.1016/j.procs.2017.08.009

Rawat, S., Rawat, A., Kumar, D., & Sabitha, A. S. (2021). Application of machine learning and data visualization techniques for decision support in the insurance sector. International Journal of Information Management Data Insights, 1(2), 100012. https://doi.org/10.1016/j.jjimei.2021.100012

Rukhsar, L., Haider Bangyal, W., Nisar, K., Nisar, S. (2022). Prediction of insurance fraud detection using machine learning algorithms. Mehran University research journal of engineering and technology, 41(1), 33–40. https://doi.org/10.22581/muet1982.2201.04

The challenge of health care fraud – NHCAA. (2019). Nhcaa.org. Recuperado el 19 de junio de 2023, de https://www.nhcaa.org/tools-insights/about-health-care-fraud/the-challenge-of-health-care-fraud/

Settipalli, L., & Gangadharan, G. R. (2023). WMTDBC: An unsupervised multivariate analysis model for fraud detection in health insurance claims. Expert Systems with Applications, 215(119259), 119259. https://doi.org/10.1016/j.eswa.2022.119259

US Government Accountability Office (2020) Payment integrity federal agencies' estimates of FY 2019 improper payments. Recuperado el 25 de junio de 2023, de https://www.gao.gov/assets/gao-20-344.pdf.

Vijayakumar, V., Nallam Sri Divya, Sarojini, P. & Sonika, K. (2020). Isolation Forest and Local Outlier Factor for credit card fraud detection system. International Journal of Engineering and Advanced Technology, 9(4), 261–265. https://doi.org/10.35940/ijeat.d6815.049420

Zareapoor, M., & Shamsolmoali, P. (2015). Application of credit card fraud detection: Based on bagging ensemble classifier. *Procedia Computer Science*, *48*, 679–685. https://doi.org/10.1016/j.procs.2015.04.201

Anexos

Anexo 1

Resultados de fraudes comprobados identificados con el modelo isolation forest sin utilizar la variable de diagnóstico codificada.

LR02.CruceLR02	PR07.nombre-beneficio	LR04.valor-presentado	LR04.cantidad-presentada	MONTO DE COBERTURA	LR02.lugar-atencion	LR13.Nombre Cabecera	LR02.Edad beneficiario	anomaly	score
27996802-0	HONORARIOS MEDICOS	2520.0	12.0	30000	CONSULTA EXTERNA	VERRUGAS VIRICAS	22	-1	-0.002693
598199111-2	HONORARIOS MEDICOS	1425.0	75.0	45000	CONSULTA EXTERNA	VERRUGAS VIRICAS	3	-1	-0.023178
598411404-0	HONORARIOS MEDICOS	2520.0	150.0	15000	CONSULTA EXTERNA	VERRUGAS VIRICAS	53	-1	-0.024912
598458271-0	HONORARIOS MEDICOS	2520.0	12.0	30000	CONSULTA EXTERNA	VERRUGAS VIRICAS	42	-1	-0.001723

Anexo 2

Resultados de fraudes comprobados identificados con el modelo isolation forest utilizando la variable diagnóstico-vectorizada.

LR02.Cruce LR02	PR07.nombre-beneficio	LR04.valor-presentado	LR04.cantidad-presentada	MONTO DE COBERTURA	LR02.lugar-atencion	LR13.Nombre Cabecera	LR02.Edad beneficiario	diagnostico_codificado	anomaly	score
27996802-0	HONORARIOS MEDICOS	2520.00	12.0	30000	CONSULTA EXTERNA	VERRUGAS VIRICAS	22	[2.0266256332397 46, -0.0970416069030 7617, 3.33...	-1	-0.006236
28109702-0	HONORARIOS MEDICOS	3875.00	5.0	30000	CONSULTA EXTERNA	VERRUGAS VIRICAS	33	[2.0266256332397 46, -0.0970416069030 7617, 3.33...	-1	-0.001068
28113552-0	HONORARIOS MEDICOS	200.00	20.0	1000000	CONSULTA EXTERNA	BRONQUITIS AGUDA	34	[-4.6740311384201 05, 2.7726617455482 483, 0.962...	-1	-0.000030
598199111-2	HONORARIOS MEDICOS	1425.00	75.0	45000	CONSULTA EXTERNA	VERRUGAS VIRICAS	3	[2.0266256332397 46, -0.0970416069030 7617, 3.33...	-1	-0.021780
598411404-0	HONORARIOS MEDICOS	2520.00	150.0	15000	CONSULTA EXTERNA	VERRUGAS VIRICAS	53	[2.0266256332397 46, -0.0970416069030 7617, 3.33...	-1	-0.012812
598458271-0	HONORARIOS MEDICOS	2520.00	12.0	30000	CONSULTA EXTERNA	VERRUGAS VIRICAS	42	[2.0266256332397 46, -0.0970416069030 7617, 3.33...	-1	-0.007112
598610240-2	HONORARIOS MEDICOS	514.52	1.0	15000	HOSPITAL	POLIPO DEL TRACTO GENITAL FEMENINO	43	[-2.7772019505500 793, -3.4414150118827 82, -1.9...	-1	-0.007397
598744102-0	HONORARIOS MEDICOS	500.00	2.0	30000	HOSPITAL	TRASTORNOS NO INFLAMATORIOS DEL OVARIO, DE LA ...	32	[1.1454779846327 645, -5.5624573911939 35, -3.33...	-1	-0.001971