Discusión - Lectura del artículo

**Módulo 4. Aplicaciones de la analítica en las finanzas**

# 2023-11-01

# Juan Diego Correa Noreña – Andres Quintero Haddad – Santiago Bustos Pianda

Como parte de la actividad de seguimiento, por los equipos de trabajo deberán responder mínimo 3 preguntas para cada una de las siguientes secciones:

# Introducción

1. ¿Cuál fue el problema abordado en el artículo?
2. ¿Por qué es relevante este problema en el sector de seguros?
3. ¿Cuál el vacío de conocimiento de este artículo? ¿Cuál es el propósito del articulo?

# Revisión de la literatura

1. ¿Qué ha permitido la tecnificación de los seguros?
2. ¿Qué explica el margen de mejora en las metodologías?
3. ¿Cuáles procesos se ven directamente impactados?
4. ¿Existe algún impacto económico cuantificado de cara al cliente externo?
5. ¿Existe algún impacto económico cuantificado de cara al cliente interno?
6. ¿Qué porcentaje de datos se estima que son utilizados en esta industrial?
7. ¿Por qué es relevante el análisis predictivo de los siniestros?

# Metodología

1. ¿Qué es Crowdsourcing?
2. ¿Qué técnicas conoce para la generación de datos sintéticos?
3. ¿Qué problema soluciona la generación de datos sintéticos?
4. ¿Es trivial la preparación de datos? ¿Por qué?

# Resultados

1. ¿Qué tipo de variables usan en el caso de estudio 1 (Seguro de Salud)?
2. ¿Cuál es común denominador entre las variables usadas en el caso de estudio 1 (Seguro de Salud)?
3. ¿Qué tipo de población es la que busca un seguro de salud?
4. ¿Existen barreras de acceso a los seguros? ¿Enuncie un ejemplo?
5. ¿Cuántos software cree que se usaron para este trabajo?
6. ¿Cambiaria alguna grafica? ¿Qué propone y como mejoraría la interpretación?
7. ¿Se usaron herramientas “tradiciones” o todo fue se enmarca dentro del “ML”?

# Discusión - Conclusiones

1. ¿Qué bondades se evidencian?
2. ¿Qué trabajo futuro se identifica?

# Artículo: Rawat, S., Rawat, A., Kumar, D., & Sabitha, A. S. (2021). Application of ma- chine learning and data visualization techniques for decision support in the insurance sector. International Journal of Information Management Data Insights, 1(2)

# Respuestas

# Introducción

1. ¿Cuál fue el problema abordado en el artículo?

El problema abordado en el artículo es la implementación y la eficacia de los algoritmos de aprendizaje automático (Machine Learning, ML) en la industria de seguros. Se enfoca específicamente en cómo estos algoritmos pueden mejorar procesos como la verificación de conocimiento del cliente, la evaluación de comportamientos de pólizas, y el asentamiento de reclamaciones, mejorando así el enfoque predictivo y la resolución de reclamaciones en la industria de seguros.

1. ¿Por qué es relevante este problema en el sector de seguros?

La relevancia de este problema en el sector de seguros radica en que, al utilizar aprendizaje automático, las compañías de seguros pueden analizar y procesar reclamaciones más eficientemente, identificar reclamaciones atípicas o fraudulentas, y hacer que el proceso sea menos problemático tanto para la compañía como para el beneficiario. Esto no solo mejora la eficiencia y efectividad en el manejo de reclamaciones y la toma de decisiones, sino que también puede llevar a un servicio al cliente mejorado y a una reducción de costos operativos.

1. ¿Cuál el vacío de conocimiento de este artículo? ¿Cuál es el propósito del articulo?

El vacío de conocimiento que el artículo busca llenar es entender cómo los algoritmos de aprendizaje automático pueden ayudar a las compañías de seguros a deducir patrones en diferentes segmentos y ramas de la industria, como es el análisis de reclamos.

Y el propósito del artículo es presentar cómo se puede utilizar el aprendizaje automático para analizar datos y aplicar técnicas de visualización de datos para mejorar el soporte de decisiones en el sector de seguros. Pretende explorar y evaluar el rendimiento de estos algoritmos utilizando diferentes métricas de rendimiento como la precisión, el recall, y la puntuación f1-score. Además, el análisis y la discusión se orientan hacia la mejora de las fases de decisión en la aceptación y liquidación final de reclamos, la implementación de resultados de trabajo durante la suscripción y la aceptación/denegación de reclamos, y la mejora en la aceptación de reclamos a través de un análisis exhaustivo de los datos y la selección de características relevantes.

# Revisión de la literatura

1. ¿Qué ha permitido la tecnificación de los seguros?

La tecnificación de los seguros, a menudo referida como "InsurTech", ha permitido la incorporación de tecnología en el mundo de los seguros para mejorar y optimizar los procesos de gestión de reclamaciones y la experiencia del cliente. Lo que ha facilitado a las compañías de seguros ofrecer productos y servicios más claros y estructurados, y ha ayudado a los clientes a comprender mejor sus procesos mediante el uso de innovaciones tecnológicas. Estas tecnologías y la innovación han permitido llegar a todos los sectores de la economía, especialmente a los mercados emergentes, ofreciendo a las startups y empresas la oportunidad de desarrollar nuevos modelos de negocio y ofrecer servicios a medida a través de soluciones digitales que mejoran la eficiencia operativa, la recopilación de datos, y la regulación del mercado, manteniendo al mismo tiempo un enfoque en la satisfacción del cliente.

1. ¿Cuáles procesos se ven directamente impactados?

Los procesos que se ven directamente impactados gracias a la implementación de técnicas de machine learning y visualización de datos empiezan por la optimización del proceso de suscripción de pólizas que se ve favorecida por un mejor entendimiento del patrón de reclamaciones de los beneficiarios, permitiendo así una personalización más efectiva y una fijación de precios más precisa. Paralelamente, la mejora la detección y prevención de fraudes, que aumenta la eficiencia y reduce la necesidad de intervención manual en estos casos. Esta tecnología también enriquece el análisis de reclamaciones, facilitando una toma de decisiones más rápida y fundamentada que beneficia tanto a la aseguradora como al cliente, lo que puede traducirse en una mayor satisfacción y lealtad del cliente, así como en una reducción de los costos asociados con el manejo de reclamaciones.

1. ¿Por qué es relevante el análisis predictivo de los siniestros?

El análisis predictivo de siniestros es relevante porque aproximadamente el 80% de los ingresos de las compañías de seguros provienen de las primas renovadas, lo que indica la importancia de mantener y mejorar las relaciones con los clientes actuales. Un análisis predictivo efectivo de los siniestros permite a las aseguradoras comprender mejor los riesgos y posibles reclamaciones, lo cual puede conducir a decisiones de tarificación más precisas, una mejor asignación de recursos y una respuesta más rápida a las reclamaciones. Esto no solo mejora la eficiencia operativa y reduce los costos, sino que también puede resultar en una mejor experiencia del cliente al minimizar el

tiempo de respuesta a las reclamaciones y mejorar la precisión en la predicción y prevención del fraude, lo cual puede a su vez influir en la satisfacción y retención del cliente.

# Metodología

1. ¿Qué es Crowdsourcing?

El crowdsourcing es una técnica de obtención de servicios, ideas, contenido o soluciones solicitando contribuciones de un gran grupo de personas, y especialmente de una comunidad en línea, en lugar de empleados o proveedores tradicionales. Este concepto aprovecha la inteligencia colectiva y la capacidad de trabajo de un amplio grupo de individuos para realizar tareas variadas, desde la resolución de problemas hasta la recopilación de datos.

1. ¿Qué técnicas conoce para la generación de datos sintéticos?

Técnica de Sobremuestreo de Minorías Sintéticas (SMOTE): SMOTE es una técnica popular para el sobremuestreo que crea ejemplos sintéticos de la clase minoritaria utilizando una combinación lineal de los vecinos más cercanos de esos ejemplos. Mejora el equilibrio de clases en un conjunto de datos.

Generación de Adaboost: Adaboost es un algoritmo de potenciación que puede usarse para mejorar la clasificación, especialmente en situaciones de desequilibrio de clases, ajustando repetidamente los pesos de los ejemplos en el conjunto de datos y dando más peso a las instancias mal clasificadas en iteraciones anteriores.

XGBoost: Aunque XGBoost es más un algoritmo de aprendizaje potenciado que una técnica directa de generación de datos sintéticos, puede manejar el desequilibrio de clases al asignar más peso a diferentes clases durante el entrenamiento.

Métodos de Clusterización: Este método implica la agrupación de datos para identificar estructuras o grupos en los datos y puede generar datos más representativos para las clases minoritarias.

Ampliación de datos (Data Augmentation): En el contexto del procesamiento de imágenes y visión por computadora, la ampliación de datos se refiere a la creación de nuevas imágenes a través de técnicas como rotar, cambiar la escala, recortar o alterar el color de las imágenes existentes.

Generación de Datos mediante Modelos Generativos: Los modelos generativos, como las Redes Generativas Adversarias (GANs), se pueden entrenar para generar datos nuevos y sintéticos que son indistinguibles de los reales.

1. ¿Qué problema soluciona la generación de datos sintéticos?

Desequilibrio de Clases: En muchos conjuntos de datos de clasificación, especialmente en contextos como fraude financiero, diagnóstico médico y detección de fallos en sistemas, las clases están altamente desequilibradas. La generación de datos sintéticos puede aumentar la representación de la clase minoritaria para mejorar el rendimiento del modelo.

Privacidad de los Datos: Cuando no es posible o deseable utilizar datos reales debido a restricciones de privacidad, los datos sintéticos pueden proporcionar una alternativa útil que refleja las propiedades estadísticas de los datos reales sin comprometer la información individual.

Mejora de Modelos de Aprendizaje Automático: Los datos sintéticos pueden ayudar a mejorar los modelos de aprendizaje automático proporcionando más variabilidad y cantidad de datos para entrenamiento, lo que es especialmente útil cuando los datos reales son limitados o costosos de obtener.

Pruebas y Validación: La generación de datos sintéticos permite a los ingenieros y científicos de datos probar algoritmos y sistemas en un rango más amplio de escenarios simulados que pueden no estar representados en los datos reales pero que son posibles o probable que ocurran.

# Resultados

1. ¿Qué tipo de variables usan en el caso de estudio 1 (Seguro de Salud)?

Las variables que se usaron para el caso de estudio 1 antes de aplicar los algoritmos de selección de características fueron la edad del beneficiario, el sexo del beneficiario (0 = femenino, 1 = masculino), el Índice de Masa Corporal (IMC), el promedio de pasos por día de cada beneficiario, la cantidad de hijos o dependientes del beneficiario, si el beneficiario es fumador o no ( 0 = no es fumador, 1 = sí es fumador), la región o el lugar de residencia de cada beneficiario (0 = noreste, 1 = noroeste, 2 = sureste, 3 = suroeste), los cargos de cada individuo por facturas medicas y la variable objetivo que si la persona hace un reclamo a la aseguradora o no (0 = no, 1 = sí).

Después de aplicar los algoritmos de selección de características, se puede concluir que las variables continuas son más importantes que las variables categóricas, las mejores características para el dataset fueron: si es del noreste, la edad, el IMC, los cargos, los hijos, si es fumador y los pasos diarios.

1. ¿Cuál es común denominador entre las variables usadas en el caso de estudio 1 (Seguro de Salud)?

Todas las variables están relacionadas con la información y las características de los beneficiarios de seguros de salud y que todas estas pueden afectar de manera significativa la decisión del beneficiario en abrir un reclamo a la aseguradora o no.

1. ¿Qué tipo de población es la que busca un seguro de salud?

Los individuos que deseen asegurarse contra costos médicos asociados a las enfermedades y que puedan acceder a servicios de calidad cuando lo necesiten, las personas que más buscan un seguro de salud son los que tienen un IMC mayor a 29, los fumadores independientemente de su sexo, las personas con cargos mayores a 39999, las personas que caminan entre 3000 y 6000 pasos al día, las personas sin hijos, al igual que las personas que residen en el sureste.

# Discusión - Conclusiones

1. ¿Qué bondades se evidencian?

El estudio demuestra que la introducción de tecnologías como el Machine Learning en la industria del seguro puede ser muy beneficiosa. Estas tecnologías ayudan a comprender mejor a los clientes, sus patrones de reclamación y demografía, así mejorando las políticas y la fijación de las primas de seguros.

El uso de técnicas de selección de características es muy beneficioso antes de clasificar datos con algoritmos de clasificación. Ayuda a elegir el mejor subconjunto de atributos para obtener resultados óptimos, reduciendo el sobreajuste de datos, aumentando la precisión y reduciendo el tiempo de cómputo.

1. ¿Qué trabajo futuro se identifica?

Se menciona la necesidad de abordar el problema del alto desequilibrio de datos en el análisis de seguros, esto podría lograrse mediante técnicas de remuestreo de datos, clustering de la clase abundante o la aplicación de técnicas como SMOTE, XGBoost y AdaBoost para equilibrar los datos y mejorar las predicciones.

Se plantea una pregunta sobre la sostenibilidad del modelo InsurTech para el escalamiento. Se sugiere que la investigación futura podría abordar este problema y explorar cómo las nuevas compañías de InsurTech pueden sobrevivir a largo plazo, especialmente considerando la falta de asociaciones con las grandes empresas de seguros.

Se menciona que la mayoría de los fondos de inicio se destinan a la distribución de pólizas, lo que dificulta la inversión en otras unidades de seguros, como suscripción y atención a reclamaciones. Futuros estudios podrían investigar formas de equilibrar la inversión en diferentes áreas de la industria de seguros.