PROYECTO DE CIENCIA DE DATOS

APROVISIONAMIENTO DE RESERVAS PARA UNA ASEGURADORA DE LA LÍNEA DE COMPENSACIÓN LABORAL

Presentado por:

WILLIAM ANDRÉS SÁNCHEZ SÁNCHEZ

Estudiante Maestría en Actuaría y Finanzas

Presentado a:

FRANCISCO GÓMEZ JARAMILLO

Aplicación del Aprendizaje de Máquinas para Actuaría y Finanzas

UNIVERSIDAD NACIONAL DE COLOMBIA - UNAL

Maestría en Actuaría y Finanzas

Bogotá D.C.

Diciembre, 2023

TABLA DE CONTENIDO

PROYECTO APROVISIONAMIENTO DE RESERVAS
LÍNEA DE SEGUROS: COMPENSACIÓN LABORAL

INTRODUCCIÓN

El presente proyecto supone el negocio de una aseguradora que cubre la línea de "Compensación Laboral", el cual se centra en proporcionar cobertura de seguro para lesiones y enfermedades relacionadas con el trabajo que puedan afectar a los empleados de una empresa. Esta forma de seguro está diseñada para proteger tanto a los trabajadores como a los empleadores en caso de accidentes laborales o enfermedades ocupacionales.

La metodología utilizada para guiar el desarrollo del presente proyecto es la Cross-Industry Standard Process for Data Mining (IBM, 1994)

El proyecto de ciencia de datos se llevará a cabo para desarrollar un modelo que mejore la precisión en la estimación de las reservas y ayude a la aseguradora a tomar decisiones informadas en términos de provisiones financieras. El proyecto se centrará en el desarrollo de un modelo predictivo utilizando técnicas de ciencia de datos y aprendizaje automático para estimar las reservas requeridas para las reclamaciones de compensación de trabajadores.

1. Entendimiento del Negocio

1.1. Descripción del negocio

El presente proyecto supone el negocio de una aseguradora que cubre la línea de "Compensación Laboral", el cual se centra en proporcionar cobertura de seguro para lesiones y enfermedades relacionadas con el trabajo que puedan afectar a los empleados de una empresa. Esta forma de seguro está diseñada para proteger tanto a los trabajadores como a los empleadores en caso de accidentes laborales o enfermedades ocupacionales.

En general, el proceso funciona de la siguiente manera (Actuarial Community, 2021):

- Cobertura del seguro: Las empresas compran pólizas de seguro de compensación de trabajadores a través de una aseguradora. Esta póliza cubre los gastos médicos, la pérdida de salarios y otras compensaciones relacionadas con lesiones o enfermedades que los empleados puedan sufrir mientras están en el trabajo.
- Primas: Las empresas pagan primas regulares a la aseguradora en función de varios factores, como el tipo de industria, la nómina total de la empresa y las tasas de riesgo asociadas con las tareas laborales. Las tasas pueden variar según la ubicación y el historial de seguridad de la empresa.
- Reclamaciones: Si un empleado se lesiona en el trabajo o sufre una enfermedad relacionada con el trabajo, tiene derecho a presentar una reclamación ante la aseguradora de compensación de trabajadores. La aseguradora evaluará la reclamación y proporcionará los beneficios correspondientes, que pueden incluir atención médica, rehabilitación y compensación por salarios perdidos.

Beneficios para los empleados: Los beneficios proporcionados por el seguro de compensación de trabajadores pueden incluir:

- Pago de gastos médicos relacionados con la lesión o enfermedad.
- Compensación por salarios perdidos debido a la incapacidad temporal o permanente.
- Rehabilitación y terapia ocupacional para ayudar a los empleados a recuperarse y regresar al trabajo.
- Beneficios por discapacidad permanente o parcial.
- Beneficios por fallecimiento o incapacidad total y permanente.

Beneficios para los empleadores: El seguro de compensación de trabajadores no solo protege a los empleados, sino que también beneficia a los empleadores al reducir la exposición a demandas legales relacionadas con lesiones laborales. Además, la cobertura puede ayudar a mantener una relación laboral positiva al demostrar el compromiso de la empresa con la seguridad y el bienestar de sus empleados.

1.2. Objetivos de negocio

Proporcionar cobertura de seguro que proteja a los empleados y a los empleadores en caso de lesiones y enfermedades relacionadas con el trabajo. Esto ayuda a garantizar que los trabajadores reciban la atención médica y las compensaciones necesarias, al tiempo que limita la responsabilidad legal y financiera de las empresas.

1.2.1. Objetivos Específicos

Mejorar la precisión de reservas (aprovisionamiento): Desarrollar modelos de análisis de datos y aprendizaje automático para mejorar la precisión en la estimación de las reservas necesarias para cubrir las reclamaciones futuras. Esto puede ayudar a reducir la variabilidad en las estimaciones y garantizar una mejor gestión financiera.

- Reducir los riesgos financieros: Implementar estrategias y procesos que minimicen el riesgo financiero asociado con las reclamaciones de compensación de trabajadores. Esto podría incluir el desarrollo de modelos predictivos de riesgo y la identificación temprana de patrones de reclamaciones inusuales.
- Mejorar la eficiencia operativa: Optimizar los procesos de gestión de reclamaciones y administración de pólizas para garantizar una mayor eficiencia en la operación diaria. Esto podría incluir la automatización de tareas repetitivas y la implementación de sistemas de información más eficaces.
- Cumplimiento Normativo: Asegurarse de cumplir con todas las regulaciones y leyes pertinentes relacionadas con la compensación de trabajadores. Mantenerse al tanto de los cambios normativos y ajustar las políticas y procedimientos en consecuencia.
- Desarrollo de relaciones empresariales: Establecer y fortalecer relaciones sólidas con empresas y empleadores que requieran cobertura de compensación de trabajadores. Brindar asesoramiento y soluciones adaptadas a sus necesidades específicas.

- Investigación y análisis de tendencias: Realizar análisis en profundidad de las tendencias en reclamaciones, costos médicos, tasas de accidentes y otros factores relevantes. Utilizar estos análisis para tomar decisiones informadas sobre precios de primas y políticas.
- Promoción de la seguridad laboral: Ofrecer programas y recursos que fomenten la seguridad laboral en las empresas aseguradas. Esto puede incluir capacitación en seguridad, evaluación de riesgos y recomendaciones para mejorar las condiciones laborales.
- Gestión de proveedores y red de prestadores de servicios: Establecer y mantener relaciones con proveedores médicos y de rehabilitación para garantizar que los trabajadores lesionados reciban la atención adecuada y oportuna.
- Innovación y tecnología: Adoptar tecnologías emergentes, como la automatización de procesos, análisis de big data y herramientas de visualización, para mejorar la toma de decisiones y la eficiencia operativa.

1.2.2. Criterios de éxito

- Precisión en la Estimación de Reservas: La aseguradora logra una mayor precisión en la estimación de las reservas necesarias para cubrir las reclamaciones futuras. Esto se reflejaría en una reducción de las variaciones entre las estimaciones y los montos reales de las reservas necesarias.
- Índice de Siniestralidad: El índice de siniestralidad, que compara el monto total de las reclamaciones pagadas con las primas ganadas, se mantiene en un nivel óptimo. Un índice más bajo indica que la aseguradora está gestionando eficazmente los costos de las reclamaciones en relación con sus ingresos.
- Retención de Clientes: La aseguradora mantiene una alta tasa de retención de clientes, lo que demuestra la satisfacción de las empresas aseguradas con sus servicios y políticas. Una alta retención indica confianza en la aseguradora y en su capacidad para cumplir con las necesidades de los clientes.
- Cumplimiento Normativo: La aseguradora se adhiere y cumple de manera constante con las regulaciones y leyes relacionadas con la compensación de trabajadores. La falta de sanciones regulatorias y disputas legales indica un buen cumplimiento.
- Margen de Beneficio: La aseguradora logra un margen de beneficio saludable al equilibrar las primas cobradas con los costos de reclamaciones y gastos operativos. Un margen estable y sostenible es un indicador de una gestión financiera sólida.

- Innovación y Tecnología: La aseguradora implementa con éxito tecnologías innovadoras para mejorar la gestión de datos, el análisis de riesgos y la toma de decisiones. La adopción exitosa de tecnologías emergentes puede mejorar la eficiencia y la competitividad.
- Participación en el Mercado: La aseguradora logra un aumento en su participación en el mercado de Compensación Laboral. Este indicador puede evaluarse mediante el crecimiento de la cartera de clientes y el aumento de la cuota de mercado.

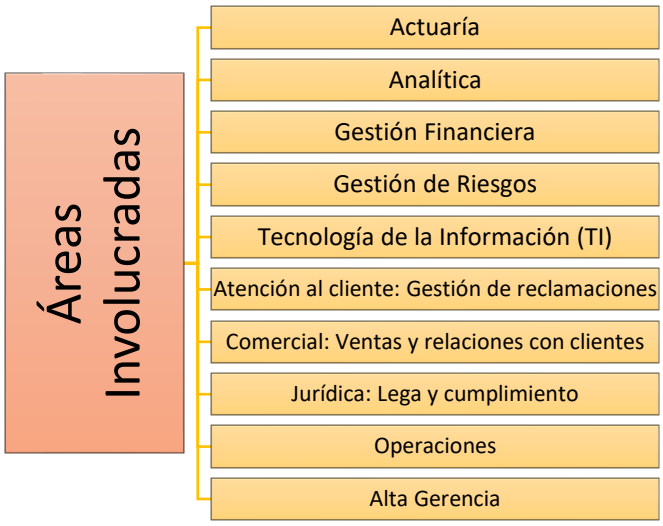
1.3. Evaluación de la situación

El objetivo principal es abordar el problema de aprovisionamiento de reservas que enfrenta la aseguradora en la línea de Compensación Laboral. La aseguradora está experimentando dificultades en la estimación precisa de las reservas necesarias para cubrir las reclamaciones futuras de compensación de trabajadores. Esto puede conducir a una mala gestión financiera y riesgo operativo. El proyecto de ciencia de datos se llevará a cabo para desarrollar un modelo que mejore la precisión en la estimación de las reservas y ayude a la aseguradora a tomar decisiones informadas en términos de provisiones financieras.

Para esto hay que tener presente las siguientes consideraciones:

1.3.1. Inventario de Recursos

- Datos históricos de reclamaciones de compensación de trabajadores.
- Expertos en seguros y especialistas en compensación de trabajadores, con un profundo conocimiento del core de negocio y de los clientes, entre los cuales se deberían incluir profesionales de las siguientes áreas:



- Herramientas y plataformas de análisis de datos y aprendizaje automático. El análisis de los datos y el modelo se construirán en el lenguaje Python en la interfaz de Google Collaboratory.

1.3.2. Restricciones y supuestos

- Se supone que se tendrá disponibilidad y calidad de los datos históricos de reclamaciones.
- Existen limitaciones de tiempo y recursos para el desarrollo del modelo.
- No se abordará la implementación completa del modelo en el entorno operativo en esta fase.

1.4. Objetivos del análisis de datos

El proyecto se centrará en el desarrollo de un modelo predictivo utilizando técnicas de ciencia de datos y aprendizaje automático para estimar las reservas requeridas para las reclamaciones de compensación de trabajadores. El alcance incluirá la limpieza y análisis de los datos históricos de reclamaciones, la identificación de variables relevantes y la construcción del modelo predictivo. La implementación final del modelo en el sistema de la aseguradora no está dentro del alcance inmediato, pero se considerará en etapas futuras.

1.4.1. Objetivos específicos del análisis de datos

- Analizar y comprender en detalle los datos históricos de reclamaciones de compensación de trabajadores.
- Identificar los factores clave que influyen en la magnitud de las reclamaciones y su impacto financiero.
- Desarrollar un modelo predictivo que estime las reservas necesarias en función de los diferentes atributos y características de las reclamaciones.
- Evaluar la precisión del modelo utilizando métricas relevantes y validación cruzada.
- Proporcionar recomendaciones basadas en los resultados del modelo para mejorar la gestión de reservas y la toma de decisiones.

1.4.2. Criterios de éxito

- Reducción de la variabilidad en la estimación de las reservas y mejora en la precisión general.
- Mejora en la gestión financiera de la aseguradora al contar con estimaciones más confiables de las provisiones necesarias.
- Mayor confianza en la toma de decisiones relacionadas con la reserva de fondos para futuras reclamaciones.

- Implementación exitosa del modelo en el entorno operativo de la aseguradora.

1.5. Plan del proyecto

El presente proyecto debe llevarse a cabo en un periodo de tiempo de tres (3) meses, para lo cual se propone el siguiente plan de trabajo:

Fase	**Tiempo**	**Recursos**	**Riesgos**
Entendimiento del Negocio	½ mes	Todas las áreas involucradas	No tener una visión completa del negocio.
Entendimiento de los datos	¼ mes	Todas las áreas involucradas	Problemas de datos y tecnología
Preparación de los datos	¼ mes	Actuaría, analítica	Problemas de datos y tecnología
Modelación	1 ½ mes	Actuaría, analítica, Riesgos y Gestión Financiera	Problemas de tecnología y modelos adecuados
Evaluación del modelo	½ mes	Todas las áreas involucradas	Inhabilidad para presentar los resultados
Desarrollo	Por definir	Todas las áreas involucradas	Inhabilidad para implementar los resultados

2. Entendimiento de los datos

2.1. Recolección de los datos iniciales

Los datos utilizados en este proyecto fueron obtenidos de la página oficial de la CAS (Casualty Actuarial Society) (Meyers, 2011). Fueron elaborados por Glenn G. Meyers, PhD, FCAS, con la intención de crear un conjunto de datos organizado y limpio que contenga información sobre las pérdidas (reclamaciones) a lo largo del tiempo, dispuestas en una estructura conocida como "loss triangle" o "triángulo de pérdidas". Este conjunto de datos se creó con el propósito de ser utilizado en estudios de reservas de reclamaciones.

Un "***loss triangle***" (triángulo de pérdidas) es una estructura de datos utilizada en la ciencia actuarial para representar y analizar los patrones de reclamaciones de seguros a lo largo del tiempo. Cada celda de la matriz representa el número de reclamaciones o el monto de las reclamaciones en un período específico, generalmente en función de la edad de la póliza y la duración de la reclamación.

Los datos de reclamaciones provienen del Anexo P: Análisis de pérdidas y gastos de pérdidas en la base de datos de la National Association of Insurance Commissioners (NAIC). Se tuvo el permiso de la NAIC para poner a disposición pública estos datos.

El Anexo P de la NAIC contiene información sobre reclamaciones de las principales líneas personales y comerciales para todas las aseguradoras de propiedad y accidentes que realizan negocios en EE. UU. Algunas partes tienen secciones que separan las coberturas de ocurrencia de reclamos realizados. Las seis líneas incluidas en esta base de datos son: (1) responsabilidad civil/médica de automóviles de pasajeros privados; (2) responsabilidad médica/de automóviles comerciales/camiones; (3) compensación laboral; (4) negligencia médica – reclamaciones realizadas; (5) otra responsabilidad – ocurrencia; (6) responsabilidad del producto – ocurrencia.

2.1.1. Consideraciones en la preparación de los datos

Los triángulos consisten en pérdidas netas de reaseguro y, muy a menudo, los grupos aseguradores tienen acuerdos mutuos de reaseguro entre las empresas del grupo. En consecuencia, la preparación se centró en los registros de entidades individuales en la preparación de datos, ya sean grupos de aseguradores o aseguradores individuales verdaderos. El proceso de preparación de datos tomó tres pasos (Meyers, 2011):

Paso I: Extraer los datos del triángulo del Anexo P del año 1997. Cada triángulo incluye reclamaciones de 10 años de accidentes (1988-1997) y 10 retrasos en el desarrollo. Estos datos son los datos de entrenamiento que se pueden utilizar para el desarrollo del modelo.

Paso II: Cuadrar los triángulos del Anexo P del año 1997 con los resultados del Anexo P de años posteriores. Específicamente, los datos del año de accidente 1989 se extrajeron del Anexo P del año 1998, los datos del año de accidente 1990 se extrajeron del Anexo P del año 1999, ……, los datos del año de accidente 1997 se extrajeron del Anexo P del año 2006. Los datos de los triángulos inferiores se pueden utilizar con fines de validación del modelo.

Paso III: Muestreo. Realizamos un análisis preliminar para garantizar la calidad del conjunto de datos. Se retuvo a una aseguradora en el conjunto de datos final si se cumplen los siguientes criterios: (1) la aseguradora está disponible tanto en el Anexo P del año 1997 como en años posteriores; (2) las observaciones (10 años de accidente y 10 retrasos en el desarrollo) están completas para el asegurador; (3) los reclamos del Anexo P del año 1997 coinciden con los de años posteriores; (4) Las primas netas emitidas no son cero para todos los años.

El producto final es un conjunto de datos que contiene triángulos de seis líneas de negocio para todas las aseguradoras de accidentes de propiedad de EE. UU.

Para el caso de este proyecto se seleccionó únicamente la línea "***(3) compensación laboral***".

2.2. Descripción de los datos

El conjunto de datos para este proyecto contiene triángulos de la línea de negocio *Compensación Laboral* para todas las aseguradoras de accidentes de propiedad de EE. UU. Los datos del triángulo corresponden a reclamaciones de accidentes del año 1988 – 1997 con un retraso de 10 años en el desarrollo. Se incluyen los triángulos superior e inferior para que se puedan utilizar los datos para desarrollar un modelo y luego probar su desempeño retrospectivamente. Esta segmentación de los datos será tenida en cuenta en la sección de ***Modelación***.

2.2.1. Descripción de las variables

La base de datos está compuesta por un total 13,200 filas y 13 columnas. Cada fila corresponde al registro por cada aseguradora de las pérdidas, reservas y primas ganadas en cada año. Por su parte las columnas contienen la siguiente información respectivamente:

- "*GRCODE*": Código de la compañía NAIC (National Association of Insurance Commissioners), que incluye tanto grupos de aseguradoras como aseguradoras individuales.

- "*GRNAME*": Nombre de la compañía NAIC, que también incluye grupos de aseguradoras y aseguradoras individuales.
- "*AccidentYear*": Año de ocurrencia del accidente. Cubre el período de 1988 a 1997.
- "*DevelopmentYear*": Año de desarrollo de la reclamación. También abarca el período de 1988 a 1997. Es una medida de cuánto tiempo ha pasado desde que una reclamación fue reportada y se utiliza para evaluar cómo evolucionan las reclamaciones a lo largo de varios años después de su ocurrencia inicial.
- "*DevelopmentLag*": Lag de desarrollo, calculado como "$(AccidentYear - 1987) + (DevelopmentYear - 1987) - 1$". Representa el tiempo transcurrido desde el año de ocurrencia del accidente y el año de desarrollo. Esta medida es esencial en el análisis de las reclamaciones de seguros para comprender cómo evolucionan los costos y los patrones de reclamaciones a lo largo del tiempo.
- "IncurLoss_D": Pérdidas incurridas y gastos asignados reportados al final del año. Representa la suma de las pérdidas incurridas (lo que espera la aseguradora pagar para cubrir las reclamaciones que ocurrieron durante el año) y los gastos asignados al final del año en cuestión. Esta cifra es una parte importante en la evaluación de la cantidad total de dinero que la aseguradora espera gastar para cubrir las reclamaciones y los costos asociados en ese año específico.
- "*CumPaidLoss_D*": Pérdidas acumuladas pagadas y gastos asignados al final del año. Esto representa la suma total de los pagos realizados por la aseguradora hasta ese momento para cubrir las reclamaciones y los costos asociados. Incluye pagos realizados hasta el momento presente.

  Entonces, "IncurLoss_D" tiene en cuenta las pérdidas estimadas que se esperan pagar por las reclamaciones ocurridas durante el año, pero que aún no se han pagado en su totalidad. "CumPaidLoss_D" refleja los pagos reales ya realizados por la aseguradora hasta ese momento, que pueden ser por reclamaciones que ocurrieron en años anteriores.
- "BulkLoss_D": Reservas a granel y IBNR (Incurred But Not Reported) en pérdidas netas y gastos de defensa y costos de contención reportados al final del año. Las reservas a granel son fondos establecidos por una aseguradora para cubrir futuras reclamaciones y costos que aún no se han liquidado completamente. Por otro lado, los IBNR son estimaciones de reclamaciones que se cree que han ocurrido, pero aún no han sido reportadas. Finalmente, los gastos de Defensa y Costos de Contención son los costos asociados con la defensa legal y la gestión de las reclamaciones, incluyendo los honorarios legales, los costos de abogados y otros gastos relacionados con la administración y el manejo de reclamaciones.

- "*EarnedPremDIR_D*": Primas devengadas en el año de ocurrencia: directas y asumidas. Son las primas que la aseguradora espera recibir por proporcionar cobertura durante un período específico. Primas recibidas por seguros emitidos directamente por la aseguradora o asumidas (reaseguro).
- "*EarnedPremCeded_D*": Primas devengadas en el año de ocurrencia: cedidas. Cedidas por la aseguradora a otras entidades a través del proceso de reaseguro.
- "*EarnedPremNet_*": Primas devengadas en el año de ocurrencia: netas. Las primas devengadas netas representan los ingresos totales generados por las pólizas de seguros emitidas por la aseguradora en el año en que ocurrieron las reclamaciones, teniendo en cuenta tanto las primas directas como las primas cedidas y asumidas.
- "*Single*": Indica si se trata de una entidad única (1) o una aseguradora de grupo (0). Aseguradoras independientes o parte de un grupo o conglomerado. Esta división es importante a la hora de realizar una evaluación estrategica de las reservas y la gestión de riesgos.
- "*PostedReserve97_D*": Reservas publicadas en el año 1997 tomadas de la Exhibición de Suscripción e Inversión - Parte 2A, incluyendo pérdidas no pagadas netas y gastos no pagados de ajuste de pérdida.

2.3. Exploración de los datos

Del total de aseguradoras que componen la base de datos de la línea de compensación laboral (132), 96 son aseguradoras del tipo individual o independientes y 36 pertenecen a un conglomerado.

2.3.1. Estadísticas Descriptivas

Se calculan estadísticas descriptivas para comprender la distribución de los valores en cada variable. Esto incluye la media, la mediana, la desviación estándar y los percentiles.

Variables de marco temporal y clasificación

La base de datos está compuesta por un total de 100 registros para cada una de las 132 aseguradoras (13,200 registros en total), para un periodo de tiempo de ocurrencia de siniestros desde 1988 hasta 1997 (10 años). Cada año de ocurrencia de siniestros tiene seguimiento por 10 años más, por lo cual se tienen la evolución de las reclamaciones en el seguro de compensación laboral para el periodo de desarrollo comprendido entre 1988 hasta 2006. Por otro lado, es importante destacar que las aseguradoras se encuentran clasificadas en dos categorías: la primera, aseguradoras de tipo independiente, y la segunda, aseguradoras pertenecientes a un conglomerado. Esta segmentación es

un aspecto importante al momento de realizar el modelo de predicción debido a la posible diferencia de tamaño entre las aseguradoras.

Ilustración 1 - Estadísticas descriptivas del conjunto de aseguradoras en la línea de negocio de compensación laboral. Elaboración Propia

	mean	std	min	25%	50%	75%	max
IncurLoss_D	11532.05	35595.56	-59.0	0.0	544.0	6526.50	367404.0
CumPaidLoss_D	8215.74	25714.08	-338.0	0.0	351.5	4565.00	325322.0
BulkLoss_D	1570.13	7259.02	-4621.0	0.0	5.0	259.25	145296.0
EarnedPremDIR_D	18438.47	51830.70	-6518.0	0.0	1419.0	11354.25	421223.0
EarnedPremCeded_D	1812.34	6666.66	-3522.0	0.0	144.5	1141.00	78730.0
EarnedPremNet_D	16626.13	48941.72	-9731.0	0.0	827.0	9180.50	418755.0
PostedReserve97_D	39714.40	130130.68	0.0	411.0	2732.0	19265.75	1090093.0

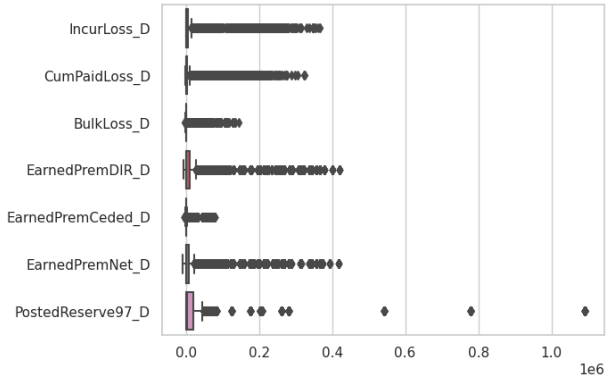
De estas estadísticas es importante resaltar la presencia de valores en cero y valores negativos para las variables correspondientes a pérdidas, primas y reservas. De las 132 aseguradoras que conforman el conjunto de aseguradoras de la línea de Compensación Laboral, un total de 70 aseguradoras (53%) tienen valores menores o iguales a cero en las pérdidas incurridas (IncurLoss_D), 74 aseguradoras (56%) tienen valores menores o iguales a cero en la variable "CumPaidLoss_D". Adicionalmente, 67 aseguradoras (50%) tienen valores en la prima neta ganada ("EarnedPremNet_D") menores o iguales a cero.

Por el lado de las variables que representan pérdidas (IncurLoss_D, CumPaidLoss_D, BulkLoss_D) los valores en cero en pérdidas podrían indicar que no se han reportado pérdidas en ese período o año específico, lo cual podría ocurrir si no ha habido reclamaciones. Por otro lado, las pérdidas no deberían ser negativas, ya que representan costos incurridos. Sin embargo, en algunos casos, podrían ocurrir valores negativos debido a errores en la contabilidad o la entrada de datos. Por tal motivo estos valores deberían ser analizados con más detalle, para evaluar si se deben corregir, eliminar o incluir en el análisis.

Para el caso de las variables representativas de primas (EarnedPremDIR_D, EarnedPremCeded_D y EarnedPremNet_D) también se observan valores mínimos en cero y negativos. Un valor de cero en primas podría indicar que no se han emitido pólizas o que no se ha generado ingreso por primas en ese período específico. Esto podría ocurrir en situaciones donde no se ha vendido ninguna póliza de seguros. Respecto a los valores negativos en las primas, al igual que con las pérdidas, las primas no deberían ser negativas en condiciones normales. Valores negativos podrían ser el resultado de errores en la contabilidad o ingreso de datos incorrectos.

Finalmente, para el caso de la variable de reservas (PostedReserve97_D), se observa un valor mínimo de cero, lo cual podría interpretarse como que no se ha establecido ninguna reserva para futuros pagos de reclamaciones. Esto indica que la compañía de seguros no anticipa la necesidad de reservar fondos para futuras reclamaciones. De las 132 aseguradoras que conforman la data de Compensación Laboral, hay un total de 11 aseguradoras con reservas en cero.

Ilustración 2 - Diagramas de caja para las variables de todo el conjunto de aseguradoras de la línea de seguros de compensación laboral. Elaboración propia.



Respecto a la dispersión en la distribución de los datos de las diferentes variables de las aseguradoras se observa que los datos siguen una distribución asimétrica positiva, lo que significa, por ejemplo, que hay una concentración de reclamaciones más pequeñas y una cola larga de reclamaciones grandes o extremas. Este comportamiento es acorde con la línea de negocio de compensación laboral, donde la mayoría de las reclamaciones son pequeñas, pero un pequeño número de reclamaciones puede ser extremadamente costoso.

Ilustración 3 - Estadísticas descriptivas de las aseguradoras independientes en la línea de negocio de Compensación Laboral. Elaboración propia.

	mean	std	min	25%	50%	75%	max
IncurLoss_D	4329.24	9720.74	-59.0	0.0	242.0	3292.75	87182.0
CumPaidLoss_D	3166.02	7386.27	-70.0	0.0	158.0	2292.00	75655.0
BulkLoss_D	553.36	2363.82	-4621.0	0.0	1.0	141.25	55176.0
EarnedPremDIR_D	7555.88	16328.13	-6518.0	0.0	618.5	5600.25	117225.0
EarnedPremCeded_D	1252.42	4395.23	-3522.0	0.0	70.0	723.00	72731.0
EarnedPremNet_D	6303.46	14245.89	-9731.0	0.0	399.5	4449.25	110693.0
PostedReserve97_D	14400.82	30521.17	0.0	276.0	1945.0	10477.25	203128.0

Ilustración 4 - Diagramas de caja para las aseguradoras independientes. Elaboración Propia

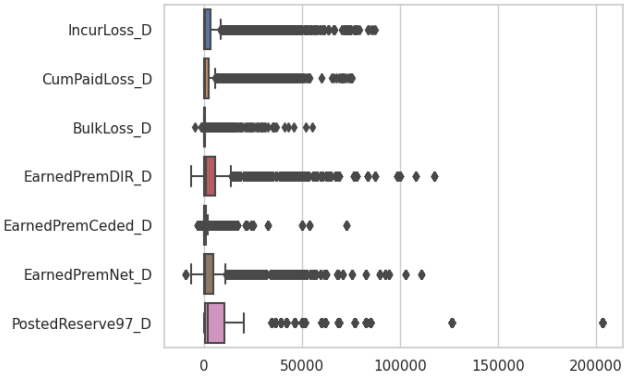


Ilustración 5 - Estadísticas descriptivas de las aseguradoras parte de un conglomerado en la línea de negocio de Compensación Laboral. Elaboración propia.

	mean	std	min	25%	50%	75%	max
IncurLoss_D	30739.54	62348.42	0.0	6.00	5122.5	18987.50	367404.0
CumPaidLoss_D	21681.65	45055.87	-338.0	0.00	3436.0	14626.25	325322.0
BulkLoss_D	4281.49	12970.51	-1467.0	0.00	51.0	1318.50	145296.0
EarnedPremDIR_D	47458.71	89346.60	-67.0	28.50	9522.0	35509.75	421223.0
EarnedPremCeded_D	3305.46	10411.80	-200.0	0.00	644.0	2257.75	78730.0
EarnedPremNet_D	44153.24	84858.99	-399.0	41.75	7633.0	31751.50	418755.0
PostedReserve97_D	107217.28	230980.95	0.0	1956.50	14486.0	51726.75	1090093.0

Respecto a la diferencia entre las aseguradoras de tipo independiente y las aseguradoras parte de un conglomerado se puede observar en los gráficos de dispersión y en las estadísticas descriptivas que hay una notoria diferencia entre el tamaño promedio de reclamaciones y primas ganadas. Sin embargo, se observa una distribución similar, compartiendo la característica de asimetría positiva en los datos.

Ilustración 6 - Diagramas de caja para las aseguradoras de conglomerado. Elaboración propia.

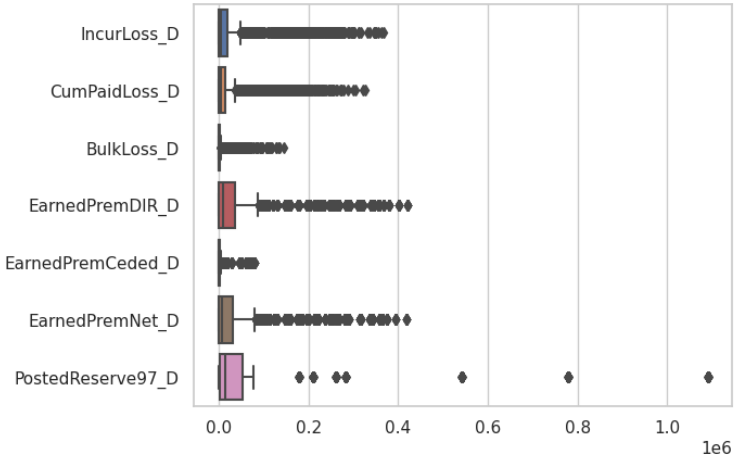
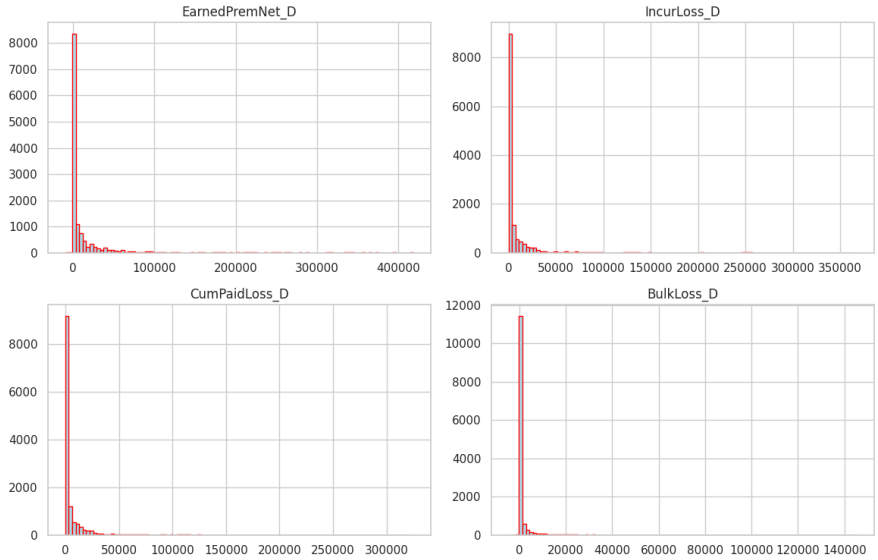


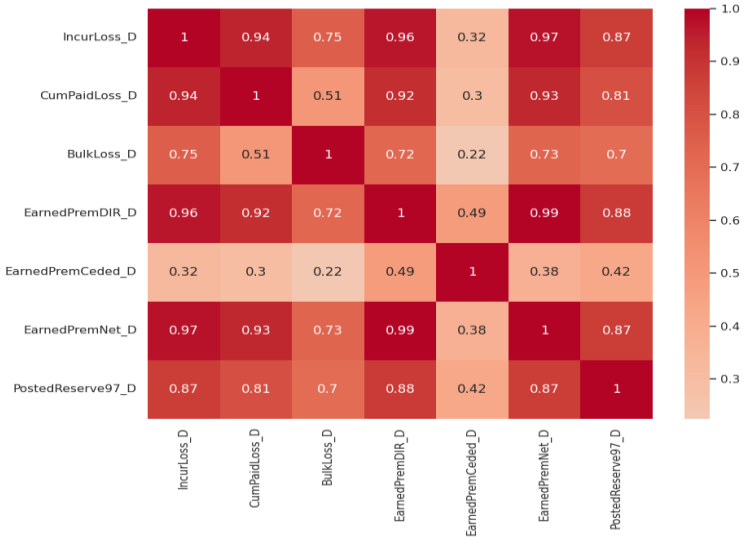
Ilustración 7 - Histogramas de las primas y las pérdidas. Elaboración propia.



De los histogramas en la Ilustración 7 se puede confirmar la asimetría positiva en la distribución de las variables que conforman las primas y pérdidas incurridas de las aseguradoras. Las reclamaciones con menor severidad son las más frecuentes (mayor probabilidad de ocurrencia), pero hay un número significativo de reclamaciones que tienen una severidad extrema o muy superior al valor esperado de los datos observados.

De estos gráficos podemos decir que todas las variables tienen una distribución sesgada a la derecha, por lo que para su modelamiento se pueden suponer distribuciones de probabilidad como la Pareto, Log-normal, Gamma o Weibull.

Ilustración 8 - Matriz de Correlación entre las variables. Elaboración propia.



De estos datos podemos concluir que hay una fuerte relación lineal entre las variables 'IncurLoss_D', 'CumPaidLoss_D', 'EarnedPremDIR_D', 'EarnedPremNet_D' y 'PostedReserve97_D'. Es decir, hay una fuerte dependencia entre las Pérdidas incurridas, las pérdidas acumuladas pagadas, las primas ganadas y las reservas de las aseguradoras.

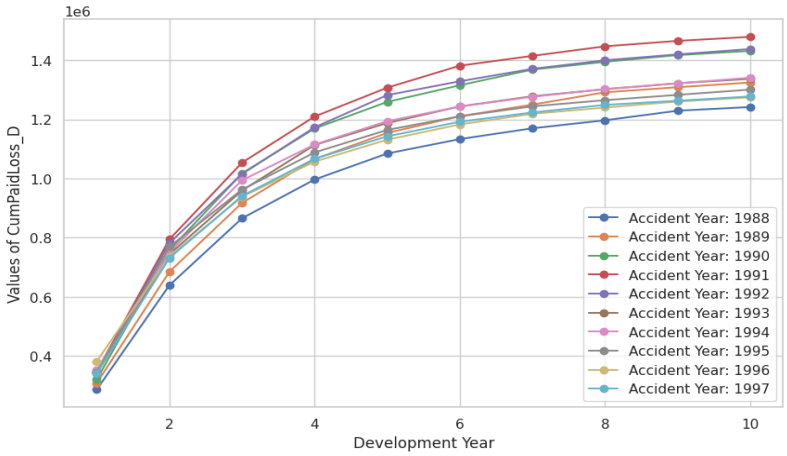
2.4. Verificar la calidad de los datos

Se evalúa la calidad de los datos en términos de integridad, consistencia y posibles problemas como valores faltantes o datos incorrectos. No hay presencia de valores nulos en la base de datos. Por otro lado, como observamos anteriormente, hay una proporción significativa de aseguradoras que tienen en sus variables de pérdidas, primas y reservas, valores menores o iguales a cero, para los cuales la interpretación requiere evaluar si fueron registrados de forma correcta. Por otro lado, de acuerdo a los gráficos anteriores observamos que hay una dispersión grande en los datos de todas las variables, distribuidos de forma asimétrica a la derecha, lo cual es justificado y tiene relación por la concentración de reclamaciones pequeñas y menos frecuentes reclamaciones grandes. Esto es común en seguros, donde un pequeño número de reclamaciones grandes puede tener un impacto significativo en las reservas.

2.5. Exploración de las estructuras de datos para el análisis - Triángulos de run-off o de desarrollo

Para el cálculo de reservas de siniestros en la industria de seguros una de las metodologías más usadas es el triángulo de desarrollo. Un triángulo de desarrollo es una tabla que muestra cambios en el valor de varias cohortes de siniestros a través del tiempo (Torres, 2022).

Ilustración 9 - Evolución de la variable CumPaidLoss_D para la industria de seguros de compensación laboral. Elaboración propia.



Para los ejemplos y el análisis desarrollado en esta sección se trabajará con la variable **CumPaidLoss_D**, que como ya se había mencionado anteriormente corresponde a las pérdidas acumuladas pagadas y gastos asignados al final del año; lo cual representa la suma total de los pagos realizados por la aseguradora hasta ese momento para cubrir las reclamaciones y los costos asociados, incluyendo pagos realizados hasta el momento presente.

Un triángulo de desarrollo es una representación de la evolución a través de los años (normalmente 10 o 15 años) de las reclamaciones por los siniestros ocurridos en cada año de negocio, de forma que se muestra el valor de las obligaciones e ingresos que recibe una aseguradora con el paso de los años por la ocurrencia de siniestros.

En la Ilustración 10 se observa la evolución histórica de las pérdidas incurridas de la industria de seguros de compensación laboral entre los años 1988 a 1997.

Ilustración 10 - Matriz de desarrollo para la variable IncurLoss_D de la industria de seguros de compensación laboral. Elaboración propia.

DevelopmentLag	1	2	3	4	5	6	7	8	9	10
AccidentYear										
1988	285804	638532	865100	996363	1084351	1133188	1169749	1196917	1229203	1241715
1989	307720	684140	916996	1065674	1154072	1210479	1249886	1291512	1308706	1324671
1990	320124	757479	1017144	1169014	1258975	1315368	1368374	1394675	1417384	1431483
1991	347417	793749	1053414	1209556	1307164	1381645	1414747	1447121	1465508	1479177
1992	342982	781402	1014982	1172915	1281864	1328801	1370935	1399901	1419809	1437891
1993	342385	743433	959147	1113314	1187581	1243689	1278194	1301968	1322101	1337171
1994	351060	750392	993751	1114842	1193861	1243285	1276145	1302620	1321778	1340950
1995	343841	768575	962081	1087925	1164217	1210269	1243983	1264903	1282868	1300678
1996	381484	736040	937936	1056949	1131168	1182716	1218813	1239839	1260284	1274282
1997	340132	730838	940850	1066652	1142476	1191886	1223282	1248906	1262903	1277512

Ilustración 11 - Triángulo de desarrollo para la variable CumPaidLoss_D de la aseguradora "Allstate Ins Co Grp". Elaboración propia.

	DevelopmentLag	1	2	3	4	5	6	7	8	9	10
GRCODE	AccidentYear										
86	**1988**	70571	155905.0	220744.0	251595.0	274156.0	287676.0	298499.0	304873.0	321808.0	325322.0
	1989	66547	136447.0	179142.0	211343.0	231430.0	244750.0	254557.0	270059.0	273873.0	NaN
	1990	52233	133370.0	178444.0	204442.0	222193.0	232940.0	253337.0	256788.0	NaN	NaN
	1991	59315	128051.0	169793.0	196685.0	213165.0	234676.0	239195.0	NaN	NaN	NaN
	1992	39991	89873.0	114117.0	133003.0	154362.0	159496.0	NaN	NaN	NaN	NaN
	1993	19744	47229.0	61909.0	85099.0	87215.0	NaN	NaN	NaN	NaN	NaN
	1994	20379	46773.0	88636.0	91077.0	NaN	NaN	NaN	NaN	NaN	NaN
	1995	18756	84712.0	87311.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
	1996	42609	44916.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
	1997	691	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

La metodología utilizada o el propósito de estas estructuras de datos consiste en tomar la parte superior de la matriz de desarrollo y a partir del análisis del comportamiento de esta parte determinar o estimar la parte inferior, lo cual correspondería a los valores o cifras que deberían esperar las aseguradoras en los años venideros. Por ejemplo, supongamos que nos encontramos en finalizando el año 1997, por lo cual la información disponible sobre las pérdidas acumuladas incurridas de la aseguradora “Allstate Ins Co Grp” se vería de la forma como se observa en la Ilustración 11.

3. Preparación de los datos

El objetivo de esta fase es asegurar que los datos estén en la forma adecuada y calidad para su análisis posterior. Esta preparación de los datos es esencial para garantizar que los resultados del análisis sean precisos y significativos.

3.1. Seleccionar los datos

Se quiere realizar un modelo de predicción de las reservas que debería tener una aseguradora de la línea de seguros de compensación laboral, siguiendo la metodología de Chain Ladder con triángulos de desarrollo con aprendizaje de máquinas para predecir el monto de las reclamaciones futuras. Para este propósito se trabajará con las siguientes variables del conjunto de datos:

- Línea de negocio ('Single'): Indica si se trata de una entidad única (1) o una aseguradora de grupo (0). Esto puede ser relevante para considerar la estructura de la empresa aseguradora.
- Código de empresa('GRCODE')
- Año del accidente ('AccidentYear'): El año en que ocurrió el evento o accidente que dio lugar a la reclamación.
- Retraso en el desarrollo ('DevelopmentLag'): Representa cuántos años han pasado desde el año de ocurrencia del evento hasta el año actual. Esta variable es fundamental para construir los triángulos de desarrollo.
- Pérdida incurrida ('IncurLoss_D'): Pérdidas incurridas y gastos asignados reportados al final del año. Esta cifra es una parte importante en la evaluación de la cantidad total de dinero que la aseguradora espera gastar para cubrir las reclamaciones y los costos asociados en ese año específico.

- Pérdida pagada acumulada ('CumPaidLoss_D'): El monto acumulado pagado en reclamaciones hasta el año de desarrollo correspondiente. Este es un valor crítico para calcular las reservas.

Para el análisis se seleccionaron las variables 'IncurLoss_D' y 'CumPaidLoss_D', las cuales muestran las pérdidas en las que incurren las aseguradoras por los reclamos por Compensación Laboral.

3.2. Limpiar los Datos

Se decidió sacar del análisis aquellas aseguradoras que presentan en las variables de pérdidas seleccionadas valores en cero o negativo, con la finalidad de no generar inconsistencias en la modelización. Para le caso de 'CumPaidLoss_D' se tiene un total de 59 aseguradoras que contienen en su data histórica valores superiores a cero. Por el lado de 'IncurLoss_D' se tienen 62 aseguradoras que cumple esta misma característica.

3.3. Construir las nuevas estructuras de datos

Para la construcción del modelo de predicción se requiere de la construcción del triángulo de desarrollo o run-off, la cual muestra la evolución de los siniestros y reclamaciones para una aseguradora. Primero, teniendo en cuenta la Ilustración 12 tenemos los componentes que forman o componen la vida de un siniestro, en la cual se puede resaltar que los pagos para cubrir el siniestro se realizan en periodos de tiempo posteriores al momento de ocurrencia del evento.

Ilustración 12 Vida o run-off de un siniestro. Fuente: (Actuarial Community, 2021)

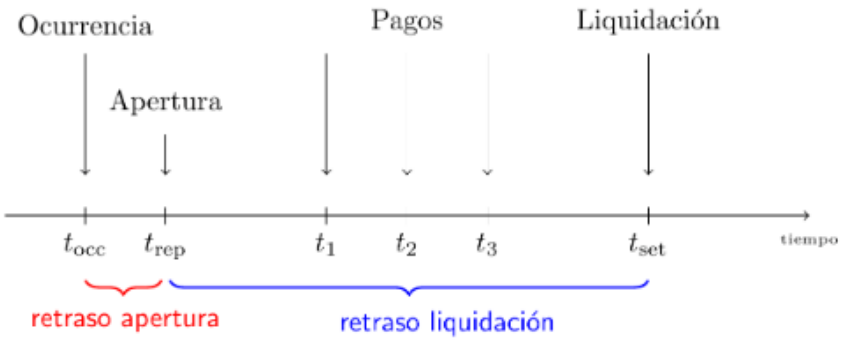
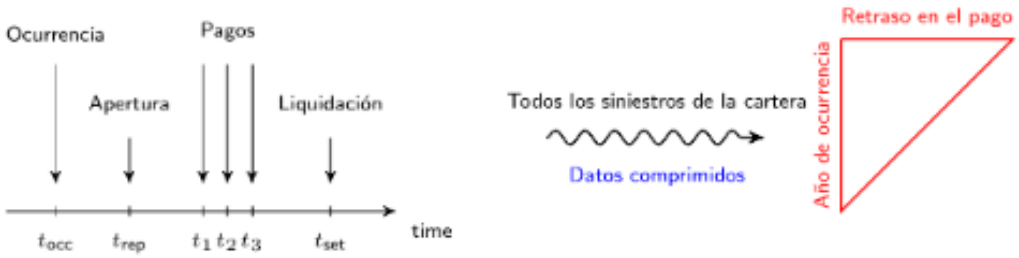


Ilustración 13 De datos granulares a triángulos de desarrollo. Fuente: (Actuarial Community, 2021)



De esta manera, de acuerdo a la información contenida en los datos iniciales, apartir de los cuales se tienen la evolución anterior para cada aseguradora en un periodo de 10 años, se construyen los triángulos de desarrollo superiores para el conjunto de aseguradoras seleccionado, los cuales serán utilizados en el modelo de predicción. Se realizará la modelización sobre las dos variables mencionadas anteriormente, para resaltar la conveniencia o mejor desempeño de un modelo respecto a otro, según la variable seleccionada para el análisis.

Ilustración 14 Triángulo de desarrollo. Elaboración propia.

	DevelopmentLag	1	2	3	4	5	6	7	8	9	10
GRCODE	AccidentYear										
86	**1988**	70571	155905.0	220744.0	251595.0	274156.0	287676.0	298499.0	304873.0	321808.0	325322.0
	1989	66547	136447.0	179142.0	211343.0	231430.0	244750.0	254557.0	270059.0	273873.0	NaN
	1990	52233	133370.0	178444.0	204442.0	222193.0	232940.0	253337.0	256788.0	NaN	NaN
	1991	59315	128051.0	169793.0	196685.0	213165.0	234676.0	239195.0	NaN	NaN	NaN
	1992	39991	89873.0	114117.0	133003.0	154362.0	159496.0	NaN	NaN	NaN	NaN
	1993	19744	47229.0	61909.0	85099.0	87215.0	NaN	NaN	NaN	NaN	NaN
	1994	20379	46773.0	88636.0	91077.0	NaN	NaN	NaN	NaN	NaN	NaN
	1995	18756	84712.0	87311.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
	1996	42609	44916.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
	1997	691	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

En la Ilustración 14 se tiene un ejemplo del triangulo de desarrollo o la estructura de datos que será utilizada para el modelo de regresión. Está estructura contiene la evolución de las pérdidas o reclamaciones para la aseguradora con código 86. En las filas se presentan los años de negocio y en las columnas se muestra la evolución o el tiempo en el que se realizan las reclamaciones de forma acumulada de cada año de negocio respectivamente. De manera que, si se observa la diagonal de la estructura de datos, está representa las pérdidas o reclamaciones que tiene la aseguradora en el año 1997, teniendo en cuenta que estas corresponden en gran parte a reclamaciones de años de negocio anteriores.

4. Modelización

Se quiere realizar un modelo predictivo para determinar el monto de dinero que una aseguradora debería reservar para responder por las reclamaciones que le serán reportadas en el futuro. Para esto es importante tener en contexto, partiendo del problema de negocio inicial, que la aseguradora actualmente está realizando la estimación de sus reservas con el modelo de Chain Ladder Determinístico.

De acuerdo con (Actuarial Community, 2021) el método de Chain Ladder es el método más usado en la estimación de reservas y el análisis de provisiones para aseguradoras. El método chain-ladder determinista se centra en el triángulo de desarrollo en forma acumulada. Recordemos que una celda (i, j) en este triángulo muestra la cantidad acumulada pagada hasta el período de desarrollo j por siniestros que ocurrieron en el año i. El método chain-ladder asume que los factores de desarrollo f_j existen de tal manera que

$$C_{i,j+1} = f_j \times C_{i,j}$$

Por tanto, el factor de desarrollo indica cómo la cantidad acumulada en el año de desarrollo j crece hasta la cantidad acumulada en el año $j + 1$.

Ilustración 15 Triángulo de desarrollo con datos de pagos acumulados que muestra la cantidad acumulada en el período 0 en azul y la cantidad acumulada en el período 1 en rojo. Fuente: Wüthrich and Merz (2008), Tabla 2.2. (Actuarial Community, 2021)

año accidente	retardo en pagos (en años)									
	0	1	2	3	4	5	6	7	8	9
1	5.947	9.668	10.564	10.772	10.978	11.041	11.106	11.121	11.132	11.148
2	6.347	9.593	10.316	10.468	10.536	10.573	10.625	10.637	10.648	
3	6.269	9.245	10.092	10.355	10.508	10.573	10.627	10.636		
4	5.863	8.546	9.269	9.459	9.592	9.681	9.724			
5	5.779	8.524	9.178	9.451	9.682	9.787				
6	6.185	9.013	9.586	9.831	9.936					
7	5.600	8.493	9.057	9.282						
8	5.288	7.728	8.256							
9	5.291	7.649								
10	5.676									

El método chain-ladder presenta una forma intuitiva para estimar o calcular estos factores de desarrollo. Dado que el primer factor de desarrollo f_0 describe el desarrollo de la cantidad acumulada de los siniestros desde el período de desarrollo 0 hasta el período de desarrollo 1, se puede estimar como la proporción de las cantidades acumuladas en rojo y las cantidades acumuladas en azul, resaltadas en la Ilustración 15 (Actuarial Community, 2021).

4.1. Técnicas de modelización seleccionadas

Dentro de las técnicas de modelación seleccionadas para la estimación de las reservas de la aseguradora se tuvieron en cuenta modelos de regresión ajustados por regularización y también se consideró un modelo de redes neuronales.

4.1.1. Modelo de regresión lineal

La regresión lineal normal busca una relación lineal entre una variable dependiente (objetivo) y una o más variables independientes (predictores). Para nuestro caso los predictores serán las reclamaciones históricas y la variable objetivo serán los reclamos futuros, los cuales nos darán un indicativo del valor que debe aprovisionar la aseguradora.

Un modelo de regresión normal tiene la siguiente forma, en función de las variables independientes para describir el comportamiento de la variable objetivo (dependiente):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$$

Donde y es la variable dependiente, x_i son las variables independientes, β_0 es la intersección, β_i son los coeficientes de las variables independientes y ε es el error.

Algunas de las consideraciones que se deben tener con este modelo según (Kobak, 2023) son:

- Minimiza la suma de los cuadrados de las diferencias entre las predicciones y los valores reales. Para este modelo se tiene la siguiente función de costo:

$$L(\beta) = \frac{1}{n}\sum_{i=1}^{n}(yi - \beta^T x_i)^2$$

- Puede sufrir de sobreajuste (overfitting) si hay multicolinealidad o demasiadas características.
- No incluye términos de penalización para reducir el sobreajuste.

Regularización del modelo de regresión lineal

Para penalizar los coeficientes de la regresión lineal se agrega un término de penalización a la función de costo así:

$$L(\beta) = \frac{1}{n}\sum_{i=1}^{n}(yi - \beta^T x_i)^2 + \lambda R(\beta)$$

Donde $\lambda R(\beta)$ es el parámetro de penalización y λ es llamado el parámetro de regularización. (Kobak, 2023). De esta forma, para el modelo de regresión propuesto se consideran los métodos de regularización de Ridge y Lasso.

4.1.2. Modelo de regresión lineal con regularización de Ridge

El modelo de regresión lineal con regularización de Ridge introduce una penalización en la función de costo para reducir la complejidad del modelo y prevenir el sobreajuste.

La función de costo de la regresión lineal con el ajuste de regularización de Ridge es (Kobak, 2023):

$$L(\beta) = \frac{1}{n}\sum_{i=1}^{n}(yi - \beta^T x_i)^2 + \lambda \sum_{i=1}^{n} {\beta_i}^2$$

El término de penalización $R(\beta) = \|\beta\|^2 = L2\left(\sum_{i=1}^{n} {\beta_i}^2\right)$ penaliza los coeficientes grandes, forzándolos a ser más pequeños. Por este motivo este modelo es útil cuando se tiene multicolinealidad y se quiere reducir el impacto de características irrelevantes.

4.1.3. Modelo de regresión lineal con regularización de Lasso

Por su lado, la regularización de Lasso, similar a Ridge, utiliza una penalización diferente para reducir el número de características significativas.

La función de costo queda de la forma (Kobak, 2023):

$$L(\beta) = \frac{1}{n}\sum_{i=1}^{n}(yi - \beta^T x_i)^2 + \lambda \sum_{i=1}^{n} |\beta_i|$$

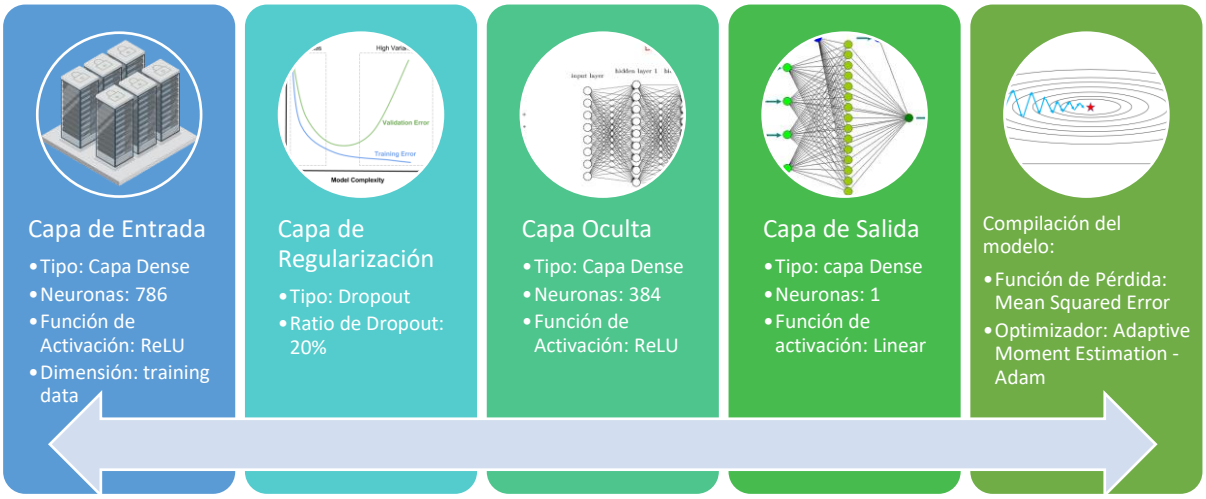
El término de penalización $R(\beta) = \|\beta\|^1 = L1(\sum_{i=1}^{n}|\beta_i|)$ tiende a hacer que algunos coeficientes sean exactamente cero, eliminando características menos importantes. Por tanto, este método es útil para la selección automática de características al descartar características irrelevantes.

Por otro lado, es importante mencionar que para el diseño del modelo de regresión se utilizó la metodología planteada en el artículo "*Statistical Methods for the Chain Ladder Technique*" de (Verall, 1994), en la cual se convierte la estructura de los triángulos de desarrollo al esquema requerido para utilizar el modelo de regresión lineal.

4.1.4. Red Neuronal

Se utilizó un esquema de red neuronal secuencial para intentar predecir las pérdidas de la aseguradora, compuesto por las siguientes capas:

Ilustración 16 Modelo Red Neuronal - Elaboración Propia



a) Capa de entrada:
   i) Tipo: capa Dense (Fully Connected) es decir que cada neurona está conectada a todas las neuronas de la siguiente capa.
   ii) Neuronas: 768 cantidad de neuronas determinada en el entrenamiento para obtener un desempeño aceptable en un tiempo razonable de procesamiento.
   iii) Función de Activación RELU (Rectified Linear Unit): en el entrenamiento del modelo se probaron las funciones de activación Sigmoid, Tanh y RELU, obteniendo con RELU el mejor desempeño en el modelo. Esta función de activación cuenta con algunas características deseables (Vakanski, 2021):
      (1) Solamente considera valores mayores a cero $f(x) = \max(0, x)$
      (2) Su tiempo de procesamiento o computo es más rápido, comparada a las funciones de activación mencionadas anteriormente.
      (3) Acelera la convergencia del gradiente descendente.
      (4) Previene o evita el problema de desvanecimiento del gradiente (Gradient Vanishing), el cual es un inconveniente común en el entrenamiento en redes neuronales muy profundas.
   iv) Input de dimensión determinado por la cantidad de características (features) en los datos de entrenamiento.

b) Capa de Regularización:
   i) Tipo Dropout: durante el entrenamiento se apagan aleatoriamente una proporción de neuronas junto con sus conexiones. Este método tiene un hiperparámetro p que indica la

proporción de neuronas a apagarse. Usualmente este valor se escoge entre 20% y 50% (Vakanski, 2021).

ii) Ratio de Dropout: Para el modelo se escogió una proporción de 20% de las conexiones que se apagan aleatoriamente para evitar overfitting o sobreajuste en el modelo.

c) Capa oculta:
   i) Tipo: Capa Dense
   ii) Neuronas: 384 (cantidad determinada en el entrenamiento)
   iii) Función de Activación RELU

d) Capa de salida:
   i) Tipo: capa Dense
   ii) Neuronas: 1 (salida unidimensional)
   iii) Función de activación: Linear, activación lineal para problemas de regresión.

e) Compilación del modelo:
   i) Función de pérdida: mean squared error para evaluar la diferencia entre las predicciones y los valores reales.
   ii) Optimizador: Adaptive Moment Estimation - "Adam", el cual es una variante del gradiente en descenso estocástico, el cual es reconocido por su eficiencia y rapidez. Este optimizador funciona de la siguiente manera (Vakanski, 2021):
      (1) De forma similar a gradiente en descenso con momento, ADAM realiza un promedio ponderado con los gradientes pasados:
         (a) Primer momento: $V^t = \beta_1 V^{t-1} + (1-\beta_2)\nabla L(\theta^{t-1})$
         (b) Segundo momento: $U^t = \beta_2 U^{t-1} + (1-\beta_2)\left(\nabla L(\theta^{t-1})\right)^2$
      (2) El parámetro actualizado es: $\theta^t = \theta^{t-1} - \propto \frac{\hat{V}^t}{\sqrt{\hat{U}^t}+\varepsilon}$
         (a) Donde $\hat{V}^t = \frac{V^t}{1-\beta_1}$ y $\hat{U}^t = \frac{U^t}{1-\beta_2}$
         (b) Con los siguientes valores propuestos por default:
            (i) $\beta_1 = 0.9$, $\beta_2 = 0.999$ y $\varepsilon = 10^{-8}$

f) Entrenamiento del modelo (learning rate scheduling):
   i) Se utiliza la partición de datos para entrenamiento
   ii) Se realizan 5,000 epochs, épocas o iteraciones completas a través de los datos.
   iii) Lotes de tamaño 32.

g) Predicción y cálculo del error:
   i) El modelo entrenado se utiliza para realizar predicciones sobre los datos de prueba.

ii) Se calcula el MSE (Mean Squared Error) entre las predicciones y los valores reales para comparar los resultados entre los modelos.

4.2. Diseño experimental

Para la selección del mejor modelo se seleccionó la métrica de desempeño de Error Cuadrático Medio y se validará bajo un esquema especial de validación cruzada como se describe a continuación.

4.2.1. Métrica de Desempeño

Se utilizarán las siguientes medidas de desempeño para evaluar el modelo sobre la industria aseguradora. Se proponen dos enfoques para este análisis: uno, las métricas de desempeño de cada aseguradora y una métrica ponderada para el desempeño del modelo en toda la industria.

- Error Cuadrático Medio (MSE - Mean Squared Error): El MSE mide el promedio de los errores al cuadrado entre las predicciones del modelo y los valores reales.

$$MSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

  Donde y_i son los valores reales, $\hat{y}_i$ son las predicciones del modelo, y n es el número de observaciones.

- Porcentaje de Error Absoluto Promedio (MAPE - Mean Absolute Percentage Error): El MAPE mide la precisión de un modelo o un método de pronóstico al calcular el porcentaje promedio de error absoluto en relación con los valores reales. Esta métrica es útil para evaluar qué tan cerca están las predicciones del modelo de los valores reales en términos porcentuales.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{|\hat{y}_i|} \times 100\%$$

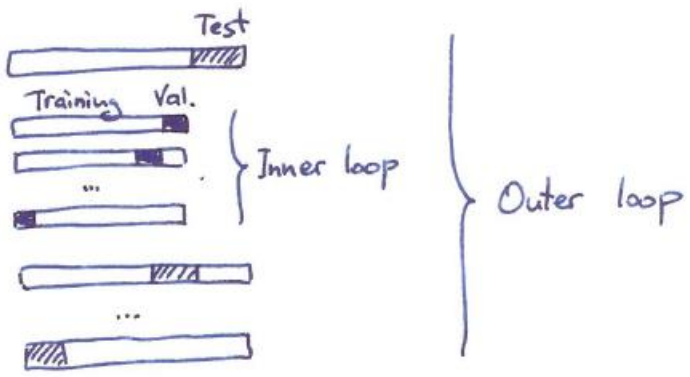
  Donde y_i son los valores reales, $\hat{y}_i$ son las predicciones del modelo, y n es el número de observaciones.

4.2.2. Validación Cruzada

Se aplica un esquema de validación cruzada sobre el conjunto de datos con el objetivo de evitar el sobreajuste (overfitting) o Debido a que se cuenta con un número pequeño de aseguradoras para realizar el modelo de predicción se seleccionó la técnica “Leave-one-out Cross Validation” (Kobak,

2023) para evaluar el desempeño de los modelos. Está técnica consiste en particionar el conjunto de datos así:

Ilustración 17 Nested Cross-Validation. Fuente: (Kobak, 2023)



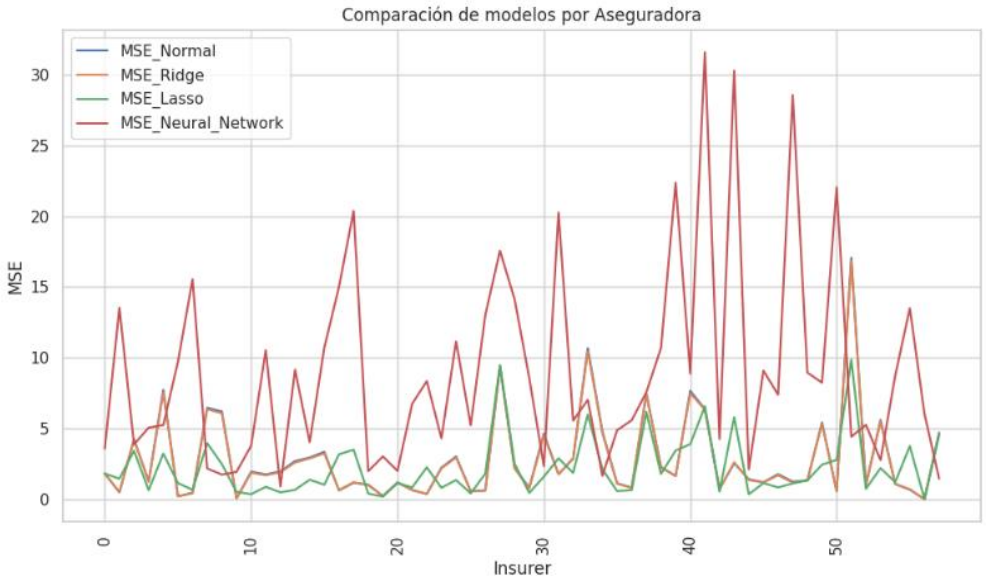
- Training:
  - Training: Conjunto para realizar el ajuste de los modelos, de forma que muestren las características que se quieren predecir.
  - Validation: Conjunto utilizado para seleccionar los hiper-parámetros de los modelos utilizados, y escoger entre ellos el mejor modelo.
- Test: conjunto utilizado para evaluar la precisión del modelo seleccionado en la predicción de las características deseadas.

En cada iteración se deja una aseguradora para test, del conjunto de entrenamiento se realiza una validación cruzada realizando iteraciones entre todas las aseguradoras de entrenamiento, dejando una por fuera en cada iteración para validar el desempeño de la combinación de parámetros seleccionada. De esta forma se llega al modelo con el mejor desempeño. Es importante mencionar que este proceso requiere de bastante tiempo de cómputo, por lo cual esta etapa demanda bastantes recursos para obtener un modelo con un buen ajuste.

4.3. Selección del mejor modelo

Para la selección del mejor modelo se realizó la validación cruzada con los cuatro modelos seleccionados (regresión normal, regresión de Ridge, regresión de Lasso y la red neuronal secuencial) y se obtuvo la métrica de desempeño definida con el error cuadrático medio. De esta manera se seleccionó el modelo con el mejor ajuste o menor porcentaje de error en la predicción.

Ilustración 18 Comparación de la Métrica de Desempeño (MSE) para los cuatro modelos seleccionados. Los valores están dados en puntos porcentuales. Elaboración Propia.



De acuerdo con los resultados obtenidos en la validación cruzada se evidenció que el modelo con el mejor desempeño de acuerdo con la métrica de Error Cuadrático Medio fue el modelo de regresión con el ajuste de regularización de Lasso, teniendo un MSE promedio de 2.21%. En la *Ilustración 19* se presenta las medidas de dispersión del error de cada modelo.

Ilustración 19 Mean Squared Error - MSE de los modelos seleccionados (en puntos porcentuales). Elaboración Propia.

	mean	std	min	25%	50%	75%	max
MSE_Normal	2.861	3.155	0.000	0.812	1.754	4.015	17.063
MSE_Ridge	2.816	3.112	0.000	0.767	1.688	3.960	16.861
MSE_Lasso	2.214	2.142	0.004	0.681	1.413	3.101	9.867
MSE_Neural_Network	9.041	7.361	0.893	3.912	7.201	11.041	31.589

De esta manera, se tiene los siguientes coeficientes para el modelo de regresión seleccionado con el mejor desempeño para la predicción de las reclamaciones de las aseguradoras:

	u	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_10	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_10
Parámetro	1.0	1.27	1.37	0.73	1.22	1.24	1.34	1.31	1.13	1.03	1.94	2.95	3.98	5.08	6.22	7.42	8.6	9.67	9.88

5. Evaluación del modelo

En esta etapa se evalúan los resultados obtenidos con el nuevo modelo de predicción propuesto respecto a los objetivos planteados inicialmente en el planteamiento del problema. Teniendo en contexto, con el desarrollo de este modelo de predicción se buscaba inicialmente mejorar la precisión en la estimación de las reservas necesarias para cubrir las reclamaciones futuras, lo cual ayuda a reducir la variabilidad en las estimaciones y garantizar una mejor gestión financiera de la compañía aseguradora.

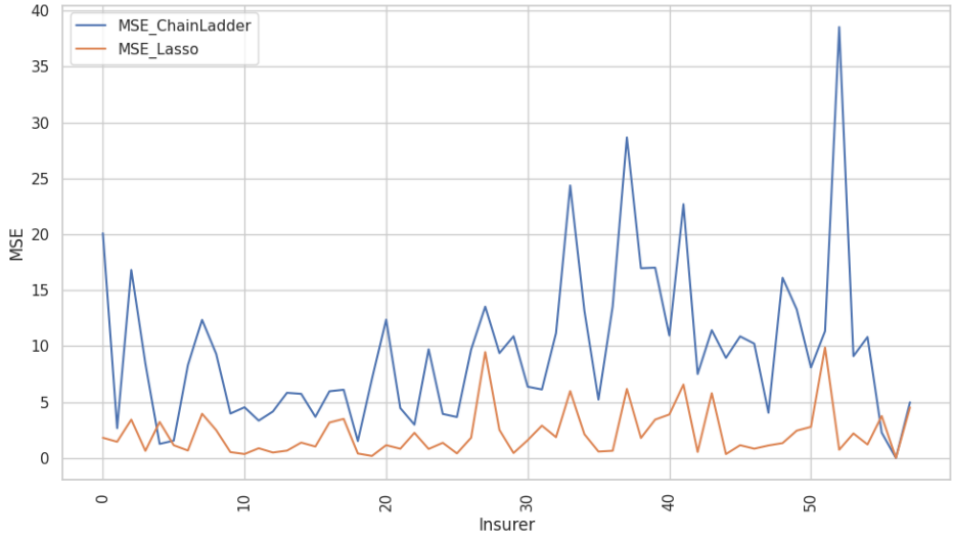
5.1. Evaluación de los resultados

El modelo implementado actualmente para la estimación de las reservas de la aseguradora es el modelo de Chain Ladder Determinístico, como se mencionó anteriormente. Con este modelo, la aseguradora tiene una margen de error del 9% aproximadamente. Por otro lado, en el nuevo modelo propuesto, el cual consiste en una regresión lineal con regularización de Lasso, se presenta una reducción en la tasa de error y por tanto una mejora en la precisión para la estimación de las reservas de la aseguradora como se observa en las siguientes ilustraciones.

Ilustración 20 Mejora en desempeño del modelo de predicción de la aseguradora. Elaboración Propia.

	mean	std	min	25%	50%	75%	max
MSE_ChainLadder	9.605	7.072	0.000	4.476	8.690	12.118	38.539
MSE_Lasso	2.214	2.142	0.004	0.681	1.413	3.101	9.867

Ilustración 21 Comparación del MSE del modelo de predicción de la aseguradora. Elaboración Propia.



5.2. Evaluación de los resultados frente a los objetivos y criterios de éxito de la organización

Consideramos el impacto que se tiene con el modelo propuesto sobre la estimación de reservas para para la cobertura de reclamaciones por lesiones laborales. En la Ilustración 22 se observa que se tiene una mejora en la precisión de la estimación de las reservas significativa, presentando una reducción o uso más óptimo de los recursos reservados en un 7.3%, tanto para las aseguradoras de tipo independiente como para aquellas que hacen parte de un conglomerado.

Ilustración 22 Impacto del modelo propuesto en la estimación de las reservas. Elaboración propia.

	Compañias	Reserva promedio	Error actual	Error propuesto	Mejora monto	Mejora %
Conglomerados	36.0	107217.0	10298.0	2374.0	7924.0	7.391
Independientes	96.0	14401.0	1383.0	319.0	1064.0	7.388
Total Industria	132.0	39714.0	3815.0	879.0	2936.0	7.393

5.2.1. Conclusiones respecto a los Objetivos de negocio:

El enfoque adoptado para mejorar la precisión en las estimaciones de reservas en la línea de Compensación Laboral resulta significativamente exitoso, logrando mejoras sustanciales en múltiples áreas clave identificadas:

- ✓ *Precisión en la estimación de reservas*: El nuevo modelo de regresión lineal con ajuste de Lasso demuestra una mejora significativa en la precisión de las estimaciones de reservas. Se reduce el MSE de 9.6% a 2.2%, lo que indica una reducción notable en la variabilidad en las estimaciones y una gestión financiera más sólida.
- ✓ *Reducción de riesgos financieros*: La implementación del nuevo modelo permite minimizar el riesgo financiero asociado con las reclamaciones de compensación de trabajadores. La capacidad predictiva mejorada facilita la identificación temprana de patrones de reclamaciones inusuales y una mejor gestión de los costos asociados.
- ✓ *Mejora en la eficiencia operativa*: La implementación de estrategias basadas en análisis de datos y modelos de aprendizaje automático optimiza los procesos de gestión de reclamaciones y administración de pólizas, lo que resulta en una mayor eficiencia operativa.
- ✓ *Cumplimiento Normativo*: Se mantiene un buen cumplimiento con las regulaciones y leyes pertinentes en el ámbito de la compensación de trabajadores. La precisión mejorada en las reservas contribuye a evitar sanciones regulatorias y disputas legales.

5.2.2. Evaluación de los Criterios de Éxito

Respecto a los criterios de éxito planteados inicialmente, tenemos las siguientes conclusiones:

- ✓ *Precisión en la estimación de reservas*: La disminución sustancial en el MSE refleja una mayor precisión en las estimaciones de reservas, demostrando la capacidad del nuevo modelo para reducir las variaciones entre estimaciones y reservas reales.
- ✓ *Índice de siniestralidad*: Aunque no se ha cuantificado directamente, la mejora en la precisión de las estimaciones de reclamaciones debería reflejarse en un índice de siniestralidad más bajo, indicando una gestión más efectiva de los costos de las reclamaciones.
- ✓ *Retención de clientes*: La satisfacción de los clientes está respaldada por la precisión mejorada en las reservas, lo que se traduce en una retención alta y constante de los clientes.
- ✓ *Cumplimiento Normativo*: El buen cumplimiento con las regulaciones y leyes regulatorias se mantiene, lo que refuerza la credibilidad y solidez de la aseguradora.
- ✓ *Evaluación de la situación*: La implementación exitosa del modelo de regresión lineal con ajuste de Lasso aborda efectivamente el desafío de aprovisionamiento de reservas en el ámbito de Compensación Laboral. El modelo supera las expectativas al mejorar significativamente la precisión en las estimaciones de reservas, permitiendo a la aseguradora tomar decisiones más informadas y garantizar una mejor gestión financiera.

Estas conclusiones reflejan el cumplimiento exitoso de los objetivos y criterios de éxito planteados inicialmente, destacando los beneficios del nuevo enfoque basado en el modelo de regresión lineal con ajuste de Lasso para la aseguradora en la línea de Compensación Laboral.

6. Despliegue del modelo en el ambiente operativo

Dentro del alcance de este proyecto de estimación de reservas para la línea de Compensación Laboral, no se llevará a cabo la implementación operativa del modelo desarrollado. A pesar de no desplegar directamente el modelo en un entorno productivo, se han logrado avances significativos en el desarrollo de un nuevo enfoque de estimación que demuestra mejoras sustanciales en la precisión de las reservas.

Aunque no se procederá con el despliegue operativo del modelo, es esencial destacar que el modelo de regresión lineal con ajuste de Lasso ha sido validado y probado en términos de su capacidad para mejorar la precisión en la estimación de reservas. Los resultados obtenidos se encuentran documentados y pueden ser utilizados como referencia en futuros análisis o proyectos relacionados con la gestión financiera y de riesgos en la línea de Compensación Laboral.

Referencias

Actuarial Community. (28 de Agosto de 2021). *Loss Data Analytics*. Obtenido de https://ewfrees.github.io/Loss-Data-Analytics-Spanish/

IBM. (1994). *IBM SPSS Modeler CRISP-DM Guide.* IBM Corporation.

Kobak, D. (2023). *Introduction to Machine Learning.* Obtenido de Dmitry Kobak: https://dkobak.github.io/

Meyers, G. G. (1 de September de 2011). *Loss Reserving Data Pulled from NAIC Schedule P*. Obtenido de Casualty Actuarial Society: https://www.casact.org/publications-research/research/research-resources/loss-reserving-data-pulled-naic-schedule-p

Torres, D. D. (2022). Aspectos Estadísticos y Actuariales. *Instituto Nacional de Seguros - FASECOLDA*. Bogotá: Instituto Nacional de Seguros.

Vakanski, A. (2021). Special Topics - Adversarial Machine Learning. University of Idaho, Idaho.

Verall, R. J. (1994). Statistical Methods for the Chain Ladder Technique. *Casualty Actuarial Society Forum*, 393 - 446.