Taller 1: Estimación Chain-Ladder

Brayan David Rincón Piñeros[a,c],

Francisco Albeiro Gomez Jaramillo[b,c]

[a]*Estudiante de Maestría en Actuaría y Finanzas*
[b]*Profesor, Departamento de Matemáticas*
[c]*Universidad Nacional de Colombia, Bogotá, Colombia*

1 ENTEDIMIENTO DEL NEGOCIO

1.1 Contexto

Históricamente, los seguros han acompañado el desarrollo de la humanidad. Los primeros surgieron en la antigüedad, cuando pequeños grupos de personas se unían para protegerse colectivamente de eventos fortuitos e inesperados. Con el tiempo, los seguros han evolucionado y se han convertido en herramientas que ofrecen protección frente a una amplia variedad de riesgos en distintos contextos.

Este avance ha sido impulsado por el desarrollo de la estadística y las matemáticas, lo que permite que los seguros sean objeto de estudios rigurosos. Esta rigurosidad confiere a los seguros un carácter formal y sólido, aspectos especialmente importantes dada su relevancia global. Este nivel de análisis riguroso posibilita que las compañías aseguradoras mantengan la solidez financiera y dispongan de los recursos necesarios para cubrir pérdidas o siniestros, aspectos que tienen un impacto directo en sus asegurados.

Con este contexto en mente, el presente proyecto busca estudiar y proponer una metodología alternativa a la técnica Chain-Ladder. El objetivo de este enfoque es estimar los flujos futuros de dinero necesarios para cubrir pérdidas que han ocurrido, pero aún no se han reportado a la compañía aseguradora.

1.1.1 Sector asegurador

Es importante entender que el sector asegurador está compuesto por diferentes entidades, entre las cuales se encuentran las compañías aseguradoras, los reguladores y otros organismos que participan de forma directa o indirecta en el mercado asegurador, como por ejemplo los reaseguradores. Para cada una de estas entidades, es crucial contar con solidez financiera para cubrir futuras pérdidas. Además, es fundamental conocer la solidez financiera de los otros actores involucrados, ya que la falta de solidez en alguno de ellos podría tener un efecto negativo en toda la red financiera del sector asegurador.

Por otro lado, es importante considerar la estructura interna de las compañías aseguradoras, la cual se organiza en lo que se conoce como 'ramos'. Un ramo es un conjunto de seguros que agrupa riesgos de características similares. Algunos ejemplos de ramos son: ramo de vida, ramo de automóviles, ramo de terremoto o ramo de responsabilidad civil.

En particular este proyecto se concentra en estudiar y analizar las perdidas del ramo de responsabilidad civil (Liability).

1.1.2 Responsabilidad civil

Los seguros de responsabilidad civil permiten a una persona, conocida como tomador, trasladar a otra entidad, en este caso la aseguradora, el riesgo de ser considerada responsable civilmente por causar daños a un tercero. Por ejemplo, el dueño de una vivienda sería responsable de los daños causados por objetos que pudieran caer o ser arrojados desde la vivienda al exterior, en caso de que esto ocurra el riesgo es cubierto por la compañía aseguradora.

Desde el enfoque del proyecto, un aspecto importante a considerar en el ramo de responsabilidad civil es expuesto por Nieto y Tamayo (2018), donde se indica: 'En particular, en negocios como vida individual, gastos médicos, responsabilidad civil, etc., la evolución del reporte de los siniestros es estacional'. Teniendo en cuenta esta característica de estacionalidad, es importante considerar aplicar técnicas y modelos que permitan capturar de forma precisa estos patrones estacionales, con el fin de realizar estimaciones más acertadas de las futuras pérdidas.

1.1.3 Metodología Chain-Ladder

Chain-Ladder es una técnica utilizada con frecuencia en la industria aseguradora. Esta metodología se emplea para prever la cantidad de dinero que una compañía de seguros necesitará reservar para cubrir futuras reclamaciones, basándose en el comportamiento histórico de los datos disponibles relacionados con reclamaciones pasadas. Como descripción general de la técnica, Chain-Ladder toma como referencia los datos asociados a reclamaciones de períodos anteriores. Con base en estos datos, se realiza la estimación de las reclamaciones futuras que podrían surgir. Es importante tener en cuenta que estas reclamaciones futuras pueden provenir de dos fuentes:

a. Siniestros incurridos, pero no reportados (Incurred But Not Reported, IBNR): En este caso, se intenta estimar las pérdidas asociadas a siniestros que ya han ocurrido pero que aún no han sido reportados a la compañía aseguradora.

b. Siniestros que aún no se han resuelto: Corresponde a siniestros que ya están bajo el dominio de la compañía aseguradora, pero que, debido a su complejidad, aún se encuentran en estudio y eventualmente se deberá incurrir en pagos futuros.

La esencia de la metodología Chain-Ladder está fundamentada en la idea de que las reclamaciones de un año en particular suelen desarrollarse de forma similar a las de años anteriores.

Pasos básicos para implementar Chain-Ladder:

1. **Crear un triángulo de pérdidas:** Los datos históricos se organizan en un triángulo en el que cada fila representa un año de origen de las reclamaciones, y cada columna representa un período de tiempo acumulativo desde ese año de origen.

Pagos acumulados de reclamaciones		**Año de desarrollo**			
		0	**1**	**2**	**3**
Año de accidente	**2011**	600	680	720	740
	2012	620	695	730	
	2013	680	760		
	2014	720			

Tabla 1. Triangulo de desarrollo

2. **Calcular factores de desarrollo:** Para cada par de años consecutivos dentro de una fila, se calcula un factor de desarrollo, que es la razón entre las pérdidas acumuladas en el año más reciente y las pérdidas acumuladas en el año anterior.

   Desarrollo año 2 al año 3: $\frac{740}{720} = 1.0278$

   Desarrollo año 1 al año 2: $\frac{720+730}{680+695} = 1.0545$

   Desarrollo año 0 al año 1: $\frac{680+695+760}{600+620+680} = 1.1237$

3. **Promediar factores de desarrollo:** Se toma un promedio de los factores de desarrollo para cada columna con el fin de estimar cómo se desarrollarán las pérdidas en el futuro.

4. **Proyectar pérdidas futuras:** Utilizando los factores de desarrollo promedio, se proyectan las pérdidas futuras para cada año de origen y se suman para obtener el total de la reserva necesaria.

La metodología Chain-Ladder es popular y está ampliamente adoptada por compañías aseguradoras; sin embargo, tiene ciertas limitaciones. Por ejemplo, asume que los patrones de desarrollo de las pérdidas en el pasado son un buen predictor de los patrones futuros, lo cual no es necesariamente correcto. Además, es sensible a variaciones en los datos, lo que puede afectar las estimaciones.

1.2 Objetivos y criterios de éxito del negocio

Objetivo general

Optimizar la asignación de recursos financieros en la compañía aseguradora mediante la mejora de la precisión en la estimación de factores de reserva para siniestros incurridos, pero no avisados en el ramo de responsabilidad civil.

Objetivos específicos

i. Identificar las limitaciones y oportunidades de mejora en el proceso actual de estimación de reservas para siniestros incurridos, pero no avisados, de tal forma que permita reducir la incertidumbre y riesgos financieros asociados.

ii. Incrementar la eficiencia operativa en la gestión de siniestros a través de un sistema de estimación de reservas más preciso.

1.3 Objetivos y criterios de éxito de la minería de datos

Objetivo general

Desarrollar y validar un modelo de machine learning que mejore el error asociado al método Chain-Ladder en la proyección de pagos futuros en el ramo de responsabilidad civil.

Objetivos específicos

i. Examinar los datos históricos de siniestros para identificar patrones, correlaciones y posibles outliers que puedan afectar la proyección.

ii. Utilizar técnicas de modelado estadístico y machine learning para crear modelos que proyecten con mayor precisión los pagos futuros de siniestros.

iii. Validar los modelos construidos mediante validación cruzada y comparar su rendimiento versus Chain Ladder.

2 ENTENDIMIENTO DE LOS DATOS

2.1 Recolección de los datos

Los datos asociados a este proyecto se han obtenido desde la web de la CAS (Casualty Actuarial Society), para esto la CAS obtuvo datos a través de la base de datos de la NAIC (National Association of Insurance Commissioners).

Para este proyecto se hace uso del conjunto de datos asociado a Responsabilidad civil de producto (Product Liability Data Set), este conjunto de datos se encuentra disponible en el repositorio del proyecto en github.

Estos datos se encuentran almacenados un archivo .csv y no poseen relación con otra fuente de datos.

2.2 Descripción y metadatos

El conjunto de datos esta formado por un total de 7000 observaciones y 13 variables. Estas variables son:

VARIABLE	DESCRIPCION
GRCODE	Código de la compañía según la NAIC (incluye grupos aseguradores e individuales).
GRNAME	Nombre de la compañía según la NAIC (incluye grupos aseguradores e individuales).
AccidentYear	Año del accidente (de 1988 a 1997).
DevelopmentYear	Año de desarrollo (de 1988 a 1997).
DevelopmentLag	Desfase en el año de desarrollo (AY-1987 + DY-1987 - 1).
IncurLoss	Perdidas incurridas y gastos asignados reportados al final del año.
CumPaidLoss	Perdidas pagadas acumuladas y gastos asignados al final del año.
BulkLoss	Reservas por pérdidas en bloque e IBNR (Reservas para siniestros ocurridos, pero no reportados) y gastos de defensa y contención de costos reportados al final del año.
EarnedPremDIR	Primas ganadas en el año de incurrencia: directas y asumidas.
EarnedPremCeded	Primas ganadas en el año de incurrencia: cedidas.
EarnedPremNet	Primas ganadas en el año de incurrencia: netas.
Single	1 indica una Única entidad, 0 indica un asegurador de grupo.
PostedReserve97	Reservas publicadas en el año 1997 tomadas del Underwriting and Investment Exhibit -Part 2A que incluye pérdidas netas no pagadas y gastos no pagados por ajuste de perdidas.

Tabla 2. Variables

2.3 Calidad de los datos

Se realiza la verificación de valores únicos y la existencia de datos faltantes (valores nulos) por variable, donde se obtienen los siguientes resultados:

VARIABLE	VALORES UNICOS	VALORES NULOS	TIPO
GRCODE	70	0	Cualitativa
GRNAME	70	0	Cualitativa
AccidentYear	10	0	Cualitativa
DevelopmentYear	19	0	Cualitativa
DevelopmentLag	10	0	Cualitativa
IncurLoss_R1	1417	0	Cuantitativa
CumPaidLoss_R1	1122	0	Cuantitativa
BulkLoss_R1	821	0	Cuantitativa
EarnedPremDIR_R1	425	0	Cuantitativa
EarnedPremCeded_R1	240	0	Cuantitativa
EarnedPremNet_R1	392	0	Cuantitativa
Single	2	0	Cualitativa
PostedReserve97_R1	50	0	Cuantitativa

Tabla 3. Resumen análisis de calidad

De este análisis de identifican variables cualitativas y cuantitativas, también se observa no existen datos faltantes. Para las variables cuantitativas se obtiene el análisis de medidas de tendencia central, dispersión y posición:

	IncurLoss	CumPaidLoss	BulkLoss	EarnedPremDIR	EarnedPremCeded	EarnedPremNet
Conteo	7000	7000	7000	7000	7000	7000
Media	2133,64	1262,71	463,16	3908,3	618,72	3289,57
Desviación estándar	10837,22	7437,6	3433,73	18108,13	3307,02	15112,64
Mínimo	-504	-862	-423	-12	-90	-22
P25%	0	0	0	5	0	3
P50%	5	1	0	120	7	84,5
P75%	249	121,25	8	1031	92	650,25
Máximo	142234	129300	87677	185016	39845	150180

Tabla 4. Medidas estadísticas variables cuantitativas

Es importante mencionar que la presencia de valores negativos en los registros de reserva es inherente a la naturaleza de la gestión de reservas en una compañía aseguradora. En casos en los que aparecen valores negativos, se ha efectuado una liberación de la reserva. Por ejemplo, supongamos que, en un determinado año, la aseguradora reserva $100 para un siniestro específico. Sin embargo, tras un análisis más detenido, se determina que el costo final del siniestro será de $70. En este caso, los $30 restantes previamente reservados deben ser liberados de la cuenta de reserva. Este monto se registra como un valor negativo, en este caso, -$30, para indicar su retirada de la reserva. De forma similar, puede ocurrir con las primas dados los movimientos contables que existen al realizar devoluciones, cambios en la vigencia de las pólizas o los valores asegurados.

2.4 Exploración inicial

Inicialmente se presenta un pequeño análisis descriptivo asociado a las variables cualitativas identificadas, donde se revisa el conteo de datos por cada uno de los niveles internos de cada una de estas variables.

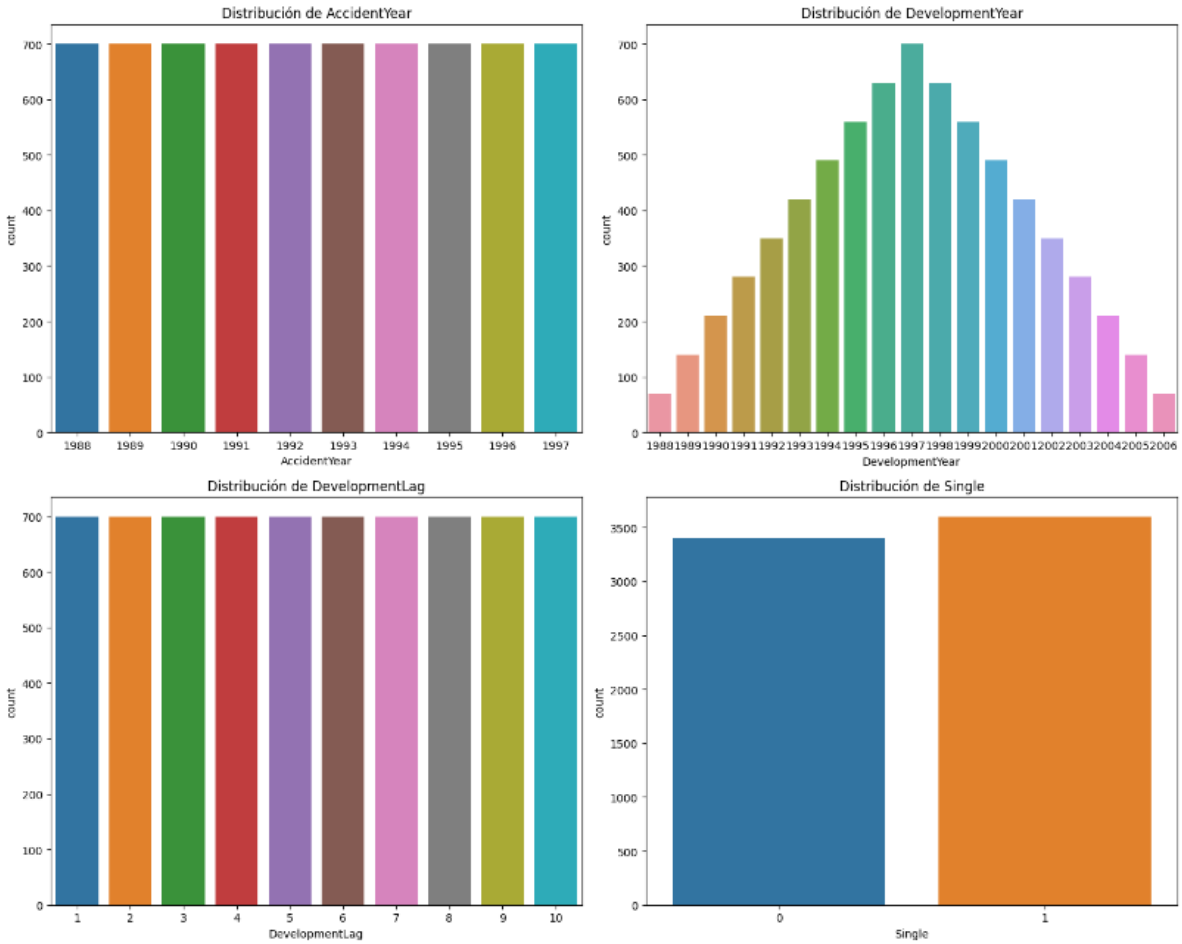

Imagen 1. Número de observaciones variables cualitativas

Se identifica que para cada año de accidente se cuenta con 700 casos. Estos casos están distribuidos en diferentes 'lags' de desarrollo, y para cada 'lag' también se observan 700 casos. Para ilustrar, supongamos que un accidente se reportó en el año 1990 y su 'lag' de desarrollo es 2. Esto significa que el siniestro se resolvió en el año siguiente al de su reporte, por lo que su año de desarrollo sería 1991. Dado que los datos consideran un desarrollo hasta un máximo de 10 años después del año de ocurrencia del accidente, observamos una alta concentración de casos que se resolvieron en 1997. Esta distribución se detalla mejor en la tabla siguiente:

A. Desarrollo / A. Accidente	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	Total
1988	70	70	70	70	70	70	70	70	70	70										700
1989		70	70	70	70	70	70	70	70	70	70									700
1990			70	70	70	70	70	70	70	70	70	70								700
1991				70	70	70	70	70	70	70	70	70	70							700
1992					70	70	70	70	70	70	70	70	70	70						700
1993						70	70	70	70	70	70	70	70	70	70					700
1994							70	70	70	70	70	70	70	70	70	70				700
1995								70	70	70	70	70	70	70	70	70	70			700
1996									70	70	70	70	70	70	70	70	70	70		700
1997										70	70	70	70	70	70	70	70	70	70	700
Total	**70**	**140**	**210**	**280**	**350**	**420**	**490**	**560**	**630**	**700**	**630**	**560**	**490**	**420**	**350**	**280**	**210**	**140**	**70**	**7000**

Tabla 5. Distribución de casos por año de accidente y desarrollo

Otro aspecto importante a tener en cuenta corresponder con la distinción entre pérdidas incurridas (IncurLoss_R1) y pérdidas pagadas (CumPaidLoss_R1). En general, las pérdidas incurridas suelen ser superiores a las pérdidas pagadas. Esto se debe a que las pérdidas incurridas no solo incluyen lo que ya se ha pagado, sino también lo que se espera pagar en el futuro. Por otro lado, las pérdidas pagadas son aquellas que ya han sido efectivamente desembolsadas.

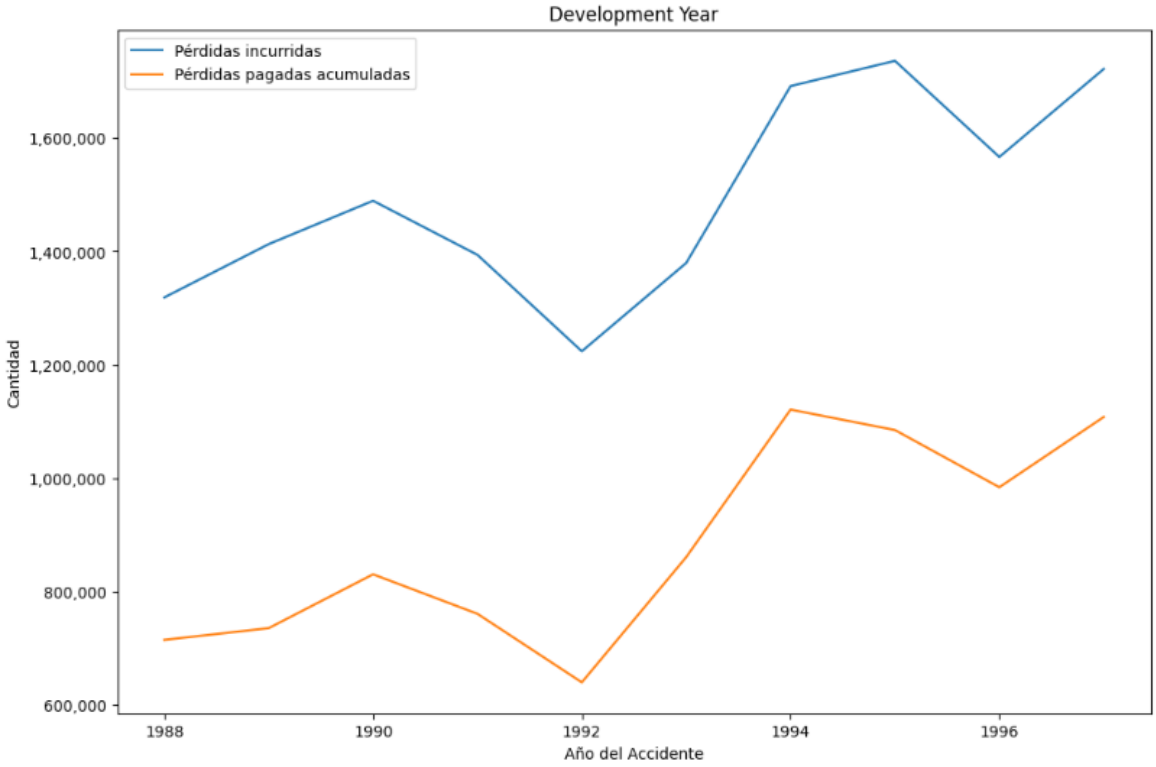

Imagen 2. Pérdidas incurridas y pagadas acumuladas

Adicionalmente, se observa que desde el año 1992 existe una tendencia creciente en los montos incurridos y pagados. También es importante precisar que la compañía Federal Ins Co Grp posee una alta participación debido al volumen histórico de sus reservas. A continuación, se presenta el 'top 10' de las compañías que tienen el mayor volumen de reservas incurridas en el conjunto de datos.

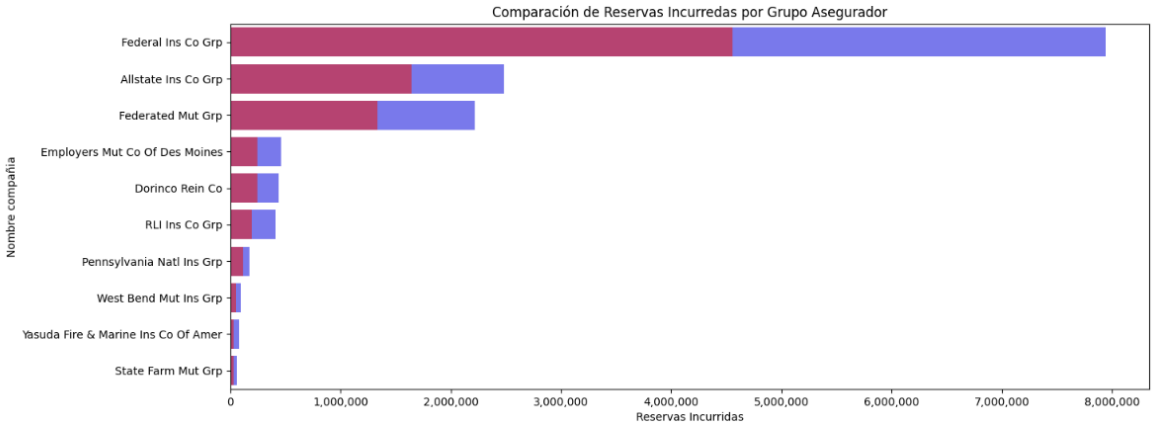

Imagen 3. Top 10 compañías con reservas incurridas

El total de reserva incurridas esta fuertemente relacionada con el volumen de primas ingresado por la aseguradora durante este tiempo, lo cual tiene sentido, dado que un mayor volumen de primas implica mayor cantidad de riesgo asumido por parte de la aseguradora.

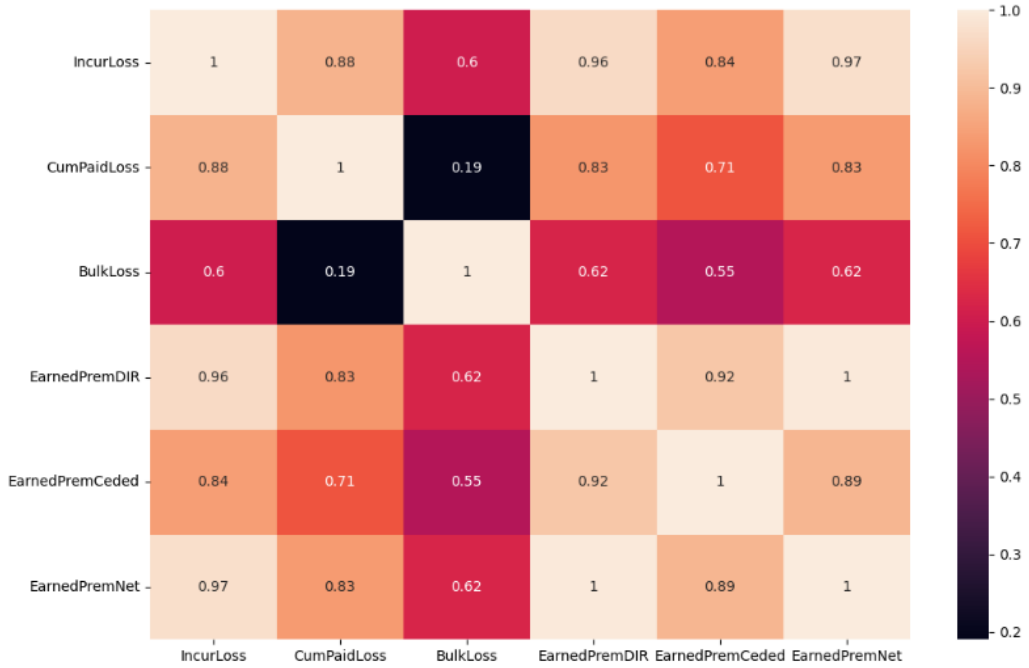

Imagen 4. Correlación

3 PREPARACIÓN DE LOS DATOS

Anteriormente se ha realizado una revisión del conjunto de datos disponible para el proyecto. Sin embargo, es importante precisar que estos datos son insumo para la construcción de los datos finales con lo que estaremos trabajando.

3.1 Construcción de los datos a trabajar

Con el objetivo de ilustrar y presentar la construcción de los datos, a continuación, se introduce el concepto de 'triángulo de desarrollo' o 'run-off triangle'. Un triángulo de desarrollo es una forma de organizar la información y es utilizado particularmente en la industria de seguros para analizar cómo evolucionan las pérdidas en el tiempo.

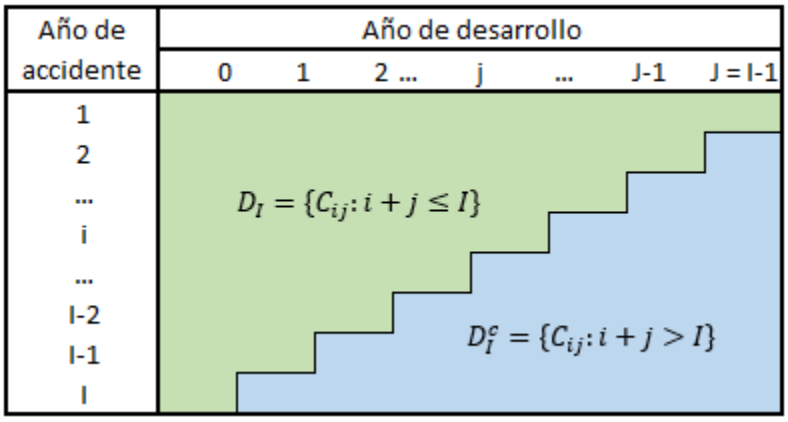

Imagen 5. Triangulo de desarrollo

La zona de color verde corresponde con el triángulo superior (upper triangle) y está vinculada a las observaciones pasadas. Por otro lado, la zona azul corresponde con el triángulo inferior (lower triangle) y es la parte que deseamos predecir.

En este proyecto, se construyen tanto los triángulos superiores como los inferiores para cada una de las compañías de seguros presentes en el conjunto de datos. Trabajaremos con los triángulos superiores como datos base que eventualmente utilizaremos para predecir el comportamiento del triángulo inferior. Los triángulos de desarrollo se presentan en dos versiones: por un lado, tenemos los triángulos de pérdidas incrementales (incremental claims); y por otro, los triángulos de pérdidas acumuladas (cumulative claims).

AccidentYear	DevelopmentLag									
	1	2	3	4	5	6	7	8	9	10
1988	152977	134765	142678	132822	127537	130700	124585	124234	125157	123424
1989	156329	147046	164934	151579	143446	135096	131342	129196	129623	124336
1990	153967	163464	157356	155191	144908	139400	144752	143904	139821	146666
1991	161047	155210	154054	138139	130849	126986	128977	133818	132789	132101
1992	150926	150118	140933	128428	112521	104923	103631	103756	122548	106058
1993	151363	138155	140392	137466	136748	137029	127915	130570	135247	144790
1994	162162	164685	151126	166509	182685	173018	165471	173014	175598	177498
1995	186276	170009	165786	171196	164834	164255	176005	173604	181154	183228
1996	165772	160219	165564	149057	148751	152526	154238	156323	157149	157095
1997	155152	173165	169897	165162	162096	172188	176946	177985	185136	184220

Tabla 6. Triangulo de perdidas incrementales general (todas las compañías)

AccidentYear	DevelopmentLag									
	1	2	3	4	5	6	7	8	9	10
1988	152977	287742	430420	563242	690779	821479	946064	1070298	1195455	1318879
1989	156329	303375	468309	619888	763334	898430	1029772	1158968	1288591	1412927
1990	153967	317431	474787	629978	774886	914286	1059038	1202942	1342763	1489429
1991	161047	316257	470311	608450	739299	866285	995262	1129080	1261869	1393970
1992	150926	301044	441977	570405	682926	787849	891480	995236	1117784	1223842
1993	151363	289518	429910	567376	704124	841153	969068	1099638	1234885	1379675
1994	162162	326847	477973	644482	827167	1000185	1165656	1338670	1514268	1691766
1995	186276	356285	522071	693267	858101	1022356	1198361	1371965	1553119	1736347
1996	165772	325991	491555	640612	789363	941889	1096127	1252450	1409599	1566694
1997	155152	328317	498214	663376	825472	997660	1174606	1352591	1537727	1721947

Tabla 7. Triangulo de pérdidas acumuladas general (todas las compañías)

Es importante tener en cuenta la siguiente notación y consideraciones.

i. Incremental claims X_{ij}:

- Pagos de reclamaciones.
- Número de reclamaciones reportadas con retraso (lag) j.
- Cambios en las reclamaciones incurridas.

ii. Cumulative claims C_{ij}:

- Pagos acumulados.
- Número total de reclamaciones reportadas.
- Perdidas incurridas (totales) reportadas.

Bajo esta notación nuestro objetivo se centra en predecir $C_{i,I-1}$ lo cual corresponde con el monto final de reclamación de año de accidente i.

Con base en lo anterior y teniendo en cuenta que nuestro objetivo es hacer uso de modelos de aprendizaje automático, los cuales no están diseñados para soportar la estructura de datos triangular típica de los siniestros, se procede a la construcción de un dataframe denominado “df_model”. Este dataframe incluye las siguientes columnas: GRCODE (código de grupo de riesgo), GRNAME (nombre del grupo de riesgo), AccidentYear (año del accidente), DevelopmentYear (año de desarrollo) y DevelopmentLag (retraso de desarrollo). Se efectúa también una limpieza de datos, excluyendo aquellas aseguradoras que reportan reservas incurridas (IncurrLoss) negativas, lo cual resulta en la selección de 17 aseguradoras de un total de 70. A pesar de esta reducción, es relevante señalar que las 17 aseguradoras seleccionadas representan el 90% del valor incurrido total, lo que indica que el conjunto de datos resultante sigue siendo representativo para el análisis.

También es importante precisar que los valores negativos no implican necesariamente errores en los datos. En la actividad habitual de las aseguradoras, estos valores negativos pueden representar liberaciones de reserva, es decir, situaciones en las cuales fondos son reintegrados a la contabilidad de la compañía.

4 MODELAMIENTO

En la etapa de modelado, nos enfocamos en la aplicación de regresiones lineal múltiple, ridge y lasso para predecir las reservas futuras. Estos modelos aprovechan como variables predictoras el lapso desde la ocurrencia del evento, conocido como lag de desarrollo, y el año en que ocurrió el accidente. La inclusión de estos factores es crucial para entender cómo los eventos pasados influyen en las estimaciones actuales y futuras, permitiendo una aproximación más ajustada y realista de las reservas necesarias para cubrir los siniestros pendientes.

Para realizar esto se cuento con la clase *Reserva_Regresion_lineal* la cual permite incorporar los modelos indicados anteriormente y obtener métricas de desempeño que permiten evaluar el rendimiento del modelo. Se cuenta con las siguientes métricas:

- MSE (mean square error)

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

- MAPE (mean absolute percentage error)

$$MAPE = \frac{1}{n}\sum\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

4.1 Estructura de la evaluación

Con el objetivo de evaluar el desempeño de los modelos ajustados, se emplea la técnica de validación cruzada. Dado el bajo número de observaciones, se recurre a la variante Leave-One-Out Cross Validation (LOOCV).

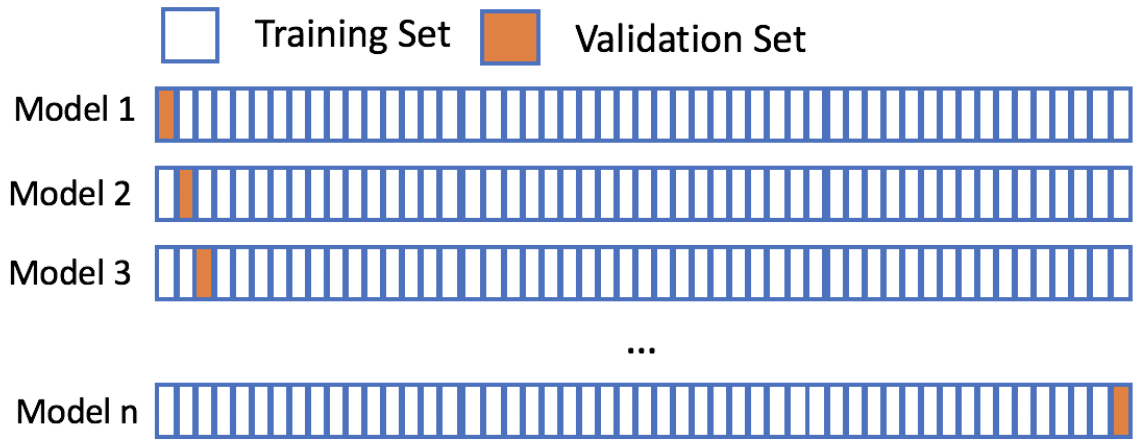

Imagen 6. Leave-One-Out Cross Validation

La imagen ilustra el procedimiento utilizado para realizar los experimentos de validación cruzada, con el fin de evaluar el rendimiento de los modelos. En cada iteración, se calcula el error asociado a la muestra de validación (validation set). Este proceso se repite 17 veces, correspondientes al número de aseguradoras en nuestro conjunto de datos. Al concluir todas las iteraciones, el modelo con el menor error medio cuadrático (MSE) o error absoluto porcentual medio (MAPE) promedio se considerará como el óptimo.

4.2 Resultados de la evaluación

Con el objetivo de constraster el mejor modelo usamos como método de referencia el Chain-ladder clásico, es decir, los modelos compiten entre ellos y seleccionamos el de mejor métricas (menor error) y luego contrastamos con el método Chain-ladder, donde obtenemos el siguiente resultado:

- Mejor modelo: regresión lineal múltiple
- Comparativa vs Chain-Ladder

```
Métrica MAPE con el método Chain-Ladder: 21.696563855753446
Métrica MAPE con el modelo final: 5.66920927155591
```

5 RECURSOS Y REFERENCIAS

1. Chain Ladder Method and Example
2. Explained Chain Ladder
3. Nieto y Tamayo 2018
4. Aplicación de los métodos Chain-Ladder
5. Liability Definition