Informe final chain Ladder usando el método CRISP-DM

1. Entendimiento del problema

1.1 Contexto

El grupo Sura es un grupo de inversionistas que manejan diferentes líneas de negocios, a estas pertenece Seguros Sura que ofrece varios tipos de seguros entre estos un tipo de seguros tanto para profesionales de la salud como para empresas prestadoras de Salud. La aseguradora cubre en el caso de ser demandados por mala praxis médica, que puede incluir un mal diagnóstico, negligencia médica o error humano. Estos seguros pueden ser de dos tipos: basado en ocurrencia o en reclamos realizados. En el primero la cobertura funciona durante el tiempo de la póliza, sin importar si los reclamos son presentados después de la vigencia de la póliza, importa cuándo ocurrió el evento; en el segundo caso la cobertura solo aplica si durante el tiempo de póliza se realiza la reclamación.

Las provisiones son un aspecto importante de las empresas de seguros, pues son necesarias para evitar problemas legales y para no afectar la liquidez de la empresa. Se realiza un estudio de provisiones dependiendo de la línea de seguro, en este caso del seguro por mala práctica médica

1.2 Método de Chain Ladder

El método de "chain Ladder" es una técnica utilizada en seguros y análisis actuarial para estimar las reservas de siniestros futuros. Aquí tienes una explicación simplificada:

-Datos Iniciales:
Comienza con datos históricos de siniestros pasados, generalmente organizados en una matriz triangular. Cada celda de la matriz representa los siniestros ocurridos en un año específico y se clasifican por año de ocurrencia y año de desarrollo.

-Desarrollo de los Siniestros:
La matriz triangular muestra cómo los siniestros desarrollan con el tiempo. Por ejemplo, la primera columna puede representar los siniestros ocurridos en el primer año, la segunda columna en el segundo año, y así sucesivamente.

-Factor de Desarrollo:
Se calculan los factores de desarrollo, que son las tasas de cambio entre los años sucesivos. Estos factores se aplicarán a los siniestros ya conocidos para estimar los siniestros futuros.

-"Chain Linking" (Encadenamiento):
Los factores de desarrollo se aplican de manera secuencial a lo largo de las columnas (a lo largo de las cadenas), de ahí el nombre "chain ladder". Cada factor se multiplica por el valor conocido en la celda correspondiente para obtener una estimación del año siguiente.

-Proyección Futura:
Repites este proceso hasta llegar al final de la matriz triangular, proyectando así los siniestros futuros para cada año de ocurrencia.
Este método asume que las tendencias históricas en el desarrollo de siniestros se mantendrán en el futuro, por lo que es más efectivo en situaciones donde las condiciones no han cambiado significativamente. También es importante considerar la calidad de los datos y ajustar el modelo según sea necesario.

2. **Objetivos**

- Mejorar la precisión de provisiones estimadas para la línea de seguro Medica F2
- Encontrar patrones en los datos que permitan entender mejor los factores que podrían influenciar las reclamaciones

El estudio se declara un éxito si

- Se logra reducir el consumo computacional por parte del método
- Se mejora la relación entre departamentos

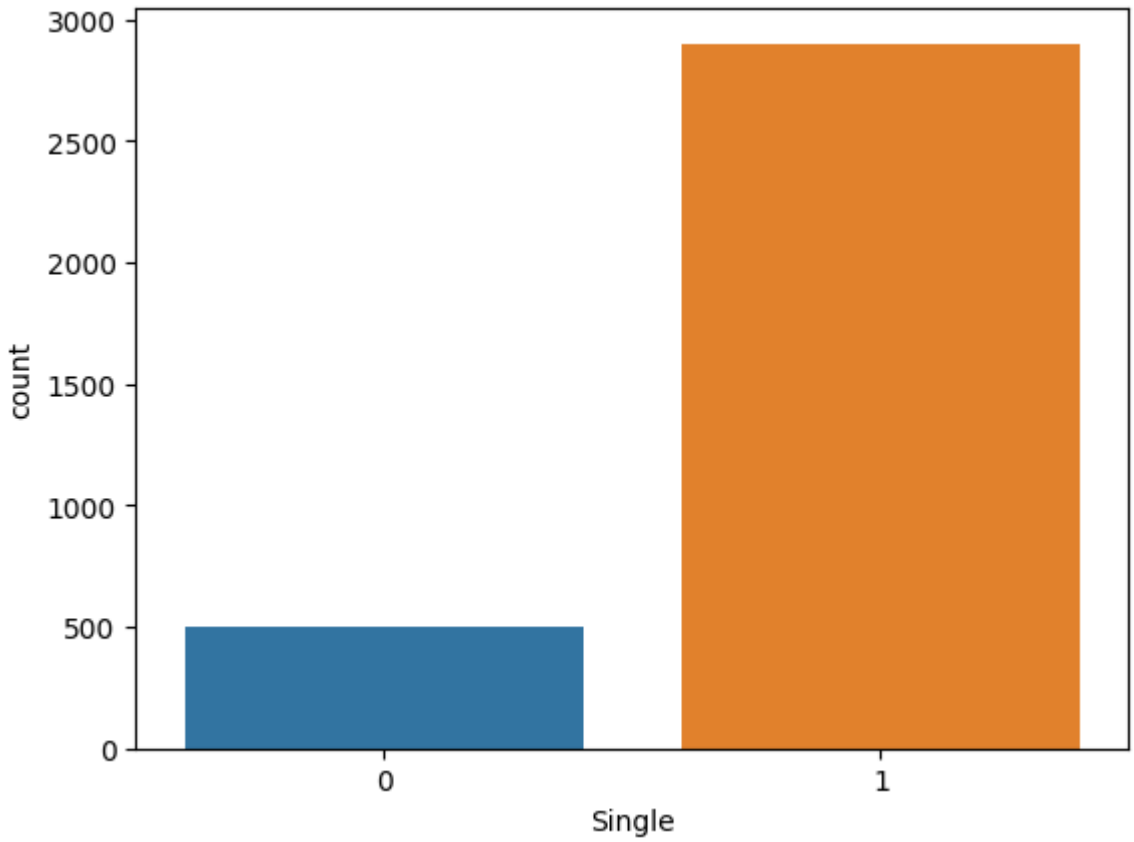
3. **Entendimiento de los datos**

Descripción del Conjunto de Datos
Este conjunto de datos contiene reclamaciones históricas sobre seguros de compensación laboral durante varios años. Desde la base de datos de la empresa, tuvimos acceso a las siguientes variables:
- GRCODE: Código de la empresa según la NAIC (Asociación Nacional de Comisionados de Seguros), que es un identificador único para cada compañía o grupo de seguros.
- GRNAME: Nombre de la empresa según la NAIC, el nombre de la compañía o grupo de seguros correspondiente al código NAIC.
- AccidentYear: El año en el que ocurrieron accidentes o reclamaciones, que va desde 1988 hasta 1997.
- DevelopmentYear: El año en el que se desarrolla o informa la reclamación, y puede ir desde 10 años después del año del accidente.
- DevelopmentLag: El rezago de desarrollo, que parece calcularse como (AY-1987 + DY-1987 - 1). Probablemente representa el período de tiempo entre cuando ocurrió el accidente y cuando se informaron o resolvieron las pérdidas.

- IncurLoss: Pérdidas incurridas y gastos asignados informados al final del año especificado. Esta columna probablemente contiene datos financieros relacionados con los costos incurridos por el asegurador debido a reclamaciones y gastos.
- CumPaidLoss_: Pérdidas acumulativas pagadas y gastos asignados al final del año especificado. Esto representa la cantidad total de dinero pagado por el asegurador por reclamaciones y gastos hasta ese año.
- BulkLoss_: Reservas de pérdidas netas y gastos de defensa y contención de costos informadas al final del año. Esta columna podría contener datos relacionados con reservas apartadas para reclamaciones y gastos futuros.
- PostedReserve97_: 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 pérdidas. Esto probablemente representa una cantidad de reserva específica para el año 1997.
- EarnedPremDIR_: Primas devengadas en el año de incurrimiento - directas y asumidas. Esta columna podría contener datos relacionados con las primas ganadas por el asegurador para pólizas emitidas en el año especificado.
- EarnedPremCeded_: Primas devengadas en el año de incurrimiento - cedidas. Esto podría representar primas devengadas pero luego cedidas a compañías de reaseguro.
- EarnedPremNet_: Primas devengadas en el año de incurrimiento - netas. Esta columna probablemente representa las primas netas devengadas después de tener en cuenta tanto las primas directas como las cedidas.
- Single: Un indicador binario donde 1 indica una entidad aseguradora única y 0 indica una aseguradora de grupo. Esta columna podría usarse para clasificar a las compañías de seguros como entidades independientes o parte de un grupo.

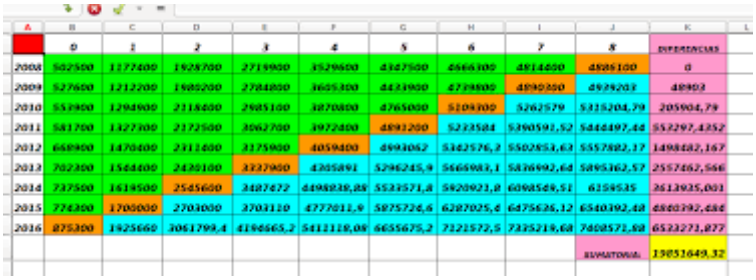
CumPaidLoss_	IncurLoss_F2	BulkLoss_F2	EarnedPremDIR_F2	EarnedPremCeded_F2	EarnedPremNet_F2	
count	3.400.000.000	3.400.000.000	3.400.000.000	3.400.000.000	3.400.000.000	3.400.000.000
mean	6.706.067.059	11.609.344.412	1.095.803.235	14.111.605.882	1.803.497.059	12.308.108.824
std	17.121.815.066	26.802.819.463	7.612.672.277	26.399.284.476	3.893.424.584	24.824.225.795
min	-1.190.000.000	-17.000.000	-32.101.000.000	-781.000.000	-6.214.000.000	-728.000.000
25%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	187.000.000	645.000.000	0.000000	1.500.000.000	106.500.000	1.302.000.000
75%	4.385.500.000	9.050.500.000	107.250.000	18.094.500.000	1.473.500.000	13.490.000.000
max	113.189.000.000	179.425.000.000	104.402.000.000	131.948.000.000	25.553.000.000	135.318.000.000

Revisamos máximos y mínimos y datos faltantes, revisamos distribución de la variable single, pues es la única que notaba diferencia interesante



4. **Preparación de los datos**

   Luego de que se realiza una exhaustiva verificación de las variables, se pueden revisar los elementos con los que se va a trabajar, en este caso los triángulos de desarrollo



	0	1	2	3	4	5	6	7	8	DIFERENCIAS
2008	502500	1177400	1928700	2719900	3529600	4347500	4666300	4814400	4886100	0
2009	527600	1212200	1980200	2784800	3605300	4423900	4729800	4890850	4929203	48903
2010	553900	1294600	2118400	2985100	3870800	4765000	5209300	5262579	5315204,79	205904,79
2011	581700	1327300	2172500	3062700	3972400	4893200	5233584	5390591,52	5444497,44	552297,4352
2012	668900	1470400	2311400	3175900	4059400	4993062	5342576,2	5502853,63	5557882,17	1498882,167
2013	702300	1544400	2420100	3327900	4205891	5296245,9	5665983,1	5836692,64	5895362,57	2557462,566
2014	737500	1619500	2545600	3487672	4498838,88	5533571,8	5920921,8	6098546,51	6159535	2613935,001
2015	774300	1700600	2703000	3703110	4777011,9	5875724,6	6287025,4	6475626,12	6540392,48	4860392,484
2016	875300	1925660	3061799,4	4194665,2	5411118,08	6655675,2	7121572,5	7335219,68	7408571,88	6523273,877
									SUMATORIA	19051649,32

   Es importante precisar que para el caso de este ejercicio se encontró que solo existían 34 aseguradoras que ofrecían este tipo de pólizas lo que reduce la cantidad de datos que puedan ser utilizados en el posterior modelamiento.

5. **Modelamiento**

   En este caso se decidieron tomar 3 tipos de modelos diferentes, el primero que se eligió fue el chain Ladder determinístico en el que realizamos un cálculo de

factores de incremento y a partir de esto predecimos los valores, un método de regresión lineal y por último una sencilla red neuronal.

El método de evaluación que se escogió fue la medida MSE pues Penaliza más los errores grandes y es matemáticamente conveniente para cálculos y optimización.

6. Resultados del modelamiento

Al probar los 3 modelos mencionados anteriormente, encontramos que el que mejor se comporta teniendo en cuenta que nuestra medida es MSE es la red neuronal, pues es el valor más bajo y considerando los datos que tenemos también es eficiente pues no tarda en cargar

7. Conclusiones

En conclusión, después de evaluar y comparar diversos modelos para el análisis de estimación de siniestros en seguros, se evidencia que la red neuronal emerge como la opción más prometedora. Su capacidad para capturar patrones complejos y no lineales en los datos, así como su capacidad de adaptación a cambios en las condiciones subyacentes, la posiciona como una herramienta poderosa en este escenario.

Si bien es crucial considerar aspectos como la complejidad computacional y la interpretabilidad del modelo, la red neuronal se destaca como la elección más efectiva para abordar la complejidad inherente en los datos de siniestros y proporcionar proyecciones precisas.