Oscar Pulido

Aplicaciones del aprendizaje de máquinas en Actuaría y Finanzas

Aplicación de la metodología CRISP-DM para la estimación de la reserva IBNR para el ramo médico

Control de versiones

Versión	Autor	Fecha	Aprobado por
1	Oscar Pulido	31/11/2023	Francisco Gómez

El objetivo de este documento es presentar el desarrollo del proyecto de estimación de reservas con base en la metodología CRISP-DM. Este documento se ve complementado por dos Notebooks en Jupyter que se encuentran en el siguiente enlace: https://github.com/osebasp/ML-Applications-for-Actuarial-Science bajo el nombre de ***Estimación reservas de siniestros V01 Parte 1*** para la parte inicial de datos y ***Estimación reservas de siniestros V01 Parte 2*** para la parte de modelación, y donde se lleva a cabo la parte práctica del trabajo. Particularmente, se desea ahondar en el uso de la metodología Chain-Ladder para la constitución de la reserva de siniestros IBNR (Incurred But Not Reported), que compone la Reserva de Siniestros Pendientes, de acuerdo al capítulo IV del decreto 2973 de 2013 de la legislación colombiana con respecto a la actualización del régimen de reservas para las compañías aseguradoras. (Estatuto Orgánico del sistema financiero, Decreto 2973 de 2013, diciembre 20 de 2013). También es deseable ahondar en distintas metodologías de Machine Learning para la solución del problema en cuestión, como la regresión de Ridge y de Lasso. Lo que se pretende en el sentido del negocio es lograr montos por reserva menores pero razonables, de manera que el resultado del ejercicio permita maximizar las ganancias de las compañías y reducir el impacto de las constituciones en los fondos propios de la compañía, ya que habría un menor pasivo en el balance contable de las compañías.

Contenido

Determinación de objetivos de negocio

Antecedentes empresariales

Determinación de la estructura organizacional

En primer lugar, se mostrará la estructura organizacional de una compañía aseguradora:

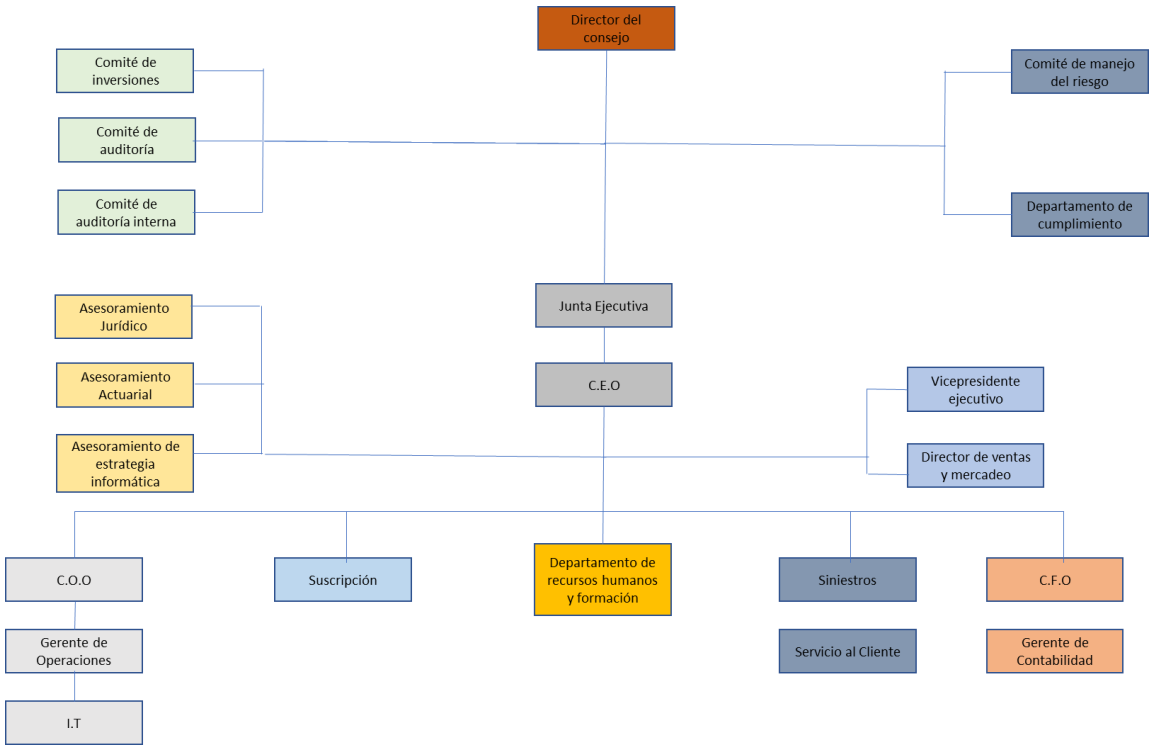


Diagrama realizado con base en https://www.pacificcrosshealth.com/en/company-information/organization-structure/

Para los propósitos de este proyecto, se determina que las áreas más importantes son: Asesoramiento actuarial, Asesoramiento de estrategia informática, I.T, Siniestros, Contabilidad y el Comité de Manejo del Riesgo, además de los directivos. De manera que, los miembros claves del proyecto se encuentran en estas áreas. Los stakeholders del proyecto no se muestran en el gráfico, pero el grupo está conformado por aquellos accionistas que financiarán el proyecto, a estos participantes les interesa que el ejercicio conduzca a una menor cantidad de dinero a reservar. Las áreas que estarán afectadas en mayor medida por este proyecto son la de siniestros, actuaría, contabilidad y I.T.

Determinación del área de problema

Una vez identificada la estructura organizacional de la compañía se pasa determinar el área del problema. El área de problema en este proyecto es la de siniestros, que en esta compañía aseguradora es la encargada de constituir la Reserva de Siniestros Pendientes, y, por lo tanto, la reserva IBNR. Si bien existen más áreas involucradas, el área que verá mayores transformaciones es esta.

El problema que enfrenta actualmente el área de Siniestros de la compañía aseguradora (y las compañías en general) es que los métodos actualmente utilizados para estimar el IBNR parecen ser insuficientes para determinar la reserva que realmente se debe constituir y así poder cumplir con los requisitos del regulador, de los accionistas y de las áreas que requieren de esta información para cumplir con sus objetivos de negocio. Por lo tanto, se piensa que se puede estar haciendo una sobreestimación, lo cual se traduce en un mayor pasivo para la compañía, es decir un mayor impacto en fondos propios. Actualmente, la compañía no contempla metodologías diferentes a las tradicionales, pues se ha contemplado lo mismo desde que se tiene conocimiento y el regulador lo aprueba así, sin embargo, valdría la pena pensar en metodologías que involucren minería de datos para abordar esta problemática.

Determinación de la solución actual al problema

Actualmente, se utiliza la metodología de triángulos run-off, cuyo objetivo es estimar de manera determinística los montos de dinero pagados por la compañía aseguradora en el futuro, gráficamente, se espera completar un triángulo de información conocida hasta obtener un rectángulo con montos estimados de dinero en pago de siniestros, esto se ve de la siguiente manera:

Accident year	Development year								
	0	*1*	⋯	*k*	⋯	*n-i*	⋯	*n-1*	*n*
0	$Y_{0,0}$	$Y_{0,1}$	⋯	$Y_{0,k}$	⋯	$Y_{0,n-i}$	⋯	$Y_{0,n-1}$	$Y_{0,}$
1	$Y_{1,0}$	$Y_{1,1}$	⋯	$Y_{1,k}$	⋯	$Y_{1,n-i}$	⋯	$Y_{1,n-1}$	$Y_{1,}$
⋮	⋮	⋮		⋮		⋮			
i	$Y_{i,0}$	$Y_{i,1}$	⋯	$Y_{i,k}$	⋯	$Y_{i,n-i}$	⋯	$Y_{i,n-1}$	$Y_{i,}$
⋮	⋮	⋮		⋮					
n-k	$Y_{n-k,0}$	$Y_{n-k,1}$	⋯	$Y_{n-k,k}$	⋯	$Y_{n-k,n-i}$	⋯	$Y_{n-k,n-1}$	Y_{n-}
⋮	⋮	⋮							
n-1	$Y_{n-1,0}$	$Y_{n-1,1}$	⋯	$Y_{n-1,k}$	⋯	$Y_{n-1,n-i}$	⋯	$Y_{n-1,n-1}$	Y_{n-}
n	$Y_{n,0}$	$Y_{n,1}$	⋯	$Y_{n,k}$	⋯	$Y_{n,n-i}$	⋯	$Y_{n,n-1}$	$Y_{n,}$

$Y_{i,n-i}$~ cumulative losses of the present calendar year n or current (cumulative) losses.
$Y_{i.n}$ ~ ultimate (cumulative) losses.

Imagen obtenida de https://www.researchgate.net/figure/Run-off-triangle-in-the-form-of-cumulative-losses_tbl1_348451688

La reserva IBNR se obtiene a partir del rectángulo mencionado hace unos momentos, particularmente interesa la última columna, normalmente llamada “Ultimate” y la diagonal principal, es decir, la que contiene aquellos montos pagados por siniestros a la última fecha de información conocida, asimismo, es la que divide la información conocida de la proyectada. Estos dos elementos se restan y el resultado obtenido es la reserva IBNR de un ramo determinado de una compañía de seguros.

Para obtener el triángulo run-off de información conocida se parte de un conjunto de datos de siniestros ocurridos y pagados de un determinado ramo de seguros, como puede ser Autos o Responsabilidad Civil, entre otros. Posteriormente se constituyen triángulos acumulados de pagos por periodicidad mensual. Después entra en juego la metodología Chain-Ladder para la estimación de los pagos futuros (la lógica de ese cálculo se muestra en secciones posteriores de este

documento). El problema que la compañía enfrenta es que la metodología es relativamente sencilla de aplicar y usar, por lo tanto, es de esperar que los resultados no sean los deseados en muchos casos y que exista poca posibilidad de calibración de parámetros, no es muy flexible.

Como se mencionó, la metodología parece ser insuficiente, por lo que la compañía se ve en la necesidad de combinar el enfoque tradicional con algunas técnicas habituales que requieren insumos como el monto de primas cobradas por la compañía a unas fechas determinadas para la estimación de la reserva IBNR, tales como Bornhuetter-Ferguson, Loss Ratio, entre otras. Esta mezcla de metodologías es aceptada por la compañía y el regulador ya que es efectiva, sin embargo, esto implica manualidades adicionales en el proceso, lo cual puede representar riesgos para la solvencia de la compañía, los intereses de los accionistas y retrasos en la presentación de los informes mensuales a la junta directiva de la compañía, esto repercute en otras áreas como la de contabilidad, que debe generar los estados financieros basándose en los montos establecidos de la reserva IBNR.

Definición de objetivos de negocio

Con base en la información anterior, se puede identificar que el problema a resolver en términos generales es lograr estimar la reserva IBNR utilizando metodologías robustas de minería de datos en el ramo de salud. En ese sentido, se esperan lograr resultados acertados basados en ejercicios de comparación con las metodologías tradicionales usando métricas que satisfagan los requisitos del regulador y los intereses económicos de la compañía. El objetivo principal que desean los accionistas, es que el resultado del ejercicio de un menor monto de reservas constituidas, lo cual se traduciría en un menor impacto en fondos propios. Como se ve a continuación, las reservas técnicas representan el mayor pasivo de las compañías aseguradoras, por lo tanto, es de interés que se reduzcan la medida de lo posible.

Liberty Seguros S.A.

Estados de Situación Financiera

	Notas	A 31 de diciembre de 2022	2021
		(Expresados en millones de pesos colombianos)	
Activos			
Efectivo y equivalentes de efectivo	6	$ 101.952	$ 102.767
Activos Financieros de Inversión			
Inversiones a valor razonable a través de resultados		1.000.200	1.125.701
Inversiones a valor razonable a través del otro resultado integral		1.079	951
Inversiones a costo amortizado		370.750	94.756
Total Activos Financieros de Inversión	7	1.372.029	1.221.408
Inversiones en asociadas	8	31.001	28.636
Otros activos financieros		9.266	22.298
Cuentas por cobrar de seguros neto	9	175.849	144.901
Cuentas por cobrar de reaseguros neto	10	138.427	117.604
Otras cuentas por cobrar	11	23.052	18.023
Propiedad y equipo, neto de depreciación	12	29.564	34.120
Propiedad de Inversión, neto de depreciación	13	606	617
Propiedad planta y equipo por derechos de uso	14	935	1.730
Activos intangibles	15	10.758	9.357
Activo por impuestos Corrientes	16	18.301	29.575
Activo por impuesto Diferido	16	96.657	38.922
Otros activos no financieros	17	108.384	96.298
Total Activos		$ 2.116.781	$ 1.866.256
Pasivos y Patrimonio de los Accionistas			
Pasivos			
Reservas Técnicas de Seguros	18	$ 1.515.290	$ 1.241.315
Cuentas por pagar actividad aseguradora	19	80.560	47.814
Otras cuentas por pagar	20	52.396	49.181
Otros pasivos financieros	21	25.313	11.591
Pasivo por impuestos Corrientes	16	28.117	15.712
Provisiones por beneficios a empleados		2.677	2.702
Otros pasivos no financieros	22	84.909	80.864
Pasivos por arrendamientos	14	1.069	1.921
Total Pasivos		1.790.331	1.451.100
Patrimonio de los accionistas			
Capital Suscrito y Pagado	23	118.113	118.113
Reservas patrimonio	24	186.922	303.397
Prima en colocación de acciones		51.982	51.982
Adopción por primera vez de las NIIF		73.648	74.709
Resultado del ejercicio		(87.756)	(116.476)
Otro resultado Integral		(16.459)	(16.569)
Total Patrimonio de los Accionistas		326.450	415.156
Total pasivo y patrimonio de los accionistas		$ 2.116.781	$ 1.866.256

Véase las notas adjuntas que forman parte integral de estos estados financieros.

Imagen de uso libre obtenida de https://www.libertyseguros.co/sites/default/files/2023-04/AS-3818-23%20Liberty%20Seguros%20S.A%20EF%20al%2031%20dic%202022%2031MAR_.pdf

Criterio de éxito de negocio

El criterio principal de éxito será obtener resultados razonables de reserva IBNR comparando contra las metodologías tradicionales que son aceptadas por el regulador y la compañía. Este objetivo se verá acompañado por uno adicional que determine que se desea reducir la reserva constituida, lo cual permita incrementar las ganancias de los accionistas. En términos de modelación, lo razonable del ejercicio se determinará con la ayuda de algunas métricas de evaluación como el error cuadrático medio y el error ajustado porcentual medio. Estas metodologías son ampliamente aceptadas en la práctica, y de ser exitosas permiten concluir que el modelo nuevo es razonable en el sentido comercial del negocio.

Evaluación de la situación

Inventario de recursos

Tarea 1: Investigación de recursos de Hardware

Se estima que el software a utilizar no es muy complejo, basta con aplicaciones de Machine Learning de Python, donde las respectivas librerías son de libre acceso al público, adicionalmente, se requerirá un manejo de grandes volúmenes de información, lo que requiere inversiones en sistemas de almacenamiento, sin embargo, no son costos que no estén dentro del presupuesto de la compañía.

Tarea 2: Identificación de fuentes de datos

La compañía aseguradora cuenta con fuentes de información relativas a la ocurrencia de siniestros, es decir, se cuenta con distinción por ramos, fechas de pago, fechas de reporte, montos pagados, montos asignados al reasegurador. También hay información relevante con respecto al cobro de las primas, como las fechas de pago, montos correspondientes al intermediario de seguros, vigencia de las polizas, entre otros. La aseguradora también cuenta con información estándar como rendimientos de las inversiones, gastos incurridos por la compañía, entre otros. Todas las fuentes son conocidas por el área de Actuaría de la compañía y se supondrá que la versión final de los datos para la modelación es la del ramo de salud.

En esta compañía, la información se almacena en una bodega de datos con históricos “útiles” que se ha aprovisionado cumplidamente desde el año 1987 hasta 1997 con base en información comercial y de las áreas de la compañía, la calidad de la información es en términos generales buena, pero hay ciertos montos en cero que parecen indicar datos faltantes en vez de pagos en cero. La compañía tiene los recursos necesarios para comprar fuentes de información que permitan llevar a cabo el proyecto de minería de datos.

Tarea 3: Identificación de recursos personales

La compañía cuenta con limitados expertos de modelación, si bien existen los suficientes actuarios que conocen la información, no se cuenta con muchos científicos de datos relevantes en la etapa de modelación. Los administradores de las bases de datos tienen gran conocimiento de la información y tienen disponibilidad para ayudar al proyecto. De cara a la implementación, es necesario contratar

los servicios de una firma consultora con el conocimiento necesario para llevar a cabo el proceso. El objetivo final es que los consultores capaciten a los actuarios y personal pertinente para lograr la operación de los modelos.

Requerimientos, supuestos y restricciones

Tarea 1: Determinación de requerimientos

No existen problemas de seguridad y todos están alineados con los requisitos del proyecto, sin embargo, es muy importante que la solución final cumpla con los requisitos de los entes reguladores.

Tarea 2: Aclaración de supuestos

Las regulaciones que dispongan los entes encargados juegan un papel importante en el proyecto y se deben contemplar en el desarrollo del proyecto. Como se comentó, es necesario incurrir en costos adicionales para contar con los especialistas en la modelación de la solución y en minería de datos. La dirección desea tomar un rol ejecutivo, de manera que está interesada en que los resultados representen un beneficio para la compañía y cumplan los resultados del regulador, no tienen interés en los tecnicismos del asunto.

Tarea 3: Verificación de restricciones

Las restricciones económicas se cumplen, el presupuesto es amplio y permite un nivel de maniobra aceptable.

Riesgos y contingencias

Principalmente hay riesgos asociados a los requisitos cambiantes del regulador, financieros, si la tarea de modelación implica aumentos en costos por la necesidad de un especialista de datos se afectaría el presupuesto. La respuesta para el primer riesgo incluye desarrollos adicionales en la etapa post-proyecto para los que se está planeando, en el marco colombiano se podría pensar actualmente en la implementación de metodologías contables que afecten el cálculo de las reservas, como es el caso de NIIF 17. El segundo riesgo se mitigaría con la contratación adecuada de firmas de consultoría que ayuden a complementar la labor en secciones específicas siempre y cuando no se exceda el presupuesto establecido.

Nota: Los análisis de costo beneficio y terminología o se incluyen en este documento

Determinación de metas de la minería de datos

Metas de la minería de datos

Como se ha podido evidenciar, el tipo de problema de minería de datos es predictivo, pues se desea estimar el monto de reserva IBNR a partir de ciertos insumos históricos del ramo de salud. Se desea obtener un modelo que obtenga resultados consistentes con el negocio y ayude a reducir la intervención humana y los errores en este proceso. Cómo se ha venido comentando, el objetivo

final es lograr una disminución razonable en los montos estimados de reserva IBNR de manera mensual, lo cuál se traduce en mayores ganancias para la compañía.

Criterios de éxito de la minería de datos

Los criterios iniciales a tener en cuenta, son los propios de los modelos estadísticos como pruebas de significancia estadística, de residuos, entre otros. Adicionalmente, las métricas de error cuadrático medio y error porcentual ajustado medio ayudarán a determinar el éxito de las predicciones de los modelos. Los criterios de éxito comercial consisten en la disminución de las cantidades de dinero a reservar por concepto de siniestros incurridos pero no reportados (IBNR). Adicionalmente, se establecerá una métrica de éxito que deben cumplir las estimaciones, donde los valores deben estar razonablemente cerca de los estimados de manera determinística.

Plan de proyecto

Fase	Tiempo	Recursos	Riesgos
Entendimiento del negocio	15 días	Todas las áreas	Cambios dispuestos por el regulador
Entendimiento de los datos	21 días	Todas las áreas	Problemas de calidad de la información y aprovisionamiento
Preparación de los datos	60 días	Actuaría, Siniestros, I.T, científico de datos	Falta de científicos de datos que satisfagan las necesidades del proyecto
Modelación	25 días	Actuaría, Siniestros, I.T, científico de datos	Falta de científicos de datos que satisfagan las necesidades del proyecto
Evaluación	15 días	Todas las áreas	Cambios dispuestos por el regulador
Lanzamiento	15 días	Actuaría, Siniestros, I.T, científico de datos	Cambios dispuestos por el regulador

Entendimiento de los datos

Recolección inicial de los datos

La fuente de los datos para este proyecto proviene del histórico de siniestros de distintas compañías para el ramo de salud, se puede encontrar en el siguiente enlace: https://www.casact.org/sites/default/files/2021-04/medmal_pos.csv. Esta fuente de información contiene información sobre la fecha de ocurrencia de siniestros, fecha de desarrollo de los pagos, montos pagados, información de la prima ganada y monto de reserva contabilizada para un conjunto de compañías. Se ha determinado que con los datos existentes es suficiente para seguir con la parte restante del proyecto. Y de cara al despliegue de los modelos finales, se supondrá que los actuarios obtienen la información de los siniestros de una bodega de datos y realizan consultas SQL desde Excel para obtener los datos de los siniestros.

Reporte de descripción de los datos

Formato: Los datos están en formato CSV y son de acceso libre.

Descripción de variables: A continuación, se muestra la composición por variables del conjunto de datos.

GRCODE: Código NAIC de la empresa (incluidos grupos de aseguradoras y aseguradoras individuales)

GRNAME: Nombre de la empresa NAIC (incluidos grupos de aseguradoras y aseguradoras individuales)

AccidentYear: Año del accidente (1988-1997)

DevelopmentYear: Año de desarrollo (1988 a 1997)

DevelopmentLag: Año de desarrollo (AY-1987 + DY-1987 - 1)

IncurLoss_F2: Siniestros ocurridos y gastos imputados declarados al final del año

CumPaidLoss_F2: Pérdidas pagadas acumuladas y gastos asignados al final del año

BulkLoss_F2: Reservas a granel e IBNR sobre pérdidas netas y gastos de defensa y contención de costes comunicados al final del año

PostedReserve97_F2: Reservas contabilizadas en el año 1997 tomadas del Anexo de Suscripción e Inversión - Parte 2A, incluidas las pérdidas netas no pagadas y los gastos de ajuste de pérdidas no pagados

EarnedPremDIR_F2: Primas devengadas en el ejercicio - directas y asumidas

EarnedPremCeded_F2: Primas devengadas en el ejercicio - cedidas

EarnedPremNet_F2: Primas devengadas en el ejercicio - netas

Individual 1 indica una entidad individual, 0 indica un asegurador de grupo

Reporte de campos relevantes: Las columnas de la base de datos que parecen de mayor relevancia para el ejercicio son: Año del siniestro (AccidentYear), Año de desarrollo (DevelopmentYear, Pérdida

incurrida (IncurLoss_C) y Perdida pagada acumulada (CumPaidLoss_C), ya que por lo general se suelen usar para la estimación de la reserva IBNR usando el método de Chain-Ladder, pues permiten generar el triángulo run off de pagos acumulados. Asimismo, hay un campo identificador de compañía, lo cual permite ver que se puede calcular la reserva para las distintas compañías (GRNAME).

Esta información parece la más relevante, sin embargo, es posible que las columnas restantes puedan complementar el análisis, de manera que, no se descartan variables por el momento. Por ejemplo, las columnas asociadas a las primas se usan por lo general para usar un método diferente a Chain-Ladder para la proyección de pagos, estos son los métodos Loss Ratio y Bornhuetter-Ferguson que no aplican por el momento. Asimismo, vale mencionar que este ejercicio no requiere la presencia de variables tipo "Primary Key"

Análisis descriptivo de los datos: El análisis exploratorio de los datos se hizo por medio de RStudio y muestra que el conjunto de datos de siniestros para las compañías cuenta con 3400 observaciones y 13 variables. Las variables presentan los siguientes formatos:

Variable	Formato
GRCODE_format	numeric
GRNAME_format	character
AccidentYear_format	numeric
DevelopmentYear_format	numeric
DevelopmentLag_format	numeric
IncurLoss_C_format	numeric
CumPaidLoss_C_format	numeric
BulkLoss_C_format	numeric
EarnedPremDIR_C_format	numeric
EarnedPremCeded_C_format	numeric
EarnedPremNet_C_format	numeric
Single_format	numeric
PostedReserve97_C_format	numeric

Las compañías en cuestión son 34, acá se muestran 10:

Scpie Indemnity Co
Promutual Grp
Great Amer Grp
Nationwide Grp
Markel Corp Grp
Overseas Partners Us Reins Co
Eastern Dentists Ins Co RRG
American Assoc Of Othodontists RRG
Controlled Risk Ins Co Of VT Inc
Texas Medical Ins Co

El conjunto de datos no presenta codificaciones especiales que interfieran con el análisis, aunque se presenta una segmentación por grupos que puede permitir estimar la reserva IBNR y aplicar el método Chain-Ladder.

A continuación, se muestra un análisis de estadísticas descriptivas del conjunto de datos por variable:

```
     GRCODE         GRNAME          AccidentYear  DevelopmentYear
 Min.   :  669   Length:3400        Min.   :1988   Min.   :1988
 1st Qu.:10341   Class :character   1st Qu.:1990   1st Qu.:1994
 Median :19764   Mode  :character   Median :1992   Median :1997
 Mean   :22810                      Mean   :1992   Mean   :1997
 3rd Qu.:36234                      3rd Qu.:1995   3rd Qu.:2000
 Max.   :44504                      Max.   :1997   Max.   :2006
 DevelopmentLag  IncurLoss_F2    CumPaidLoss_F2    BulkLoss_F2
 Min.   : 1.0   Min.   :   -17   Min.   : -1190   Min.   :-32101.0
 1st Qu.: 3.0   1st Qu.:     0   1st Qu.:     0   1st Qu.:     0.0
 Median : 5.5   Median :   645   Median :   187   Median :     0.0
 Mean   : 5.5   Mean   : 11609   Mean   :  6706   Mean   :  1095.8
 3rd Qu.: 8.0   3rd Qu.:  9050   3rd Qu.:  4386   3rd Qu.:   107.2
 Max.   :10.0   Max.   :179425   Max.   :113189   Max.   :104402.0
 EarnedPremDIR_F2 EarnedPremCeded_F2 EarnedPremNet_F2     Single
 Min.   :  -781   Min.   :-6214.0    Min.   :  -728   Min.   :0.0000
 1st Qu.:     0   1st Qu.:    0.0    1st Qu.:     0   1st Qu.:1.0000
 Median :  1500   Median :  106.5    Median :  1302   Median :1.0000
 Mean   : 14112   Mean   : 1803.5    Mean   : 12308   Mean   :0.8529
 3rd Qu.: 18094   3rd Qu.: 1473.5    3rd Qu.: 13490   3rd Qu.:1.0000
 Max.   :131948   Max.   :25553.0    Max.   :135318   Max.   :1.0000
 PostedReserve97_F2
 Min.   :     0
 1st Qu.:   629
 Median :  5875
 Mean   : 57066
 3rd Qu.: 46762
 Max.   :702246
```

Este análisis descriptivo permite entender que los siniestros más antiguos ocurren en 1988, y el desarrollo máximo va hasta 2006, para propósitos de este reporte, se asumirá que los montos de pagos conocidos van hasta 1997 por lo tanto, se buscará pronosticar desarrollos mayores a 1997 y hasta 2006.

Si se ven las variables de interés, es decir, IncurLoss_F2 y CumPaidLoss_F2, se pueden evidenciar algunos problemas de calidad de la información iniciales, aunque este punto se abordará más en detalle en el informe de calidad, se puede evidenciar la presencia de valores atípicos al ver la diferencia entre los montos máximos de cada variable y la media. Asimismo, es posible ver valores negativos, los cuáles pueden tener un tratamiento especial, como se verá en el reporte de calidad más adelante.

En conclusión, el reporte permite ver que los datos desde un enfoque descriptivo son aptos para el cálculo de la reserva IBNR usando el método Chain-Ladder.

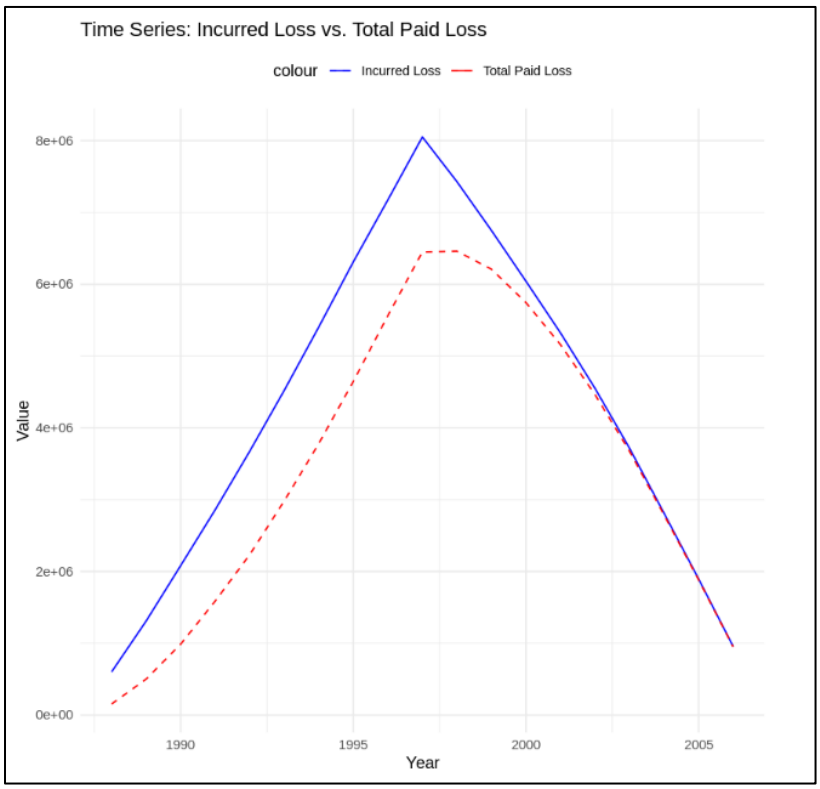
Reporte de exploración de los datos

Dada la naturaleza de los datos, y del ejercicio de este proyecto, no se requiere un ejercicio con gran robustez, ya que básicamente las variables de mayor interés se comportan como series de tiempo con temporalidad para la ocurrencia de los siniestros y el desarrollo de los pagos.

Análisis exploratorio: Las variables de interés permiten ver la magnitud que tendrá la reserva a estimar. A continuación, se muestra un resumen de los montos incurridos y pagados por año de desarrollo, lo cual permite tener un punto de partida para la construcción de los triángulos run-off

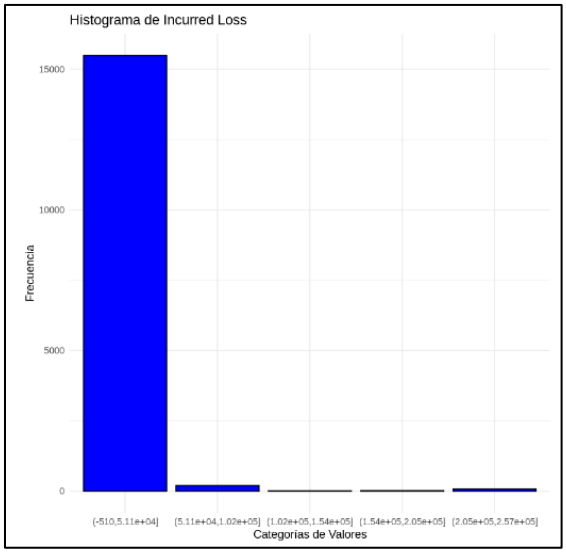
DevelopmentYear	Incurred_Loss	Cummulate_Paid_Loss
<dbl>	<dbl>	<dbl>
1988	380154	5934
1989	708400	56691
1990	1109363	158905
1991	1536151	294851
1992	1976508	519149
1993	2339316	764008
1994	2713949	1048700
1995	3142552	1364851
1996	3585381	1705949
1997	3937189	2084334

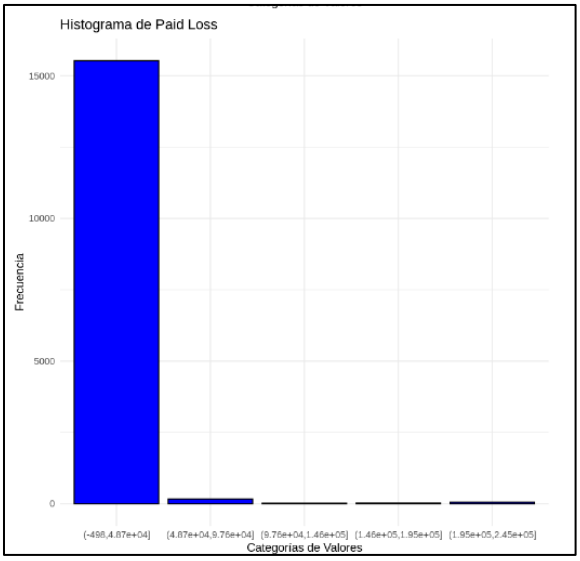
El cuadro permite ver que los años de desarrollo se concentran hacia los años de 1994 hasta 1997, lo cual indica que los pagos por lo general no son inmediatos. Esto se ve mejor con una gráfica de la perdida incurrida y la perdida pagada con respecto al año de desarrollo, la cual permite ver que los mayores montos incurridos se tienen en 1996, sin embargo, no es un monto mucho mayor en comparación con los otros años, a excepción de los años de 1988 a 1992, donde los montos incurridos son menores, esto se puede explicar porque en general los pagos no son inmediatos:



Una conclusión que se puede obtener de estos gráficos es que en general, las proyecciones se basarán en la información disponible, por lo tanto, aquellos años de desarrollo con mayores montos concentrarán la mayor cantidad de proyección, es decir, para estos segmentos en particular se espera aprovisionar mayor cantidad de reserva.

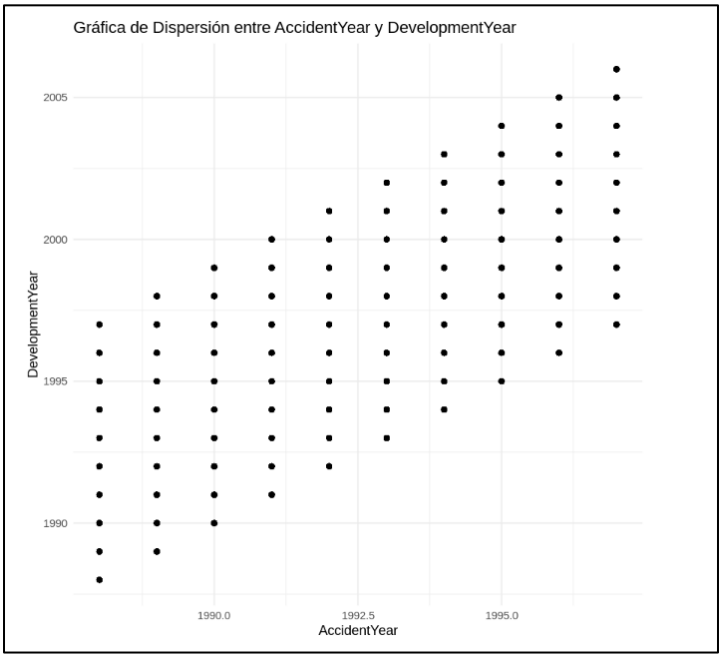
Ahora bien, un histograma permite entender la composición de la información:





Las gráficas anteriores permiten concluir que la distribución de las observaciones se concentra en aquellos siniestros de monto bajo, por lo tanto, los otros se consideran atípicos.

Otra gráfica de interés muestra la relación entre los históricos de ocurrencia y desarrollo para las compañías, esto se puede ver de la siguiente manera:



Esto permite ver que el desarrollo de los siniestros tiene una duración similar independiente del año de ocurrencia. Esto parece una característica deseable, ya que se puede pensar en una mejor proyección basada en el método de Chain-Ladder cuando el desarrollo de los siniestros es similar.

Reporte de calidad de los datos

Dada la fuente de origen de los datos, se sabe que la base de datos contó con un proceso de limpieza y verificación de la calidad previo.

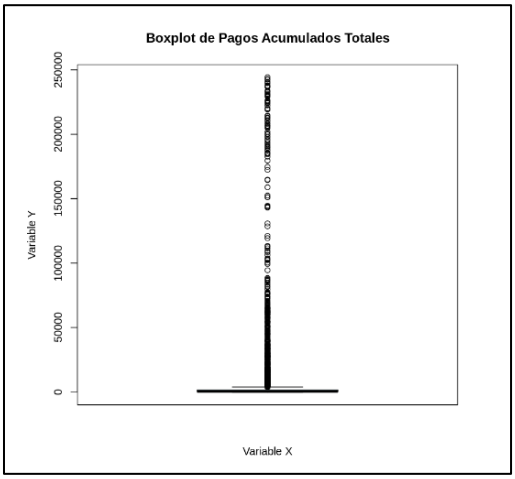
Análisis de valores nulos: El análisis de valores nulos de los datos permite ver que la calidad de los datos parece estar garantizada en ese sentido, dado que no se evidencian valores nulos:

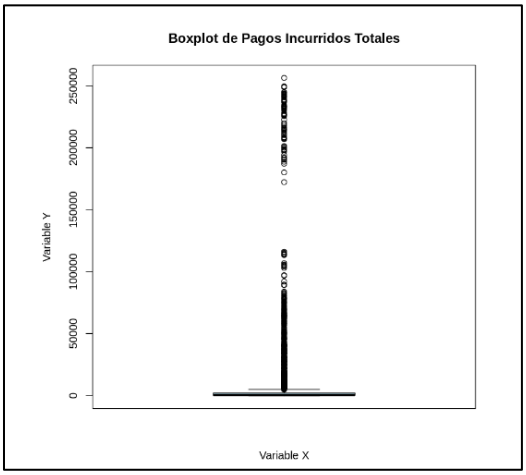
	Variable	ValoresNulos
GRCODE	GRCODE	0
GRNAME	GRNAME	0
AccidentYear	AccidentYear	0
DevelopmentYear	DevelopmentYear	0
DevelopmentLag	DevelopmentLag	0
IncurLoss_F2	IncurLoss_F2	0
CumPaidLoss_F2	CumPaidLoss_F2	0
BulkLoss_F2	BulkLoss_F2	0
EarnedPremDIR_F2	EarnedPremDIR_F2	0
EarnedPremCeded_F2	EarnedPremCeded_F2	0
EarnedPremNet_F2	EarnedPremNet_F2	0
Single	Single	0
PostedReserve97_F2	PostedReserve97_F2	0

Análisis de errores en los datos o atípicos: Los errores tipográficos aplicarían a las variables numéricas, ya que las categóricas no son relevantes dentro del análisis, pues indican una segmentación por grupos. En ese sentido, interesa buscar aquellos valores que sean negativos o que sean muy altos, dado que indicarían siniestros atípicos, los cuales podrían ser excluidos al momento de estimar la reserva IBNR. Para ello vale la pena revisar nuevamente el cuadro de estadísticas descriptivas:

```
     GRCODE         GRNAME           AccidentYear  DevelopmentYear
 Min.   :  669   Length:3400        Min.   :1988   Min.   :1988
 1st Qu.:10341   Class :character   1st Qu.:1990   1st Qu.:1994
 Median :19764   Mode  :character   Median :1992   Median :1997
 Mean   :22810                      Mean   :1992   Mean   :1997
 3rd Qu.:36234                      3rd Qu.:1995   3rd Qu.:2000
 Max.   :44504                      Max.   :1997   Max.   :2006
 DevelopmentLag  IncurLoss_F2    CumPaidLoss_F2    BulkLoss_F2
 Min.   : 1.0   Min.   :   -17   Min.   : -1190   Min.   :-32101.0
 1st Qu.: 3.0   1st Qu.:     0   1st Qu.:     0   1st Qu.:     0.0
 Median : 5.5   Median :   645   Median :   187   Median :     0.0
 Mean   : 5.5   Mean   : 11609   Mean   :  6706   Mean   :  1095.8
 3rd Qu.: 8.0   3rd Qu.:  9050   3rd Qu.:  4386   3rd Qu.:   107.2
 Max.   :10.0   Max.   :179425   Max.   :113189   Max.   :104402.0
 EarnedPremDIR_F2 EarnedPremCeded_F2 EarnedPremNet_F2     Single
 Min.   :  -781   Min.   :-6214.0    Min.   :  -728   Min.   :0.0000
 1st Qu.:     0   1st Qu.:    0.0    1st Qu.:     0   1st Qu.:1.0000
 Median :  1500   Median :  106.5    Median :  1302   Median :1.0000
 Mean   : 14112   Mean   : 1803.5    Mean   : 12308   Mean   :0.8529
 3rd Qu.: 18094   3rd Qu.: 1473.5    3rd Qu.: 13490   3rd Qu.:1.0000
 Max.   :131948   Max.   :25553.0    Max.   :135318   Max.   :1.0000
 PostedReserve97_F2
 Min.   :     0
 1st Qu.:   629
 Median :  5875
 Mean   : 57066
 3rd Qu.: 46762
 Max.   :702246
```

Esto permite ver la presencia de valores negativos y de valores lejanos a la media de las variables, por lo tanto, valdría la pena considerar excluir estos siniestros en la fase de preparación de los datos. El punto anterior se verifica con los siguientes boxplots:





En el caso de siniestros por montos negativos se puede pensar en casos específicos como cuando tras un siniestro, el beneficiario paga un deducible o copago mayor al monto del sinestro. Por otro lado, un siniestro atípico puede presentar un periodo de desarrollo muy largo que influya negativamente sobre el patrón de pagos obtenido con el triángulo run-off. Es por eso que se considera pertinente tratar de excluir los siniestros atípicos en las partes siguientes del proyecto.

Preparación de los datos

Es un proceso que consta de los siguientes pasos:

- Selección de los datos
- Limpieza de los datos
- Construcción de los datos
- Integración de los datos
- Formateo de los datos

Se evidenciará que no todos los pasos aplican a este proyecto en particular, sin embargo, se enunciaran las razones del porqué se toman las decisiones de no entrar a alguna de las fases. Asimismo, se detallarán los pasos que apliquen cuando sea pertinente.

Selección de los datos

El punto de partida es el conjunto de datos que se encuentra en el siguiente link sobre el ramo de salud y que se explicó en la sección anterior: https://www.casact.org/sites/default/files/2021-04/medmal_pos.csv. Los datos seleccionados provienen del histórico de siniestros de para el ramo de salud de distintas compañías. Esta fuente de información contiene información sobre la fecha de ocurrencia de siniestros, fecha de desarrollo de los pagos, montos pagados, información de la prima ganada por la compañía bajo categorización de neto, bruto y cedido. Adicionalmente hay un monto de reserva contabilizada para cada conjunto de compañías. Para abordar el tema de selección de los datos desde la perspectiva de selección de ítems y atributos, se parte de algunos de los descubrimientos de los datos mencionados en los reportes anteriores.

Selección de ítems

Con respecto a este rubro, un aspecto está relacionado con que hay un supuesto de información importante sobre la disponibilidad de información. A pesar de que se conoce la historicidad de los siniestros hasta 2007 aproximadamente, se asumirá que la fecha de pagos máxima conocida es de 1997. De esta forma, se puede afirmar que se seleccionaran aquellos registros de información donde el año de accidente sea menor o igual a 1997. De igual manera, se seleccionarán aquellos años de desarrollo hasta 1997. Por lo tanto, de los accidentes ocurridos en 1997, sólo se conocen los pagos ocurridos durante el mismo año, en términos prácticos equivale a afirmar que el proyecto se está desarrollando en 1997 con el fin de conocer el dinero que debe reservar la compañía aseguradora por concepto de siniestros incurridos, pero no reportados.

Selección de características

Con respecto a este rubro, el proceso es más directo dado que para el cálculo de la reserva IBNR sólo se requieren tres insumos en términos generales. Se requiere una fecha de ocurrencia de siniestro, una fecha de movimiento del siniestro y un monto pagado por el siniestro. En las etapas anteriores se determinó que la base de datos tiene estas variables representadas por AccidentYear, DevelopmentYear, DevelpmentLag y CumPaidLoss, también es deseable una variable para identificar las compañías. Adicionalmente hay una variable que puede complementar el análisis que es la de IncurLoss, por lo que se mantiene. La estructura de las características de la base final sería la siguiente:

GRCODE	GRNAME	AccidentYear	DevelopmentYear	CumPaidLoss_F2	IncurLoss_F2	DevelopmentLag
<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
669	Scpie Indemnity Co	1988	1988	55091.33	121905	1
669	Scpie Indemnity Co	1988	1989	55091.33	112211	2
669	Scpie Indemnity Co	1988	1990	55091.33	103226	3
669	Scpie Indemnity Co	1988	1991	55091.33	99599	4
669	Scpie Indemnity Co	1988	1992	55091.33	96006	5
669	Scpie Indemnity Co	1988	1993	55091.33	90487	6
669	Scpie Indemnity Co	1988	1994	55091.33	82640	7
669	Scpie Indemnity Co	1988	1995	55091.33	80406	8
669	Scpie Indemnity Co	1988	1996	55091.33	78920	9
669	Scpie Indemnity Co	1988	1997	55091.33	78511	10
669	Scpie Indemnity Co	1989	1989	55091.33	122679	1
669	Scpie Indemnity Co	1989	1990	55091.33	113165	2

Limpieza de los datos

Reporte de limpieza de datos

Valores negativos: Conviene eliminar aquellos montos negativos. La justificación para esto es que por lo general representan casos específicos como cuando tras un siniestro, el beneficiario paga un deducible o copago mayor al monto del sinestro, lo cual en la práctica representaría un comportamiento atípico. En este proceso se usó un filtro en la base de datos para excluir esos montos. El mayor detalle se encuentra en el cuaderno con el procedimiento de tratamiento de datos.

Valores atípicos alejados de la media: Complementando este punto, también se opta por reemplazar por el valor promedio aquellos siniestros cuya severidad excede considerablemente las medias de cada compañía, esto es deseable dado el impacto sobre las proyecciones deterministas de siniestros con grandes diferencias sobre los valores medios. A continuación, se muestra un extracto de la tabla de medias por compañía.

GRNAME	Promedio_CumPaidLoss_F2
<chr>	<dbl>
American Assoc Of Othodontists RRG	4.109091e+00
California Healthcare Ins Co Inc	4.205636e+02
Campmed Cas & Ind Co Inc MD	1.963636e+00
Clinic Mut Ins Co RRG	9.845818e+02
Community Blood Cntr Exch RRG	8.018182e+00
Controlled Risk Ins Co Of VT Inc	1.109091e+02
Dentists Ins Co	5.681927e+03
Eastern Dentists Ins Co RRG	4.709091e+00

La idea es reemplazar por la media aquellos outliers de los datos para cada una de las compañías. Esto se hizo usando el método z-score que mide cuántas desviaciones estándar un valor está por encima o por debajo de la media. De manera que, valores z-score muy altos o muy bajos pueden considerarse atípicos y se reemplazan por la media de cada una de las compañías (el mayor detalle se encuentra en el cuaderno de este proyecto referenciado en GitHub). Estos atípicos dependen de un umbral definido en el desarrollo práctico de este trabajo. Con esto concluye el reporte de limpieza de los datos. Dado el origen de los datos, se pudo evidenciar que el proceso no constó de muchas etapas.

Fases de construcción de nuevos datos, integración y formateo de nuevos datos

Para la parte de limpieza de datos se generaron algunas variables auxiliares para ayudar a la detección de outliers, sin embargo, de cara al resultado final no se requieren nuevos datos construidos. De igual manera, se usaron joins para la integración de una tabla auxiliar y la principal con el fin de tratar los outliers, sin embargo, no hubo integraciones de cara a la preparación de los datos para el proyecto. Finalmente, tampoco hubo aplicaciones de nuevos formatos a los datos.

Modelación

Selección de técnicas de modelación

Cómo se comentó, el propósito de este ejercicio es implementar un modelo en la compañía aseguradora que ayude a mejorar la estimación de la reserva IBNR, dado que el método habitual de Chain Ladder puede derivar en ciertas imprecisiones a la hora de la estimación de las reservas. La perspectiva de los accionistas consiste en lograr reducir los montos de dinero reservados para el ramo de salud y disminuir el impacto en fondos propios en cada mes, lo que ayuda a generar ganancias para la compañía. En esta sub sección se pretenden mostrar el proceso detrás de las técnicas de modelación planteadas para la realización del ejercicio.

Escogiendo las técnicas de modelación apropiadas

Para el problema de estimación de reservas IBNR, se suelen usar métodos determinísticos como Chain Ladder, en este ejercicio se pretende implementar dicha técnica de modelación para comparar resultados. Se sabe que esta técnica no requiere una separación entre datos de prueba y entrenamiento, sin embargo, para validaciones posteriores se hará la segmentación mencionada, también, dada la revisión de la información en las secciones anteriores, se sabe que la información a disposición es más que suficiente y cuenta con la calidad necesaria para llevar a cabo el desarrollo del modelo.

Por otro lado, se probarán modelos con un enfoque diferente, uno de ellos es un modelo de regresión log lineal basado en el siguiente artículo sobre métodos estadísticos para la técnica de Chain Ladder. Este se encuentra en el siguiente enlace, y la teoría detrás de su implementación en este proyecto de muestra en la sección correspondiente: https://www.casact.org/sites/default/files/2021-02/pubs_forum_94spforum_94spf393.pdf. Al igual que en la técnica anterior, no existen problemas de calidad de información, sin embargo, hay distinción entre datos de prueba y de entrenamiento. Finalmente se usan dos métodos muy comunes en Machine Learning, que son la regresión de Ridge y de Lasso, los detalles de la implementación se muestran con detalle más adelante. Para estas dos técnicas también se cuenta con la información necesaria y se requiere de la partición en datos de prueba y de entrenamiento. A continuación, se muestran los modelos candidatos para la resolución del problema.

Modelo determinístico

En esta fase del proyecto se lleva a cabo la estimación de las reservas. Como se mencionó en las fases anteriores de este documento, en la práctica es muy utilizado el método Chain Ladder para la estimación de la reserva IBNR. Un punto de partida adecuado tiene que ver con la construcción de triángulos de desarrollo. Este puede ser visto como una matriz con dos ejes, uno de años de ocurrencia y otro de años de desarrollo, o periodos de desarrollo. Es relativamente sencillo llevar a cabo la construcción de esta matriz y no requiere de grandes recursos informáticos para su completitud. En la práctica existen dos tipos de triángulos, un formato incremental puede tener la variación en las cantidades incurridas, donde estas son la suma de los siniestros pagados acumulados y las estimaciones individuales. La estimación individual consiste en el ejercicio realizado por el tramitador de los siniestros sobre la cantidad pendiente de pago de un siniestro. Por otro lado, el formato acumulado puede mostrar la cantidad pagada acumulada, el número total

de siniestros en un año determinado de ocurrencia y las cantidades incurridas de los siniestros. Los triángulos de desarrollo tienen la siguiente estructura:

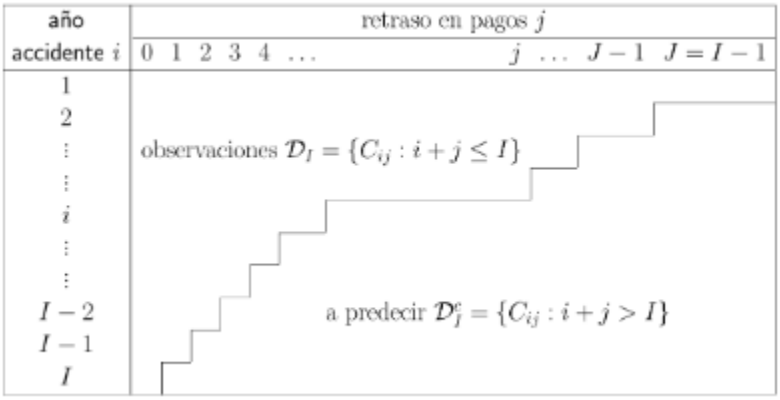


Imagen obtenida de: https://ewfrees.github.io/Loss-Data-Analytics-Spanish/C-LossReserves.html

Desde una perspectiva más teórica, la variable aleatoria X_{ij} denotará las reclamaciones incrementales pagadas en el periodo de desarrollo j de los siniestros del año de accidente i. De manera que, la variable aleatoria representa la cantidad total pagada en el año de desarrollo j por todos los siniestros ocurridos en i. Particularmente, interesa la perspectiva acumulada, donde se tendrá una variable aleatoria C_{ij} que es la cantidad acumulada pagada hasta el año de desarrollo j por los accidentes ocurridos en el año i.

Este método trae a la luz unos factores de desarrollo F_{ij} que se pueden representar mediante la siguiente relación:

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

En ese orden de ideas, el factor de desarrollo indica como la cantidad acumulada en el año de desarrollo, crece hasta la cantidad acumulada en el año j+1. Estos factores de desarrollo son vitales para la predicción determinística de los pagos desconocidos y por la tanto de la reserva IBNR. Vale la pena recordar que el IBNR se calcula como la última columna de predicción (ultimate) menos la diagonal principal, que en términos prácticos equivale a los pagos acumulados conocidos a la fecha de análisis. La notación sería la siguiente:

$$\begin{aligned} \hat{C}_{ij}^{CL} &= C_{i,I-i} \cdot \prod_{l=I-i}^{j-1} \hat{f}_l^{CL} \\ \hat{f}_j^{CL} &= \frac{\sum_{i=1}^{I-j-1} C_{i,j+1}}{\sum_{i=1}^{I-j-1} C_{ij}}, \end{aligned}$$

Con base en este marco teórico, se procede replicando el proceso en una clase de Python que permitirá determinar estos pagos desconocidos, cálculo de factores de desarrollo y comparación con datos reales (recordar que se asume que no se conocen los pagos por siniestros a partir de 1997). El código se encuentra con mayor detalle en el Notebook adjunto de apoyo. Las validaciones para este ejercicio no serán muy rigurosas, dado que el foco está centrado en la estimación de las reservas con métodos usados en Machine Learning, como la regresión de Ridge, regresión de Lasso y una regresión con modelo log lineales. A continuación, se revisa el desarrollo detrás de este último modelo.

Modelo multiplicativo

El punto de partida en términos de los datos es un triángulo de pagos, tal y como en el modelo determinístico. El razonamiento detrás de este método también es similar al del determinístico, es decir se piensa en una variable aleatoria C_{ij} que permite representar la cantidad acumulada pagada hasta el año de desarrollo j por los accidentes ocurridos en el año i. La diferencia radica en que se pensará en la aparición de unos efectos fila y unos efectos columna, esto en conjunto con la variable aleatoria de reclamaciones no acumuladas permite establecer la siguiente relación

$$E(Z_{ij}) = U_i S_j$$

Donde U_i representa el parámetro de las filas i y es el monto total esperado por siniestros ultimate para el año e negocio i y S_j representa el parámetro de las columnas y es la proporción esperada de reclamos ultimate que ocurren en el J-esimo año de desarrollo.

El modelo multiplicativo mostrado arriba se puede transformar usando logaritmos, por lo que el resultado sería un modelo de elasticidades que puede ser estimado usando regresión lineal, esto quedaría de la siguiente manera:

$$\ln\left(E(Z_{ij})\right) = \ln(U_i) + \ln(S_j)$$

En el Notebook adjunto se ve con detalle la estimación de este modelo, así como la validación de los resultados.

Regresión de Ridge

Antes de abarcar esta parte, cabe aclarar que la regresión de Ridge es una técnica de regularización para abordar problemas de regresión cuando hay multicolinealidad entre las variables predictoras. Y se sabe que la multicolinealidad ocurre cuando algunas de las variables predictoras están altamente correlacionadas entre sí. En la regresión de Ridge, se agrega una penalización a los coeficientes de la regresión para evitar que tomen valores extremadamente grandes. Esta penalización se controla mediante un parámetro llamado α que ajusta la fuerza de la penalización. Cuanto mayor sea el valor de α, mayor será la penalización y más se reducirán los coeficientes.

La función objetivo de la regresión de Ridge es una combinación de la función de pérdida de mínimos cuadrados ordinarios y un término de penalización de norma L2 de los coeficientes. La inclusión del término de penalización de esta norma puede ayudar a suavizar los coeficientes, lo que reduce la magnitud de los valores individuales. Esto resulta beneficioso en situaciones en las que hay

multicolinealidad, ya que evita que un conjunto pequeño de variables tenga coeficientes muy grandes, lo que podría llevar a una mayor variabilidad en las predicciones.

Para implementar esto en Python existen diversas librerías como sklearn.linear_model. Posteriormente, un proceso general indicaría que es necesario hacer una preparación de los datos, lo cual consiste en asegurar que se tenga una matriz de diseño X con datos de entrada, así como la variable de respuesta Y. A continuación, se considera dividir los datos en particiones de entrenamiento y de prueba. Una vez se tiene lo anterior, el paso sería implementar el modelo con ayuda de la librería mencionada y realizar predicciones sobre los datos de prueba. Generalmente se usan métricas de evaluación como el error cuadrático medio, estas se verán en la siguiente sección.

Regresión de Lasso

El último modelo que se tendrá es uno que contemple la regresión de Lasso, la cual es otra técnica que se utiliza en problemas de regresión para abordar la multicolinealidad y prevenir el sobreajuste de modelos lineales. En este caso, la función objetivo incluye un término de penalización de norma L1 de los coeficientes, además de la función de pérdida de mínimos cuadrados ordinarios.

La adición del término de penalización tiene el efecto de forzar algunos coeficientes a ser exactamente cero, lo que puede llevar a la selección automática de características. Por lo tanto, la regresión de Lasso no solo penaliza la magnitud de los coeficientes, sino que también puede eliminar ciertos predictores, volviendo sus coeficientes cero. La regresión de Lasso es especialmente útil en situaciones donde hay un gran número de características y se desea realizar una selección automática de características para mejorar la interpretación del modelo y reducir la complejidad. La implementación en Python se puede realizar utilizando la clase Lasso de scikit-learn.

Construcción de modelos

Modelo determinístico

Para el desarrollo del modelo determinístico, se crea una clase que permite obtener distintos escenarios a partir de la información dada. Un triángulo obtenido a partir de los años de ocurrencia sin acumular, un caso acumulando los pagos correspondientes y un caso obteniendo la parte inferior del triángulo acumulado, es decir, se tienen los pagos estimados asumiendo que no se conoce la información real. Adicional a esto, se puede ver la estimación de los factores de desarrollo y la reserva IBNR.

El método implementado para tratar los triángulos dentro de esta clase itera sobre las aseguradoras únicas en la columna "GRCODE" y se apoya de la función pivot_table de Python para la construcción de una matriz completa de datos dada la información, con base en esto, es fácil transformar la información para obtener las mitades inferiores y superiores, es decir, la información conocida de siniestralidad y la esperada al hacer las proyecciones, suponiendo que sólo hay información conocida hasta 1997. Posteriormente, se calculan los factores de desarrollo mencionados en la sección anterior de este documento para las predicciones determinísticas y se obtienen los montos acumulados futuros. Para estimar la reserva interesa el monto total del último año de predicción, es decir, el ultimate, también interesa el total del monto pagado por reclamaciones en el último periodo conocido de información, eso correspondería a la diagonal principal del triángulo de

reclamaciones conocidas. Al restar los dos montos mencionados se obtiene la reserva total. La matriz de información conocida luce de la siguiente manera:

```
triangulos_resultados[codigo_aseguradora]["Triangulo_full_acumulado"]
```

DevelopmentLag	1	2	3	4	5	6	7	8	9	10
AccidentYear										
1988	121905	234116	337342	436941	532947	623434	706074	786480	865400	943911
1989	122679	235844	345881	447023	537840	619759	697250	770827	843543	915860
1990	118157	235654	352031	451926	541178	623094	702228	778561	854173	929523
1991	117981	240424	361480	475275	578105	676176	771046	862108	952601	1042946
1992	131059	261214	385409	499383	606200	705382	797970	888970	978226	1067477
1993	134700	265457	390710	505427	616721	714735	811607	907321	1003338	1099385
1994	136749	264941	386296	498173	594325	685827	776325	868195	960043	1051981
1995	140962	273367	391699	491749	580558	662918	744904	826791	908587	990369
1996	134473	263453	377098	481371	580647	678429	775711	873449	971050	1068301
1997	137944	265671	379728	486729	588872	688537	788479	888447	988037	1087415

Los resultados estimados son los siguientes:

```
triangulos_resultados[codigo_aseguradora]["Triangulo_estimado"]
```

DevelopmentLag	1	2	3	4	5	6	7	8	9	10
AccidentYear										
1988	121905	234116.00000	337342.000000	436941.000000	532947.000000	623434.000000	706074.000000	786480.000000	865400.000000	9.439110e+05
1989	122679	235844.00000	345881.000000	447023.000000	537840.000000	619759.000000	697250.000000	770827.000000	843543.000000	9.200711e+05
1990	118157	235654.00000	352031.000000	451926.000000	541178.000000	623094.000000	702228.000000	778561.000000	854369.993215	9.318803e+05
1991	117981	240424.00000	361480.000000	475275.000000	578105.000000	676176.000000	771046.000000	855386.937928	938676.522910	1.023835e+06
1992	131059	261214.00000	385409.000000	499383.000000	606200.000000	705382.000000	798084.554401	885383.107099	971593.438671	1.059739e+06
1993	134700	265457.00000	390710.000000	505427.000000	616721.000000	716316.455938	810456.035945	899107.844295	986654.562494	1.076166e+06
1994	136749	264941.00000	386296.000000	498173.000000	603790.859451	701298.202217	793464.056665	880257.195792	965968.414031	1.053603e+06
1995	140962	273367.00000	391699.000000	507257.864803	614801.810119	714087.332406	807933.957043	896309.887559	983584.166882	1.072817e+06
1996	134473	263453.00000	386575.428325	500622.739291	606759.968961	704746.798910	797365.874057	884585.813839	970718.512445	1.058784e+06
1997	137944	270785.33457	397334.464575	514555.901631	623647.106690	724361.072868	819557.890711	909205.306285	997735.220954	1.088252e+06

Modelo multiplicativo

Hace unos momentos se explicó la teoría detrás del modelo multiplicativo. En términos prácticos se busca la creación de varias funciones. Esta función toma una lista de valores, extrae dígitos de esos valores, y crea un DataFrame donde cada categoría única se representa mediante columnas binarias, indicando la presencia o ausencia de esa categoría para cada elemento en la lista original. Posteriormente, se desarrollan unas funciones para la construcción de una matriz de diseño y de una matriz con la variable de respuesta como insumos del modelo de regresión, asimismo, se desarrolla otra función para la creación de la matriz que permite organizar los montos acumulados por reclamaciones de manera triangular. Finalmente, se hace la partición de los datos en prueba y entrenamiento, el conjunto de entrenamiento tiene la siguiente forma:

	Y	b0	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_10	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_10
0	9.283591	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	10.010322	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
2	10.420792	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
3	10.634195	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
4	10.778310	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0

Con los insumos anteriores, ya es posible usar el modelo de regresión lineal y mostrar las predicciones, los resultados son los siguientes:

	y_pred
0	10.955414
1	11.008417
2	11.095115
3	10.925601
4	10.958487
5	11.045186
6	11.032278
7	11.120948
8	11.153835
9	11.240533
10	11.308487
11	11.388227
12	11.476897
13	11.509784
14	11.596482
15	11.437528
16	11.556048
17	11.635788
18	11.724458
19	11.757344
20	11.844043

Modelo de regresión Ridge y Lasso

Dada la naturaleza de las regresiones Ridge y Lasso, la modelación se lleva a cabo en el mismo sector. En términos simples, la modelación consistió en crear una clase compuesta por varias funciones. En primer lugar, se define un dataframe con algunos insumos esenciales como el origen, desarrollo, entre otros. Otra función de relevancia usa las librerías de Python mencionadas en las secciones anteriores, separa los datos de la matriz de diseño y la variable de respuesta para posteriormente aplicar las regresiones de Ridge y Lasso. Posteriormente, se tiene una función que sirve para realizar las predicciones en un conjunto de validación utilizando los modelos entrenados. También se calculan algunas métricas de evaluación de los modelos como el error cuadrático medio y el error porcentual absoluto medio. Finalmente, se realizan predicciones en el conjunto de prueba utilizando el mejor modelo producto del proceso anterior de validación y evaluación.

El resultado final será un conjunto de coeficientes para los casos de regresión lineal estándar. Al aplicar lo anterior se obtienen resultados con la siguiente estructura:

```
{'coeficientes': {'Coef_normal': array([ 0.        , -0.05366319, -0.12604845, -0.22685027, -0.28797468,
        -0.35457203, -0.39991772, -0.49217904, -0.49546519, -0.51181813,
         0.29862739,  0.39832096,  0.53248843,  0.60872173,  0.63867745,
         0.75294324,  0.81777038,  0.84446757,  0.69437918]),
  'Coef_ridge1': array([ 0.        , -0.05364488, -0.12603211, -0.22683501, -0.28796046,
        -0.35455841, -0.39990494, -0.49216591, -0.49545364, -0.51180904,
         0.29858475,  0.39827697,  0.53244259,  0.60867406,  0.63862804,
         0.75288984,  0.8177119 ,  0.84440078,  0.69430212]),
  'Coef_lasso': array([ 0.        , -0.        , -0.07173795, -0.17254086, -0.23366693,
        -0.30026734, -0.34561888, -0.4378919 , -0.4412045 , -0.45764135,
         0.24350776,  0.34320307,  0.47737344,  0.55360948,  0.5835672 ,
         0.69783375,  0.76265966,  0.7893516 ,  0.6385957 ])},
```

Para el mayor detalle técnico, se recomienda revisar simultáneamente el Notebook titulado ***Estimación reservas de siniestros V01 Parte 2*** en el siguiente enlace: https://github.com/osebasp/ML-Applications-for-Actuarial-Science

Generación de un diseño de prueba

Escribiendo un diseño de prueba

Un diseño de prueba en machine learning se refiere a la estrategia planificada para evaluar el rendimiento y la generalización de un modelo. Involucra la división adecuada del conjunto de datos en conjuntos de entrenamiento y prueba para permitir la evaluación imparcial de la capacidad predictiva del modelo en datos no vistos. Métodos comunes incluyen la validación cruzada, donde el conjunto de datos se divide en varios subconjuntos para entrenar y validar iterativamente el modelo, y la división simple de datos, donde una fracción se utiliza para el entrenamiento y otra para las pruebas. El diseño de prueba es crucial para asegurar que el modelo sea robusto y pueda generalizar bien a datos nuevos, lo que ayuda a evitar el sobreajuste y a proporcionar una estimación realista de su desempeño en situaciones del mundo real.

En este diseño de prueba se usarán los modelos de Chain Ladder y de Regresión Ridge y Lasso. Como se comentó anteriormente, para todos los casos se está haciendo la partición correspondiente entre conjunto de datos de prueba y conjunto de datos de entrenamiento. El éxito de los modelos se verificará usando la métrica de error absoluto porcentual medio dentro del marco de un ejercicio de Leave One Out Cross Validation. La idea final es comparar las métricas de error de las regresiones de Ridge y Lasso con las de Chain Ladder.

De acuerdo con la literatura moderna, Leave-One-Out Cross-Validation es una técnica de validación cruzada en la que se utiliza un solo dato como conjunto de prueba y el resto de los datos se utilizan como conjunto de entrenamiento. La manera de seguir es que en este proceso se repite para cada dato en el conjunto de datos, de manera que cada dato se usa exactamente una vez como conjunto de prueba. El método resulta particularmente útil cuando el tamaño del conjunto de datos es pequeño y se desea aprovechar al máximo la información disponible para evaluar el rendimiento del modelo

Ahora bien, La implementación del diseño de prueba para este caso requiere obtener un subconjunto de la información de 10 aseguradoras. La configuración inicial del proceso consta de obtener una lista de códigos de la aseguradora a partir del subconjunto anterior. A continuación,

sigue el proceso de validación cruzada a través de un bucle externo, donde para cada aseguradora en una lista definida se realiza un proceso de validación correspondiente, luego se separa la aseguradora seleccionado como conjunto de prueba y se crea un conjunto de entrenamiento y validación excluyendo la aseguradora mencionada. Para la siguiente etapa se realiza un bucle interno orientado a los modelos de regresión lineal. Por lo tanto, a cada aseguradora del conjunto de entrenamiento y validación se le ajusta un modelo de regresión lineal con base en los modelo enunciados en la sub sección anterior de este documento, de manera que, el modelo se entrena con el conjunto de entrenamiento y se valida con el conjunto definido de validación, posteriormente, se evalúa el modelo en el conjunto de prueba usado la métrica del error porecentual absoluto medio, lo cual permite elegir los mejores modelos y posteriormente se almacenan en un diccionario.

Tras aplicar lo anterior al conjunto de datos, se determina que el mejor modelo es la regresión de Lasso con los siguientes coeficientes y parámetros:

```
array([ 0.        , -0.        , -0.07488313, -0.16424896, -0.2327577 ,
       -0.31436918, -0.37158089, -0.49418081, -0.50544788, -0.55232072,
        0.21127437,  0.29961793,  0.26306526,  0.36098568,  0.42668986,
        0.50299664,  0.58388548,  0.63958113,  0.5572926 ])
```

Lasso

```
Lasso(alpha=0.001)
```

La comparación luce de la siguiente forma:

	Y_test	Y_ajustado modelo final	GRCODE
0	9.852194	9.944153	36676
1	9.798127	9.951559	36676
2	9.827794	9.869270	36676
3	9.740969	9.806497	36676
4	9.740969	9.862193	36676
5	9.698920	9.779904	36676
6	9.740969	9.657099	36676
7	9.758462	9.737988	36676
8	9.605755	9.793684	36676
9	9.698920	9.711395	36676
10	9.546813	9.499181	36676
11	9.711116	9.575488	36676
12	9.698920	9.656377	36676

Finalmente, al comparar la métrica MAPE de este modelo, con la del modelo Chain Ladder se puede ver que las diferencias saltan a la vista sobre la elección del modelo, y por lo tanto el modelo Lasso parece ser el indicado para mejorar la estimación de reservas y lograr reducir el impacto en los fondos propios de las compañías sin comprometer la solvencia de las mismas.

```
Métrica MAPE con el método Chain-Ladder: 7.689302482799469
Métrica MAPE con el modelo final: 1.5342914600687543
```

Evaluación

Vista general de la evaluación

Se asumirá que todos los modelos vistos son buenos, por lo tanto, se muestra una evaluación para el modelo Chain Ladder determinístico, regresión lineal del modelo multiplicativo y regresiones lineales de Ridge y Lasso.

Evaluación de los resultados

Modelo determinístico

Dado que se conoce la información real (supuesto información conocida hasta 1997) es posible obtener las siguientes diferencias para las predicciones con una métrica sencilla se variación absoluta porcentual:

DevelopmentLag	1	2	3	4	5	6	7	8	9	10
AccidentYear										
1988	0.0	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1989	0.0	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.018257
1990	0.0	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	2.477673	4.489351
1991	0.0	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	1.827310	4.269424	5.834606
1992	0.0	0.00000	0.000000	0.000000	0.000000	0.000000	3.522644	6.783678	10.196513	12.777367
1993	0.0	0.00000	0.000000	0.000000	0.000000	4.126469	8.042790	11.204964	14.356125	16.895881
1994	0.0	0.00000	0.000000	0.000000	0.010265	1.114509	3.361387	4.865206	6.981588	8.600395
1995	0.0	0.00000	0.000000	2.306833	4.918728	7.681654	10.422485	12.906863	15.641882	17.840191
1996	0.0	0.00000	2.513417	4.090412	4.752501	4.875502	3.449271	1.894176	0.276049	1.948717
1997	0.0	2.45936	5.909221	5.034678	3.708036	1.230057	1.535361	4.114687	7.057029	9.421511

Aunque las diferencias no son grandes porcentualmente, en términos de dinero es dinero que puede ser reasignado como un pasivo para la compañía, lo cual permite pensar en un mayor impacto en fondos propios y una menor ganancia para los accionistas de la compañía.

Modelo de regresión bases modelo multiplicativo

La evaluación del desempeño para el modelo multiplicativo se hace por error cuadrático medio, comparando la data de prueba y la data de predicción mostrada en la sección anterior. El error cuadrático medio se calcula de manera habitual y es de 4.47%. Se implementa de la siguiente manera:

```
mse = mean_squared_error(y_prueba, y_pred)
print("El error cuadrático medio es ", round(mse*100,3), " %")
```

```
El error cuadrático medio es  4.447  %
```

Regresión Ridge y Lasso

La evaluación de esta parte usará la métrica de error porcentual ajustado medio. Vale la pena recordar los coeficientes obtenidos del proceso de modelación:

```
resultados_regresion["coeficientes"]
```

```
{'Coef_normal': array([ 0.        , -0.05366319, -0.12604845, -0.22685027, -0.28797468,
        -0.35457203, -0.39991772, -0.49217904, -0.49546519, -0.51181813,
         0.29862739,  0.39832096,  0.53248843,  0.60872173,  0.63867745,
         0.75294324,  0.81777038,  0.84446757,  0.69437918]),
 'Coef_ridge1': array([ 0.        , -0.05364488, -0.12603211, -0.22683501, -0.28796046,
        -0.35455841, -0.39990494, -0.49216591, -0.49545364, -0.51180904,
         0.29858475,  0.39827697,  0.53244259,  0.60867406,  0.63862804,
         0.75288984,  0.8177119 ,  0.84440078,  0.69430212]),
 'Coef_lasso': array([ 0.        , -0.        , -0.07173795, -0.17254086, -0.23366693,
        -0.30026734, -0.34561888, -0.4378919 , -0.4412045 , -0.45764135,
         0.24350776,  0.34320307,  0.47737344,  0.55360948,  0.5835672 ,
         0.69783375,  0.76265966,  0.7893516 ,  0.6385957 ])}
```

Dentro de la clase previamente creada para la modelación de esta sección se estableció el método de evaluación, acá se tiene el error porcentual ajustado medio para determinar el rendimiento del modelo, en este caso se observará la métrica para regresión lineal, regresión de Ridge y Regresión de Lasso.

```
'MAPE': 17.862565448352345,
'MAPE_ridge_1': 17.862732438617133,
'MAPE_lasso': 17.86437860509435}
```

La comparación de muestra de la siguiente manera:

	Y_test	Y_ajustado modelo final	GRCODE
0	10.293365	9.751004	683
1	10.238637	9.828707	683
2	10.172522	9.678619	683
3	10.339837	9.701208	683
4	10.164274	9.727906	683
5	10.126511	9.577817	683
6	10.464873	9.575257	683
7	10.273464	9.640084	683
8	10.297993	9.666781	683
9	10.174926	9.516693	683
10	10.489495	9.394394	683
11	10.506491	9.508660	683
12	10.289566	9.573487	683
13	10.370361	9.600184	683

Valores en escala logarítmica

Finalmente, la métrica de error ajustado porcentual medio del modelo muestra el siguiente detalle:

```
resultados_test["Metricas MAPE"]

{'MAPE modelo final': 8.848337795010787}
```

Se puede concluir entonces, que los resultados de implementar las nuevas técnicas son eficientes. Los resultados de la imagen superior muestran que el valor predicho y ajustado por el modelo de regresión Lasso cuenta con un error razonable de evaluación. Adicionalmente, se puede evidenciar que se lograron menores valores con respecto a los valores de prueba. Por lo tanto, es válido afirmar que los resultados del proyecto cumplen con los criterios de rendimiento comercial en el marco de la estimación de la reserva IBNR para la base de datos del ramo de salud, particularmente, para los resultados de la regresión Lasso. Las predicciones para la variable de respuesta muestran que los montos por dinero en pagos de siniestros están razonablemente por debajo de los valores de prueba, por lo tanto, al implementar este proyecto es de esperar que el pasivo destinado a reservar por concepto de IBNR por este método que por Chain Ladder y en comparación con la información real de la compañía. En ese sentido, los accionistas de la compañía percibirán mayores ganancias mensualmente debido a la reducción del impacto en los fondos propios de la compañía

Despliegue

Como se vio, el modelo propuesto genera resultados satisfactorios para la organización, lo cual se traduce en menores estimaciones de siniestros futuros. Por la tanto, se piensa en una menor reserva a constituir y en un menor pasivo que la compañía debe asignar en su balance contable y a sus informes de pérdidas y ganancias, con respecto a las estimaciones usando el método tradicional de Chain Ladder. Lo anterior es de gran interés para los accionistas y socios de la compañía, porque un menor pasivo implica un menor impacto sobre los fondos propios de la compañía y una menor afectación de sus inversiones en el negocio asegurador, lo cual siempre es deseable.

Planificación de despliegue

Un breve contexto del modelo

Antes de entrar en detalle al proceso de despliegue, el primer paso es resumir los resultados de los modelos y descubrimientos, como se vio anteriormente, la metodología de validación cruzada para el diseño de prueba indicó que el mejor modelo es la regresión de Lasso. Esa técnica de modelación muestra las siguientes predicciones con respecto a los valores de respuesta y comparando con los valores de prueba.

	Y_test	Y_ajustado modelo final	GRCODE
0	10.293365	9.751004	683
1	10.238637	9.828707	683
2	10.172522	9.678619	683
3	10.339837	9.701208	683
4	10.164274	9.727906	683
5	10.126511	9.577817	683
6	10.464873	9.575257	683
7	10.273464	9.640084	683
8	10.297993	9.666781	683
9	10.174926	9.516693	683
10	10.489495	9.394394	683
11	10.506491	9.508660	683
12	10.289566	9.573487	683
13	10.370361	9.600184	683

El ajuste del modelo final muestra disminuciones en la cantidad de dinero proyectado para siniestros por la compañía. Aunque el resto de modelos presentaron un buen desempeño, el presente obtuvo mejores resultados comerciales, además de lo mostrado durante la fase del diseño de experimento. Para poner en contexto a los lectores, el proceso de modelación se estructuró gracias a la creación de una clase compuesta por varias funciones. El primer paso consistió, en definir un dataframe con algunos insumos esenciales de cara a la estimación de reservas por metodología Chain Ladder como el origen, desarrollo, nombre de la compañía. entre otros. Otra función de relevancia empleada en

la modelación usa las librerías de Python como scikit-learn, entre otros, así como aquella que separa los datos de la matriz de diseño y la variable de respuesta para posteriormente aplicar la técnica de modelación en cuestión. Posteriormente, se tiene una función que sirve para realizar las predicciones en un conjunto de validación utilizando los modelos entrenados.

Planificación para el despliegue e integración

Paso 1: La firma consultora, con el apoyo de los actuarios designados, llevarán a cabo la implementación del mejor modelo en un entorno de pruebas, posteriormente se hará la implementación en el entorno de desarrollo.

Paso 2: Como se mencionaba al inicio de este documento, aunque los actuarios conocen la información, no tienen los conocimientos necesarios dado el tema a trabajar en cuestión. La firma consultora impartirá capacitaciones con respecto al uso y operación del modelo, así como calibración de parámetros e hipótesis de modelación. En adición ofrece sus servicios de asesoría y apoyo para realizar ajustes en caso de ser necesario.

Paso 3: Como se comentaba al inicio del proyecto, los requerimientos informáticos de este proyecto no suponen un impacto significativo para la compañía, de manera que un primer cambio en términos de integración, requiere que se haga un cambio en la metodología manual que implica el uso de Excel, R Studio y conexiones con SQL para la extracción de la información por una de mayor programación en lenguaje Python con un entorno apto para el trabajo. En ese orden de ideas, es necesario implementar un conector que enlace el entorno SQL con el entorno destinado de Python, lo cual no es muy costoso y es relativamente rápido de hacer, la mayor dificultad puede ser el proceso de empalme como tal. Por lo demás, los formatos y demás aspectos no requieren cambios.

Paso 4: Una vez se tiene la conexión se procede a realizar la programación respectiva en los ambientes de trabajo definitivos, como se vio, el código no es muy largo y además en un proceso que se debe realizar de manera mensual, no es algo que se deba hacer todos los días. Sin embargo, requiere cierto entendimiento, por lo tanto, se espera que el personal esté capacitado adecuadamente.

Bibliografía

IBM (2011). IBM SPSS Modeler CRISP-DM Guide. https://www.ibm.com/docs/es/spss-modeler/

Amin, Z., Antonio, K., Beirlant, J., & Carpentier, A. (2021). Loss Data Analytics (Versión en español). GitHub. https://ewfrees.github.io/Loss-Data-Analytics-Spanish/

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Ministerio de Hacienda y Crédito Público. (2013). Reservas técnicas de las entidades aseguradoras, Decreto 2973, 20 de diciembre de 2013. https://www.fasecolda.com/cms/wp-content/uploads/2019/08/dec-2973-2013.pdf

James, G., Witten, D., & Hastie, T. (2021). An Introduction to Statistical Learning with Applications in R.

Verrall, R. (Obtenido en 2023). Statistical Methods for the Chain Ladder Technique. https://www.casact.org/sites/default/files/2021-02/pubs_forum_94spforum_94spf393.pdf