Oscar Pulido

Aplicaciones del aprendizaje de máquinas en Actuaría y Finanzas

Primera versión Fase de preparación de los datos para el problema de estimación de provisiones

Control de versiones

Versión	Autor	Fecha	Aprobado por
0.1	Oscar Pulido	7/09/2023	Francisco Gómez

Este documento presentará los pasos seguidos para la fase de 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: https://www.casact.org/sites/default/files/2021-04/comauto_pos.csv. Esto es de fuente propia y proviene del histórico de siniestros de distintas compañías estadounidenses para el ramo de Auto Comercial. 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. En un inicio, se pretende escoger a todas las compañías en el proceso de modelación, pero se verá más adelante en el reporte de limpieza de

datos que esto no es posible. Adicionalmente, se mostrará el tratamiento de los outliers en el reporte.

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 y CumPaidLoss_C, también es deseable una variable para identificar las compañías. Adicionalmente hay una variable que puede complementar el análisis que es la de IncurLoss_C, por lo que se mantiene. La estructura de las características de la base final sería la siguiente:

Limpieza de los datos

Reporte de limpieza de datos

Como se mencionó en los puntos anteriores, en esta fase se tiene como objetivo tratar los outliers.

Valores en 0: Se decide excluir aquellos montos pagados de 0 para mejorar el rendimiento del modelo, y en dado caso que se encuentren estos valores en partes esenciales de los triángulos se sustituirían por 0 nuevamente usando métodos de imputación durante la fase de modelación. Esto permite descartar periodos de accidentalidad completos de algunas compañías aseguradoras como el caso de Public Underwritter Gro y Yel Co Ins. Eventualmente, hay casos donde se elimina la totalidad de una compañía, pues no hay pagos para calcular la reserva IBNR.

GRNAME	CumPaidLoss_C	AccidentYear	DevelopmentYear
<chr>	<dbl>	<dbl>	<dbl>
Public Underwriters Grp	0	1988	1988
Public Underwriters Grp	0	1988	1989
Public Underwriters Grp	0	1988	1990
Public Underwriters Grp	0	1988	1991
Public Underwriters Grp	0	1988	1992
Public Underwriters Grp	0	1988	1993
Public Underwriters Grp	0	1988	1994
Public Underwriters Grp	0	1988	1995

Yel Co Ins	0	1989	1989
Yel Co Ins	0	1989	1990
Yel Co Ins	0	1989	1991
Yel Co Ins	0	1989	1992
Yel Co Ins	0	1989	1993
Yel Co Ins	0	1989	1994
Yel Co Ins	0	1989	1995
Yel Co Ins	0	1989	1996

Lo anterior se vería de la siguiente manera en el triángulo de pagos:

	DevelopmentYear									
AccidentYear	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997
1988	0	0	0	0	0	0	0	0	0	0
1989	NA	6	20	21	23	24	24	24	24	24
1990	NA	NA	52	95	105	117	128	128	128	128
1991	NA	NA	NA	127	258	301	332	372	376	377
1992	NA	NA	NA	NA	120	225	277	339	361	370
1993	NA	NA	NA	NA	NA	427	1005	1111	1263	1306
1994	NA	NA	NA	NA	NA	NA	584	1511	1716	1817
1995	NA	NA	NA	NA	NA	NA	NA	369	741	931
1996	NA	NA	NA	NA	NA	NA	NA	NA	294	595
1997	NA	NA	NA	NA	NA	NA	NA	NA	NA	312

Valores negativos: También 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 uso 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_C
<chr>	<dbl>
Aegis Grp	19.142857
Agway Ins Co	352.181818
Alaska Nat Ins Co	1787.181818
Alliance Mut Ins Co	51.750000
American Contractors Ins Grp	432.339623
American Hallmark Ins Co Of TX	9.947368
American Mining Ins Co Inc	119.750000
American Modern Ins Grp Inc	488.733333
American Resources Ins Co Inc	244.636364
American Vehicle Ins Co	366.088889
Amerisafe Grp	46.909091
Amguard Norguard & Eastguard Grp	157.909091
Antilles Ins Co	416.854545
Atlantic Cas Ins Co	28.571429
Badger Mut Ins Co	106.360000
Bancinsure Inc	3.000000

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.