Problema de modelación de reservas de seguros comerciales de salud

Diciembre 4 de 2023

Universidad Nacional de Colombia
Escrito por: Carlos Alberto Orozco Merchan

Metodologia de trabajo CRISP-DM

Introducción

En el día de día los seres humanos nos vemos expuestos a múltiples riesgos que pueden atentar contra nuestra integridad y más específicamente hablando de la salud como individuos, estos factores pueden ser tanto de índole de nuestro entorno como de índole genética de enfermedades que pueden ser hereditarias o que se encuentran codificado en nuestro ADN.

Teniendo en cuenta lo anterior es evidente que siendo una persona viviendo en comunidad no tenemos control sobre nuestro entono o al menos tenemos control limitado y sobre las enfermedades genéticas no son de nuestro control, teniendo en cuenta por ello estamos contantemente expuestos aun riesgo latente en el cual podamos sufrir de una enfermedad accidente entre otras cosas que afecten nuestra integridad, estos incidentes para cuestiones de facilidad y no se requerido la discriminación de ello los llamaremos ***siniestro.***

Teniendo en cuenta estos riesgos a los que nos exponemos diario y que pueden desembocar en siniestros, a través del tiempo se empezaron a crear empresas que tienen como principal producto o objetivo social el cubrir ese riesgo al que se encuentran expuestas las personas, en la actualidad existen múltiples empresas de esta índole que prestan sus servicios para cubrir este riesgo de las personas en el sector salud y las cuales conocernos como seguradoras.

Mas específicamente estas empresas ofrecen un portafolio de seguros y soluciones en coberturas para el publico en general que busque cubrirse frente a un riesgo de tipo salud, estos seguros que ofrecen estas compañías corresponden a contratos de contraprestación en el cual el individuo que adquiere el bien adquiere el derecho de que la empresa aseguradora cubra los gastos y/o costos asociados a la ocurrencia de un sinestro del cliente en un periodo de tiempo a cambio de un pago ya se mensual, anual, trimestral, semestral o un único pago anual o según lo pactado en la suscripción del contrato.

Existen múltiples tipos de seguros para múltiples ramas de riesgo al que se encuentra expuesto el ser humano y para cada uno existe un tipo de seguro ya se bien: carros, vida, educativos, agropecuarios entre otros. Sin embargo, como se ha hablado anteriormente en este caso buscaremos enfocarnos en seguros de salud siendo la motivación para nuestro estudio el ver los costos incurridos en la materialización del siniestro de salud para múltiples aseguradoras.

Lo primero a considerar en el presente apartado es explorar de manera macro como funciona el negocio de las aseguradoras.

Las seguradoras como bien se dijo ofrece sus servicios de cobertura al publico sin embargo no todos los seguros y productos ofrecidos tiene el mismo costo o las mismas barreras de acceso para todas las personas, estas empresas a la hora de vender sus productos tienen en cuenta probabilidades de riesgo de cada cliente para dar un servicio muchas más personalizado, es decir las empresas tiene en cuenta las variables generales y medibles de sus clientes para así cuantificar el riesgo de un evento y/o siniestro y así poder asignarle un producto acorde a las necesidades del cliente.

Sin embargo, cabe aclarar que aunque la empresa buscar brindarle un servicio a sus cliente no siempre es deseable para la empresa y por ello los costos o requerimientos que solicita la empresa para brindar el servicio crece y se convierte en barreras de ingreso al servicio, lo anterior teniendo en cuenta que la empresa va a cubrir el riesgo al que el cliente se encuentra expuesto quiere decir que se transfiere el riesgo a la empresa y si las probabilidades de materialización del riesgo y costos asociados al mismo es muy alto no es deseable para la empresa asumir esos riesgo a no ser de que el cobre al cliente ese costo asociado el cual podría se muy alto y haría que el cliente puede querer asumir este riesgo en vez de transferirlo a la empresa.

Cuando la empresa realiza los cálculos y estimación de materialización del riesgo el otro calculo que realizan es el de la reserva, la reserva es el porcentaje de dinero y/o principal (monto solicitado al cliente para cubrir su riesgo)que deben guardar las aseguradoras para poder responder por las diferentes y posibles materializaciones de riesgo de sus cliente, comúnmente esta reserva se calcula de manera anual, es decir se calcula cuánto dinero se debe guardar para responder por los siniestros que pasen en el siguiente año. La forma de cálculo de la reserva es por medio del método Chain-Ladder, el cual es un método cuyo objetivo es calcular las reservas de los siniestros futuros por medio de los históricos y bajo la suposición de que los patrones del pasado seguirán pasando en el futuro.

El cálculo de la reserva es un tema de extremo cuidado puesto que una mala estimación puede traer fuertes consecuencias para la compañía, en primer lugar, si el cálculo de la reserva subestima los gastos la aseguradora presentara falta de liquidez y se tendrá que endeudar para cumplir sus obligaciones, en segundo lugar, si el cálculo de la reserva sobreestima los gastos entonces la aseguradora pierden la oportunidad de invertir ese dinero y obtener nuevas ganancias. Así pues, el tema central de este trabajo es plantear un nuevo cálculo de la reservas para mejorar la estimación futura teniendo en cuenta los históricos de la empresa y nuevos modelos teóricos estadísticos o de machine Learning, este trabajo se realizara para una asegurador cuyo producto en un seguro sobre responsabilidad comercial salud, Los seguros de responsabilidad comercial de salud se ofrecen a empresas o individuos que usan los automóviles en sus operaciones comerciales, como empresas construcción, transporte, entre otros. El seguro cubre daños a terceros que pueden suceder en el momento en que las personas o las empresas laboran, estos daños que cubre el seguro pueden ser lesiones corporales o daños a estructuras.

Es de vital importancia para la empresa saber cuánto dinero tienen la probabilidad de pagar o perder en un futuro determinado por eso es importante hacer un buen estudio sobre cuánto dinero necesitan reservar para hacer frente a la materialización del riesgo ya que si subestiman esa cantidad pueden perder dinero porque pueden tener problemas de liquidez para pagar la cantidad que aseguraron, y si la empresa sobreestima esa cantidad pueden perder beneficios de posibles inversiones que no se realizaron.

Haciendo énfasis en lo anterior deriva en que el tema central de este trabajo es realizar un ejercicio de cálculo de la reservas para mejorar la estimación futura de los riesgo, teniendo en cuenta los históricos de la empresa y nuevos modelos teóricos estadísticos o de machine Learning, este trabajo teórico se realizara suponiendo que se hace el presente ejercicio para una aseguradora que busca entrar en el mercado de los seguros de salud y para ver como deberían empezar a operar se hará un estudio histórico sobre los datos disponibles de otras aseguradoras.

Comprensión del Negocio.

- La comprensión empresarial: siendo el objetivo una descripción general de estructura empresarial, para conseguirlo se debe tener en cuenta y plantear diferentes pasos y fases que deben servir como base para definir las personas y/o profesionales clave de la organización para la solución de nuestro problema, los diferentes objetivos que tiene la compañía, definir el personal asociado, recursos, riesgos, contingencias, costos y beneficios, entre otros, para luego finalizar con una visión general del proyecto y así definir los tiempos de las fases el proyecto.
- Determinación de objetivos empresariales: gracias a la globalización se evidencia que siempre existen nuevas tecnologías y nuevos conocimientos disponibles varias de estas en la actualidad sin barreras de acceso, siendo así mas deseable para las compañías mantenerse a la vanguardia para así afrontar los problemas que se pueden presentar con los mejores y más eficientes conocimientos. Teniendo en cuenta lo anterior las compañías de seguros en la actualidad cada día se interesan mas y se actualizan en las tecnologías de aprendizaje de máquinas, siendo esta tecnología una innovación en cuanto a estimaciones más versátiles y eficientes sobre cómo se debe llevar el negocio ya se en un futuro cercano o lejano.
  Enfocándonos en nuestro problema sabemos que para una empresa aseguradora es de sus mayores intereses estimar la reserva y minimizar los errores o la distancia entre la reserva estimada y la que afrontan realmente, lo cual puede ser indicativo de eficiencia del negocio y cumplimiento de sus metas, al mismo tiempo generando confianza del asegurado, cumplimiento de metas e indicadores económicos, cumplimiento normativo y atracción de inversionistas.

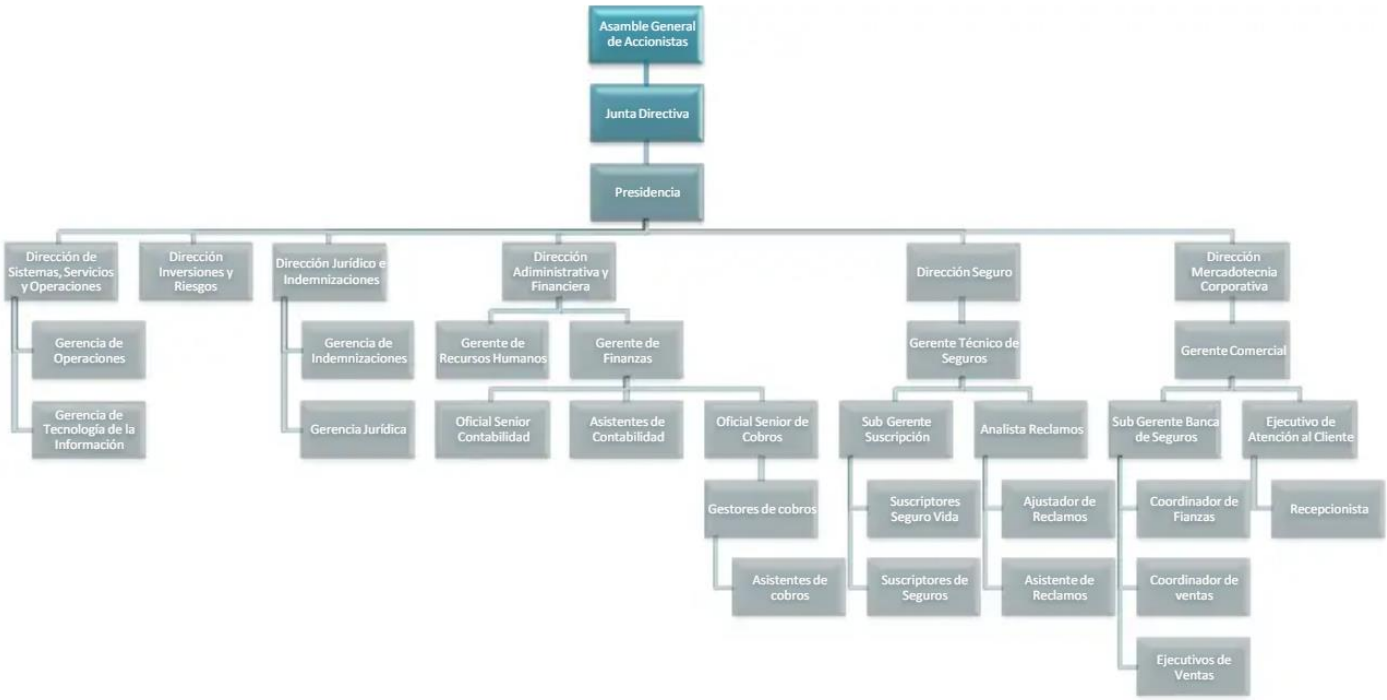
A través de la tecnología de aprendizaje de máquinas se busca el objetivo de minimizar los errores de la estimación de la reserva, siendo así se procede a aplicar todo un proceso de minería de datos del cual se evaluará si su desempeño es favorable o no si las estimaciones de la reserva con el método utilizado mejora sustancialmente, concretamente el

desarrollo del proyecto se considera un éxito si el rendimiento en la estimación es mejor que con la técnica de Chain-Ladder.

1. Compilación de antecedentes comerciales

La compañía durante el tiempo que lleva en el mercado a mantenidos grandes estándares de calidad respecto a los pagos de sus obligaciones con sus clientes y en el pago de sus obligaciones con terceros, lo cual lo ha llevado mantener sus estándares reputacionales tanto en la industria como en sectores vinculados a su negocio ya sea calificadoras de riesgo como con el sector bancario, sumado a ello la estructura empresarial no a presentado cambios significativos a través del tiempo, aun con todo lo descrito siendo un sector económico con cambios contantes se cree hay un espacio para la mejora del rendimiento en cuento a las estimaciones de las reservas de la compañía.

Sin embargo, para realizar nuestra evaluación de reservas bajo metodologías de aprendizaje de máquinas se requiere contar con toda la informacion disponible para saber con qué áreas de la compañía se debe contar y de cuales se puede obtener la informacion para la solución y enfoque de nuestro problema.



Personas Clave en la organización

De los puestos y/o personas que tiene relación directa y poseen informacion relevante para la solución de nuestro problema de reservas.

-Presidencia.

-Dirección de inversiones y riesgos

-Gerencia de Finanzas y Contabilidad

-Gerencia de Tecnología y la Información

-Gerencia Técnica de seguros

Dependencias Internas para apoyo financiero y con conocimientos para el desarrollo del proyecto de minería de datos requeridos

-La presidencia corresponde a un patrocinador financiero del proyecto.

-Gerencia de Finanzas y Contabilidad es un patrocinador financiero que adicional cuenta con experiencia en el tema de la estimación de reservas en el rubro financiero.
- Gerencia Técnica de seguros ofrecen su experiencia en el cálculo de la reserva del rubro asegurador.
-Gerente de Tecnología y la Información la cual tiene el acceso a los datos de la compañía y por ende de las operaciones necesarias para el estudio
-Grupo de informáticos ofrecen experiencia en sistemas, programación y modelos estadísticos y de aprendizaje de máquinas.

Unidades de negocio que verán el impacto y el cambio generado por el proyecto a implementar.

Se prevé que la unidad de negocios que verán el cambio y el impacto del proyecto desarrollados serán la gerencia de Finanzas y Contabilidad puesto que implementará una nueva forma de estimación de las reservas de la compañía, a su vez si esta mejora da resultados generará mayor solvencia para múltiples dependencias al igual que se verá una mejora en la Gerencia General debido a una mejora para los intereses generales de la empresa.

Descripción del área con oportunidades de mejora

La oportunidad de mejora en el desempeño y eficiencia de la empresa está enfocada a la estimación de la reserva de seguros en salud para los años de operación subsecuentes, por medio de un proyecto de minería de datos, cuyas áreas encargadas de esta labor es la gerencia de Actuaria/Seguros y Finanzas/Contabilidad.

Para desarrollar la minería de datos que ayudará a solucionar el proyecto propuesto se necesita previamente una revisión bibliográfica y sistemática sobre metodologías de minería de datos aplicada por el área de actuaria de la organización siendo este grupo de trabajo los encargados de la estimación de la reserva de seguros.

Descripción del método trabajo para la solución del problema.

El método implementado hasta el momento para la estimación del cálculo de la reserva para años posteriores se hacía por medio del método Chain-Ladder basada en datos históricos de los siniestros los cuales se organizan en una matriz que denominamos “triángulo de siniestros” a la cual se le aplica el método de la cadena basado en el supuesto de que la relación entre las reservas actuales o evaluadas en el presente y las reservas subsecuentes se mantiene constante a medida que se avanza en el tiempo, ello para extrapolar los resultados de las reservas futuras en la cartera de seguros.

Esta método siendo el común denominador en el sector asegurador debido a la facilidad de su uso, sin embargo presenta espacios para mejoras en su estimación teniendo en cuenta este método se basa en suposiciones las cuales no tienen siempre por qué cumplirse, por ello la revisión literaria recomienda aplicar este método en conjunto con otras herramientas de modelación o estimación, siendo este el móvil del proyecto de minería de datos para así mejorar el actual método implementado para la estimación de la reserva y el cual que no dependa de supuestos tan subjetivos que puedan no tener fundamentos, desembocando así e n un modelo y predicciones mas precisos para la compañía.

2. Definición de objetivos comerciales

- Mejorar la solvencia económica de la compañía.
- Aumento en la competitividad en el rubro asegurador
- Brindar mejor y mayor servicio a sus clientes cumplimiento con su misión y visión.
- Mejora en eficiencia y procesos internos de la compañía.

3. Criterios de mejora de procesos de reserva de la empresa

- Objetivo

Mejora y eficiencia en la estimación de la reserva de los seguros, a través del uso de nuevas metodologías de la minería de datos aplicados a la metodología actual, aumentando rendimiento de esta a largo de los años de operación de la compañía.

- Objetivo secundario

Incrementar la solvencia económica de la empresa mostrando de esa forma una redistribución y eficiencia en la asignación y uso de los recursos de la compañía.

- Evaluación General

Para el presente proyecto la compañía tiene a su disposición el histórico de los datos agregados con variables de relevancia ya sea bien valores de prima, valores asegurados, así como en caso de materialización de riesgos los valores y fechas en los que la compañía asumió los mismos, entre otros datos relacionados a la actividad aseguradora.

La compañía tiene a su disposición al capital humano necesario para realizar el proyecto de minería de datos, puesto que cuenta con los grupos de trabajo que tienen como responsabilidad el cálculo de reservas y personal calificado en metodologías de minería de datos.

Una de las contingencias que podría presentar el proyecto es implementar un método que resulte en una desmejora de eficiencia y estimación que el método utilizado actualmente, dejando como resultado que la estimación de las reservas de la compañía no tenga una mejora, siendo que su implementación genere problemas de solvencia económica que puede generar detrimento patrimonial al interior de la compañía que podrían generar una pérdida de eficiencia financiera de la misma. Siendo así el plan de contingencia el desistir en la implementación del modelo desarrollado si no se evidencia una mejora en los indicadores de eficiencia y cumplimiento del negocio de la compañía, así volviendo a la implementación tradicional que se utilizaba en el proceso de reservas.

1. Inventario de Recursos

Requerimientos tecnológicos para el desarrollo del proyecto

Teniendo en cuenta que para el proyecto actual se trabajará con bases de datos históricas de la compañía con demasiados datos se requerirá:

- Computadoras y Servidores.
- CPU (Unidad Central de Procesamiento).
- GPU (Unidad de Procesamiento Gráfico).
- Memoria RAM que soporte la cantidad de datos utilizados.
- Unidades de memoria para el almacenamiento y archivado de los datos (SSD o HDD).
- Conexiones de banda ancha a internet con velocidad estable de subida o bajada.
- Cluster de Servidores con acceso a copias de seguridad y/o para realización de las mismas.

Identificar fuentes de datos y almacenes de conocimiento

Los datos en la compañía se encuentran almacenados en los servidores de esta y para el cual es acceso para los grupos de trabajo de la compañía y del proyecto serán gestionados y se otorgara la autorización correspondiente por el área de informática de la compañia. Por otro lado, se pone a disposición del equipo de trabajo del proyecto un presupuesto para la adquisición de bases de datos externas, en el caso actual de la compañía se evidencia que la línea de seguros de salud sobre la que quieren realizar el estudio no tiene la cantidad de datos requerida siendo así se hará un ejercicio con las bases de datos de compañías de uso publico las cuales contienen la información de primas emitidas reservas entre otros para un periodo de tiempo especificado, por ello se usara la base de datos de aseguradoras externas y los de la compañía del estudio, de tal forma que se cuenta con estos datos históricos de la compañía como datos históricos de otras compañías.

Identificar recursos de personal

El grupo de trabajo del Proyecto cuenta con capital humano con previa y amplia experiencia en el sector asegurador y en el negocio de la compañia, grupo de trabajo el cual estará conformado por profesionales de: actuaria, programación y datos.
Adicional se contará con profesionales en informática que tiene conocimientos en administración de bases de datos, ETL, modelado y análisis de datos.

2. Requerimientos, supuestos y restricciones

Requisitos

Los requisitos legales que competen al desarrollo y al grupo desarrollador del proyecto, en el cual se construirá el nuevo método para el cálculo de reservas deben cumplir con ciertas normativas en la manipulación de los datos y estar alineados con los requisitos de seguridad y confidencialidad de los datos implementado por los estatutos de la compañía, los cuales comprenderán:

- No debe acceder a datos sin autorización
- No debe compartir datos sin consentimiento
- No debe recopilar datos innecesarios
- No debe retener datos más tiempo del necesario
- No debe usar datos para fines no autorizados
- No debe dejar datos desprotegidos
- No debe ignorar las solicitudes de los titulares de datos
- No debe falsificar datos
- No debe compartir contraseñas o credenciales
- No debe dejar dispositivos sin protección
- No debe usar datos para discriminación
- No debe evadir la notificación de violaciones de datos
- No debe dar acceso a los datos a terceros no vinculados al proyecto o a la compañía.

Aclaración de suposiciones

El Proyecto no cuenta con objetivos desconocidos para la competencia del mercado no representando así competencia desleal en el sector y no siendo causal de infracciones de códigos de conducta en el mercado, sin embargo, no se desconoce que no existan otras compañías que buscarían conseguir el mismo objetivo del proyecto planteado,
Aunado a lo anterior, se consideran que los datos utilizados en el estudio son de calidad si son: completos, correctos, coherentes, precisos y actualizados.

Finalmente, el proceso y resultados del proyecto serán mostrados a estos entes los entes directivos de la compañia, enfocándose en la visualización de los resultados por medio de presentaciones y representación visual de los datos de manera dinámica y entendible para todos los interesados al proyecto.

Restricciones

El personal involucrado en el proyecto cuenta con las credenciales y contraseñas requeridas para llevar a cabo el desarrollo del proyecto.

No existen restricciones legales para utilizar los datos de los clientes y de la compañía teniendo en cuenta las autorizaciones de tratamiento de datos a la cual se acogen los clientes y terceros, pero siempre manteniendo las buenas prácticas sobre los datos mencionadas anteriormente.

3. Riesgos y contingencias

A continuación, se presenta una lista de riesgos con las contingencias respectivas.

- Programación: el tiempo previsto para el desarrollo del proyecto puede ampliarse frente a la propuesta original, el plan de contingencia frente a este es ampliar la capacidad de cómputo y de recursos requeridos para así mejorar el desempeño y disminuir el tiempo de desarrollo de este, otro plan es alargar el tiempo de terminación del proyecto en caso de que no contar con los recursos para ampliar la capacidad computacional instalada.
- Financiero: siempre y cuando se sigan los lineamientos y el plan de acción propuesto para el proyecto y la calidad y cantidad de los datos sea la adecuada o necesaria no se debería presentar un inconveniente de falta de musculo financiero para la culminación del proyecto, sin embargo, frente a este escenario se planea una refinanciación del proyecto y así ampliar los recursos disponibles o en el caso de no ser posible el refinanciamiento de este culminarlo con los recursos disponibles.
- Mala calidad de los datos: Los datos pueden contener errores, pueden estar corrompidos, pueden presentar valores y o datos faltantes, para este caso se busca contar con el capital humano necesario y con conocimiento y experiencia en la limpieza e imputación de datos.
- Resultados: si los resultados iniciales del proyecto sobre el método a implementar para el cálculo de las reservas no generan un impacto representativo en mejoras dentro de la actividad de la compañía, seria una reafirmación que el método tradicional es igualmente eficiente y eficaz que el nuevo y por lo tanto no hay motivo para realizar el cambio teniendo en cuenta que genere gastos adicionales para la compañía, o se puede pensar en otra metodología de minería de datos que pueda generar mejores resultados.

4. Terminología

Para los individuos involucrados en el proyecto minería de datos se vean involucrados y utilicen la misma terminología se pone a disposición un glosario con las palabras, frases y términos, de mayor más relevancia y del cual su utilización es indispensable en un proyecto de minería de datos para la estimación de las reservas de la compañía de seguros en el tramo de salud. El siguiente glosario también puede ser encontrado en la intranet de la empresa.

- Minería de Datos (Data Mining): El proceso de descubrir patrones, tendencias y conocimientos ocultos en grandes conjuntos de datos utilizando técnicas estadísticas y de aprendizaje automático.
- Reserva de Siniestros: Una estimación de la cantidad de dinero que una compañía de seguros debe reservar para cubrir reclamaciones de seguros pendientes y futuras.
- Datos de Siniestros: Información detallada sobre los siniestros reportados, que incluye fechas, descripciones, costos estimados y pagos realizados.
- Desarrollo de Siniestros: El proceso por el cual los costos de un siniestro aumentan o disminuyen con el tiempo a medida que se investigan, se procesan y se resuelven las reclamaciones.
- Triángulo de Siniestros: Una representación tabular de los siniestros a lo largo del tiempo, que muestra cuándo se reportaron, cuánto se pagó en cada período y cuánto queda pendiente de pago.
- Tasa de Desarrollo: La tasa promedio a la que los costos de los siniestros se incrementan o disminuyen con el tiempo en función del análisis de datos históricos.
- Modelo de Reservas: Un modelo matemático o estadístico que se utiliza para prever los costos futuros de siniestros y, en última instancia, calcular las reservas necesarias.

- Ajuste de Reservas: Los cambios que se realizan en las estimaciones de reserva a medida que se obtienen más datos o se actualiza el modelo de reserva.
-
- Análisis de Pérdida Triangular: Una técnica que se utiliza para estimar las reservas basadas en la información contenida en el triángulo de siniestros.
- Exceso de Pérdida (Excess Loss): La cantidad de dinero que una aseguradora está dispuesta a pagar por encima de un cierto límite antes de que se active la cobertura de reaseguro.
- Reaseguro (Reinsurance): Un acuerdo en el que una compañía de seguros transfiere parte de sus riesgos a otra compañía de seguros o reaseguradora para limitar sus pérdidas potenciales.
- Cobertura de Responsabilidad Comercial: Un tipo de seguro que proporciona protección contra reclamaciones por lesiones corporales o daños a la propiedad que puedan surgir en el curso de las operaciones comerciales.
- Modelo de Aprendizaje Automático: Un enfoque que utiliza algoritmos y técnicas de aprendizaje automático para analizar datos históricos y hacer predicciones sobre futuros siniestros y costos.
- Validación Cruzada (Cross-Validation): Una técnica que se utiliza para evaluar la precisión y la eficacia de un modelo de aprendizaje automático mediante la división de los datos en conjuntos de entrenamiento y prueba.
- Clasificación de Riesgo: El proceso de categorizar diferentes tipos de riesgos y evaluar su probabilidad y gravedad.
- Primas de Seguro: Los pagos periódicos que los asegurados realizan a la compañía de seguros a cambio de la cobertura de seguro.
- Seguro de Automóviles Comerciales: Un tipo de seguro que proporciona cobertura para vehículos utilizados con fines comerciales, como camiones, furgonetas y flotas de vehículos
- Póliza: Un contrato legal que establece los términos y condiciones de la cobertura de seguro, incluyendo los riesgos cubiertos, las primas y otros detalles.
- Primas: Pagos regulares realizados por el asegurado a la compañía de seguros a cambio de la cobertura de seguro.
- Riesgo: La probabilidad de que ocurra un evento adverso que pueda dar lugar a una reclamación de seguro.
- Reclamación: Una solicitud presentada por un asegurado para recibir compensación por un evento cubierto por la póliza de seguro.
- Estadísticas: El análisis de datos numéricos y la aplicación de métodos estadísticos para obtener información sobre patrones y tendencias.
- Segmentación: La división de un conjunto de datos en grupos más pequeños o segmentos para un análisis más detallado.
- Ajuste: La modificación de las estimaciones de reservas de seguros en función de nuevos datos o información actualizada.
- Cobertura: El alcance y los términos de protección proporcionados por una póliza de seguro.
- Fraude: La presentación de información falsa o engañosa con el propósito de obtener beneficios indebidos del seguro.
- Exceso: El monto que una compañía de seguros no cubrirá y que debe ser asumido por el asegurado o por otra forma de seguro.
- Estimación: Una aproximación calculada o proyectada de un valor o cantidad, como la estimación de las reservas de siniestros.
- Modelo: Un conjunto de algoritmos y reglas matemáticas utilizado para predecir o analizar datos en función de patrones históricos.
- Experiencia: El historial de siniestros y reclamaciones de una compañía de seguros, utilizado para hacer estimaciones futuras.
- Aprendizaje: La capacidad de un sistema informático para mejorar su rendimiento a través de la experiencia y la adaptación a nuevos datos.
- Validación: El proceso de confirmar la precisión y eficacia de un modelo o método de estimación mediante la comparación con datos reales.
- Regulaciones: Las leyes y normativas gubernamentales que rigen la industria de seguros y establecen estándares para la conducta y las prácticas.

- Reserva: La cantidad de dinero que una compañía de seguros establece para cubrir futuras reclamaciones y obligaciones.
- Siniestro: Un incidente o evento adverso que da lugar a una reclamación de seguro.
- Historial: Un registro de eventos pasados y datos relacionados, como el historial de siniestros de un asegurado.
- Pérdida: La cantidad de dinero que una compañía de seguros paga como resultado de una reclamación realizada por un asegurado.

5. Análisis costo/beneficio

Dependiendo del musculo financiero de la compañía y del capital humano con el que cuente para el proyecto se requerirá hacer una evaluación de costos implícitos del proyecto sin embargo este análisis de costos debe ser gerenciado y analizado en conjunto con la gerencia general y de costos de la compañía.

Sin embargo, algunos costos enunciados y asociados al proyecto pueden corresponder a:

- Costo por Recopilación de Datos
- Despliegue de Resultados
- Costos de Operación
- Costos Laborales

Y finalmente una estimación y balanceo general de costos destinados a cada fase del proyecto.

Beneficios:

- Se cumple el objetivo del proyecto lo cual traerá mejoras en el cálculo de la reserva y a largo plazo mejores rentabilidades para la empresa.
- Se logra una mayor organización de los datos para el cálculo de reservas.
- Se toman de Decisiones Basada en Datos reales de la compañía
- Ventaja Competitiva
- Satisfacción del Cliente
- Cumplimiento Regulatorio
- Innovación en Productos y Servicios

Determinación de los objetivos de la minería de datos

1. Objetivos de minería de datos

Utilizando datos históricos de la actividad en marcha de la compañía y sus reservas, se generará un modelo de aprendizaje de máquinas o estadístico para realizar el cálculo y estimación de triángulos de reservas anuales, siendo un punto de comparación la validación cruzada con las reservas reales de cada año.

2. Criterios de éxito de la minería de datos

Se desea obtener un modelo que tenga un mejor rendimiento que el método usual de Chain ladder y tenga bueno valores en las diferentes medidas de rendimiento, este rendimiento será calculado con diferentes métricas.

- Error Cuadrático Medio
- Error Absoluto Medio
- Coeficiente de Determinación
- Error porcentual absoluto medio

Estas métricas se utilizan porque se van a pronosticar reservas, es decir valores de números continuos.

- Plan del Proyecto

Fase	Tiempo	Recursos	Riesgos
Comprensión Empresarial	1 semana	Corresponde a los recursos humanos y tecnologías puesto a disposición por parte de la compañía para el desarrollo del proyecto,	Cambios en directrices de la compañía o enfoques del negocio, cambio en los productos relacionados con el cálculo de la reserva que se desea estimar.
Comprensión de los datos	1 semana	Todos los involucrados en el proyecto pertenecientes a la compañía, especialmente analistas de datos, mineros de datos, ingenieros de datos y actuarios. Lenguaje de programación Python, adquisición de bases de datos y almacenamiento.	Problemas de limpieza de datos, origen de la información, problemas de hardware o software, licencias, versiones de librerías del lenguaje de programación.
Preparación de los datos	2 semana	Analistas de datos, mineros de datos, ingenieros de datos y actuarios. Lenguaje de programación Python, adquisición de bases de datos y almacenamiento	Problemas de hardware o software, licencias, versiones de librerías del lenguaje de programación, definir y crear la unidad de análisis.
Modelado	3 semanas	Analistas de datos, mineros de datos, ingenieros de datos y actuarios. Lenguaje de programación Python, adquisición de bases de datos y almacenamiento.	Incapacidad para escoger el modelo adecuado, falta de de hardware o software, licencias, versiones de librerías del lenguaje de programación
Evaluación	1 semana	Analistas de datos, mineros de datos, ingenieros de datos, actuarios y personas de la compañía que están involucradas y les interesa el proyecto. Lenguaje de programación Python.	Cambios en la dirección de la compañía, incapacidad para implementar los resultados, mal escogencia de evaluación.
Despliegue	1 semana	Analistas de datos, mineros de datos, ingenieros de datos, actuarios y personas de la compañía que están involucradas y les interesa el proyecto.	incapacidad para implementar los resultados porque no se tienen los recursos necesarios

1. Evaluación de herramientas y técnicas

La herramienta a utilizar para lograr el éxito en la minería de datos es el lenguaje de programación Python por medio de Notebook de Jupyter el cual es idóneo para presentar algoritmos, texto y lenguaje matemático sobre un modelo de minería de datos. El Lenguaje de programación Python tienen diferentes ventajas.

- Legibilidad y Simplicidad
- Amplia Comunidad y Soporte

- Multiplataforma
- Librerías y Ecosistema
- Aprendizaje Automático y Ciencia de Datos
- Desarrollo Web
- Integración

Comprensión de los datos

1. Recopilación de datos iniciales

Para la elaboración del proyecto de minería de datos se cuenta con una base de datos históricos de la compañía que le ayudará a diseñar un nuevo modelo diferente al usual, según la base de datos se considera no es requerido la adquisición de información o bases de datos adicionales, además se cuenta con el diccionario de datos que muestra información detallada de cada variable en de la base de datos, concretamente muestra el Nombre del atributo, Tipo de atributo, Datos de referencia, Reglas para validación, esquema o calidad de datos, propiedades detalladas de los elementos de datos e Información física sobre dónde se almacenan los datos.

2. Preguntas

- ¿Qué atributos (columnas) de la base de datos son las que se deberían utilizar para resolver el proyecto a desarrollar?
Los atributos más importantes son las variables, GRCODE, GRNAME, AccidentYear, DevelopmentYear, DevelopmentLag y IncurLoss_F2.

- ¿Qué atributos parecen irrelevantes y pueden excluirse?
Los atributos que se pueden excluir son CumPaidLoss_F2, BulkLoss_F2, EarnedPremDIR_F2, EarnedPremCeded_F2, EarnedPremNet_F2, Single, PostedReserve97_F2.

- ¿Existen datos suficientes para sacar conclusiones generalizables o hacer predicciones precisas?
Para este caso existen pocos datos, pero aun así puede funcionar la generalización, aun asi en caso de que mientras el desarrollo del proyecto se disponga de más datos pueden ser implementados.

- ¿Hay demasiados atributos para el método de modelado que elija?
Realmente no hay muchos atributos para el modelado a no ser que se desee modelar teniendo en cuenta las diferentes aseguradoras que hay en la base de datos

- ¿Ha considerado cómo se manejan los valores faltantes en cada una de sus fuentes de datos?
Si, los valores faltantes son por cuestiones del negocio y se pueden eliminar esos registros no tendrán mayor incidencia en los resultados.

3. Descripción de los datos

La base fue tomada de Casualty Actuarial Society (CAS) de la seccion https://www.casact.org/publications-research/research/research-resources/loss-reserving-data-pulled-naic-schedule-p. esta base de datos corresponde a la información sobre siniestros de los principales ramos personales y comerciales de todas las aseguradoras de daños materiales que operan en Estados Unidos. Es de aclarar que estos archivos son públicos y no tienen ninguna restricción sobre su uso. Este formato se encuentra completo y no tiene problemas con el tratamiento de los datos De la base anterior se va a trabajar con solamente la data de Negligencia médica - Reclamaciones realizadas.

Se evidencia que la base de datos cuenta con 13 variables, de las cuales todas tienen datos completos y no tienen datos faltantes o vacíos y por ende no tendríamos problemas al trabajar con la base ya que no hay que realizar una limpieza profunda de los datos, más solamente tendríamos que definir con que variables es deseable trabajar.

Las variables con las que contamos son:

GRCODE: código de empresa (incluidos grupos de aseguradores y aseguradores individuales).

GRNAME: nombre de la empresa (incluidos grupos de aseguradores y aseguradores individuales).

AccidentYear: Año del accidente va desde 1988 a 1997.

DevelopmentYear: Año del desarrollo y o de pago.

DevelopmentLag: cantidad de años que tardo en hacerse el pago.

IncurLoss_F2: pérdidas y gastos asignados reportados al final del año.

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

BulkLoss_F2: Reservas y IBNR (Incurridas, pero no reportadas) sobre pérdidas netas y gastos de defensa y contención de costos reportados al final del año.

EarnedPremDIR_F2: Primas devengadas en el año incurrido: directas y asumidas.

EarnedPremCeded_F2: Primas devengadas en el año incurrido - cedidas.

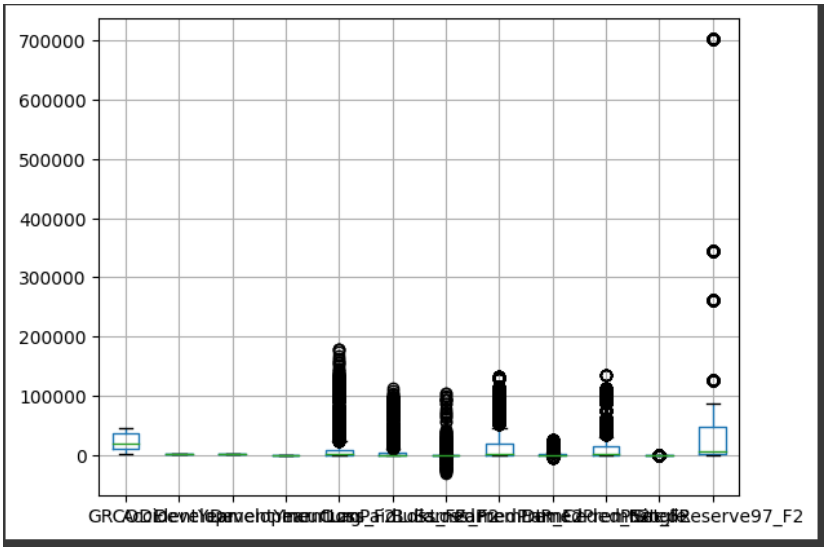
EarnedPremNet_F2: Primas devengadas en el año en que se incurrió - netas.

Single: 1 indica una sola entidad, 0 indica una aseguradora grupal.

PostedReserve97_F2: Reservas contabilizadas en el año 1997 tomadas del Anexo de Suscripción e Inversiones

4. Explorando los datos

Se evidencia si hay valores atípicos o outliers en la información que se tiene, se realiza una grafica de los datos y sus valores en los cuales evidenciamos que en la variable de PostedReserve97_F2 presenta montos de reservas atípicas, las cuales podrían representar la suscripción de un seguro de vida el cual podría presentar un riesgo muy alto y por lo mismo se decidió realizar esa reserva.



Aun cuando se evidencian valores outliers en los datos que se tiene se decide no eliminarlos teniendo en cuenta que son un riesgo al cual la compañía se puede ver expuesta por esta misma razón se decide no eliminarlos de la muestra.

Posterior se hace una descriptiva básica de cada uno de los datos que tenemos en nuestra base e datos, la cual corresponderá al promedio de cada uno de los datos, su desviación estándar y su variancia, sus percentiles y sus valores máximos y mínimos.

	GRCODE	AccidentYear	DevelopmentYear	DevelopmentLag	IncurLoss_F2	CumPaidLoss_F2	BulkLoss_F2	EarnedPremDIR_F2	EarnedPremCeded_F2	EarnedPremNet_F2	Single	PostedReserve97_F2
count	3400.000000	3400.000000	3400.000000	3400.000000	3400.000000	3400.000000	3400.000000	3400.000000	3400.000000	3400.000000	3400.000000	3400.000000
mean	22809.764706	1992.500000	1997.000000	5.500000	11609.344412	6706.067059	1095.803235	14111.605882	1803.497059	12308.108824	0.852941	57065.529412
std	14708.377001	2.872704	4.062617	2.872704	26802.819463	17121.815066	7612.672277	26399.284476	3893.424584	24824.225795	0.354217	134355.533990
min	669.000000	1988.000000	1988.000000	1.000000	-17.000000	-1190.000000	-32101.000000	-781.000000	-6214.000000	-728.000000	0.000000	0.000000
25%	10341.000000	1990.000000	1994.000000	3.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	629.000000
50%	19764.000000	1992.500000	1997.000000	5.500000	645.000000	187.000000	0.000000	1500.000000	106.500000	1302.000000	1.000000	5875.000000
75%	36234.000000	1995.000000	2000.000000	8.000000	9050.500000	4385.500000	107.250000	18094.500000	1473.500000	13490.000000	1.000000	46762.000000
max	44504.000000	1997.000000	2006.000000	10.000000	179425.000000	113189.000000	104402.000000	131948.000000	25553.000000	135318.000000	1.000000	702246.000000

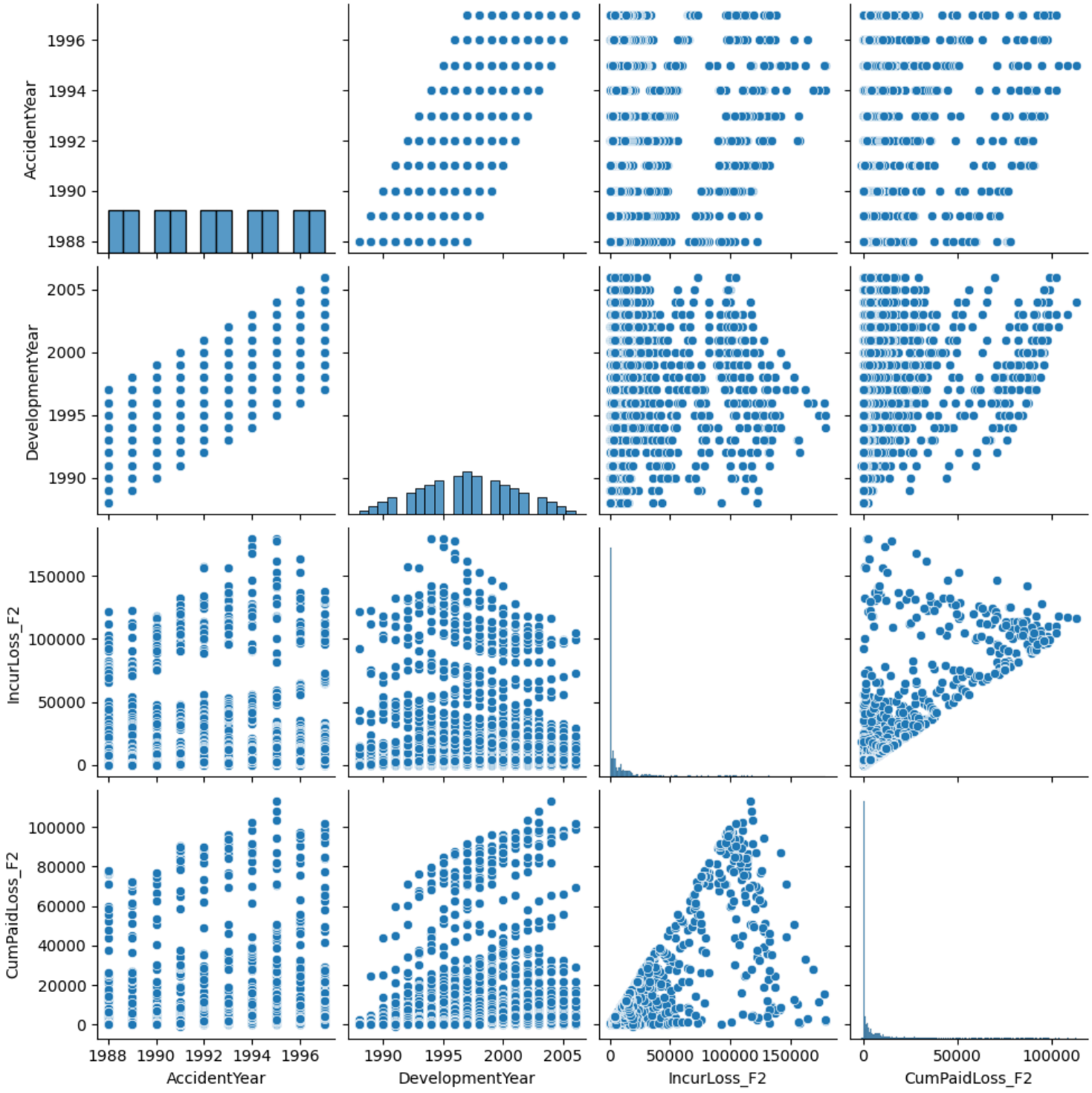
Se valida cuantas aseguradoras y con cuentas observaciones cuenta cada una en nuestra base de datos

```
Scpie Indemnity Co                  100
Preferred Professional Ins Co       100
Nichido Fire & Marine Ins Co Ltd    100
Texas Hospital Ins Exch             100
State Volunteer Mut Ins Co          100
MHA Ins Co                          100
Health Care Ind Inc                 100
National Guardian RRG Inc           100
Medical Mut Ins Co Of ME            100
Promutual Grp                       100
Utah Medical Ins Assoc              100
Seguros Triples Inc                 100
Dentists Ins Co                     100
Physicians Recip Insurers           100
Louisiana Med Mut Ins Co            100
Clinic Mut Ins Co RRG               100
Michigan Professional Ins Exch      100
National American Ins Co            100
NCMIC Ins Co                        100
Underwriters At Lloyds London       100
Community Blood Cntr Exch RRG       100
Campmed Cas & Ind Co Inc MD         100
Homestead Ins Co                    100
Franklin Cas Ins Co RRG             100
MCIC VT Inc RRG                     100
Texas Medical Ins Co                100
Controlled Risk Ins Co Of VT Inc    100
American Assoc Of Othodontists RRG  100
Eastern Dentists Ins Co RRG         100
Overseas Partners Us Reins Co       100
Markel Corp Grp                     100
Nationwide Grp                      100
Great Amer Grp                      100
California Healthcare Ins Co Inc    100
```

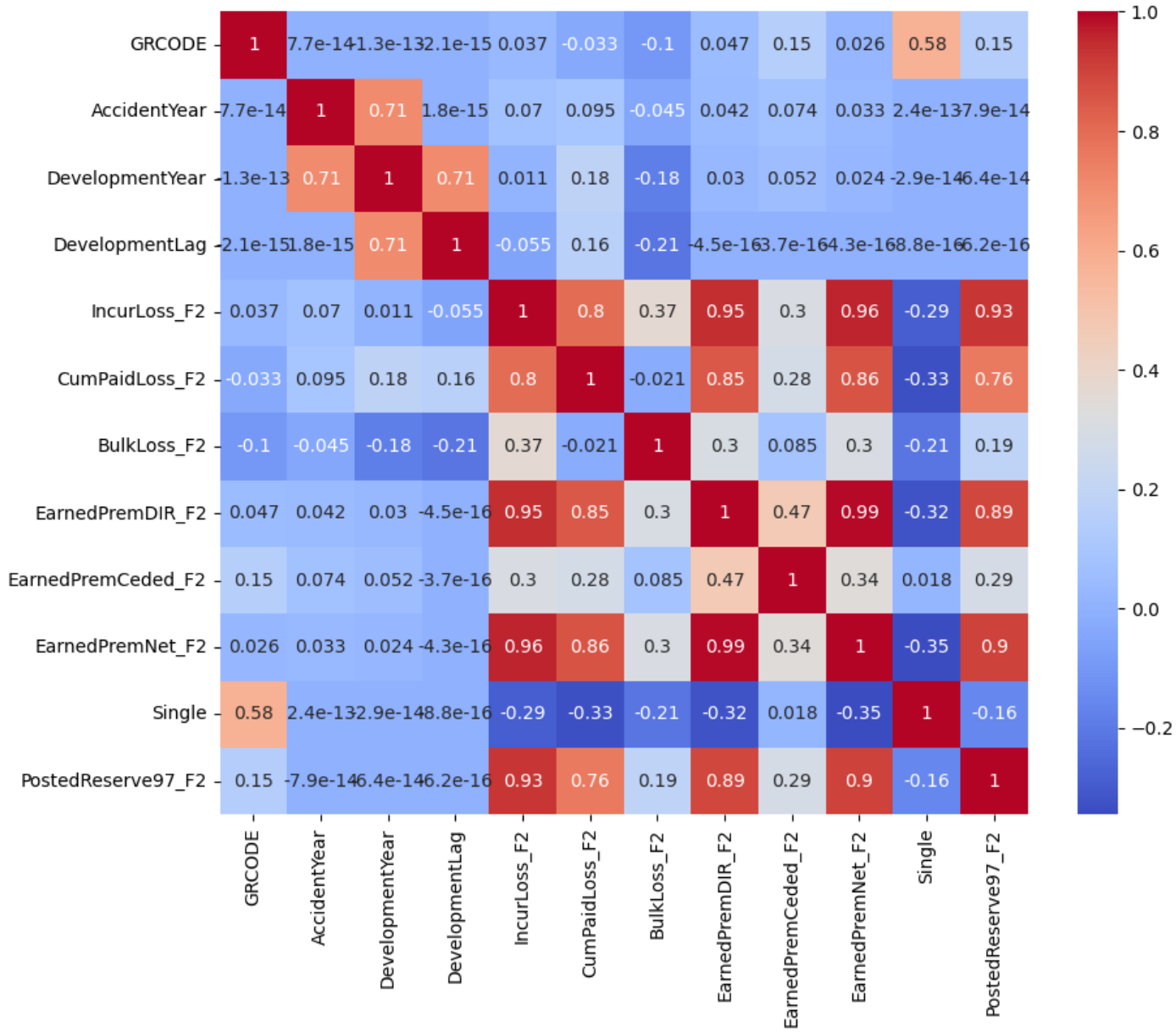
Se evidencia que la base de datos con la que estamos trabajando es una base de datos de pólizas de salud emitidas por 34 compañías de salud y cada una cuenta con 100 observaciones.

Se intenta ver las graficas descriptivas de las relaciones de las variables que queremos trabajar se realiza una serie de gráficos los cuales corresponden a un gráfico de dispersión de pares utilizando la librería Seaborn de phyton para

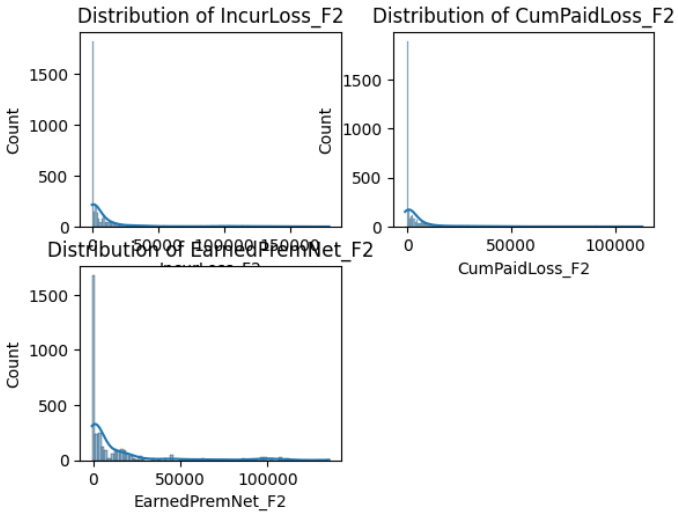
visualizar las relaciones entre las variables 'AccidentYear', 'DevelopmentYear', 'IncurLoss_F2', y 'CumPaidLoss_F2' de nuestra base de datos, Cada punto en el gráfico representa una observación en el conjunto de datos y proporciona información sobre cómo estas variables están relacionadas entre sí.



Se realiza una matriz de correlación de las variables de interés en el cual si el numero corresponde a 1 es correlación perfecta muestra a que si disminuye daría indicios de que no presenta correlación, de igual forma si los valores son negativos daría indicios de una correlación negativa entre las variables y la diagonal de la matriz corresponderá a valores de 1 ya que seria la varianza de las variables.



Posterior se buscó las gráficas de las distribuciones de variables de interés relacionadas a los pagos de la empresa y primas.



teniendo en cuenta nuestro objetivo que es realizar un trabajo para estimar por método chain ladder (método estadístico utilizado para el cálculo de la provisión de prestaciones, basado en el análisis de los triángulos de siniestros) y los triángulos de siniestros corresponden a una distribución bidimensional de la información histórica de siniestralidad. generalmente las dos dimensiones son el año de ocurrencia (eje vertical) y el año de pago (eje horizontal).

Se debe tener en cuenta la limitada base de datos con la que se cuenta es para tener en cuenta que se debe utilizar una cantidad o porcentaje de datos de los disponibles para realizar el modelo y su entrenamiento y reservar otros datos para testear el modelo y ver si se puede generalizar con datos externos a la muestra inicial por lo mismo lo primero es que se va a reservar los datos de una compañía para las pruebas el modelo y el resto para entrenamiento.

Preparación de los datos

1. Seleccionar datos

Para el caso de las compañías de seguros de pólizas de salud comerciales solo se tienen en cuenta las siguientes variables o columnas: GRCODE, GRNAME, AccidentYear, DevelopmentYear, DevelopmentLag y IncurLoss_F2.
Pues son estas las variables necesarias para crear los triángulos de reserva y aplicar los métodos de estimación planteados.

Por otro lado, se filtran filas o se eliminan registros, específicamente se eliminan de la base de datos compañías que tienen valores negativos o ceros en las variables IncurLoss_F2 pues esto no es del todo consistente con el contexto que se plantea y puede traer problemas en la etapa del modelado. Ya que si, la columna IncurLoss_F2 no se limpiase puede traer problemas en la validez de los resultados.

La eliminación o filtrado de estos datos no impide recuperar datos originales, por lo mismo al no tener en cuenta los mismos no generaría un problema relevante para nuestro estudio.

Es de relevancia tener en cuenta que la base de datos utilizada no va a presentar limitaciones o discriminación que pueda dar apertura a problemas éticos como si la realización de este estudio puede estar sesgado por el genero o identidad étnica ya que para este caso no existen campos con limitaciones para su utilización, pues no presenta segregaciones de este estilo.

2. Limpieza de datos

- Datos perdidos: Para la base de datos de las aseguradoras se tiene que no hay valores perdidos.
- Errores de datos: El único error encontrado en los datos son valores nulos o negativos en ciertas variables donde esto no puede suceder y puede afectar resultados del modelo si no se eliminan esos registros.
- Errores de medición: No hay errores de medición
- Inconsistencia de codificación: No hay inconsistencia de caracteres especiales, codificación u ortográficos
- Metadatos faltantes o incorrectos: Los metadatos son claros y confiables.
- ¿Qué tipos de ruido se produjeron en los datos?: Para este caso no hubo un ruido real, aunque se evidenciaron posibles valores atípicos el contexto del negocio pide tenerlo encuentra en el modelado de los datos, como se hico referencia anteriormente.

3. Construcción de nuevos datos

- Derivando atributos
  Para el proyecto de minería de datos no es necesario derivar o crear nuevos atributos o normalizar las variables, en otras palabras, no es necesario hacer ingeniería de características.

- Generar nuevas filas
  No es necesario utilizar métodos como la simulación para generar nuevas filas, con la cantidad que se tiene es suficiente.

4. Integrando datos

Para nuestro caso de estudio no es necesario integrar más datos o utilizar otras tablas, la tabla con la que se trabaja en el proyecto ya tiene toda la información necesaria.

5. Formato de los datos

Modelos a utilizar:
Los modelos que se escogieron para este trabajo son el método tradicional Chain-Ladder, modelo de regresión lineal normal, regresión lineal de ridge y regresión lineal de lasso.

¿Estos modelos requieren un formato u orden de datos particular?

Estos modelos no necesitan un orden en particular, pero sí que los siguientes atributos tengan cierto formato.

- GRCODE: Numérico entero
- GRNAME: String – Categorico
- AccidentYear: Numérico entero
- DevelopmentYear: Numérico entero
- DevelopmentLag: Numérico entero
- IncurLoss_F2: Numérico decimal

Estos formatos afortunadamente se tienen así desde la tabla de datos y no es necesario transformarlos.

Modelado

1. Seleccionar técnicas de modelado

- ¿Los campos de interés son de que formato?

El campo de mayor interés es IncurLoss_F2 y se necesita en un formato numérico el cual cumple con los requisitos.

- ¿Cuál es el objetivo puntual del modelo?

El objetivo es estimar la reserva o la parte inferior de un triángulo para el cálculo de reservas.

- ¿Los modelos requieren un tamaño en particular?

Los modelos no necesitan de un tamaño particular, pero entre más datos es mejor pues lo modelos de regresión lineal pueden evidenciar mejor los patrones de predicción.

- ¿Necesita modelos con resultados fácilmente presentables?

No es necesario que el modelo tenga resultados fácilmente presentables sin embargo los modelos que se plantean cumplen con esa cualidad, de igual forma la interpretación de los resultados será presentado de tal forma de que cualquier interesado pueda entenderlos.

2. Elegir las técnicas de modelado adecuadas

- ¿El modelo requiere que los datos se dividan en conjuntos de prueba y entrenamiento?

Los modelos que se implementara necesitan de tres conjuntos, entrenamiento, validación y test, el único modelo o metodología que no necesita esto es el método de Chain-Ladder al ser un modelo determinístico.

- ¿Tiene suficientes datos para producir resultados confiables para un modelo determinado?
  El modelo cuenta con los datos suficientes para obtener resultados confiables, pero si se pueden obtener más datos en un futuro es mucho mejor.

- ¿Requiere el modelo un cierto nivel de calidad de los datos? ¿Puedes alcanzar este nivel con el ¿datos actuales?
  Sí, los modelos requieren cierta calidad de datos las cual ya se alcanzó en la etapa de preparación de los datos.

- ¿Sus datos son del tipo adecuado para un modelo en particular? Si no, ¿puedes hacer lo necesario?
  Sí, los datos son del tipo adecuado para el modelo.

3. Supuestos del modelo

Supuestos

- Modelo de Chain-Ladder

El método de Chain-Ladder no tiene supuesto matemáticos, solo que la información muestre el año del siniestro, años en el que se hace la reclamación y el monto de pérdidas o Incor_loss_F2.

- Modelos de regresión lineal

Supuestos del Modelo de Regresión Lineal Normal

- Linealidad: Se asume que la relación entre las variables independientes y la variable dependiente es lineal. Esto significa que los cambios en las variables independientes se reflejan de manera proporcional en la variable dependiente.

- Independencia de errores: Los errores (residuos) deben ser independientes entre sí. Esto implica que el error en la predicción de una observación no está relacionado con el error en la predicción de otra observación.

- Homocedasticidad: La varianza de los errores debe ser constante en todos los niveles de las variables independientes. En otras palabras, la dispersión de los residuos no debe cambiar a medida que cambian los valores de las variables independientes.

- Normalidad de errores: Se supone que los errores siguen una distribución normal con una media de cero. Esto implica que la mayoría de los errores se agrupan alrededor de cero, y la distribución de errores se asemeja a una campana de Gauss.

- No multicolinealidad: Se asume que no hay multicolinealidad perfecta entre las variables independientes. Esto significa que las variables independientes no están altamente correlacionadas entre sí.

Supuestos adicionales para Ridge y Lasso:
Además de los supuestos de regresión lineal normal, Ridge y Lasso tienen algunas particularidades:

- Regularización: Ridge y Lasso introducen términos de regularización en la función objetivo. Ridge agrega una penalización L2 (norma euclidiana) a los coeficientes, mientras que Lasso agrega una penalización L1 (norma de valor absoluto). Esto se hace para evitar el sobreajuste y reducir la varianza del modelo.

- Selección de características (Lasso): Lasso, a diferencia de Ridge, tiene la capacidad de llevar a cabo la selección automática de características al forzar algunos coeficientes a cero. Esto significa que Lasso

puede eliminar variables independientes menos relevantes del modelo.

- Criterios de bondad para los modelos

Para todos los modelos se utilizará una medida para determinar la bondad de ajuste, el cual es el MAPE.

Esta métrica se aplicará al conjunto de test y es la encargada de determinar el mejor modelo. El mejor modelo es el que obtenga el MAPE más pequeño. Para el caso de regresiones lineales existen otras formas para saber si hubo un buen ajuste, como los métodos de bondad de ajuste que se basan en la determinación de distribución de probabilidad de los errores del modelo, sin embargo, para este trabajo no se tendrá en cuenta este método de verificación y es más que todo porque el método de Chain-Ladder al ser determinístico no se aplicar este método.

Diseño de prueba.

El diseño de prueba a utilizar es el Cross-Validation, para este caso este método se implementa de la siguiente manera:

1. Se escogen tres conjuntos, entrenamiento, validación y prueba, para efectos de los datos, se escoge una aseguradora como test, de las que quedan se escoge una para validar y de las que quedan se entrena el modelo.

2. Luego con de entrenar los tres modelos se escoge el mejor con el conjunto de validación aplicando la métrica MAPRE y luego se aplica el modelo al conjunto de test y se halla la métrica MAPE para el conjunto de test.

3. Luego se iteran los conjuntos, es decir, otra aseguradora pasa a ser de test, otra de validación y otras de entrenamiento hasta que todas en algún momento perteneces a los tres conjuntos.

4. Luego se escoge el modelo que haya tenido el menor MAPE en los diferentes conjuntos de test para luego compararse con el método tradicional Chain-Ladder igualmente por la métrica MAPE.

- ¿Qué datos se utilizarán para probar los modelos? ¿Ha dividido los datos en conjuntos de tren/prueba?

Para el desarrollo del proyecto se ha divido el data set en conjuntos de entrenamiento, validación y test, sin embargo, estos conjuntos no son fijo, todos los datos en algún momento son de entrenamiento, validez y de test pues se implementará.

El conjunto de validación tiene como objetivo escoger el mejor modelo de regresión de los tres planteados en cada corrida interior del Cross-Validation

El conjunto de testeo elige el mejor modelo de cada corrida exterior.

- ¿Cómo se podría medir el éxito de los modelos supervisados?

Por medio de la métrica MAPE

- ¿Cuántas veces estás dispuesto a volver a ejecutar un modelo con la configuración ajustada antes de intentar otro tipo de modelo?

El método Cross-Validation es un método iterativo completo, entonces ya se realizan todos los intentos posibles para ajustar y probar los modelos.

4. Construyendo modelos

- Configuración de parámetros

Para los modelos Ridge y Lasso se utiliza como parámetro de regularización el número 0.001 pues ya se han hecho análisis anteriores y este parámetro es el más óptimo.

- Descripción del modelo

¿Conclusiones significativas del modelo final?

El modelo final de Lasso tuvo un mejor comportamiento que el método tradicional, sin embargo, esto puede deberse a que no se trabaja con todas las aseguradoras de la base de datos, puesto que en anteriores estudios de los modelos cuando se aumenta la cantidad de aseguradoras al momento de entrenar los modelos estos pierden la posibilidad de caracterizar el comportamiento individual de cada aseguradora.

- ¿Dónde hubo problemas para la ejecución de los modelos?

La clase que calcula el Chain-Ladder es la parte que más demora en ejecutarse y como el algoritmo de los modelos de regresión dependen de esta clase y el cross-validation depende del algoritmo de los modelos de regresión está en la razón de porque el cross-validation se demora mucho en ejecutar.

- ¿Qué tan razonable fue el tiempo de procesamiento?

Se debe mejorar el entrenamiento de los modelos

- ¿El modelo tuvo problemas de calidad de datos, como datos faltantes?

No, los datos fueron de calidad

- ¿Hubo alguna inconsistencia en los cálculos?

No, los resultados son consistentes

Evaluación del modelo

- Los resultados del modelo son fáciles de implementar.
- El modelo cumple con los objetivos empresariales de encontrar una mejor forma de estimar la reserva de las compañías de seguro de automóviles comerciales.
- La opinión de otros analistas de la compañía respecto a los modelos es favorable.}

Representación gráfica del proceso:

Posteriormente a realizar la limpieza de datos, se procede a realizar el código correspondiente para hacer los distintos triángulos de siniestros con la muestra que se tiene, como ejemplo del desempeño del mismo se tomo una aseguradora con código de identificación 669, de la cual se muestran los siguientes triángulos que se generaron

Resultados para la aseguradora: 669

DevelopmentLag	1	2	3	4	5	6	7	8	9	10
AccidentYear										
1988	121905	112211	103226	99599	96006	90487	82640	80406	78920	78511
1989	122679	113165	110037	101142	90817	81919	77491	73577	72716	72317
1990	118157	117497	116377	99895	89252	81916	79134	76333	75612	75350
1991	117981	122443	121056	113795	102830	98071	94870	91062	90493	90345
1992	131059	130155	124195	113974	106817	99182	92588	91000	89256	89251
1993	134700	130757	125253	114717	111294	98014	96872	95714	96017	96047
1994	136749	128192	121355	111877	96152	91502	90498	91870	91848	91938
1995	140962	132405	118332	100050	88809	82360	81986	81887	81796	81782
1996	134473	128980	113645	104273	99276	97782	97282	97738	97601	97251
1997	137944	127727	114057	107001	102143	99665	99942	99968	99590	99378

El cuadro presentado anteriormente corresponde a el triángulo de reservas según su año y su lag y la cantidad total que se reservó, en este caso no se ha realizado la estimación de ningún tipo simplemente se realizó la representación completa de los datos.

DevelopmentLag	1	2	3	4	5	6	7	8	9	10
AccidentYear										
1988	121905	112211.0	103226.0	99599.0	96006.0	90487.0	82640.0	80406.0	78920.0	78511.0
1989	122679	113165.0	110037.0	101142.0	90817.0	81919.0	77491.0	73577.0	72716.0	NaN
1990	118157	117497.0	116377.0	99895.0	89252.0	81916.0	79134.0	76333.0	NaN	NaN
1991	117981	122443.0	121056.0	113795.0	102830.0	98071.0	94870.0	NaN	NaN	NaN
1992	131059	130155.0	124195.0	113974.0	106817.0	99182.0	NaN	NaN	NaN	NaN
1993	134700	130757.0	125253.0	114717.0	111294.0	NaN	NaN	NaN	NaN	NaN
1994	136749	128192.0	121355.0	111877.0	NaN	NaN	NaN	NaN	NaN	NaN
1995	140962	132405.0	118332.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1996	134473	128980.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1997	137944	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

La grafica presentada anteriormente corresponde a el triángulo de reservas con los datos que se tiene, pero a la mitad, ya que los valores NaN corresponden a los valores de reserva que se quieren estimar con nuestro proyecto.

Posterior a ello se procedió a realizar el calculo de los valores de los factores de cada uno de los años que se estan evaluando

0	1.963009
1	1.467341
2	1.295020
3	1.212010
4	1.161492
5	1.131422
6	1.109385
7	1.097371
8	1.090722

Posterior a realizar el calculo de los factores se procede a realizar la estimación del triangulo inferior para ver el valor que se debería tener de reservas de a cuerdo a los datos históricos con los que se tienen.

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

Y se procede a realizar un calculo de la reserva total que debería tener la aseguradora que dio un valor de 4278625.83

Se procedió a la modelación de los modelos para entrenamiento de regresión lineal, Ridge y Lasso, posterior a la modelación se procede a validar los coeficientes obtenidos por cada uno de los métodos utilizados.

```
{'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 ])}
```

Evaluación

1. Revisión de Objetivos del Negocio

- ¿Están sus resultados expresados de forma clara y en una forma que pueda presentarse fácilmente?

Sí, todos los resultados son entendibles

- ¿Hay hallazgos particularmente novedosos o únicos que deban destacarse?

No como tal, solo el hecho de que nuevas metodologías a la tradicional pueden ser una buena opción.

- ¿Puede clasificar los modelos y hallazgos según su aplicabilidad a los objetivos comerciales?

Realmente todos los modelos se pueden aplicar, pero solo se escogerá el que tenga el menor MAPE.

- En general, ¿qué tan bien responden estos resultados a los objetivos comerciales de su organización?

Muy bien, son consistentes

- ¿Qué preguntas adicionales han planteado sus resultados? ¿Cómo podría formular estas preguntas en términos comerciales?

Los resultados no han planteado nuevas preguntas debido a la sencillez y claridad del objetivo del proyecto

2. Proceso de revisión

- ¿Las etapas contribuyeron al valor de los resultados finales?

Sí, se abarco una evaluación más completa.

- ¿Existen formas de agilizar o mejorar en las etapas u operación en particular?

Si, mediante mejoras en los códigos para que corran más rápido.

- ¿Cuáles fueron los fracasos o errores en cada fase? ¿Cómo se pueden evitar la próxima vez?

Hubo una falla en la realización del Cross-Validation pues el tiempo estimado del mismo tomaba mucho, aunque posteriormente se realizo un arreglo al código para realizar la validación de forma mas eficiente y rápida.

- ¿Hubo callejones sin salida, como determinados modelos que resultaron infructuosos? ¿Existen formas de predecir esos callejones sin salida para que los esfuerzos puedan dirigirse de manera más productiva?

Solo hubo uno y fue el tiempo de procesamiento o de entreno de los modelos, para próxima se puede mejorar estos aspectos con mejor código o mejores maquinas o computadores, sin embargo, esto no impidió realizar el proyecto de minería de datos y comparar, probar y escoger el mejor modelo

- ¿Hubo sorpresas (tanto buenas como malas) durante las fases? En retrospectiva, ¿existe una manera obvia de predecir tales sucesos?

La sorpresa del largo tiempo en maquina en la etapa de diseño de prueba.

- ¿Existen decisiones o estrategias alternativas que podrían haberse utilizado en una fase determinada?

Sí, mejorar ciertos códigos para tener un procesamiento más rápido.

3. Determinar los próximos pasos

Al hacer la validación de los modelos evaluados se evidencio que el mejor modelo corresponde a un modelo Lasso, en el cual al hacer la comparación con el método tradicional de Chain-Ladder se evidenciaron los siguientes resultados del MAPE.

```
Métrica MAPE con el método Chain-Ladder: 7.036457837171299
Métrica MAPE con el modelo final: 0.9320003843622218
```

Adicional se hace una comparación entre los datos que se usaron de entrenamiento y los datos del modelo ajustado Lasso para ver que los valores estimados son muy cercanos y está dando resultados coherentes.

	Y_test	Y_ajustado modelo final
0	9.859118	9.792073
1	9.687133	9.864140
2	9.767611	9.723448
3	9.743495	9.741038
4	9.552653	9.766971
5	9.673005	9.626279
6	9.578173	9.620519
7	9.609049	9.685024
8	9.537123	9.710957
9	9.642967	9.570265
10	9.494240	9.430540

Se evidencian una mejora en la métrica del método tradicional frente a evidenciado e implementado en nuestra modelación, estos resultados pueden significar una mejora representativa en mediante el uso de nuestro modelo implementado.

Mas sin embargo el cumplimiento de los Objetivos comerciales que en este caso correspondería a una mejora en cuanto a el método de la reserva, haciendo que los montos que se establezcan sean más acordes a los escenarios reales que se pueden presentar deje como resultado una redistribución eficiente de los recursos financieros de los cuales dispone la empresa y sea así una fuente más fructífera de rendimientos tanto financieros como en el desempeño de su actividad en el mercado asegurador. Pero la mejora se verá reflejada en el transcurso, despliegue e implementación del modelo

Ahora el siguiente para finalizar el proyecto que nos atañe será implementar el modelo final, observar su comportamiento, seguir evaluándolo y esperar que sea un éxito para estimar las reservas de las pólizas de salud comerciales emitidas por la compañía.