Seguros SARR

26 de octubre de 2023

Por: Brian Felipe Sarmiento Valencia

CRISP DM – Modelamiento de reservas de seguros comerciales de autos con una metodología distinta a Chain-Ladder

Introducción

Existen un tipo de compañías que trabajan con la incertidumbre diaria que genera la propia vida o naturaleza, específicamente trabajan con eventos aleatorios inesperados que normalmente pueden ser funestos y toman como nombre de **siniestros**, estas compañías se ocupan de ofrecer el producto del seguro, el cual consiste en dar un respaldo económico a las personas, familias o entidades si algún siniestro suceda, por ejemplo, la existencia de un posible terremoto que genere daños y pérdidas es un evento inesperado y que contiene incertidumbre o aleatoriedad debido a que no se sabe en qué momento ocurre el siniestro, igualmente, también es aleatorio el valor del daño causado; para este tipo de eventos las empresas de seguros ofrecen "seguros" que protegen a las personas de estas posibles pérdidas cuando acontece el siniestro.

Específicamente, un seguro es un pago que hace un cliente a la aseguradora de acuerdo a unos intervalos de tiempo, esto a cambio de que cuando ocurra el siniestro la compañía de seguros cubra los daños causados por el evento inesperado.

Existen muchos tipos de seguros, dependiendo si son a corto o largo plazo, que siniestro cubre, como vida, accidentes automovilísticos, viajes, desempleo, entre otros, o dependiendo del cliente al que se cubra.

Otra característica que tienen los seguros es que a la hora de vender sus productos tienen en cuenta probabilidades de riesgo de cada cliente para dar un servicio muchas más específico, es decir, normalmente las aseguradoras miden características de sus clientes para estimar o cuantificar el riesgo de un evento y así poder asignarle un producto acorde a las cualidades del consumidor.

Luego de que las aseguradoras calculan estos riesgos el otro calculo que realizan es el de la reserva, la reserva es la cantidad de dinero que deben guardar las aseguradoras para poder responder por los diferentes siniestros de sus clientes que pueden ocurrir, 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, por otro lado, la forma en cómo se calcula esta 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 de autos

Los seguros de responsabilidad comercial de autos 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.

Comprensión del Negocio

- **Descripción general de la comprensión empresarial**

  Para lograr un descripción general de la comprensión empresarial exitosa se deben plantear diferentes pasos y fases que se debe argumentar como definir las personas clave de la organización, los diferentes objetivos, definir el personal, recursos, riesgos, continencias, costos, beneficios, entre otros pasos para luego finalizar con un plan del proyecto que define los tiempos de todo el proyecto de minería de datos.

- **Determinación de objetivos empresariales**

  A medida que van surgiendo nuevos métodos para la estimación y predicción de diferentes problemas por medio de modelos estadísticos y de Machine Learning las compañías de seguros se van interesando cada vez más en estas nuevas técnicas debido a los buenos resultados que ha tenido el aprendizaje de máquinas, Así mismo, las compañías que venden seguros comerciales automovilísticos desean utilizar estos nuevos métodos para poder estimar la reserva y minimizar los errores o la distancia entre la reserva estimada y la real, lo cual se traduce en una mayor eficiencia de recursos, mayores ganancias, confianza del cliente, estabilidad económica, cumplimiento normativo y atracción de inversionistas.

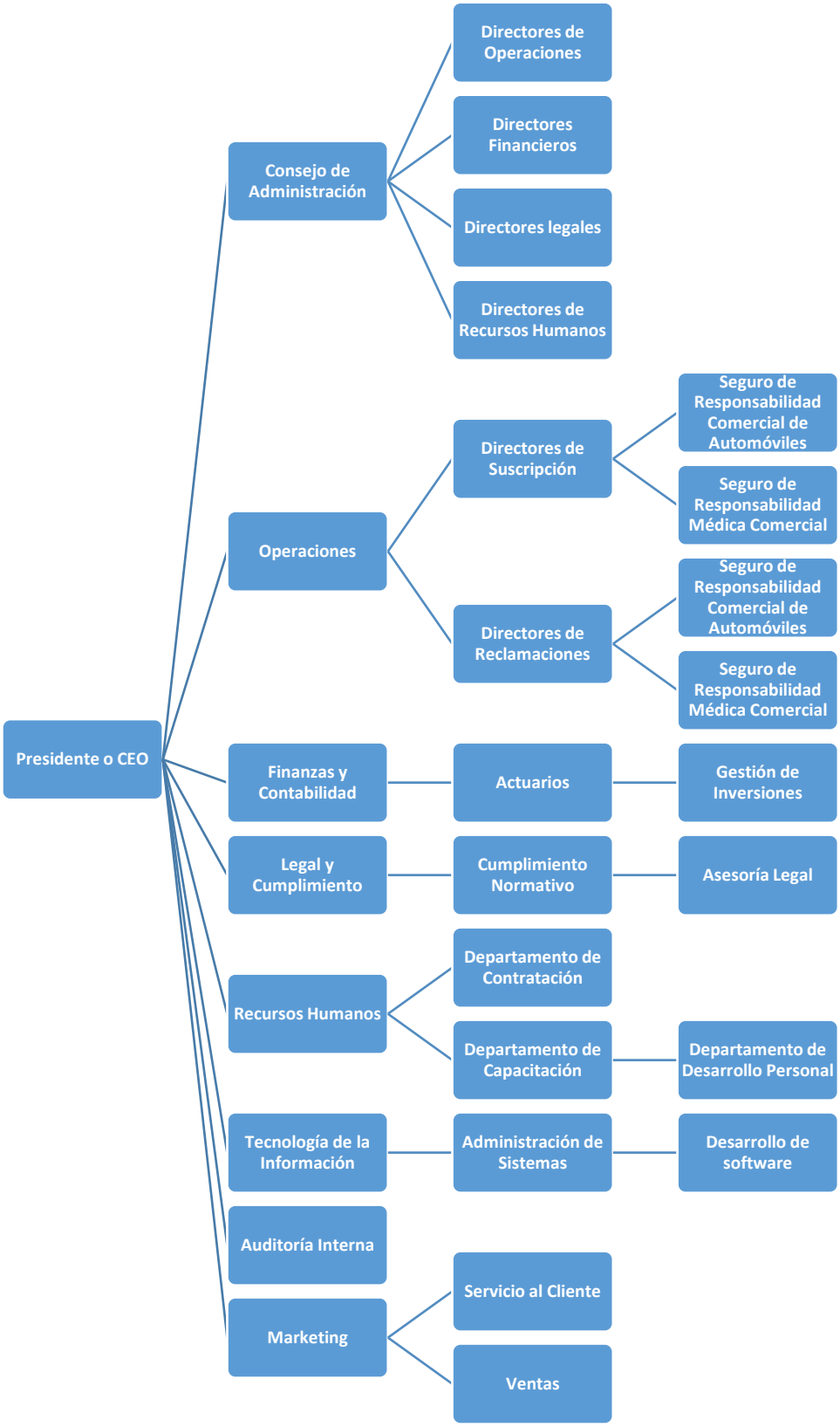
  Para lograr el objetivo de minimizar los errores de la estimación de la reserva se procede a realizar o aplicar toda una metodología completa o proyecto de minería de datos y cuyo éxito será valorado si las estimaciones de la reserva con las nuevas metodologías mejoran considerablemente, concretamente 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 aseguradora tiene un buen historial respecto a los pagos de sus seguros a sus clientes y en el pago de sus deudas, lo cual lo ha llevado a tener buenas relaciones comerciales con otras compañías del sector bancario, además la estructura empresarial ha cambiado pocas veces y su solvencia es buena aunque se quiere mejorar con la estimación de las reservas de cada año.

   En particular, la compañía de seguros ofrece una cobertura nacional de sus productos, es reconocida por la atención al cliente y la rápida respuesta cuando ocurre un siniestro lo cual lo lleva a ser una empresa con alta experiencia en el mercado y que tiene la reputación de cumplir con las regulaciones y requisitos gubernamentales**.**

- **Determinar la estructura organizacional**



Personas Clave en la organización

Para esta empresa de seguros las personas clave dentro de la organización para hacer el nuevo cálculo de reservas son:

- Ceo
- Consejo de Administración
- Gerencia de Finanzas y Contabilidad
- Gerencia de Tecnología y la Información

Patrocinadores Internos para apoyo financiero o con experiencia para el Proyecto de minería de datos

- Ceo o Gerente General es un patrocinador financiero del proyecto
- Gerente de Finanzas y Contabilidad es un patrocinador financiero y con experiencia en el tema de la estimación de la reservas
    - Grupo de actuarios ofrecen su experiencia en el cálculo de la reserva para la aseguradora
- Gerente de Tecnología y la Información ofrece su experiencia
    - Grupo de informáticos ofrecen experiencia en sistemas, programación y modelos estadísticos y de machine Learning

Unidades de negocio que se verán afectadas por el proyecto de minería de datos

Las unidades de negocios más afectadas son la gerencia de Finanzas y Contabilidad puesto que se hará un cambio en el cálculo de la estimación de la reserva y la Gerencia General debido a que el nuevo cálculo de la reserva puede generar mejores estimaciones y mayor solvencia económica.

- ### Describir el área problemática

El problema principal es mejorar la estimación de la reserva para los años posteriores por medio de un proyecto de minería de datos, cuya área encargada de este problema es la gerencia de Finanzas y Contabilidad que tiene a cargo el grupo de Actuaría de la organización.

Para el proyecto de minería de datos que ayudará a solucionar la problemática se necesita previamente una preparación académica y aplicada sobre las metodologías de la minería de datos en los profesionales de actuaría de la organización, pues son estos los que se encargaran de mejorar la estimación de la reserva.

La empresa ya antes ha utilizado proyecto de minería de datos, más que todo en el área de marketing para desarrollar modelos de Churn o tasa de cancelación de clientes, por lo tanto la aseguradora aprueba y apoya el proyecto de minería de datos para el problema en cuestión.

- ### Describir la solución actual

La solución actual a la estimación del cálculo de la reserva para años posteriores se hace por medio del método Chain-Ladder que se basa en datos históricos de los siniestros para organizarlos en un elemento llamado triángulo de siniestros que luego se le aplica el método de la cadena que se basa en el supuesto de que la relación entre las reservas de un período y las de otro período consecutivo se mantiene constante a medida que se avanza en el tiempo para último extrapolar los resultados de las reservas futuras en la cartera de seguros.
Normalmente esta técnica es muy aceptada dentro de la organización debido a la facilidad de su uso pero tiene la desventaja de que a la hora de estimar reservas se basa en suposiciones que pueden

no ser ciertos, por eso en la literatura se recomienda aplicar este método en conjunto con otras herramientas, por tal motivo, el proyecto de minería de datos desea crear un mejor método para la estimación de la reserva que no dependa de supuesto tan fuertes y sea más robusto

2. Definición de objetivos comerciales

- Mejorar la solvencia económica de la aseguradora
- Mejorar la competitividad en el mercado
- Maximizar la satisfacción del cliente
- Innovación dentro de los procesos de la empresa

3. Criterios de éxito empresarial

- **Objetivo**

  Mejorar la estimación de la reserva por medio de nuevas metodologías de la minería de datos en comparación con el método de Chain-Ladder y que produzca un rendimiento mejor a largo de los años.

- **Subjetivo**

  Mejorar la solvencia económica de la aseguradora para poder gestionar mucho mejor los recursos de la compañía.

- **Evaluación de la situación**

Para este proyecto la aseguradora tiene a su disposición todo el histórico de los datos sobre los siniestros y sobre seguros de autos comerciales como valores de las primas, valores de los siniestros, fecha del accidente, pérdidas acumuladas pagadas y gastos asignados, entre otros.

Actualmente la empresa aseguradora tiene a su disposición los profesionales y personal necesario para completar el proyecto de minería de datos puesto que cuenta con profesionales expertos en el cálculo de reservas y personal calificado en metodologías de minería de datos

Los mayores riesgos que tiene el proyecto es implementar un método que resulte peor que el método tradicional y que la estimación de la reserva no sea mejor, lo cual conlleva a que la empresa pueda caer en serios problemas de solvencia económica que a la larga puede llevar a la aseguradora a tiempos difíciles complicados de superar.

El plan de contingencia consiste es desistir en el modelo de minería de datos que se construya si los objetivos no se cumplen anualmente y volver rápidamente a utilizar el método tradicional de Chain-Ladder.

1. Inventario de Recursos

- **¿Qué hardware necesita soportar?**

  Para el proyecto de minería de datos se necesitan los siguientes hardware

  - Computadoras y Servidores
  - CPU (Unidad Central de Procesamiento)
  - GPU (Unidad de Procesamiento Gráfico)
  - Buena capacidad de memoria RAM
  - Unidades de estado sólido (SSD)
  - Almacenamiento en la Nube

- Buena red o conexión a internet
- Cluster de Servidores
- Copia de Seguridad

- **Identificar fuentes de datos y almacenes de conocimiento**

  Los datos en la compañía estan almacenados en los servideros de los que dispone la empresa y cuyo acceso para los profesionales que trabajarán en el proyecto es simple, pues tienen acceso en vivo y disponen del permiso de seguridad necesaria puesto que en el proyecto estan los personas del área de informática. Por otro lado, la compañía dispone de recursos para comprar o adquirir bases de datos externas.

- **Identificar recursos de personal**

  El personal encargada del Proyecto son expertos en el negocio de los seguros, cálculos actuariales, programación y datos, además se cuenta 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**

  Respecto a los requisitos legales, se tiene que las personas encargadas del plan para construir el nuevo método para el cálculo de reservas deben cumplir con ciertas normas a la hora de manipular los datos y estar alineados con los requisitos, estos requisitos son:

  - 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

- **Aclarar suposiciones**

  El Proyecto no cuenta con ninguna competencia dentro de la organización, sin embargo no se descarta que existan otras compañías que buscan el mismo objetivo, así mismo no contiene factores económicos que puedan afectar su ejecución. Por otro lado, se consideran que los datos son de calidad si son completos, correctos, coherentes, precisos y actualizados.

  Por último, respecto a los patrocinadores del proyecto, el proceso y resultados del proyecto serán mostrados a estos, enfocándose mucho más en la visualización de los resultados por medio de software de visualización de datos dinámica.

- **Verificas restricciones**

  El personal dedicado al proyecto cuenta con todas las credenciales y contraseñas necesarias para llevar a cabo el desarrollo del nuevo método.

No existen restricciones legales para utilizar los datos de los clientes y en general de la compañía gracias a las autorizaciones de tratamiento de datos a la cual se acogen los clientes, pero siempre recordando las buenas practicas sobre los datos mencionadas anteriormente.

Así mismo, el proyecto cuenta con el apoyo financiero para desarrollarse con éxito.

3. Riesgos y contingencias

A continuación se presenta una lista de riesgos posibles con su respectivo plan de contingencias

- Programación: el proyecto puede demorarse más de lo previsto, para este caso se planea redoblar esfuerzos para terminar de forma temprana, otra idea es alargar un poco la terminación del proyecto.
- Financiero: aunque es difícil que pase existe la posibilidad de que el proyecto quede sin recursos, para este caso se planea refinanciar el proyecto o en el peor de los casos terminarlo con los recursos disponibles.
- Mala calidad de los datos: Los datos pueden tener errores, para este caso se cuenta con un equipo especializado y con experiencia en la limpieza e imputación de datos.
- Resultados: si los resultados iniciales del proyecto sobre el nuevo método para calcular la reserva de la aseguradora no son tan impactantes esto reafirma el hecho de que el método tradicional es igualmente eficiente y eficaz que el nuevo y por lo tanto no hay motivo para cambiarlo o se puede pensar en otra metodología de minería de datos que pueda tener realmente mejores resultados.

4. Terminología

Para que las personas que trabajan en el proyecto minería de datos hablen el mismo idioma se dispone de un glosario con las palabras/frases más importantes o desconocidas para un proyecto de minería de datos para la estimación de la reserva en una empresa de seguros que ofrece seguros de automóviles comerciales. 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

- **Costo: Estimación de costos en Dólares**

1. Costo por Recopilación de Datos:

- Contratación de personal o servicios externos para recopilar datos históricos de seguros: $10,000.
- Adquisición de bases de datos externas relevantes para el análisis: $20,000.
- Gastos relacionados con la limpieza y preprocesamiento de datos: $10,000.

2. Despliegue de Resultados:

- Desarrollo de un sistema o plataforma para implementar el modelo de cálculo de triángulos y reservas en un entorno de producción: $5,000.
- Gastos asociados a la implementación de software y hardware necesarios para el despliegue: $1,000.
- Capacitación de personal en el uso del sistema y la interpretación de los resultados: $10,000.

3. Costos de Operación:

- Gastos recurrentes de mantenimiento y actualización del sistema de cálculo de triángulos y reservas: $15,000 al año.
- Costos de almacenamiento de datos a largo plazo: $5,000 al año.
- Servicios de consultoría o asesoría para ajustar el modelo de minería de datos: $10,000 al año.
- Licencias de software y herramientas utilizadas en el proyecto: $5,000 al año.

Costos Laborales:

- Salarios y beneficios para los científicos de datos, ingenieros de datos, actuarios y analistas involucrados en el proyecto: $15,000 al año.
- Posibles bonificaciones o incentivos para el personal en función de los resultados del proyecto: $5,000 al año.
- Gastos de contratación si se necesita personal adicional: $10,000.

Estimación total:

1. Costo por Recopilación de Datos: $40,000
2. Despliegue de Resultados: $11,500
3. Costos de Operación: $35,000
4. Costos Laborales: $30,000

Costos totales: $116,500 para todo el proyecto de minería de datos

- **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 evidencian nueva información y comprensión de los datos gracias a la exploración de los datos
  - Se toman de Decisiones Basada en Datos
  - 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 las reservas de la aseguradora se generará un modelo de machine Learning o estadístico para hacer cálculo y estimación de triángulos para cálculo de reservas anuales; las estimaciones se validarán 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	Todos los involucrados en el proyecto pertenecientes a la compañía	Cambios en la dirección de la compañía, 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,	Problemas de hardware o software, licencias, versiones de librerías del lenguaje de

		adquisición de bases de datos y almacenamiento.	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, problemas 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 estan 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 estan 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
- Automatización
- Comunidad Activa
- Gratuito y de Código Abierto

Comprensión de los datos

1. Recopilación de datos iniciales

Para la elaboración del proyecto de minería de datos la compañía de seguros cuenta con una base de datos propia que le ayudará a diseñar un nuevo modelo diferente al usual, por el momento no se necesitan comprar 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 parecen más prometedores?
  Los atributos más importantes son las variables, GRCODE, GRNAME, AccidentYear, DevelopmentYear, DevelopmentLag y IncurLoss_C.

- ¿Qué atributos parecen irrelevantes y pueden excluirse?
- Loa atributos que se pueden excluir son CumPaidLoss_C, BulkLoss_C, EarnedPremDIR_C, EarnedPremCeded_C, EarnedPremNet_C, Single, PostedReserve97_C.
- ¿Existen datos suficientes para sacar conclusiones generalizables o hacer predicciones precisas?
  Para este caso existen pocos datos pero aun así puede funcionar la generalización, sin embargo si en algún momento se tienen más datos sería mucho mejor
- ¿Hay demasiados atributos para el método de modelado que elija?
  Realmente no hay mucho atributo para el modelado a no ser que se desee modelar teniendo en cuenta las diferentes aseguradoras que hay en la base de datos
- ¿Está fusionando varias fuentes de datos? Si es así, ¿hay áreas que podrían plantear un problema cuando fusionándose?
  No se estan fusionado varias fuentes 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

3. Descripción de los datos

Para la realización del modelo de minería de datos se poseen una gran cantidad de registros de varias aseguradoras respecto al negocio asegurador de seguros de autos comerciales, específicamente se poseen 15800 registros y 13 variables, sin embargo son solo 6 columnas o atributos los realmente importantes.

Algunas variables son temporales, pues muestra años en los que ocurrieron ciertos sucesos, una variable que muestra el código de cada aseguradora, para las demás variables se tienen que todas son numéricas pues muestras valores de pérdidas, reservas y primas, exceptuando una variable binaria.

Respecto a los esquemas de codificación el único atributo que puede superponerse en la clave o código de cada aseguradora en las diferentes bases de datos sin embargo esto normalmente no ocurre en la compañía.

El conjunto de datos contiene las siguientes variables

- GRCODE: El código de la compañía NAIC (National Association of Insurance Commissioners), que incluye tanto grupos de aseguradoras como aseguradoras individuales.
- GRNAME: El nombre de la compañía NAIC, que incluye tanto grupos de aseguradoras como aseguradoras individuales.
- AccidentYear: El año de ocurrencia de las reclamaciones. Este va desde 1988 hasta 1997.
- DevelopmentYear: El año de desarrollo de las reclamaciones. También varía de 1988 a 1997.
- DevelopmentLag: El rezago de desarrollo de las reclamaciones, calculado como (AccidentYear - 1987) + (DevelopmentYear - 1987) - 1.
- IncurLoss_: Pérdidas incurridas y gastos asignados reportados al final del año.
- CumPaidLoss_: Pérdidas pagadas acumuladas y gastos asignados al final del año.
- BulkLoss_: Reservas masivas (Bulk) e IBNR (Incurred But Not Reported) sobre pérdidas netas y gastos de defensa y costos de contención informados al final del año.
- PostedReserve97_: Reservas publicadas en el año 1997, tomadas del "Underwriting and Investment Exhibit – Part 2A," incluyendo pérdidas netas no pagadas y gastos de ajuste no pagados.

- EarnedPremDIR_: Primas devengadas en el año de ocurrencia, tanto directas como asumidas.
- EarnedPremCeded_: Primas devengadas en el año de ocurrencia, cedidas a reaseguradores.
- EarnedPremNet_: Primas devengadas en el año de ocurrencia, netas (directas menos cedidas).
- Single: Un indicador que toma el valor 1 si se trata de una entidad única y 0 si se trata de una aseguradora del grupo.

Cantidad de datos

- **¿Cuál es el formato de los datos?**
  - GRCODE: Numérico entero, Este atributo se utiliza para almacenar un código numérico que identifica de manera única a una empresa o aseguradora. Puede ser un número entero.
  - GRNAME: String – Categorico, Nombre NAIC de la empresa (incluidos grupos de aseguradores y aseguradores individuales).
  - AccidentYear: Numérico entero, Año del accidente (1988 a 1997).
  - DevelopmentYear: Numérico entero, Año de desarrollo (1988 a 1997).
  - DevelopmentLag: Numérico entero, Año de desarrollo (AY-1987 + DY-1987 – 1
  - IncurLoss_C: Numérico decimal, Pérdidas incurridas y gastos asignados reportados al final del año
  - CumPaidLoss_C: Numérico decimal, Pérdidas pagadas acumuladas y gastos asignados al final del año.
  - BulkLoss_C: Numérico decimal, Reservas de Bulk e IBNR sobre pérdidas netas y gastos de defensa y contención de costos reportados al final del año.
  - EarnedPremDIR_C, Numérico decimal, Primas ganadas en el año incurrido: directas y asumidas.
  - EarnedPremCeded_C: Numérico decimal, Primas obtenidas en el año en que se produce - cedidas.
  - EarnedPremNet_C: Numérico decimal, Primas ganadas en el año en que se produce - netas.
  - Single: Binario, 1 indica una sola entidad, 0 indica una aseguradora grupal.
  - PostedReserve97_C: Numérico decimal, Reservas contabilizadas en el año 1997 tomadas del Anexo de Suscripción e Inversiones - Parte 2A, incluidas las pérdidas netas no pagadas y los gastos de ajuste de pérdidas no pagados.

- **Identifique el método utilizado para capturar los datos**
- La captura de los datos se hace por medio de los diferentes registros que deben hacer los clientes donde entregan toda su información personal y de los vehículos.
- **¿Qué tamaño tiene la base de datos (en número de filas y columnas)?**
  La base de datos cuenta con 15800 registros o filas y 13 columnas, variables o atributos.
- **¿Los datos incluyen características relevantes para la cuestión empresarial?**
  Si, los datos contienen información necesaria sin embargo muchos de los atributos se pueden excluir
- **¿Qué tipos de datos están presentes (simbólicos, numéricos, etc.)?**
  Los datos presentes son categóricos, binarios y numéricos enteros en su mayoría
- **¿Calculó estadísticas básicas para los atributos clave? ¿Qué información proporcionó esto sobre la pregunta de negocios?**
  Existen variables numéricas con asimetrías hacia la derecha, valores máximos que pueden ser atípicos, pero lo más importante que se observo es que algunas variables tienen como

mínimo valores negativos lo cual no es normal en el contexto de las aseguradoras cuyas pólizas son de automóviles comerciales.

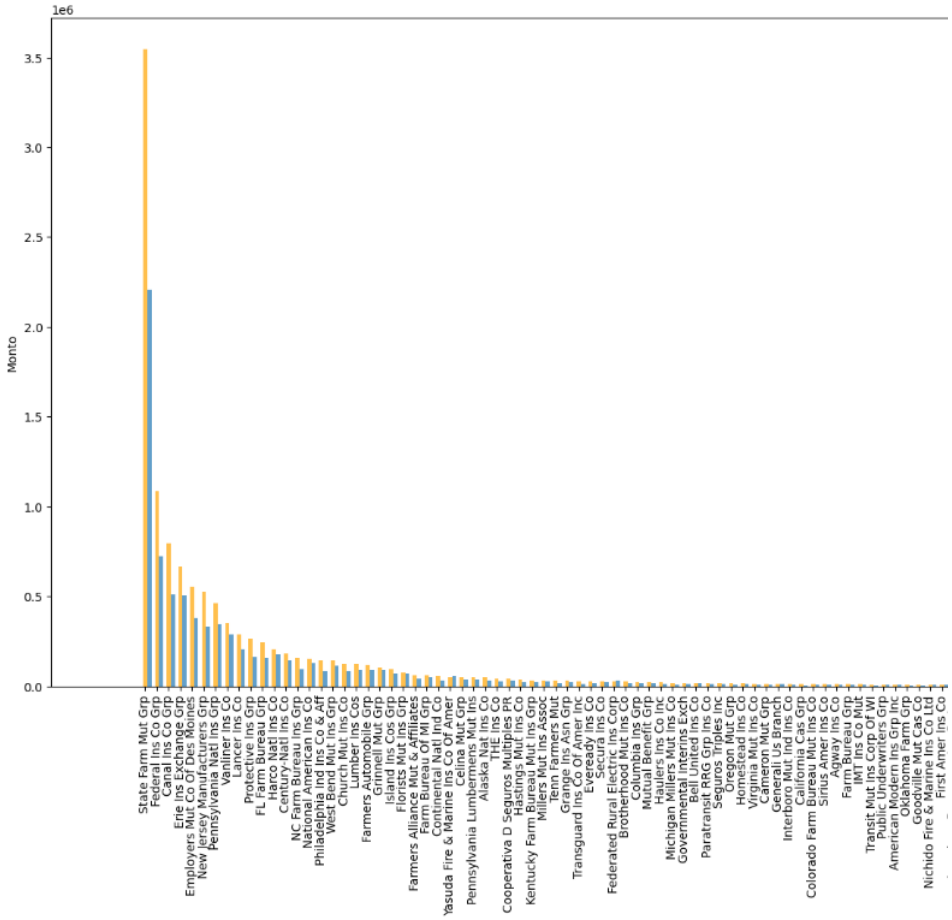
- **¿Eres capaz de priorizar atributos relevantes?**

  Se pueden priorizar los siguientes atributos:

  GRCODE, GRNAME, AccidentYear, DevelopmentYear, DevelopmentLag y IncurLoss_C

4. Explorando los datos

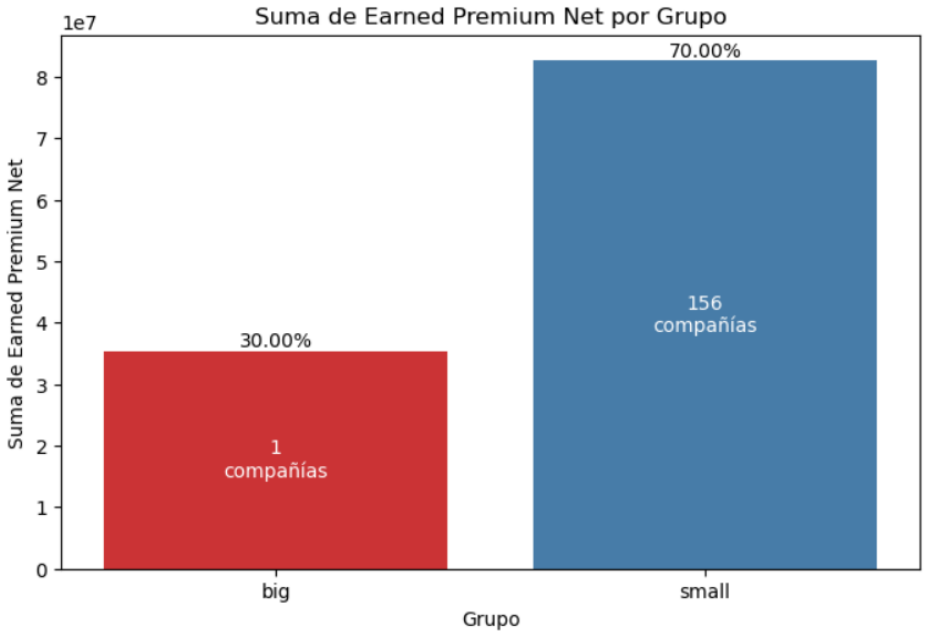
- **¿Qué tipo de hipótesis ha formado sobre los datos?**
  - ¿Qué tan rentables son las compañías?



En la gráfica, se presentan las primas emitidas durante esos 10 años en orden descendente, y se contrastan con los pagos realizados más las reservas para siniestros reportados. Se observa que algunas compañías muestran costos incurridos que superan el monto de las primas recibidas. Esto sugiere que en ese ramo de negocio específico, la rentabilidad es cuestionable, ya que los costos superan los ingresos generados por las primas emitidas. La aseguradora más grande es State Farm Mut Grp

- **¿Qué tanto acapara la compañía más grande State Farm Mut Grp el mercado?**

  Como se vio anteriormente la compañía State Farm Mut Grp es la más grande, la proxima gráfica muestra que esta compañía acapara el 30% del mercado total.



- **¿Qué compañías ceden más prima al reasegurador?**
  Las compañías de seguros suelen compartir el riesgo con reaseguradores mediante la cesión de un porcentaje de las primas. La cantidad de primas cedidas varía según varios factores, como el tipo de compañía y su percepción del riesgo, entre otros.

  A continuación, se presenta un desglose de las primas totales, distinguiendo entre primas netas y cedidas, además de clasificarlas en función del tamaño de las compañías y si estas son parte de un grupo asegurador. Se observa que las compañías más pequeñas que no pertenecen a un grupo asegurador tienden a ceder un porcentaje mayor de sus primas en comparación con aquellas que no se encuentran en esta categoría. Esto sugiere una estrategia de gestión de riesgos diferenciada entre diferentes tipos de compañías aseguradoras.

	grupo	Single	EarnedPremNet_C	EarnedPremCeded_C	Porcetaje_Cedido
0	big	0	35437960	599960	1.692987
1	small	0	60616430	11130040	18.361424
2	small	1	22075190	12110440	54.859958

- **¿Sus exploraciones han revelado nuevas características sobre los datos?**
  Sí, los datos han mostrado que hay errores en ciertas columnas como se menciona anteriormente, además muestra la compañía de seguros de autos comerciales que más acapara el mercado.
- **¿Cómo han cambiado estas exploraciones su hipótesis inicial?**
  Realmente no ha cambiado las hipótesis o ideas respecto al proyecto de minería de datos
- **¿Puede identificar subconjuntos particulares de datos para su uso posterior?**
  Por el momento se trabajará con todo el conjunto de datos, no es necesario observar subconjuntos.

5. Verificación de la calidad de los datos

Los datos es su mayoría para este proyecto son bastante buenos, no hay valores faltantes, problemas de caracteres especiales, codificación, ortográficos o errores de medición, los únicos problemas son valores

negativos o ceros en ciertas variables numéricas que no se pueden tener encuenta a la hora de realizar un modelo predictivo que estime la reserva.
Respecto al ruido en los datos se llega a la conclusión de presencia posible de valores atípicos pero que por contextos del problema y verificación de los datos se puede tener encuenta, no se normalizan las variables pues no se es necesario, por lo tanto no se realiza ningún tratamiento para el ruido.

Preparación de los datos

1. **Seleccionar datos**
   Para el caso de las compañías de seguros de pólizas de autos comerciales solo se tienen encuenta las siguientes variables o columnas GRCODE, GRNAME, AccidentYear, DevelopmentYear, DevelopmentLag y IncurLoss_C, pues son estan las necesarias para crear los triangulos 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 la variables IncurLoss_C pues esto no es del todo consistente con el contexto que se plantea y puede traer problemas en la etapa del modelado.
   - **¿Es un atributo determinado relevante para sus objetivos de minería de datos?**
     Si, los atributos importantes son GRCODE, GRNAME, AccidentYear, DevelopmentYear, DevelopmentLag y IncurLoss_C.
   - **¿La calidad de un conjunto de datos o atributo en particular impide la validez de sus resultados?**
     Sí, la columna IncurLoss_C debe limpiarse pues si eso no ocurre puede traer problemas en la validez de los resultados.
   - **¿Puedes recuperar esos datos?**
     La eliminación o filtrado de datos no impide recuperar datos, realmente ningún procedimiento impide recuperar los datos originales.
   - **¿Existen limitaciones a la hora de utilizar campos concretos como el género o la raza?**
     Para este caso no existen campos con limitaciones para su utilización.

2. **Limpieza de datos**
   - **Datos perdidos:** Para la bases de datos de las aseguradora 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 encuenta en el modelado de los datos.
   - **¿Qué enfoques utilizaste para eliminar el ruido? ¿Qué técnicas tuvieron éxito?**
     No fue necesario eliminar ruido o utiliza técnicas para eliminar el ruido como la normalización
   - ¿Hay algún caso o atributo que no se haya podido salvar?
     No, todo los atributos siempre estan a disposición
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 el caso del proyecto de minería de datos 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. **Datos de formato**
   - **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 atributo tengan cierto formato.
     - GRCODE: Numérico entero
     - GRNAME: String – Categorico
     - AccidentYear: Numérico entero
     - DevelopmentYear: Numérico entero
     - DevelopmentLag: Numérico entero
     - IncurLoss_C: 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 interes es IncurLoss_C y se necesita en un formato numérico el cual cumple
   - **¿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
   - **¿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.
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_C
     - **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, el cual se calcula de la siguiente manera

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_{i-\widehat{y_i}}}{y_i}\right|$$

Esta métrica se aplicara 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 encuenta 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 implementara.

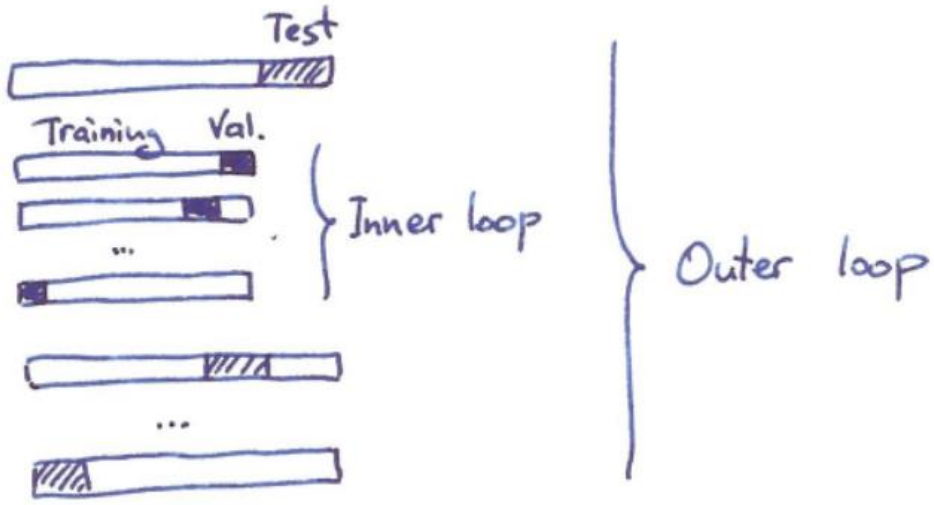
       El conjunto de validación tiene como objetivo escoger el mejor modelo de regresión de los tres planteados en cada corrida interior del Cross-Validation

       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.

     Ejemplo visual del Cross-Validation



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 toda 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**
  - Considere si los resultados de un modelo son fácilmente implementables.

    Los resultados del modelo son fáciles de implementar
  - Analice el impacto de los resultados en sus criterios de éxito. ¿Cumplen con los objetivos establecidos durante la fase de entendimiento empresarial?

    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 son favorables

5. Preguntas a tener en cuenta

- ¿Eres capaz de entender los resultados de los modelos?
  Sí
- ¿Tienen sentido para usted los resultados del modelo desde una perspectiva puramente lógica?
  Sí, los resultados del modelo son consistentes con el contexto abordado
- Desde su vistazo inicial, ¿los resultados parecen abordar la pregunta comercial de su organización?
  Los modelos respondes estas preguntas
- ¿Ha utilizado nodos de análisis y gráficos de elevación o ganancias para comparar y evaluar la precisión del modelo?
  No es necesario la métrica MAPE es suficiente
- ¿Ha explorado más de un tipo de modelo y ha comparado los resultados?
  Se exploraron en total 4 modelos y sus variaciones debido a los diferentes modelos de entrenamientos
- ¿Son implementables los resultados de su modelo?
  Todos los modelos lo son

Evaluación

1. Evaluación de los resultados

- ¿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 metodología 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 no se pudo utilizar toda la información o datos por el tiempo de procesamiento, esto se puede evitar mejorando los algoritmos para que sean más rápidos o con mejores maquinas.
- ¿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 aspecto 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

Luego de la evaluación comercial y de minería de datos se obtiene que el mejor modelo fue la regresión de Ridge cuya métrica de MAPE fue de 1.399% y coeficientes [0, 0.02791292, 0.01728219, 0.00486176, -0.01410948, -0.0270063 , -0.02865714, -0.02675069, -0.02558651, -0.02346842, -0.01723067, 0.13499618, 0.15901684, 0.14648954, 0.11313231, 0.14768033, 0.14773552, 0.16175628, -0.06189354]

Este modelo fue mejor que las demás regresiones y el método tradicional de Chain-Ladder.

Ahora el siguiente paso es implementar el modelo final, observar su comportamiento, seguir evaluándolo y esperar que sea un éxito para estimar las reservas de las pólizas de autos comerciales.