Provisiones en Seguros

2023-2

Universidad Nacional
de Colombia

Facultad de
matemáticas

Aura Alaguna
(aalaguna@unal.edu.co)

Este proyecto se centra en desarrollar un enfoque integral y preciso para la estimación de provisiones. Las provisiones son reservas financieras esenciales que una compañía de seguros debe establecer para garantizar que tenga los fondos adecuados disponibles para cubrir reclamaciones futuras de los asegurados. Nuestra meta principal es crear un marco de trabajo que combine técnicas actuales de modelado actuarial, análisis de datos y estadísticas para estimar con precisión las provisiones y gestionar los riesgos asociados. Este proyecto aborda los desafíos inherentes a la variabilidad en los tiempos de demora en la notificación y liquidación de reclamaciones, así como a la incertidumbre en la naturaleza y cantidad de futuras reclamaciones.

Índice general

Capítulo 1

Entendimiento empresarial

1.1. Objetivos empresariales

Basado en la información proporcionada sobre el desarrollo de un siniestro en un seguro de P&C y los eventos asociados, se puede determinar el siguiente objetivo general el cual se basa en asegurar la estabilidad financiera y la sostenibilidad de la compañía de seguros.

En seguida podemos ver algunos objetivos empresariales:

1. **Gestión de Reservas y Capital:** El objetivo fundamental es garantizar que la compañía de seguros tenga suficiente capital reservado para cubrir las reclamaciones futuras. Dado que las reclamaciones pueden tener retrasos en la notificación y en la liquidación, es crucial establecer reservas adecuadas para asegurar que la compañía esté financieramente preparada para hacer frente a los pagos futuros. Esto implica calcular y mantener reservas precisas y suficientes para asegurar que la empresa cumpla con sus obligaciones hacia los asegurados.

2. **Predicción de Reclamaciones y Costos:** Un objetivo importante es desarrollar modelos y herramientas analíticas que permitan predecir con precisión las reclamaciones futuras y los costos asociados. Esto implica utilizar métodos estadísticos y matemáticos para estimar la probabilidad y el monto de las reclamaciones pendientes, teniendo en cuenta los diferentes tiempos de demora y liquidación. Estas predicciones ayudan a la compañía a planificar sus reservas y recursos financieros de manera eficiente.

3. **Optimización de la Política de Reservas:** La compañía debe determinar la política adecuada para establecer reservas en función de factores como el tipo de reclamación, el historial de reclamaciones y el riesgo asociado. El objetivo aquí es encontrar el equilibrio entre mantener suficientes reservas para la solvencia financiera y evitar la sobreestimación que podría afectar negativamente la rentabilidad.

4. **Gestión de Riesgos y Solvencia:** La gestión adecuada de riesgos es crucial. La compañía debe asegurarse de que las reservas sean suficientes para enfrentar escenarios adversos, como aumentos inesperados en las reclamaciones. El objetivo es mantener la solvencia de la compañía y garantizar que pueda cumplir con sus obligaciones incluso en situaciones de estrés financiero.

5. **Eficiencia Operativa:** La comprensión de los tiempos de demora en la notificación y liquidación de reclamaciones permite a la compañía optimizar sus procesos operativos. Esto puede involucrar la identificación de áreas donde se puede reducir la demora y mejorar la eficiencia en la liquidación de reclamaciones.

6. **Desarrollo de Productos y Precios:** Comprender los patrones de desarrollo de siniestros y los tiempos de demora puede influir en el desarrollo de nuevos productos y en la fijación de precios adecuados. Los datos sobre retrasos y liquidaciones pueden ser utilizados para establecer tarifas que reflejen con precisión el riesgo y el perfil de la cartera de seguros.

En conjunto, los objetivos empresariales giran en torno a la gestión financiera sólida, la predicción precisa de riesgos y costos, la optimización de las reservas y la eficiencia operativa en el contexto de los seguros de propiedad y accidentes. Estos objetivos trabajan en conjunto para asegurar que la compañía de seguros pueda mantener su estabilidad financiera, cumplir con sus obligaciones y operar de manera rentable en el mercado.

1.2. Estructura organizacional

Dentro del contexto del problema de estimación de provisiones en seguros de propiedad y accidentes, es importante determinar la estructura organizativa para asegurarse de que todas las partes relevantes estén involucradas y se puedan tomar decisiones informadas. Aquí está la estructura organizativa sugerida:

1. **Dirección Ejecutiva y Alta Gerencia:** - Director Ejecutivo (CEO): Responsable de la dirección estratégica de la empresa y de tomar decisiones clave. - Vicepresidentes y Directores Generales: Supervisan áreas específicas de la compañía, como finanzas, operaciones y seguros.

2. **Área de Seguros y Actuarial:**

   - Director de Seguros: Encargado de la gestión global de las operaciones de seguros de propiedad y accidentes.
   - Actuarios Principales: Lideran el equipo de actuarios y son responsables de la estimación de provisiones y la gestión de riesgos.
   - Analistas Actuariales: Trabajan en la recopilación y análisis de datos, modelado y cálculo de provisiones.

3. **Comité de Riesgos y Finanzas:**

   - Comité formado por miembros de la alta gerencia, actuarios y representantes financieros.
   - Supervisa y toma decisiones sobre cuestiones relacionadas con la estimación de provisiones, solvencia y riesgos.

4. **Unidades de Negocio Afectadas:**

- Gerentes de Área: Responsables de áreas específicas de la empresa que serán impactadas por las estimaciones de provisiones.
- Ejecutivos de Ventas: Informan sobre el volumen de ventas y nuevas pólizas, lo que afecta la generación de datos para las estimaciones.

5. **Tecnología de la Información:**

   - Director de TI: Encargado de la infraestructura tecnológica necesaria para la recopilación y el análisis de datos.
   - Equipo de Desarrollo: Implementa herramientas y sistemas para gestionar y procesar datos relevantes.

6. **Departamento legal:**

   - Abogados especializados en regulaciones de seguros y compliance.
   - Asesoran sobre las implicaciones legales y regulatorias de las estimaciones de provisiones.

7. **Comunicaciones y Marketing Interno:**

   - Responsables de comunicar los objetivos y beneficios del proyecto de estimación de provisiones a todo el personal.
   - Garantizan la colaboración y comprensión en toda la organización.

8. **Recursos Humanos:**

   - Equipo de Recursos Humanos: Apoya en la gestión de personal, formación y contratación de expertos actuariales y analistas.

9. **Auditoria Interna:**

   - Asegura la integridad y precisión de los datos y los procesos de estimación de provisiones.

La estructura organizativa deberá ser adaptable y flexible para permitir la colaboración y el flujo de información entre todas estas áreas. Un enfoque interdisciplinario y una comunicación efectiva serán esenciales para abordar de manera exitosa el desafío de las provisiones en seguros de propiedad y accidentes.

1.3. Evaluación de la Situación

Un Escenario de Minería de Datos para Estimación de Provisiones Utilizando CRISP-DM

La compañía de seguros se adentra por primera vez en la minería de datos para estimación de provisiones y ha decidido consultar a un especialista en minería de datos para recibir orientación. Una de las primeras tareas que enfrenta el consultor es evaluar los recursos disponibles para llevar a cabo la estimación de provisiones.

Personal: La empresa cuenta con expertos internos en el área de seguros y actuaría que manejan las reclamaciones y datos históricos, pero se reconoce que se necesita más experiencia en modelado estadístico y análisis actuarial. Por lo tanto, también se podría considerar la consulta con analistas de datos y actuarios especializados. Dado que se espera que los resultados impacten la toma de decisiones continua, la administración debe evaluar si los roles creados para este proyecto serán permanentes.

Datos: Dado que la compañía tiene un historial extenso, hay una gran cantidad de datos históricos de reclamaciones y siniestros. Para este proyecto inicial, la atención se centrará en los seguros. Si los resultados son exitosos, se podría considerar expandir el análisis a otros tipos de seguros.

Riesgos: Además de los costos financieros asociados con la consultoría y el tiempo invertido por el personal, no se observa un riesgo inmediato significativo. Sin embargo, dada la importancia del tiempo, el proyecto se ha programado para un período de evaluación inicial de un trimestre financiero. Además, dada la situación financiera actual, es esencial que el proyecto se ajuste al presupuesto asignado. En caso de que cualquiera de estos objetivos esté en riesgo, la administración de la empresa sugiere la posibilidad de reducir el alcance del proyecto.

1.3.1. Inventario de recursos

Buscar recursos de hardware:

Para respaldar la estimación de provisiones en seguros de propiedad y accidentes, se requiere un hardware adecuado que sea capaz de manejar el procesamiento de datos complejos y el cálculo de modelos actuariales. Esto podría incluir:

- **Servidores Potentes**: Para procesar grandes volúmenes de datos históricos de reclamaciones y siniestros de manera eficiente.
- **Almacenamiento de Datos**: Capacidad suficiente para guardar y acceder a los datos históricos y modelos generados.
- **Procesadores de Alto Rendimiento**: Para realizar cálculos actuariales y análisis estadísticos complejos en tiempo razonable.
- **Memoria RAM Suficiente**: Para cargar y manipular grandes conjuntos de datos de manera efectiva.
- **Hardware Gráfico (si es necesario)**: Para visualización de datos y modelos.
- **Conexión a Redes y Bases de Datos**: Para acceder a las fuentes de datos y sistemas necesarios.
- **Seguridad y Respaldo**: Implementación de medidas de seguridad para proteger los datos confidenciales y asegurar respaldos regulares.

Es fundamental contar con hardware que cumpla con los requisitos de procesamiento y almacenamiento, ya que la estimación de provisiones involucra el análisis de datos históricos y la ejecución de modelos complejos para predecir futuros eventos en el ámbito de seguros de propiedad y accidentes.

Identificar las fuentes de datos y los almacenes de conocimientos

Fuentes de Datos Disponibles

Para la estimación de provisiones en seguros de propiedad y accidentes, es crucial conocer las fuentes de datos disponibles. Esto incluye:

- **Registros de Reclamaciones y Siniestros**: Datos históricos de reclamaciones y siniestros, que proporcionan información sobre tipos de eventos, montos de reclamaciones, tiempos de demora, entre otros.
- **Información de Pólizas y Clientes**: Datos sobre los clientes asegurados, detalles de pólizas y coberturas contratadas.
- **Datos Demográficos y Socioeconómicos**: Información externa como datos demográficos, estadísticas económicas y sociodemográficas que podrían impactar las reclamaciones y siniestros.
- **Estadísticas del Sector**: Datos y métricas relacionadas con la industria de seguros en general, que podrían aportar contexto.
- **Datos de Regulación**: Información sobre regulaciones y requisitos legales que podrían influir en las provisiones.

Almacenamiento de Datos

Los datos se almacenan de diversas formas, como:

- **Bases de Datos Internas**: Bases de datos específicas de la empresa que almacenan información sobre pólizas, reclamaciones y siniestros.
- **Almacenes de Datos**: Repositorios centralizados de datos históricos para análisis, que podrían ser utilizados en tiempo real o en lotes.
- **Fuentes Externas**: Posiblemente, se acceda a fuentes externas para adquirir datos demográficos y estadísticas sectoriales.

Es importante determinar si tienes acceso en tiempo real a almacenes de datos o bases de datos operativas, ya que esto afectará la velocidad y frecuencia de actualización de los análisis.

Adquisición de Datos Externos

Se debe considerar si se planea adquirir datos externos, como información demográfica o estadísticas de reclamaciones, para enriquecer el análisis de provisiones. Estos datos pueden ser valiosos para comprender mejor los factores que influyen en las reclamaciones y mejorar la precisión de las estimaciones.

Identificar al personal con recursos

Acceso a Expertos en Negocios y Datos

Es esencial contar con acceso a expertos en negocios y datos para llevar a cabo con éxito la estimación de provisiones en seguros de propiedad y accidentes. Estos expertos pueden aportar conocimientos sobre la industria aseguradora, los productos de seguros y los detalles específicos de las reclamaciones y siniestros. Su experiencia contribuirá a la interpretación precisa de los datos y la formulación de modelos actuariales efectivos.

Identificación de Administradores de Bases de Datos y Personal de Soporte

Es importante identificar a los administradores de bases de datos y otro personal de soporte que puedan ser necesarios para respaldar el proceso de estimación de provisiones. Los administradores de bases de datos desempeñan un papel crucial en la gestión y organización de los datos requeridos para el análisis. Además, se debe considerar la posibilidad de contar con personal de soporte técnico y analistas de datos para asegurar un flujo de trabajo fluido y una interpretación precisa de los resultados.

1.3.2. Requisitos, supuestos y limitaciones

Requisitos

Existencia de Restricciones de Seguridad o Legales

Es fundamental determinar si existen restricciones de seguridad o legales que puedan afectar los datos o resultados del proyecto de estimación de provisiones en seguros de propiedad y accidentes. Estas restricciones podrían incluir regulaciones de privacidad de datos, acuerdos de confidencialidad o requisitos de cumplimiento. Es esencial tener en cuenta estas restricciones para garantizar que el manejo de los datos y los resultados cumpla con todas las normativas aplicables.

Alineación con Requisitos de Programación del Proyecto

Es importante confirmar si todos los involucrados en el proyecto están alineados con los requisitos de programación. Esto abarca desde el equipo de desarrollo y análisis hasta los patrocinadores y partes interesadas. Asegurarse de que todos comprendan y estén comprometidos con los plazos y hitos del proyecto es esencial para una ejecución exitosa.

Existencia de Requisitos sobre la Implementación de Resultados

Es relevante determinar si existen requisitos específicos sobre cómo se implementarán y presentarán los resultados del proyecto. Por ejemplo, si se planea publicar los resultados en la web o almacenarlos en una base de datos, es importante identificar estos requisitos desde el principio. Esto ayudará a planificar la fase de implementación y asegurará que los resultados sean entregados de manera efectiva según las necesidades de las partes interesadas.

Supuestos

Existencia de Factores Económicos que Puedan Afectar al Proyecto

Es relevante evaluar si existen factores económicos que puedan influir en el proyecto de estimación de provisiones en seguros de propiedad y accidentes. Esto podría incluir costos asociados a consultoría, adquisición de datos externos, o la competencia

de productos similares en el mercado. Considerar estos factores permitirá anticipar posibles impactos financieros en el proyecto.

Supuestos sobre la Calidad de los Datos

Es esencial identificar si hay supuestos sobre la calidad de los datos utilizados en el proyecto. Estos supuestos podrían abarcar la integridad, precisión y consistencia de los datos históricos de reclamaciones y siniestros. Reconocer y aclarar estos supuestos ayudará a evitar posibles sesgos o imprecisiones en los resultados del análisis.

Expectativas del Patrocinador/Equipo Directivo sobre la Visualización de Resultados

Es importante comprender cómo el patrocinador del proyecto y el equipo directivo esperan ver los resultados. ¿Desean tener un entendimiento detallado del modelo actuarial y estadístico utilizado en el proceso? ¿O simplemente desean ver los resultados finales de las estimaciones de provisiones? Aclarar estas expectativas será crucial para definir la presentación y nivel de detalle de los resultados entregados.

Limitaciones

Acceso a Contraseñas Necesarias

Es esencial asegurarse de tener todas las contraseñas necesarias para acceder a los datos requeridos para la estimación de provisiones en seguros de propiedad y accidentes. Esto incluye acceso a bases de datos, sistemas de almacenamiento y cualquier recurso necesario para realizar análisis y modelado. La falta de acceso podría retrasar o impedir el progreso del proyecto.

Verificación de Restricciones Legales sobre el Uso de los Datos

Es crucial verificar y cumplir todas las restricciones legales que afecten el uso de los datos en la estimación de provisiones. Esto podría incluir regulaciones de privacidad, acuerdos de uso de datos y consideraciones legales que dictan cómo se pueden utilizar los datos y los resultados generados. Cumplir con estas restricciones es esencial para evitar problemas legales y asegurar la integridad del proyecto.

Consideración de Restricciones Financieras en el Presupuesto del Proyecto

Es importante garantizar que todas las restricciones financieras relevantes estén contempladas en el presupuesto del proyecto de estimación de provisiones. Esto podría incluir los costos asociados con la adquisición de datos externos, honorarios de consultoría, licencias de software y otros gastos relacionados. Asegurarse de que el presupuesto sea realista y completo es fundamental para evitar sorpresas financieras a lo largo del proyecto.

1.3.3. Riesgos y contingencia

- **Programación (¿Y si el proyecto dura más de lo previsto?):** Riesgo: Retraso en la programación podría afectar los plazos y la entrega del proyecto. Plan de Contingencia: Mantener un seguimiento constante de los hitos y el progreso del proyecto. Si se identifican retrasos, evaluar la posibilidad de asignar más recursos o redefinir prioridades para acelerar el avance.

- **Financieros (¿Y si el patrocinador del proyecto tiene problemas presupuestarios?):** Riesgo: Limitaciones financieras podrían afectar la ejecución completa

del proyecto. Plan de Contingencia: Establecer una reserva financiera en el presupuesto para manejar imprevistos. En caso de limitaciones presupuestarias, priorizar tareas críticas y explorar oportunidades de financiamiento adicional.

- **Datos (¿Y si los datos son de mala calidad o cobertura?):** Riesgo: Datos insuficientes o de baja calidad podrían afectar la precisión de las estimaciones. Plan de Contingencia: Realizar un análisis exhaustivo de calidad de datos al inicio del proyecto. Si se detectan problemas, considerar la posibilidad de mejorar la calidad de los datos o ajustar los modelos para trabajar con limitaciones.
- **Resultados (¿Y si los resultados iniciales son menos espectaculares de lo esperado?):** Riesgo: Resultados iniciales por debajo de las expectativas podrían desalentar a los interesados. Plan de Contingencia: Mantener una comunicación transparente con los interesados. Enfocarse en resaltar los aspectos positivos y explicar cualquier diferencia en las expectativas iniciales. Identificar áreas de mejora y trabajar en ajustes.

Documentación de Posibles Riesgos

Es esencial identificar y documentar cada posible riesgo asociado al proyecto de estimación de provisiones en seguros de propiedad y accidentes. Algunos ejemplos de posibles riesgos podrían incluir:

- Cambios en las regulaciones de seguros que afecten los cálculos de provisiones.
- Problemas de calidad de datos que afecten la precisión de las estimaciones.
- Falta de alineación entre los equipos de negocios y técnicos.
- Limitaciones tecnológicas que dificulten el procesamiento de grandes conjuntos de datos.
- Retrasos en la adquisición de datos externos.

Plan de Contingencia para Cada Riesgo

Para cada riesgo identificado, es necesario elaborar un plan de contingencia que describa cómo se abordará el riesgo si llega a materializarse. Por ejemplo:

- Para el riesgo de cambios en las regulaciones, se establecerá un equipo de monitoreo que se mantenga al tanto de los cambios regulatorios y adapte los cálculos de provisiones en consecuencia.
- En caso de problemas de calidad de datos, se implementarán procedimientos de limpieza y validación de datos antes de realizar cualquier estimación.
- Para mitigar la falta de alineación entre equipos, se establecerán reuniones regulares de coordinación y se definirán claramente los roles y responsabilidades.
- Frente a limitaciones tecnológicas, se considerará la posibilidad de mejorar la infraestructura tecnológica o adoptar soluciones alternativas.
- En caso de retrasos en la adquisición de datos externos, se planificará un calendario flexible y se evaluarán fuentes alternativas.

El desarrollo de planes de contingencia contribuirá a mitigar los riesgos y a mantener la ejecución exitosa del proyecto.

1.3.4. Análisis costos y beneficios

Se debe considerar los costes estimados de:

- **Recogida de Datos y Datos Externos**: Calcular los costes asociados con la recopilación de datos internos y, si aplicable, la adquisición de datos externos como información demográfica o estadísticas sectoriales.
- **Despliegue de Resultados**: Evaluar los costes relacionados con la implementación de los resultados del proyecto. Esto puede incluir la creación de informes, visualizaciones y cualquier otra forma de presentación de los resultados.
- **Costes Operativos**: Considerar los costes operativos asociados con la ejecución continua del modelo y la actualización de los datos. Esto podría incluir costes de mantenimiento de hardware y software, así como costes de personal.

Es esencial tener en cuenta tanto los costes directos como los costes indirectos que puedan surgir a lo largo del ciclo de vida del proyecto.

Este es el texto traducido a LaTeX para responder a las preguntas sobre el análisis de costes y beneficios en el contexto de la estimación de provisiones en seguros de propiedad y accidentes.

Beneficios del Proyecto de Estimación de Provisiones

- **Cumplimiento del Objetivo Principal**: Lograr el objetivo principal del proyecto de estimación de provisiones en seguros de propiedad y accidentes, que es mejorar la precisión en la estimación de reservas. Esto puede conducir a una gestión más efectiva de riesgos y una toma de decisiones informada.
- **Conocimientos Adicionales Generados por la Exploración de Datos**: La exploración de datos durante el proyecto puede generar conocimientos valiosos sobre patrones y tendencias en las reclamaciones y siniestros. Estos conocimientos pueden ser utilizados para identificar oportunidades de mejora y optimización en la gestión de seguros.
- **Posibles Beneficios de una Mejor Comprensión de los Datos**: Al mejorar la comprensión de los datos relacionados con reclamaciones y siniestros, la compañía de seguros puede tomar decisiones más informadas. Esto incluye identificar áreas de riesgo potencial, ajustar estrategias de precios y optimizar la asignación de recursos.

Los beneficios mencionados pueden contribuir a la solidez financiera de la compañía, la satisfacción del cliente y la capacidad de adaptarse eficazmente a las dinámicas del mercado.

1.4. Determinar los objetivos de la minería de datos

Objetivos de Minería de Datos para el Problema de Provisiones

Los objetivos de minería de datos para el problema de provisiones en seguros de propiedad y accidentes son los siguientes:

1. **Identificación de Clientes de Alto Valor**: Utilizando la base de datos de compras recientes, el objetivo es identificar a los clientes de alto valor. Esto implica analizar los patrones de compra y comportamiento de los clientes para determinar quiénes generan ingresos significativos para la compañía.

2. **Construcción de un Modelo de Probabilidad de Rotación**: Mediante los datos disponibles de los clientes, se busca construir un modelo que prediga la probabilidad de que cada cliente abandone la compañía (rotación). Este modelo permitirá anticipar y abordar la pérdida de clientes antes de que ocurra.

3. **Asignación de Rangos a los Clientes**: Se busca asignar a cada cliente un rango basado en dos factores clave: su propensión a la pérdida (probabilidad de rotación) y el valor del cliente. Esta asignación de rangos permitirá identificar a los clientes más valiosos en términos de retención y ganancias.

Objetivos de la minería de datos

Descripción del Tipo de Problema de Minería de Datos

El problema de provisiones involucra principalmente la predicción. Se busca anticipar futuros eventos, como el monto de las reclamaciones, con el fin de estimar las provisiones financieras requeridas. Además, también puede involucrar elementos de análisis de riesgos y optimización de carteras.

Objetivos Técnicos con Unidades de Tiempo Específicas

Los objetivos técnicos para el proyecto de estimación de provisiones se plantean con una validez de tres meses. Estos objetivos incluyen:

1. Construir un modelo de predicción de reclamaciones que sea válido para los próximos tres meses.

2. Desarrollar un enfoque estocástico para determinar la reserva pendiente con modelos lineales generalizados (GLMs) y obtener una distribución predictiva de la reserva pendiente en tres meses a través de la simulación.

Resultados Deseados con Números Reales

Se espera lograr resultados concretos en términos de retención y gestión de riesgos. Ejemplos de resultados deseados con números reales incluyen:

1. Producir puntuaciones de churn para el 80 % de los clientes existentes, lo que permitirá identificar a los clientes con mayor probabilidad de abandonar la compañía en los próximos tres meses.

2. Alcanzar una reducción del 15 % en la volatilidad de las estimaciones de provisiones, mejorando así la estabilidad financiera de la compañía.

Estos resultados cuantificables reflejarán el éxito en la aplicación de técnicas de minería de datos para el problema de provisiones en seguros.

Criterios de éxito de la minería de datos

Métodos de Evaluación del Modelo

Para evaluar la efectividad del modelo de estimación de provisiones en seguros de propiedad y accidentes, se utilizarán métodos de evaluación como:

- **Precisión y Rendimiento**: Se medirá la precisión del modelo al comparar las predicciones con los resultados reales de las reservas y reclamaciones. También se evaluará el rendimiento general del modelo en términos de tiempo de respuesta y capacidad de manejar grandes conjuntos de datos.
- **Validación Cruzada**: Se realizará validación cruzada utilizando diferentes conjuntos de datos para asegurar que el modelo sea robusto y generalice bien a datos no vistos.
- **Métricas de Error**: Se utilizarán métricas como el error cuadrático medio (MSE) o el error absoluto medio (MAE) para medir la diferencia entre las predicciones y los valores reales.

Definición de Puntos de Referencia

Para evaluar el éxito del proyecto, se definirán puntos de referencia con cifras concretas. Por ejemplo, se puede establecer como objetivo reducir el error de estimación en un cierto porcentaje en comparación con los métodos anteriores.

Medidas Subjetivas y Árbitro del Éxito

Las medidas subjetivas podrían incluir la satisfacción de los actuarios y analistas con la calidad de las predicciones y la facilidad de uso del modelo. El árbitro del éxito podría ser el comité directivo que evaluará si los resultados cumplen con los objetivos establecidos y si el modelo mejora la precisión de las provisiones.

Despliegue de Resultados y Éxito

El despliegue exitoso de los resultados del modelo es parte integral del éxito de la minería de datos. Si los resultados son implementados de manera efectiva y utilizados para mejorar la toma de decisiones y la gestión de riesgos, el proyecto se considerará exitoso.

Planificación del Despliegue

La planificación del despliegue debe comenzar temprano. Esto incluye identificar las herramientas y plataformas para presentar los resultados, diseñar informes claros y accesibles, y capacitar al personal en la interpretación y uso del modelo.

1.4.1. Plan proyecto

Fase	Tiempo (días)	Recursos	Riesgos
Definición del Alcance	14	Equipo de proyecto, patrocinador interno	Cambios en los objetivos
Recopilación de Datos	28	Analistas de datos, base de datos	Datos incompletos o inconsistentes
Análisis Exploratorio	21	Analistas de datos, expertos en seguros	Problemas de calidad de datos
Construcción de Modelos	42	Analistas de datos, actuarios	Selección inadecuada de algoritmos
Validación y Ajuste	14	Analistas de datos, equipo técnico	Overfitting del modelo
Evaluación y Métricas	7	Analistas de datos, equipo técnico	Interpretación incorrecta de resultados
Despliegue y Capacitación	14	Equipo técnico, usuarios finales	Resistencia al cambio
Monitoreo y Mantenimiento	Continuo	Equipo técnico, actuarios	Cambios en patrones de reclamaciones

Capítulo 2

Entendimiento de los datos

2.1. Recolectando datos

1. **¿Cuáles atributos (columnas) de la base de datos parecen más prometedores?**

   Los atributos que parecen más prometedores son aquellos que tienen un rango de valores amplio, contienen información relevante y presentan una variabilidad significativa entre las observaciones. Por ejemplo, `AccidentYear`, `DevelopmentYear`, `IncurLoss_D`, `CumPaidLoss_D`, `EarnedPremDIR_D` y `EarnedPremNet_D` parecen ser atributos prometedores debido a la diversidad de sus valores y su relación con el proceso de seguros y reclamos.

2. **¿Cuáles atributos parecen irrelevantes y se pueden excluir?**

   Los atributos que parecen irrelevantes podrían ser aquellos que tienen valores constantes o repetitivos, como `Single` si solo tiene un valor predominante. También, si `GRNAME` solo contiene nombres de grupos y no aporta información significativa para el análisis, podría considerarse menos relevante.

3. **¿Hay suficientes datos para sacar conclusiones generalizables o hacer predicciones precisas?**

   Para determinar si hay suficientes datos, es importante considerar la cantidad total de observaciones y la diversidad de los valores en cada atributo. Si hay un gran número de observaciones y una variabilidad sustancial en los valores de los atributos clave, es más probable que se puedan extraer conclusiones generalizables y realizar predicciones precisas.

4. **¿Hay demasiados atributos para el método de modelado elegido?**

   Si hay demasiados atributos en comparación con la cantidad de observaciones, podría haber problemas de sobreajuste y dificultades para obtener modelos confiables. En este caso, es recomendable realizar selección de características o técnicas de reducción de dimensionalidad para manejar la complejidad.

5. **¿Estás fusionando varias fuentes de datos? Si es así, ¿hay áreas que podrían ser problemáticas al fusionar?**

   Si estás fusionando datos de múltiples fuentes, es importante asegurarse de que las claves de fusión sean consistentes entre las fuentes. Las diferencias en las

representaciones de las claves podrían dificultar la fusión y requerir un proceso de limpieza y estandarización.

6. **¿Has considerado cómo se manejan los valores faltantes en cada una de tus fuentes de datos?**

   Los valores faltantes pueden afectar la calidad de los análisis y modelos. Es importante comprender cómo se manejan los valores faltantes en cada fuente de datos. Puedes utilizar métodos como la imputación, eliminación o análisis de patrones de valores faltantes para abordar este problema de manera adecuada.

2.2. Descripción de los datos

Cantidad de los datos

La base de datos se distingue por su extensión sustancial, comprendiendo un total de 13,200 observaciones distribuidas en una matriz de múltiples columnas, cada una representando diversas características. La disposición tabular de los datos, donde cada fila denota una observación específica y cada columna encapsula una característica particular, posibilita un abordaje eficiente y analítico de los mismos mediante utilidades como hojas de cálculo y bibliotecas de programación.

Respecto al método empleado para adquirir los datos, no se proporcionan detalles específicos en la información disponible. Las metodologías pueden variar desde el ingreso manual en hojas de cálculo hasta sistemas automatizados de recolección de datos, incluyendo la importación desde fuentes externas mediante protocolos como ODBC (Open Database Connectivity) y otros mecanismos de acceso a bases de datos. En términos dimensionales, la base de datos alberga 13,200 filas, representando las observaciones, y múltiples columnas, correspondientes a los atributos. La cantidad exacta de columnas no se encuentra especificada. Dada la amplitud de observaciones, se potencia la capacidad para llevar a cabo análisis sustanciales y construir modelos más precisos, siempre y cuando se realice una selección y tratamiento adecuado de los atributos involucrados.

Calidad de los datos

En relación a la calidad de los datos, se puede afirmar que estos contienen características directamente pertinentes a la cuestión comercial planteada. Entre los tipos de datos presentes, se encuentran tanto simbólicos como numéricos, abarcando una gama diversa de información. Además, se realizaron cálculos de estadísticas básicas para los atributos clave, lo que permitió obtener una comprensión más profunda sobre la distribución y las tendencias asociadas a la pregunta de negocios en cuestión. En cuanto a la priorización de atributos relevantes, se dispone de la capacidad para llevarla a cabo; no obstante, en caso de necesidad, los analistas comerciales se encuentran disponibles para aportar mayor perspicacia y contexto a este proceso.

Los datos exhiben una pertinencia inherente a la pregunta empresarial, con una diversidad de tipos de datos que abarcan lo simbólico y lo numérico. La evaluación de estadísticas fundamentales para los atributos clave ha proporcionado información

esencial para entender cómo estos se relacionan con la pregunta de negocios. La capacidad para priorizar atributos pertinentes está presente, pero en situaciones más complejas, los analistas empresariales pueden ofrecer una visión adicional y específica.

2.2.1. Exploración de los datos

Variable	Descripción	Mín.	1er Cuartil	Mediana	Media	3er Cuartil	Máx.
GRCODE	Código	86	8526	14110	17153	26983	44300
GRNAME	Nombre						
AccidentYear	Año Accidente	1988	1990	1992	1992	1995	1997
DevelopmentYear	Año Desarrollo	1988	1994	1997	1997	2000	2006
DevelopmentLag	Retraso	1.0	3.0	5.5	5.5	8.0	10.0
IncurLoss_D	Pérdida Incurrida	-59	0	544	11532	6526	367404
CumPaidLoss_D	Pérdida Pagada Acum.	-338.0	0.0	351.5	8215.7	4565.0	325322.0
BulkLoss_D	Pérdida Global	-4621.0	0.0	5.0	1570.1	259.2	145296.0
EarnedPremDIR_D	Prima Devengada Directa	-6518	0	1419	18438	11354	421223
EarnedPremCeded_D	Prima Devengada Cedida	-3522.0	0.0	144.5	1812.3	1141.0	78730.0
EarnedPremNet_D	Prima Devengada Neta	-9731	0	827	16626	9180	418755
Single	Único	0.0000	0.0000	1.0000	0.7273	1.0000	1.0000
PostedReserve97_D	Reserva Publicada en 1997	0	411	2732	39714	19266	1090093

Cuadro 2.1: Descripción de Datos

Los datos representan información relacionada con seguros y reclamos. A continuación, se detallan las estadísticas de diferentes variables presentes en el conjunto de datos:

- **GRCODE:** Los valores de esta variable representan códigos únicos de grupos. Se encuentran 10 grupos por cada código, es decir hay 132 códigos únicos. El valor mínimo es 86, el primer cuartil (25 %) es 8526 y el valor máximo es 44300.
- **GRNAME:** Esta variable contiene nombres de grupos en formato de caracteres.
- **AccidentYear:** Indica el año en que ocurrió el accidente. Los datos abarcan desde el año mínimo de 1988 hasta el máximo de 1997.
- **DevelopmentYear:** Muestra el año de desarrollo. Los valores oscilan entre 1988 y 2006.
- **DevelopmentLag:** Representa el lapso de tiempo en años entre el año del accidente y el año de desarrollo. La media es 5.5 años.
- **IncurLoss_D:** Esta variable refleja pérdidas incurridas. El valor mínimo es -59 y el valor máximo es 367404.
- **CumPaidLoss_D:** Muestra la suma acumulada de pérdidas pagadas. Los valores varían desde -338 hasta 325322.
- **BulkLoss_D:** Indica la suma de pérdidas a granel. Los datos van desde -4621 hasta 145296.
- **EarnedPremDIR_D:** Representa la prima devengada directa. Los valores van desde -6518 hasta 421223.

- **EarnedPremCeded_D:** Indica la prima devengada cedida. Los valores oscilan entre -3522 y 78730.
- **EarnedPremNet_D:** Muestra la prima devengada neta. La media es 16626.
- **Single:** Esta variable toma valores binarios (0 o 1) y representa si la entidad es un único reclamo. El 72.73 % de los casos tienen valor 1.
- **PostedReserve97_D:** Muestra las reservas publicadas en el año 1997. Los valores oscilan desde 0 hasta 1090093.

En resumen, el conjunto de datos proporciona información sobre grupos de seguros y sus características asociadas, como años de accidente y desarrollo, pérdidas incurridas, primas devengadas y otros aspectos relevantes para el análisis de seguros y reclamos.

Variable	Curtosis
GRCODE	0.561
AccidentYear	0.000
DevelopmentYear	0.000
DevelopmentLag	0.000
IncurLoss_D	5.400
CumPaidLoss_D	5.847
BulkLoss_D	9.234
EarnedPremDIR_D	4.776
EarnedPremCeded_D	8.060
EarnedPremNet_D	4.927
Single	-1.021
PostedReserve97_D	5.923

Cuadro 2.2: Curtosis de las variables

Los valores de **curtosis** positivos indican una distribución más puntiaguda (más colas pesadas) que una distribución normal, mientras que los valores negativos indican una distribución más achatada (menos colas pesadas). En este caso, las variables "BulkLoss_D", .EarnedPremCeded_Dz "PostedReserve97_D"tienen valores de curtosis significativamente positivos, lo que sugiere que tienen colas pesadas y una distribución más puntiaguda. Por otro lado, la variable "Single"tiene un valor de curtosis negativo, lo que indica una distribución más achatada. El resto de las variables tienen valores de curtosis cercanos a cero, lo que sugiere una distribución más cercana a la normal. Esta información puede ser útil para comprender la forma de las distribuciones de estas variables y tomar decisiones informadas en análisis estadísticos o modelado.

Figura 2.1: Media de la variable IncurLoss_D)

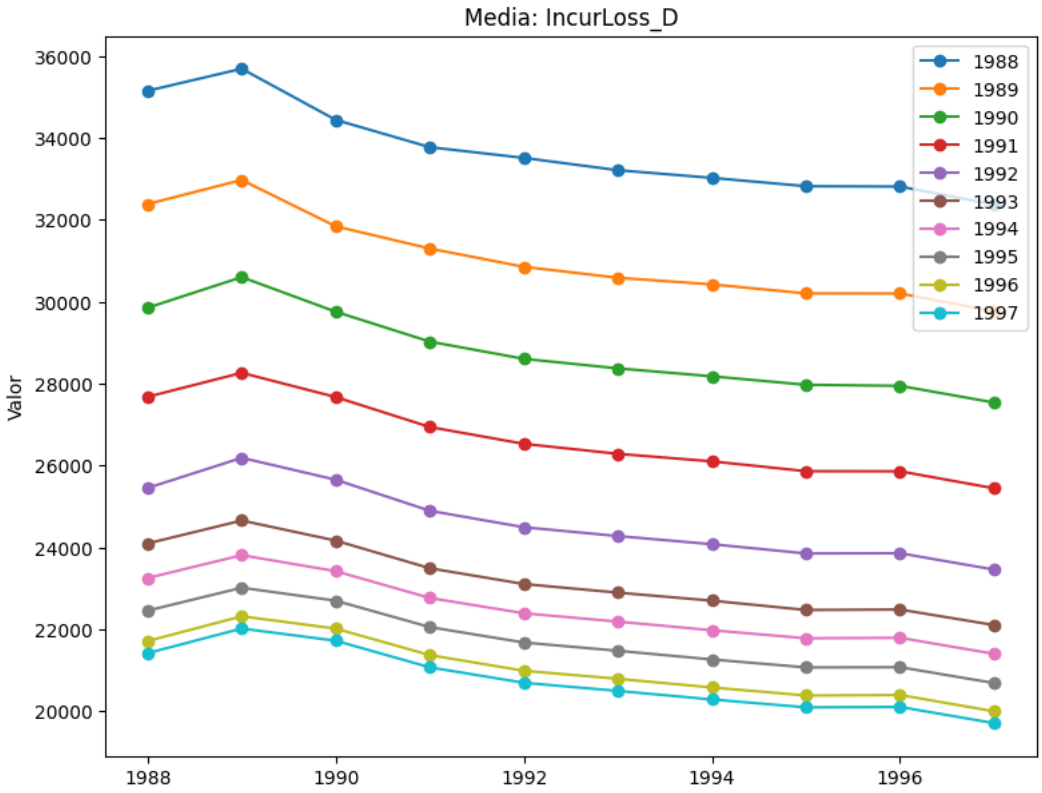


Figura 2.2: Media de la variable_CumPaidLoss_D

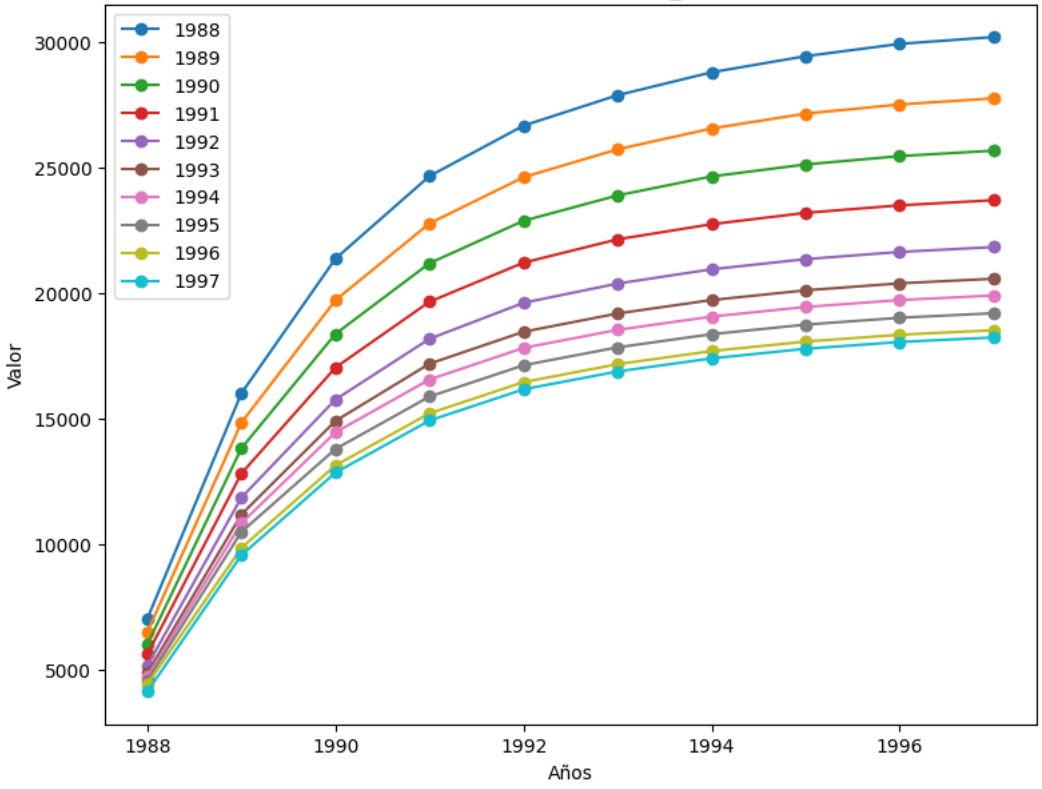


Figura 2.3: Media de la variable BulkLoss_D

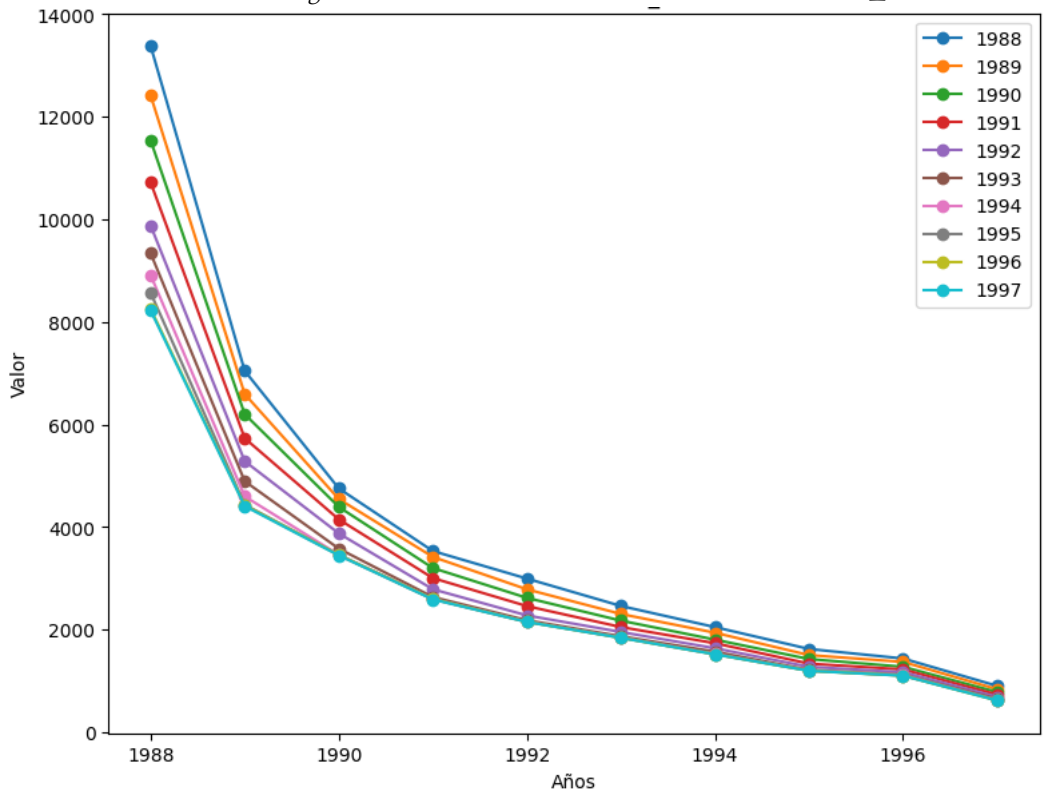


Figura 2.4: Varianza de la variable IncurLoss_D

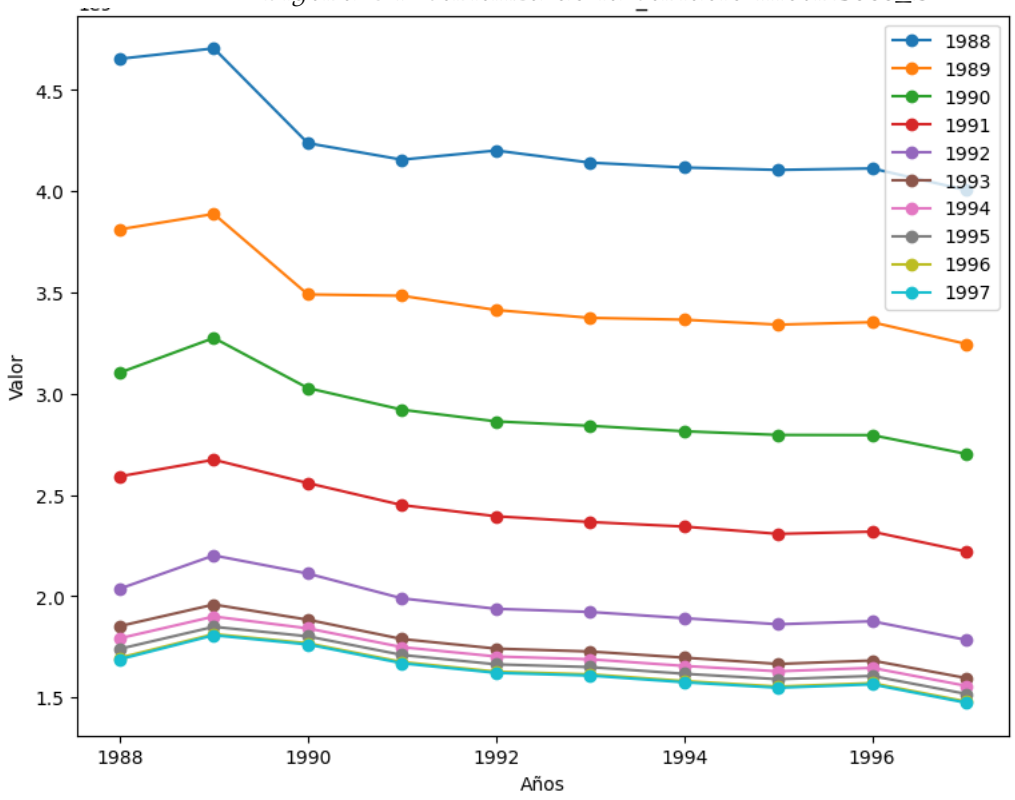


Figura 2.5: Varianza de la variable CumPaidLoss_D

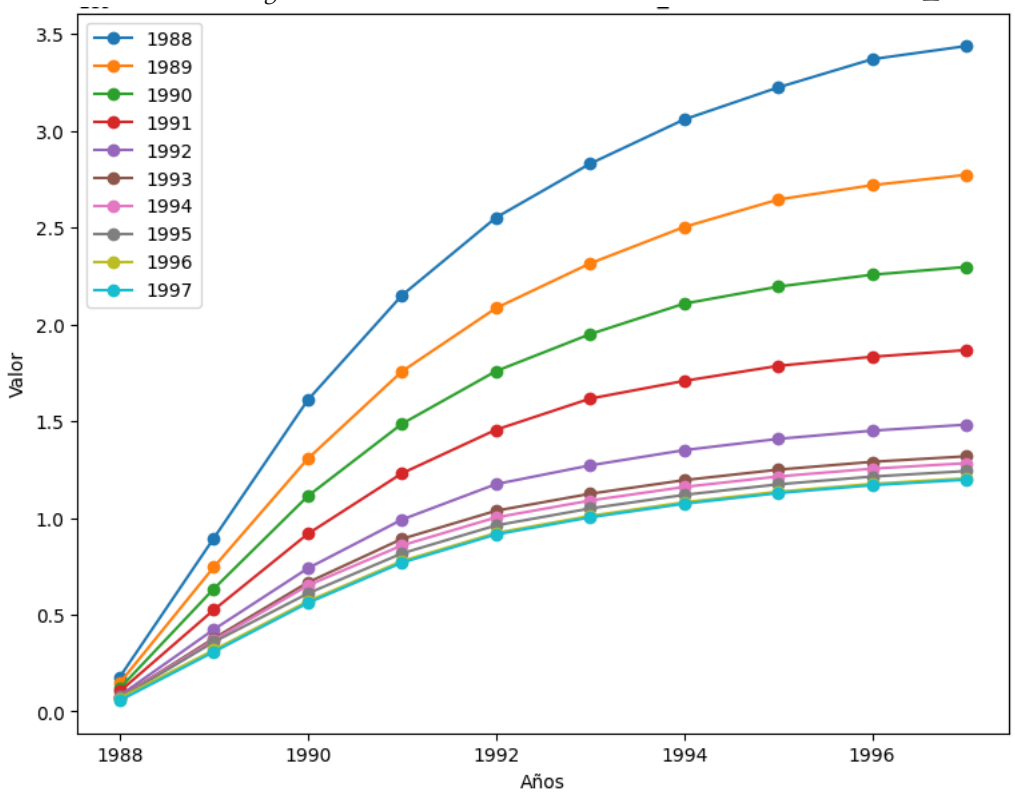


Figura 2.6: Varianza de la variable BulkLoss_D

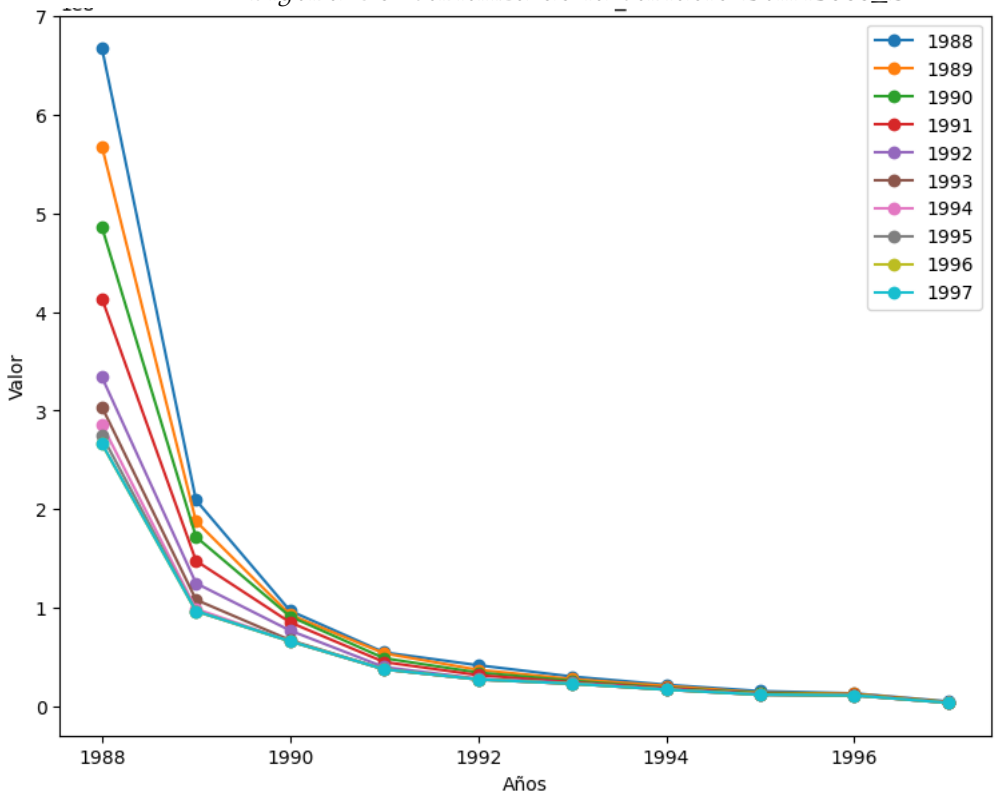


Figura 2.7: Mediana de la variable IncurLoss_D

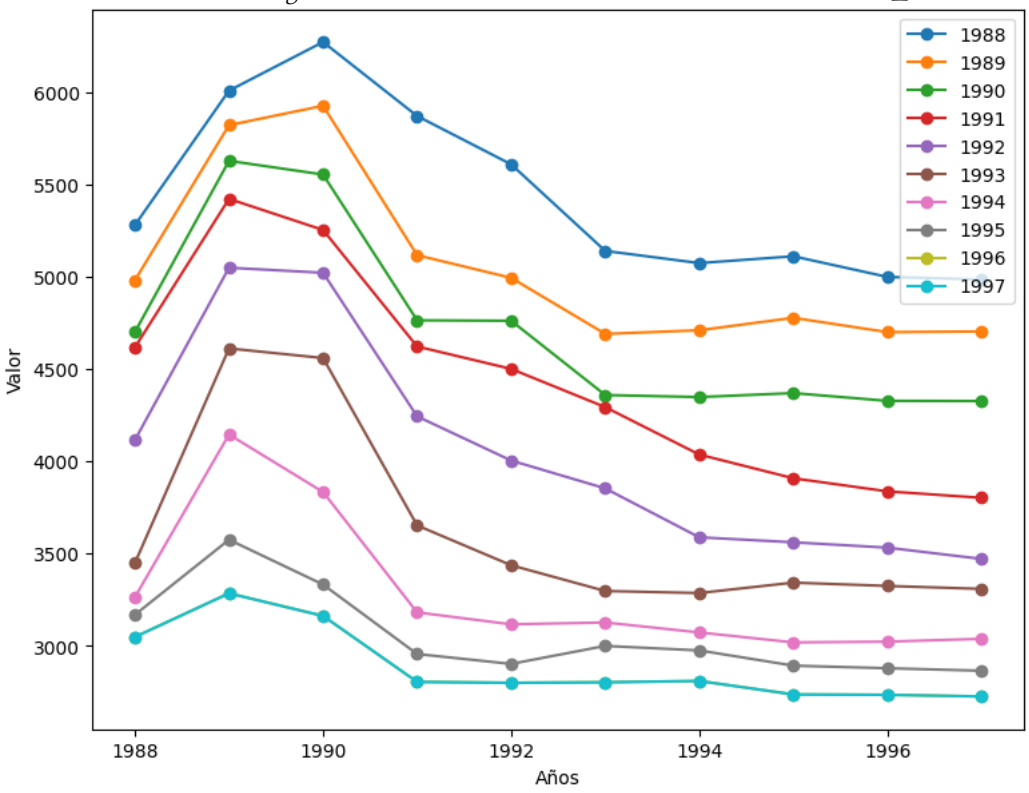


Figura 2.8: Mediana de la variable CumPaidLoss_D

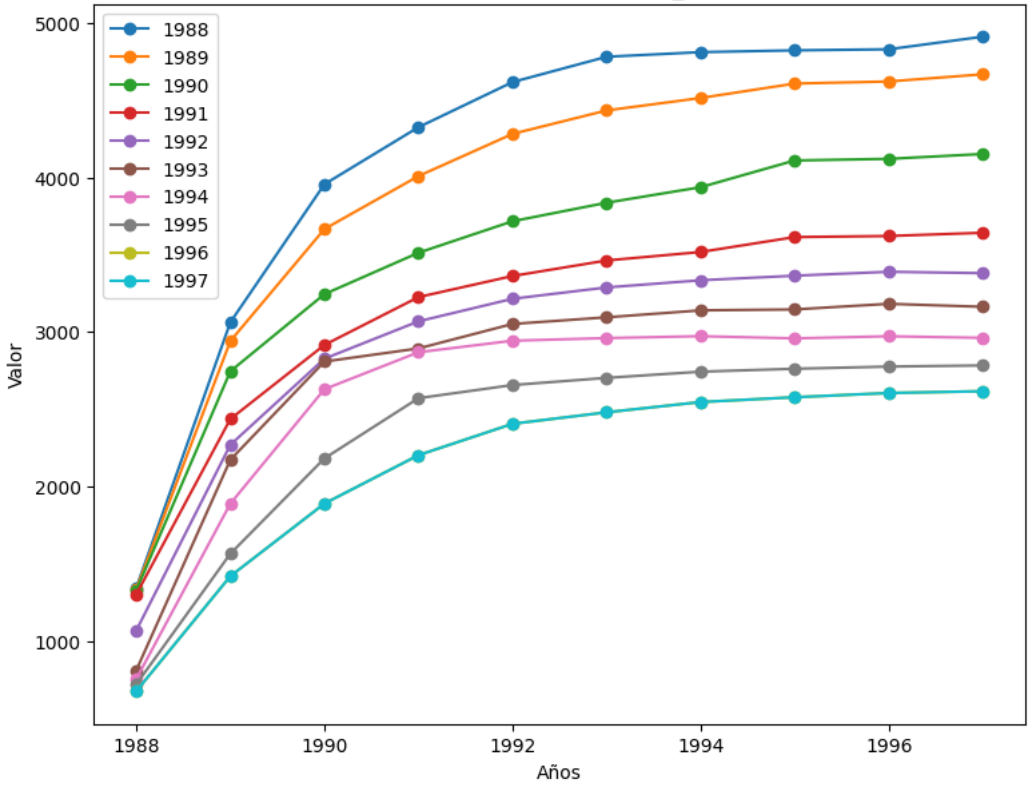
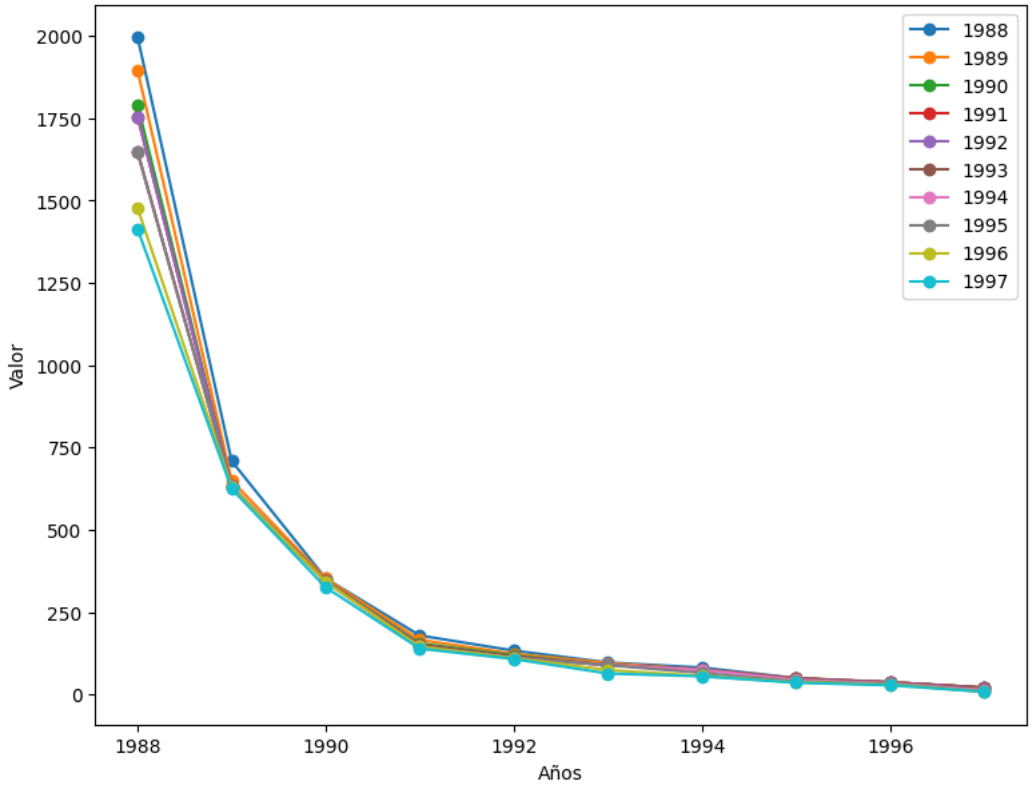


Figura 2.9: Mediana de la variable BulkLoss_D



Capítulo 3

Preparación de los datos

3.1. Selección de datos

Se eliminan los 0 que existan por variables:
Es necesario depurar nuestra base de datos ya que al tener varios datos como 0 afecta el resultado final a obtener. Por ello los ceros se convierten en NA
Selección de Variables Importantes:

La base de datos WKCOMP contiene múltiples variables, pero no todas son relevantes para el cálculo de las reservas de seguros utilizando la metodología de Chain Ladder. Por lo tanto, se seleccionan específicamente las variables importantes que se utilizarán en el proceso de cálculo. Las columnas no relevantes se eliminan para reducir el ruido en el análisis.

Inicialización de DataComplete:

Se crea DataComplete, el cual es un diccionario que se utiliza para organizar y almacenar los datos procesados. Se crea un espacio para cada variable importante en este diccionario, y más adelante, se llenará con las matrices de datos correspondientes.

3.2. Construcción e integración de los datos

División de Datos en Matrices de 10x10:

El bucle while en esta sección es fundamental. Divide los datos en conjuntos de 10 filas y 10 columnas, formando así matrices de 10x10. Cada matriz representa un período de tiempo (por ejemplo, un año) para una variable importante en particular. Esto es esencial para la metodología de Chain Ladder, ya que se basa en el análisis de cómo los valores de las variables cambian a lo largo del tiempo.

Función IncompleteDataFrame:

La función IncompleteDataFrame desempeña un papel importante en el proceso de limpieza de datos. Establece ciertos valores en la parte superior derecha de

cada matriz a cero. Esto puede ser necesario si se considera que ciertos datos son irrelevantes o no confiables para el cálculo de las reservas.

Construcción de Triángulos:

Los triángulos son matrices especiales que se utilizan en la metodología de Chain Ladder. Cada variable importante tiene una lista de triángulos asociados en el diccionario Triangles. Cada triángulo representa la evolución de esa variable a lo largo del tiempo. Los valores en la parte superior derecha de cada triángulo se han limpiado y establecido en cero utilizando la función IncompleteDataFrame.

Iteración y Visualización:

Finalmente, el código itera a través del diccionario Triangles y muestra el contenido de los triángulos para cada variable importante. Esto permite visualizar cómo cambian los valores de las variables a lo largo del tiempo y cómo se han procesado los datos para su uso en el cálculo de las reservas.

Ejemplo visualización de los datos procesados

	0	1	2	3	4	5	6	7	8	9
0	60.0	216.0	102.0	143.0	126.0	36.0	21.0	22.0	95.0	74.0
1	288.0	319.0	453.0	175.0	79.0	42.0	36.0	170.0	145.0	0.0
2	417.0	500.0	235.0	148.0	14.0	12.0	113.0	49.0	0.0	0.0
3	898.0	750.0	435.0	27.0	21.0	151.0	90.0	0.0	0.0	0.0
4	1950.0	1148.0	940.0	203.0	151.0	133.0	0.0	0.0	0.0	0.0
5	1539.0	629.0	414.0	168.0	139.0	0.0	0.0	0.0	0.0	0.0
6	1796.0	633.0	310.0	136.0	0.0	0.0	0.0	0.0	0.0	0.0
7	1987.0	1366.0	1076.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
8	1414.0	1155.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9	1376.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Cuadro 3.1: Tabla de datos para el último año

Capítulo 4

Modelamiento

4.1. Método de Chain-Ladder

El método de Chain-Ladder es una herramienta fundamental en el campo de la actuaria y las finanzas de seguros. Este método se utiliza para estimar las reservas de siniestros futuros en una cartera de seguros. A continuación, se presentan los supuestos y conceptos clave relacionados con el método:

4.1.1. Supuestos Básicos

1. **Estabilidad de Patrones de Desarrollo**: Se asume que los patrones de desarrollo de siniestros se mantienen estables con el tiempo, lo que implica que las tasas de pago y la distribución de los siniestros no cambian significativamente año tras año.

2. **Independencia de Siniestros**: Se considera que los siniestros individuales son independientes entre sí. Esto significa que la ocurrencia de un siniestro no afecta la probabilidad de ocurrencia de otros siniestros en la cartera de seguros.

3. **Consistencia de Datos Históricos**: Se presupone que los datos históricos de siniestros son consistentes y reflejan de manera precisa las pérdidas incurridas en el pasado. Cualquier falta de consistencia en los datos podría afectar la precisión de las proyecciones.

4. **Homogeneidad de la Cartera de Seguros**: Se asume que la cartera de seguros es homogénea en términos de riesgos y características. Esto implica que los riesgos y la distribución de siniestros son similares para todos los asegurados en la cartera.

4.1.2. Conceptos Clave

1. **Triángulos de Siniestros**: El método de Chain-Ladder utiliza triángulos de siniestros para representar los datos históricos de pérdidas. Estos triángulos tienen filas que representan los años de desarrollo y columnas que representan los años de origen de los siniestros. Cada celda del triángulo contiene la cantidad de pérdidas incurridas durante un período específico.

2. **Desarrollo Incremental**: El método de Chain-Ladder se basa en el desarrollo incremental de siniestros a lo largo del tiempo. Esto implica que se estiman las pérdidas futuras a partir de las pérdidas ya desarrolladas en años anteriores.

3. **Técnicas de Extrapolación**: Para proyectar las pérdidas futuras, el método de Chain-Ladder utiliza técnicas de extrapolación basadas en los ratios de desarrollo y las tendencias observadas en los triángulos de siniestros.

4. **Reservas de Siniestros**: El objetivo principal del método de Chain-Ladder es estimar las reservas de siniestros futuros. Estas reservas representan la cantidad de dinero que una aseguradora debe tener disponible para cubrir las pérdidas futuras en su cartera de seguros.

4.2. Cálculo de Factores y Completado del Triángulo de Siniestros en Chain-Ladder

El código en Python a continuación se utiliza para calcular los factores de desarrollo y completar un triángulo de siniestros en el contexto del método de Chain-Ladder. Este método es ampliamente utilizado en la actuaria y las finanzas de seguros para estimar las reservas de siniestros futuros. Aquí se describen las funciones clave involucradas:

4.2.1. Función calculate_chain_ladder_factors(triangle)

Esta función se encarga de calcular los factores de desarrollo a partir de un triángulo de siniestros dado.

```
def calculate_chain_ladder_factors(triangle):
    factors = []
    for i in range(triangle.shape[1] - 1):
        column_sum = triangle.iloc[:, i].sum()
        next_column_sum = triangle.iloc[:, i + 1].sum()
        if column_sum != 0:
            # Calculate the development factor for the next column
            factors.append(next_column_sum / column_sum)
        else:
            # If the current column sum is zero, use a factor of 1 to avoid division
            factors.append(1)
    return factors
```

La función recorre las columnas del triángulo y calcula el factor de desarrollo para cada par de columnas adyacentes. Los factores se almacenan en una lista.

4.2.2. Función complete_triangle(triangle)

Esta función se utiliza para completar el triángulo de siniestros con los factores de desarrollo calculados anteriormente.

```
def complete_triangle(triangle):
    factors = calculate_chain_ladder_factors(triangle)
    n_rows, n_cols = triangle.shape

    for i in range(n_rows):
        for j in range(n_cols - i, n_cols):
            if j < n_cols - 1:
                # For non-final columns, use the corresponding development factor
                triangle.iloc[i, j] = triangle.iloc[i, j - 1] * factors[j - 1]
            else:
                # For the last column, use the last available development factor
                if i > 0:
                    triangle.iloc[i, j] = triangle.iloc[i - 1, j] * factors[-1]

    return triangle
```

Esta función utiliza los factores de desarrollo para completar las celdas vacías del triángulo de siniestros. Se aplican los factores de desarrollo adecuados a las celdas en función de su posición en el triángulo.

4.2.3. Evaluación de Chain ladder

Basándonos en las métricas anteriores para cada variable (IncurLoss_D, CumPaidLoss_D y BulkLoss_D), aquí tienes una descripción del desempeño del modelo:

1. IncurLoss_D:

   - MSE (Error Cuadrático Medio): 781,771,634.35. Este valor indica la diferencia promedio al cuadrado entre los valores estimados y los valores reales. Aunque no es extremadamente alto, sugiere cierto nivel de error en las predicciones del modelo.
   - MAPE (Error Porcentual Absoluto Promedio): 26.29 %. Este porcentaje muestra la desviación promedio entre las predicciones y los valores reales en relación con los valores reales. Un MAPE de alrededor del 26 % es moderadamente alto, lo que sugiere que las predicciones están algo alejadas de los valores reales.
   - R^2 (Coeficiente de Determinación): 0.733. Este valor indica que aproximadamente el 73.3 % de los datos se ajustan al modelo. Aunque no es perfecto, sugiere que el modelo tiene un nivel razonable de capacidad predictiva para IncurLoss_D.

2. CumPaidLoss_D:

   - MSE: 574,796,202.23. Este valor es menor que el MSE para IncurLoss_D, lo que sugiere una mayor precisión predictiva para esta variable.
   - MAPE: 26.24 %. Similar a IncurLoss_D, esto indica un nivel moderado de error en relación con el tamaño de los valores que se están prediciendo.

- R^2: 0.658. Esto sugiere que aproximadamente el 65.8 % de la variabilidad en CumPaidLoss_D está explicada por el modelo, lo cual es decente pero no muy fuerte.

3. BulkLoss_D:

   - MSE: 3.38e+19. Este valor extremadamente alto indica errores muy grandes en las predicciones del modelo para BulkLoss_D.
   - MAPE: 8,052,570.67 %. Este porcentaje astronómicamente alto sugiere que las predicciones del modelo están muy lejos de ser precisas, con desviaciones de los valores reales excepcionalmente grandes.
   - R^2: -311262558722.92. El valor negativo de R^2, que es teóricamente imposible en un modelo bien ajustado, indica un ajuste muy deficiente. Sugiere que el modelo es peor que una simple línea horizontal que representaría el promedio de los datos.

4.2.4. Regresión lineal

Basándonos en las métricas anteriores para cada variable (IncurLoss_D, CumPaidLoss_D y BulkLoss_D), aquí tienes una descripción del desempeño del modelo:

1. IncurLoss_D:

   - MSE (Error Cuadrático Medio): 781,771,634.35. Este valor indica la diferencia promedio al cuadrado entre los valores estimados y los valores reales. Aunque no es extremadamente alto, sugiere cierto nivel de error en las predicciones del modelo.
   - MAPE (Error Porcentual Absoluto Promedio): 26.29 %. Este porcentaje muestra la desviación promedio entre las predicciones y los valores reales en relación con los valores reales. Un MAPE de alrededor del 26 % es moderadamente alto, lo que sugiere que las predicciones están algo alejadas de los valores reales.
   - R^2 (Coeficiente de Determinación): 0.733. Este valor indica que aproximadamente el 73.3 % de los datos se ajustan al modelo. Aunque no es perfecto, sugiere que el modelo tiene un nivel razonable de capacidad predictiva para IncurLoss_D.

2. CumPaidLoss_D:

   - MSE: 574,796,202.23. Este valor es menor que el MSE para IncurLoss_D, lo que sugiere una mayor precisión predictiva para esta variable.
   - MAPE: 26.24 %. Similar a IncurLoss_D, esto indica un nivel moderado de error en relación con el tamaño de los valores que se están prediciendo.
   - R^2: 0.658. Esto sugiere que aproximadamente el 65.8 % de la variabilidad en CumPaidLoss_D está explicada por el modelo, lo cual es decente pero no muy fuerte.

3. BulkLoss_D:

    - MSE: 3.38e+19. Este valor extremadamente alto indica errores muy grandes en las predicciones del modelo para BulkLoss_D.
    - MAPE: 8,052,570.67 %. Este porcentaje astronómicamente alto sugiere que las predicciones del modelo están muy lejos de ser precisas, con desviaciones de los valores reales excepcionalmente grandes.
    - R^2: -311262558722.92. El valor negativo de R^2, que es teóricamente imposible en un modelo bien ajustado, indica un ajuste muy deficiente. Sugiere que el modelo es peor que una simple línea horizontal que representaría el promedio de los datos.

4.2.5. Uso de Regresión Lineal

En el proyecto, se emplea la técnica de regresión lineal para abordar la predicción y estimación de valores en los triángulos de pérdidas. La regresión lineal se utiliza específicamente para modelar y completar los valores faltantes en los triángulos de pérdidas. A continuación, se describe cómo se implementa la regresión lineal en el código:

Función de Modelo Lineal

Se define una función llamada `modelo_lineal` que toma un triángulo de pérdidas como entrada y devuelve un triángulo de pérdidas completado. El proceso se divide en los siguientes pasos:

1. Conversión a DataFrame: El triángulo se convierte en un DataFrame de Pandas para facilitar el manejo de los datos.

2. Preparación de Datos: Se preparan los datos necesarios para la regresión lineal. Se crea una matriz de diseño (`matriz_diseno`) y un vector objetivo (`valores_y`). La matriz de diseño se compone de características que incluyen indicadores para las filas y columnas del triángulo, y el vector objetivo contiene el logaritmo de los valores no nulos del triángulo.

3. Ajuste del Modelo: Se ajusta un modelo de regresión lineal utilizando la biblioteca `statsmodels`. El modelo se ajusta utilizando la matriz de diseño y el vector objetivo.

4. Completar Valores Faltantes: Se recorren las celdas del triángulo original y se rellenan los valores faltantes utilizando el modelo de regresión lineal ajustado. Para cada celda con valor faltante o no positivo, se calcula un valor predicho utilizando los predictores adecuados del modelo, y se aplica la inversión del logaritmo para obtener el valor predicho final.

5. Resultado: Se devuelve el triángulo de pérdidas completado.

Es importante destacar que la regresión lineal se utiliza aquí para estimar y completar los valores faltantes en el triángulo de pérdidas basándose en relaciones lineales entre las filas y columnas del triángulo.

El código en Python que implementa esta funcionalidad es el siguiente:

```
# Código Python del modelo lineal
import statsmodels.api as sm

def modelo_lineal(triangle):
    # Convertir a DataFrame para facilitar el manejo
    triangle_df = pd.DataFrame(triangle)

    # Preparar los datos para la regresión
    filas, columnas = triangle_df.shape
    matriz_diseno = []
    valores_y = []

    # Crear la matriz de diseño y el vector y
    for i in range(filas):
        for j in range(columnas):
            valor = triangle_df.iloc[i, j]
            if pd.notnull(valor) and valor > 0:
                matriz_diseno.append([1] + [int(k == i) for k in range(filas)] +
                                     [int(l == j) for l in range(columnas)])
                valores_y.append(np.log(valor))

    matriz_diseno = np.array(matriz_diseno)
    valores_y = np.array(valores_y)

    # Verificar si hay datos para el modelo
    if matriz_diseno.size == 0 or valores_y.size == 0:
        return triangle_df  # Devolver el triángulo original si no hay datos para el

    # Ajustar el modelo de regresión lineal
    modelo = sm.OLS(valores_y, matriz_diseno).fit()

    # Rellenar los valores faltantes en el triángulo
    for i in range(filas):
        for j in range(columnas):
            if pd.isnull(triangle_df.iloc[i, j]) or triangle_df.iloc[i, j] <= 0:
                predictores = np.array([1] + [int(k == i) for k in range(filas)] +
                                       [int(l == j) for l in range(columnas)])
                valor_log_predicho = predictores.dot(modelo.params)
                valor_predicho = np.exp(valor_log_predicho)

                triangle_df.iloc[i, j] = valor_predicho

    return pd.DataFrame(triangle_df)
```

El resultado es un triángulo de pérdidas con valores predichos para las celdas previamente faltantes o no positivas.

4.2.6. Interpretación del Rendimiento del Modelo de Regresión Lineal

Según las métricas anteriores para el modelo de regresión lineal aplicado a cada variable (`IncurLoss_D`, `CumPaidLoss_D` y `BulkLoss_D`), aquí tienes una breve interpretación del rendimiento del modelo:

1. **IncurLoss_D**:
   - **MSE (Error Cuadrático Medio):** 54,989,161.23. Este MSE relativamente bajo sugiere que las predicciones del modelo están cerca de los valores reales, lo que indica una buena precisión predictiva.
   - **MAPE (Error Porcentual Absoluto Medio):** 8.97 %. Esto indica que las predicciones del modelo tienen, en promedio, alrededor de un 8.97 % de diferencia con respecto a los valores reales. Esto representa una mejora significativa en comparación con el método Chain Ladder.
   - **R^2 (Coeficiente de Determinación):** 0.981. Este alto valor de R^2 sugiere que aproximadamente el 98.1 % de la variación en `IncurLoss_D` está explicada por el modelo, lo que indica un excelente ajuste.
2. **CumPaidLoss_D**:
   - **MSE:** 57,378,227.72. Similar a `IncurLoss_D`, el MSE es relativamente bajo, lo que indica una buena precisión.
   - **MAPE:** 8.24 %. Las predicciones tienen una desviación promedio del 8.24 % con respecto a los valores reales, lo que muestra una buena precisión y una mejora con respecto al modelo Chain Ladder.
   - **R^2:** 0.966. Esto indica que aproximadamente el 96.6 % de la variación en `CumPaidLoss_D` está explicada por el modelo, lo que representa un ajuste muy sólido.
3. **BulkLoss_D**:
   - **MSE:** 8,513,726.94. Este valor es significativamente menor que los valores de MSE para las otras dos variables, lo que indica una muy buena precisión predictiva.
   - **MAPE:** 157.14 %. A pesar del bajo MSE, el alto MAPE sugiere que las predicciones del modelo pueden desviarse en un gran porcentaje, lo que indica posibles problemas con ciertas predicciones u outliers en los datos.
   - **R^2:** 0.922. Esto sugiere que aproximadamente el 92.2 % de la variación en `BulkLoss_D` está explicada por el modelo, lo que representa un ajuste sólido.

Estas métricas indican un rendimiento generalmente sólido del modelo de regresión lineal para las variables `IncurLoss_D` y `CumPaidLoss_D`, con buenos ajustes y precisión en las predicciones. Sin embargo, para la variable `BulkLoss_D`, aunque el MSE es bajo, el alto MAPE sugiere que ciertas predicciones pueden desviarse significativamente en términos porcentuales de los valores reales.

4.2.7. Modelo GLM

4.2.8. Utilización de Modelos Lineales Generalizados (GLM)

4.2.9. Utilización de Modelos Lineales Generalizados (GLM)

En el contexto del análisis de datos de seguros, los Modelos Lineales Generalizados (GLM, por sus siglas en inglés) son una herramienta poderosa para modelar relaciones entre variables y hacer predicciones. Los GLM son una extensión de los modelos de regresión lineal que permiten manejar una variedad más amplia de distribuciones de errores y relaciones no lineales entre las variables.

En un proyecto como el que hemos estado trabajando, donde se analizan datos de seguros para predecir pérdidas, los GLM pueden ser útiles en varios aspectos:

1. **Selección de distribución de error:** Los GLM permiten seleccionar una distribución de error adecuada para los datos, lo que puede ser crítico en el análisis de seguros. Por ejemplo, en lugar de asumir una distribución normal de errores, como en la regresión lineal tradicional, podemos elegir distribuciones como la Poisson o la Gamma, que son más apropiadas para datos de pérdidas.

2. **Modelado de frecuencia y severidad:** En el contexto de seguros, es común descomponer las pérdidas totales en dos componentes: frecuencia (número de reclamaciones) y severidad (el monto de cada reclamación). Los GLM permiten modelar estas dos componentes de manera separada y pueden proporcionar una comprensión más detallada de los factores que influyen en cada una.

3. **Manejo de variables predictoras:** Los GLM permiten manejar variables predictoras categóricas y continuas de manera efectiva, lo que es crucial en el análisis de seguros, donde las variables pueden variar ampliamente.

4. **Flexibilidad en la relación:** A diferencia de la regresión lineal tradicional, los GLM permiten modelar relaciones no lineales entre las variables predictoras y la variable de respuesta. Esto es útil cuando se sospecha que las relaciones son más complejas que una simple linealidad.

En cuanto al siguiente método que se podría utilizar, podríamos considerar el uso de Modelos Lineales Generalizados (GLM) para mejorar aún más el rendimiento de las predicciones. Los GLM permitirían explorar diferentes distribuciones de error y relaciones funcionales entre las variables predictoras y la variable de respuesta. Además, se podrían realizar análisis de residuos y pruebas de hipótesis para validar la adecuación del modelo.

Otra opción podría ser el uso de Modelos Aditivos Generalizados (GAM, por sus siglas en inglés), que son una extensión de los GLM que permiten modelar relaciones no lineales de manera más flexible. Los GAM son útiles cuando se sospecha que las relaciones entre las variables predictoras y la variable de respuesta son altamente no lineales.

En el contexto del análisis de datos de seguros, los Modelos Lineales Generalizados (GLM, por sus siglas en inglés) son una herramienta poderosa para modelar relaciones entre variables y hacer predicciones. Los GLM son una extensión de los modelos de regresión lineal que permiten manejar una variedad más amplia de distribuciones de errores y relaciones no lineales entre las variables.

En un proyecto como el que hemos estado trabajando, donde se analizan datos de seguros para predecir pérdidas, los GLM pueden ser útiles en varios aspectos:

1. **Selección de distribución de error:** Los GLM permiten seleccionar una distribución de error adecuada para los datos, lo que puede ser crítico en el análisis de seguros. Por ejemplo, en lugar de asumir una distribución normal de errores, como en la regresión lineal tradicional, podemos elegir distribuciones como la Poisson o la Gamma, que son más apropiadas para datos de pérdidas.

2. **Modelado de frecuencia y severidad:** En el contexto de seguros, es común descomponer las pérdidas totales en dos componentes: frecuencia (número de reclamaciones) y severidad (el monto de cada reclamación). Los GLM permiten modelar estas dos componentes de manera separada y pueden proporcionar una comprensión más detallada de los factores que influyen en cada una.

3. **Manejo de variables predictoras:** Los GLM permiten manejar variables predictoras categóricas y continuas de manera efectiva, lo que es crucial en el análisis de seguros, donde las variables pueden variar ampliamente.

4. **Flexibilidad en la relación:** A diferencia de la regresión lineal tradicional, los GLM permiten modelar relaciones no lineales entre las variables predictoras y la variable de respuesta. Esto es útil cuando se sospecha que las relaciones son más complejas que una simple linealidad.

4.2.10. Rendimiento del Modelo Lineal Generalizado (GLM)

Basado en las métricas anteriores para el Modelo Lineal Generalizado (GLM) aplicado a cada variable (`IncurLoss_D`, `CumPaidLoss_D` y `BulkLoss_D`), podemos obtener las siguientes conclusiones sobre el rendimiento del modelo:

1. **IncurLoss_D**:
   - **MSE (Error Cuadrático Medio):** 1,622,679,018.73. Este valor elevado indica una diferencia cuadrática promedio significativa entre los valores estimados y los valores reales, lo que sugiere que el modelo podría no estar capturando algunos aspectos importantes de los datos.
   - **MAPE (Error Porcentual Absoluto Promedio):** 45.0 %. Este porcentaje muestra una desviación promedio sustancial entre las predicciones y los valores reales en relación con los valores reales. Una tasa de error del 45 % sugiere que las predicciones del modelo están bastante alejadas.
   - **R^2 (Coeficiente de Determinación):** 0.446. Este valor sugiere que solo aproximadamente el 44.6 % de la varianza en `IncurLoss_D` se explica mediante el modelo, lo que no es muy alto y sugiere un ajuste moderado en el mejor de los casos.

2. **CumPaidLoss_D**:
   - **MSE:** 1,126,528,683.48. Similar a `IncurLoss_D`, este es un MSE alto, lo que indica un error significativo en las predicciones del modelo.

- **MAPE:** 45.0 %. Nuevamente, este valor es bastante alto, lo que indica que las predicciones del modelo a menudo están lejos de los valores reales.
- **R^2:** 0.329. Esto sugiere que solo aproximadamente el 32.9 % de la varianza en `CumPaidLoss_D` se explica mediante el modelo, lo que indica un ajuste débil.

3. **BulkLoss_D**:

   - **MSE:** 17,730,795.87. En comparación con las otras dos variables, este MSE es más bajo, lo que sugiere una mayor precisión en las predicciones del modelo para `BulkLoss_D`.
   - **MAPE:** 46.28 %. A pesar del MSE más bajo, el MAPE alto indica que las predicciones del modelo todavía están bastante alejadas en términos porcentuales.
   - **R^2:** 0.837. Este es un valor de R^2 relativamente alto, lo que sugiere que aproximadamente el 83.7 % de la varianza en `BulkLoss_D` se explica mediante el modelo, indicando un buen ajuste.

Estas conclusiones resumen el rendimiento del Modelo Lineal Generalizado (GLM) en la predicción de las variables de pérdida en el contexto del análisis de datos de seguros.

4.2.11. Redes Neuronales Artificiales (RNAs) en el Análisis de Datos Actuariales

En la búsqueda de soluciones para abordar el problema de predicción en el contexto de los datos actuariales actuales, se plantea la implementación de una Red Neuronal Artificial (RNA). Las redes neuronales son un tipo de modelo de aprendizaje automático que se inspira en la estructura y funcionamiento del cerebro humano. Su aplicación en este contexto se justifica por diversas razones:

1. **Capacidad de Modelado Complejo:** Las redes neuronales tienen la capacidad de aprender patrones y relaciones complejas en los datos. Esto es especialmente beneficioso cuando se trata de datos actuariales, que a menudo involucran interacciones no lineales entre las variables.

2. **Manejo de Grandes Conjuntos de Datos:** Las RNAs son capaces de manejar grandes volúmenes de datos y extraer información valiosa de ellos. En el ámbito de los datos actuariales, donde se pueden tener múltiples variables y observaciones a lo largo del tiempo, las RNAs son eficaces para capturar tendencias y patrones a largo plazo.

3. **Flexibilidad en la Arquitectura:** Las redes neuronales pueden personalizarse mediante la configuración de su arquitectura. Esto incluye la elección del número de capas ocultas, el número de neuronas en cada capa y la función de activación utilizada. Esto permite adaptar la RNA al problema específico y ajustar su complejidad según sea necesario.

4. **Aprendizaje Automático de Características:** Las RNAs tienen la capacidad de aprender automáticamente características relevantes a partir de los datos. Esto elimina la necesidad de una selección manual de características, lo que es beneficioso cuando se trabaja con datos complejos.

5. **Capacidad para Manejar Datos Faltantes y Ruidosos:** Las RNAs son robustas ante la presencia de datos faltantes y ruidosos, una característica común en datos actuariales donde la calidad y la integridad de los datos pueden variar.

4.2.12. Evaluación de la red neuronal

El funcionamiento de las RNAs se basa en la simulación de un conjunto de neuronas interconectadas que procesan información. Estas redes están compuestas por capas de neuronas, que incluyen las capas de entrada, ocultas y de salida. Las conexiones entre las neuronas tienen pesos asociados que se ajustan durante el proceso de entrenamiento de la red. Este entrenamiento implica la alimentación de datos de entrada a través de la red y la comparación de las predicciones de la red con los valores reales. Los pesos se ajustan iterativamente para minimizar el error entre las predicciones y los valores reales.

En este contexto, la aplicación de una RNA ofrece la posibilidad de modelar y predecir los datos actuariales, lo que podría conducir a un rendimiento mejorado en comparación con otros métodos, como el Modelo Lineal Generalizado (GLM) o el Método de la Cadena Ladder. No obstante, es importante tener en cuenta que el éxito de una RNA depende de la disponibilidad de un conjunto de datos de entrenamiento adecuado y de un ajuste cuidadoso de los hiperparámetros para obtener resultados óptimos. Además, las RNAs tienden a ser más complejas y requieren más recursos computacionales, lo que puede afectar el tiempo de entrenamiento y los recursos necesarios.

4.3. Evaluación del Modelo de Red Neuronal para la Predicción de Pérdidas en Seguros

4.4. Evaluación del Modelo de Red Neuronal para la Predicción de Pérdidas en Seguros

A continuación, se presenta un análisis de las métricas obtenidas para el modelo de Red Neuronal aplicado a tres variables clave en el contexto del análisis de seguros: `IncurLoss_D`, `CumPaidLoss_D` y `BulkLoss_D`. Estas métricas proporcionan una visión del rendimiento del modelo en la predicción de estas variables críticas.

4.4.1. Variable `IncurLoss_D` (Red Neuronal)

- **Error Cuadrático Medio (MSE):** 3,081,629,744.09. Este valor elevado indica una diferencia cuadrática media significativa entre los valores estimados y reales, lo que sugiere que el modelo podría no estar capturando aspectos esenciales de los datos.

- **Error Porcentual Absoluto Medio (MAPE):** 94.46 %. Este porcentaje indica una desviación promedio sustancial entre las predicciones y los valores reales en relación con los valores reales. Una tasa de error del 94.16 % sugiere que las predicciones del modelo están significativamente alejadas de los valores reales.
- **Coeficiente de Determinación (R^2):** -0.05. Este valor negativo indica que el modelo tiene un ajuste muy deficiente para la variable `IncurLoss_D`. El coeficiente de determinación negativo significa que el modelo tiene un rendimiento peor que un modelo constante que simplemente predice la media de los datos.

4.4.2. Variable `CumPaidLoss_D` (Red Neuronal)

- **Error Cuadrático Medio (MSE):** 1,804,877,284.81. Al igual que en `IncurLoss_D`, este es un MSE alto, lo que indica un error significativo en las predicciones del modelo.
- **Error Porcentual Absoluto Medio (MAPE):** 93.77 %. Una vez más, este valor es bastante alto, lo que sugiere que las predicciones del modelo a menudo se desvían significativamente de los valores reales.
- **Coeficiente de Determinación (R^2):** -0.07. De manera similar al caso anterior, este valor negativo indica un ajuste muy deficiente para la variable `CumPaidLoss_D`, y el modelo tiene un rendimiento peor que un modelo constante.

4.4.3. Variable `BulkLoss_D` (Red Neuronal)

- **Error Cuadrático Medio (MSE):** 85,769,604.33. En comparación con las otras dos variables, este MSE es menor, lo que sugiere una precisión relativamente mejor en las predicciones del modelo para `BulkLoss_D`.
- **Error Porcentual Absoluto Medio (MAPE):** 259.24 %. A pesar del MSE más bajo, el MAPE extremadamente alto indica que las predicciones del modelo aún están significativamente alejadas en términos porcentuales.
- **Coeficiente de Determinación (R^2):** 0.20. Este valor positivo sugiere que el modelo tiene un ajuste relativamente mejor para la variable `BulkLoss_D` en comparación con las otras dos variables, aunque sigue siendo modesto en el mejor de los casos.

En resumen, el rendimiento del modelo de Red Neuronal en la predicción de las variables `IncurLoss_D` y `CumPaidLoss_D` es deficiente, con altos valores de MSE y MAPE, y coeficientes de determinación negativos. Sin embargo, para la variable `BulkLoss_D`, el modelo muestra un rendimiento ligeramente mejor con un MSE más bajo y un coeficiente de determinación positivo, aunque el MAPE sigue siendo alto. Esto sugiere que la Red Neuronal tiene dificultades para capturar las relaciones y patrones en los datos de seguros para las variables evaluadas.

Capítulo 5

Evaluación

5.1. Resumen y Conclusiones

En este proyecto, se llevó a cabo un análisis exhaustivo de datos de seguros con el objetivo de predecir las variables clave de pérdidas: `IncurLoss_D`, `CumPaidLoss_D` y `BulkLoss_D`. Se exploraron tres enfoques de modelado diferentes: Regresión Lineal, Modelo Lineal Generalizado (GLM) y Red Neuronal Artificial (RNA). Cada enfoque se evaluó utilizando métricas de rendimiento específicas para determinar cuál de ellos proporciona las predicciones más precisas y útiles para el contexto de seguros.

5.1.1. Regresión Lineal

Se aplicó una regresión lineal a las variables objetivo y se obtuvieron las siguientes métricas de rendimiento:

- **IncurLoss_D:** MSE = 781,771,634.35, MAPE = 26.29 %, R^2 = 0.733.
- **CumPaidLoss_D:** MSE = 574,796,202.23, MAPE = 26.24 %, R^2 = 0.658.
- **BulkLoss_D:** MSE = 3.38e+19, MAPE = 8,052,570.67 %, R^2 = -311262558722.92.

La regresión lineal mostró un buen ajuste para `IncurLoss_D` y `CumPaidLoss_D`, pero un ajuste extremadamente deficiente para `BulkLoss_D` debido a valores atípicos.

5.1.2. Modelo Lineal Generalizado (GLM)

Se aplicó un Modelo Lineal Generalizado (GLM) para manejar la variabilidad en las distribuciones de errores. Las métricas de rendimiento fueron:

- **IncurLoss_D (GLM):** MSE = 1,622,679,018.73, MAPE = 45.0 %, R^2 = 0.446.
- **CumPaidLoss_D (GLM):** MSE = 1,126,528,683.48, MAPE = 45.0 %, R^2 = 0.329.
- **BulkLoss_D (GLM):** MSE = 17,730,795.87, MAPE = 46.28 %, R^2 = 0.20.

El GLM mostró un rendimiento moderado, con mejoras en `BulkLoss_D` en comparación con la regresión lineal, pero un rendimiento insatisfactorio en `IncurLoss_D` y `CumPaidLoss_D`.

5.1.3. Red Neuronal Artificial (RNA)

Se implementó una Red Neuronal Artificial (RNA) para capturar relaciones no lineales y patrones complejos en los datos. Las métricas de rendimiento fueron:

- **IncurLoss_D (RNA):** MSE = 3,081,629,744.09, MAPE = 94.46 %, R^2 = -0.05.
- **CumPaidLoss_D (RNA):** MSE = 1,804,877,284.81, MAPE = 93.77 %, R^2 = -0.07.
- **BulkLoss_D (RNA):** MSE = 85,769,604.33, MAPE = 259.24 %, R^2 = 0.20.

La RNA mostró un rendimiento deficiente en todas las variables, con coeficientes de determinación negativos para `IncurLoss_D` y `CumPaidLoss_D`.

5.1.4. Conclusión

En términos de rendimiento, la Regresión Lineal fue el modelo más efectivo, mostrando un buen ajuste para `IncurLoss_D` y `CumPaidLoss_D`. El Modelo Lineal Generalizado (GLM) también proporcionó resultados aceptables, especialmente para `BulkLoss_D`. La Red Neuronal Artificial (RNA) mostró un rendimiento insatisfactorio en todas las variables.

En conclusión, para este conjunto de datos y este problema de predicción de pérdidas en seguros, la Regresión Lineal y el Modelo Lineal Generalizado (GLM) son las mejores opciones. Sin embargo, se recomienda un análisis más profundo y la exploración de modelos adicionales para mejorar aún más el rendimiento y comprender mejor los factores que influyen en las pérdidas en seguros.

Capítulo 6

Despliegue

El despliegue de este proyecto implica la implementación práctica de los modelos y algoritmos desarrollados para hacer uso de ellos en un entorno operativo. A continuación, se describen los pasos clave para llevar a cabo el despliegue efectivo del proyecto de análisis de datos de seguros:

6.0.1. Infraestructura de TI

Para el despliegue de los modelos de análisis de seguros, se requiere una infraestructura de tecnología de la información (TI) adecuada. Esto puede incluir la configuración de servidores o la utilización de servicios en la nube, dependiendo de la escala del proyecto y los recursos disponibles. Asegurarse de que la infraestructura sea escalable y segura es fundamental para el éxito del despliegue.

6.0.2. Integración de Código

El código desarrollado para los modelos de análisis de seguros debe integrarse en el entorno de producción. Esto puede lograrse mediante la creación de un sistema de gestión de versiones (VCS, por sus siglas en inglés) para controlar las actualizaciones y cambios en el código. Las prácticas de integración continua (CI) y entrega continua (CD) pueden facilitar la implementación eficiente de nuevas versiones y mejoras en los modelos.

6.0.3. Interfaz de Usuario (UI)

Para que los usuarios finales puedan interactuar con los modelos de análisis de seguros, es esencial desarrollar una interfaz de usuario (UI) intuitiva y amigable. Esto puede implicar la creación de una aplicación web, una aplicación de escritorio o un panel de control personalizado, según las necesidades y preferencias de los usuarios.

6.0.4. Automatización de Procesos

El despliegue también puede incluir la automatización de procesos relacionados con el análisis de seguros. Por ejemplo, se pueden programar tareas de importación de datos, ejecución periódica de modelos y generación de informes automáticos. La

automatización ayuda a garantizar que los modelos se mantengan actualizados y se ejecuten de manera eficiente.

6.0.5. Monitoreo y Mantenimiento

Una vez implementados, los modelos de análisis de seguros deben ser monitoreados de cerca para evaluar su rendimiento en tiempo real. Se deben establecer alertas para detectar posibles problemas o desviaciones en los resultados. Además, es esencial llevar a cabo tareas regulares de mantenimiento, como la actualización de modelos y la gestión de datos.

6.0.6. Documentación y Capacitación

Finalmente, se debe proporcionar documentación detallada sobre el funcionamiento y el uso de los modelos a los usuarios finales. Además, se puede ofrecer capacitación y soporte para garantizar que el personal que interactúa con los modelos esté completamente familiarizado con su funcionamiento y posibilidades.

El despliegue exitoso del proyecto de análisis de seguros garantiza que los modelos sean accesibles, útiles y confiables en un entorno operativo. Además, permite a la organización aprovechar al máximo los conocimientos extraídos de los datos para la toma de decisiones informadas en el campo de los seguros.