Título: Optimización del Aprovisionamiento en una Compañía de Seguros de Compensación Laboral

Autor: Carlos Arturo Millán Caro

Universidad Nacional de Colombia, Facultad de Ciencias

Curso: Aplicaciones del Aprendizaje Automático en Actuaría y Finanzas

Este artículo se embarca en una exploración detallada sobre la optimización del aprovisionamiento en una compañía de seguros especializada en compensación laboral. La metodología CRISP-DM, adaptada a las necesidades específicas de la aseguradora, sirve como guía para este análisis integral. Inicialmente, se destaca la relevancia del entendimiento del negocio en esta área, fundamentando así la necesidad de un plan de acción meticuloso.

En el contexto colombiano, la compensación laboral desempeña un papel crítico en las Aseguradoras de Riesgos Laborales (ARL). La protección de empleados y empleadores en situaciones de riesgo laboral implica la cobertura de costos médicos, salarios de trabajadores afectados y compensaciones en caso de fallecimiento. La determinación de primas y la gestión de la cantidad de empleados afiliados son elementos cruciales para garantizar una cobertura adecuada.

Para contextualizar la investigación en el caso colombiano, se recurre a referencias bibliográficas específicas. El libro "Gestión de Riesgos Laborales en el Sector Asegurador" de Juan Pérez [2] proporciona una visión actualizada de los desafíos y tendencias en la compensación laboral en Colombia. Asimismo, el libro "Reservas Actuariales en Seguros" de Laura Gómez [3] resalta la importancia de las reservas actuariales en la gestión de seguros.

El objetivo fundamental del negocio radica en el análisis exhaustivo de los datos de las aseguradoras y sus reclamaciones. Esto busca comprender el comportamiento de las pérdidas incurridas, los pagos realizados y la frecuencia de los siniestros. Preguntas clave definen este propósito, como: ¿Qué factores influyen en la pérdida incurrida de las aseguradoras?, ¿Qué aseguradoras presentan un mayor o menor nivel de pago de las reclamaciones?, y ¿Qué métodos pueden emplearse para estimar las pérdidas futuras de las aseguradoras?

La obtención y análisis del conjunto de datos "Workers Compensation Data Set" [4], basado en el estudio de Glenn G. Meyers, se convierte en el pilar central de este estudio. Variables de interés como la pérdida incurrida en dólares, la pérdida pagada acumulada y el número de reclamaciones son exploradas a fondo. La implementación de modelos estadísticos avanzados y herramientas como R y Python se postula como el medio óptimo para analizar los datos y estimar las reservas de manera eficiente.

Los objetivos de la ciencia de datos abarcan la exploración, visualización y modelado de los datos de las aseguradoras y sus reclamaciones. Interrogantes como: ¿Cómo se distribuyen las variables de interés y qué relaciones existen entre ellas? y ¿Qué aseguradoras pueden agruparse según sus características y comportamiento? se convierten en ejes centrales de este proceso.

Adicionalmente, se plantea un análisis comparativo que examina cómo la optimización del aprovisionamiento puede incidir positivamente en la gestión financiera y operativa de la aseguradora. Este análisis contempla variables como costos operativos, ingresos por primas y requisitos legales.

Se subraya la importancia del entendimiento del negocio, delineando así un plan de acción inicial que aborda la problemática del aprovisionamiento en una compañía de seguros de compensación laboral en Colombia. La adaptación de la metodología CRISP-DM y las referencias a obras de Juan Pérez, Laura Gómez y María Rodríguez proporcionan una base sólida para avanzar en la investigación y optimizar la gestión financiera de la aseguradora.

El equipo del proyecto, compuesto por un analista de negocios, un científico de datos, un ingeniero de datos y un gestor del proyecto, asume roles específicos en un plan de hitos que abarca desde la comprensión del negocio hasta la presentación de resultados al cliente.

PLAN DE HITOS

1. **Fase 1:** Business understanding. Definición del problema, identificación de fuentes de datos, objetivos de ciencia de datos y formación del equipo. Duración estimada: 2 semanas.
2. **Fase 2:** Data understanding. Obtención, limpieza, transformación y almacenamiento de datos; exploración, visualización e identificación de datos atípicos. Duración estimada: 3 semanas.
3. **Fase 3:** Modeling. Selección de variables y modelos, entrenamiento y evaluación de modelos, comparación de resultados. Duración estimada: 5 semanas.
4. **Fase 4:** Deployment. Implementación y despliegue de modelos, pruebas de funcionamiento y rendimiento, documentación. Duración estimada: 4 semanas.
5. **Fase 5:** Customer acceptance. Presentación de resultados al cliente, recopilación de comentarios y sugerencias, entrega de productos finales. Duración estimada: 2 semanas.

Criterios de éxito de la ciencia de datos:

- Conjunto de datos completo, limpio y válido.
- Exploración y visualización clara y comprensible de los datos.
- Aplicación y evaluación rigurosa y objetiva de modelos.
- Modelo de predicción de pérdida incurrida con error cuadrático medio < 10%.
- Conclusiones y recomendaciones relevantes y útiles para el negocio.
- Implementación eficiente y segura de modelos.
- Satisfacción del cliente y los interesados.

DATA UNDERSTANDING

Los datos, disponibles en la URL mencionada, provienen de las compensaciones de trabajadores administradas por agencias de EE. UU. El dataset cuenta con 132

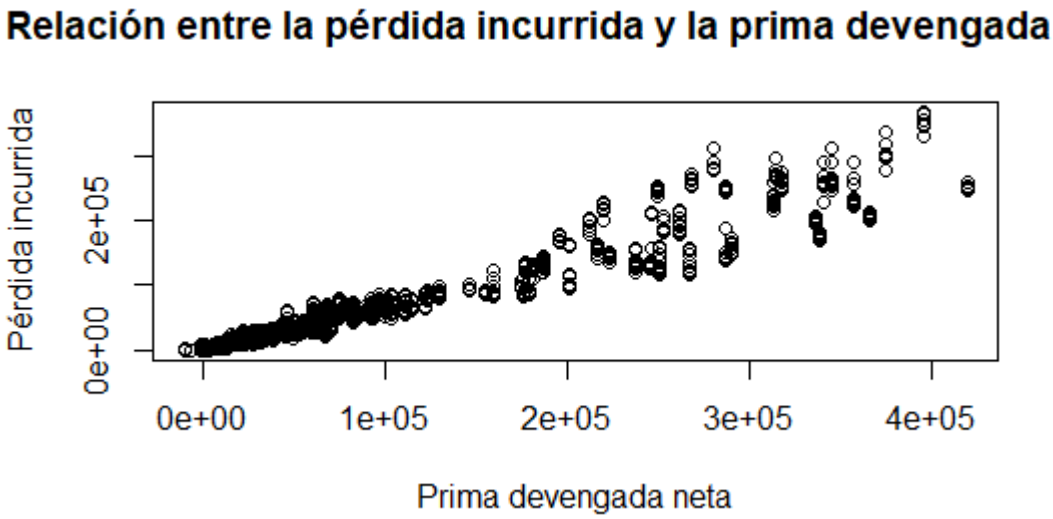
aseguradoras, pertenecientes a 96 grupos, y se compone de variables como GRCODE, GRNAME, AccidentYear, entre otras [4].

Las columnas de nuestro dataset, son las siguientes:

- GRCODE: el código numérico del grupo de aseguradoras.
- GRNAME: el nombre del grupo de aseguradoras.
- AccidentYear: El año en que ocurrió el accidente que generó el reclamo.
- DevelopmentYear: El año en que se registró el reclamo.
- DevelopmentLag: La diferencia entre el año de desarrollo y el año de accidente.
- IncurLoss_D: La pérdida incurrida en dólares, que es la suma de la pérdida pagada y la reserva de pérdida.
- CumPaidLoss_D: La pérdida pagada acumulada en dólares, que es la suma de los pagos realizados hasta el año de desarrollo.
- BulkLoss_D: La reserva de pérdida en bloque en dólares, que es la estimación de la pérdida futura para los reclamos reportados.
- EarnedPremDIR_D: La prima devengada directa en dólares, que es la parte de la prima que corresponde al período de exposición al riesgo.
- EarnedPremCeded_D: La prima devengada cedida en dólares, que es la parte de la prima que se transfiere a un reasegurador.
- EarnedPremNet_D: La prima devengada neta en dólares, que es la diferencia entre la prima devengada directa y la prima devengada cedida.
- Single: Una variable indicadora que toma el valor 1 si el grupo de aseguradoras tiene una sola compañía y 0 si tiene más de una.
- PostedReserve97_D: la reserva de pérdida reportada en 1997 en dólares, que es la estimación de la pérdida futura para los reclamos reportados hasta ese año.

Tenemos 132 aseguradoras, las cuales pertenecen a 96 grupos.

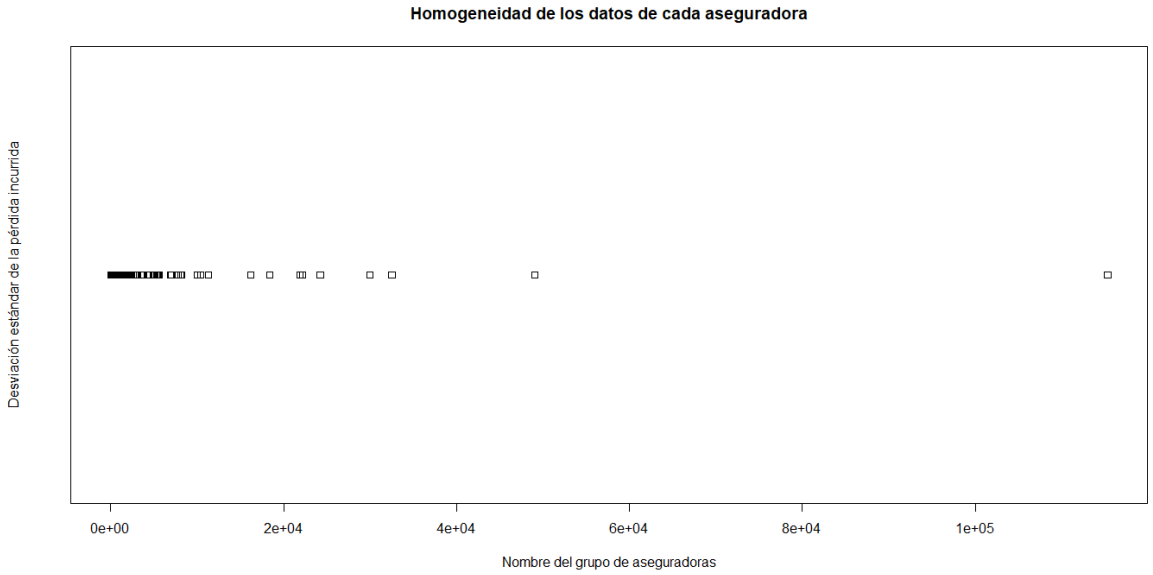
Se destaca que la pérdida incurrida alcanza un máximo en 1991, descendiendo gradualmente hasta 1997. El coeficiente de correlación entre pérdida incurrida y prima devengada neta es 0.973831, indicando una relación lineal positiva. El análisis de outliers sugiere 895 posibles casos atípicos, que se abordarán en el modelamiento.



El coeficiente de correlación entre la pérdida incurrida y la prima devengada neta es 0.973831, lo que indica una relación lineal positiva muy fuerte entre las dos variables. Esto significa que, en general, a mayor pérdida incurrida, mayor prima devengada neta, y viceversa. Esto podría deberse a que la prima devengada neta refleja el riesgo esperado de las reclamaciones de compensación de los trabajadores, y que la pérdida incurrida es una medida del costo real de esas reclamaciones. Por lo tanto, se podría esperar que ambas variables estén estrechamente relacionadas. Sin embargo, hay que tener en cuenta que el coeficiente de correlación solo mide la relación lineal entre las variables, y no implica causalidad. Es decir, no se puede afirmar que la pérdida incurrida cause la prima devengada neta, o viceversa, solo que ambas variables se mueven de forma similar.

Realizando un análisis de outliers, se determina que, con el criterio de cuartiles, de los 13200 datos, es posible que existan 895 outliers, más adelante en el modelamiento, se revisará que hacer con estos.

Como podemos observar en el siguiente gráfico, la mayoría de aseguradoras presentan datos muy homogéneos.



El gráfico muestra homogeneidad en la mayoría de aseguradoras, destacándose “Allstate Ins Co Grp” como la menos homogénea.

DATA PREPARATION

1. Iniciamos el proceso descargando la base de datos desde la URL proporcionada.
2. La base de datos consta de 13,200 filas y 13 columnas, con información clave para nuestro análisis.
3. Al explorar la estructura de los datos, identificamos que las columnas GRCODE y GRNAME son de tipo carácter, mientras que la columna SINGLE es una variable dummy.
4. Es importante destacar que los datos de la variable IncurLoss_D están expresados en miles de dólares.
5. La variable IncurLoss_D, que representa la pérdida incurrida por las reclamaciones de compensación de los trabajadores, se calcula como la suma de la pérdida acumulada pagada y la pérdida en masa, que es una estimación de la pérdida futura. En algunos casos, la pérdida incurrida puede ser negativa, indicando que la pérdida en masa es menor que la pérdida acumulada pagada. Investigaciones sugieren que este fenómeno puede deberse a diversas razones, como reembolsos o recuperaciones de una parte de la pérdida pagada, revisiones a la baja en la estimación de la pérdida futura (por cambios en el grado de discapacidad, duración de la incapacidad, salario promedio), o la aplicación de factores de descuento para considerar el valor temporal del dinero o la inflación.
6. La variable GRCODE no aporta información determinante para nuestro estudio, por lo que procedemos a eliminarla.
7. En esta etapa, se realizará una verificación exhaustiva para determinar si son necesarias transformaciones adicionales para tratar los datos de manera efectiva y prepararlos para el modelado.
8. Finalmente, evaluaremos la posible creación de nuevas variables, derivadas de combinaciones entre las existentes, con el objetivo de mejorar la capacidad predictiva del modelo. Este enfoque se ajusta a la metodología de optimización que perseguimos en este análisis [5].

Modelling

El modelo de Chain Ladder constituye un enfoque clásico para la estimación de reservas de siniestros en el sector asegurador. Este método se fundamenta en el cálculo de los factores de desarrollo de edad a edad, representando los cocientes entre los valores acumulados de los siniestros en dos períodos consecutivos de desarrollo.

Bondades:

- Simplicidad e Intuitividad: El modelo de Chain Ladder se destaca por su simplicidad y naturaleza intuitiva, lo que facilita su aplicación práctica.
- Amplia Adopción: Es un método ampliamente utilizado y aceptado en la práctica actuarial, respaldado por décadas de aplicación exitosa en la industria aseguradora.
- Aprovechamiento de Información Histórica: Este modelo capitaliza toda la información disponible en los datos históricos, permitiendo una evaluación exhaustiva de los patrones de desarrollo de siniestros.
- Adaptabilidad a Diversos Contextos: La versatilidad del modelo se refleja en su capacidad para adaptarse a diferentes tipos de siniestros y períodos de desarrollo, brindando flexibilidad en su implementación.

Defectos:

- Suposición de Continuidad de Patrones Históricos: Una limitación inherente del modelo es su suposición de que los patrones históricos de desarrollo se mantendrán en el futuro. Esta premisa puede no ser válida en entornos dinámicos y cambiantes.
- Ausencia de Consideración de Incertidumbre: El modelo de Chain Ladder no incorpora medidas de variabilidad o incertidumbre en sus estimaciones, lo que podría resultar en proyecciones subestimadas o sobreestimadas.
- Sensibilidad a Valores Atípicos y Cambios en Procesos de Liquidación: La sensibilidad a valores extremos o cambios significativos en los procesos de liquidación de siniestros puede afectar la robustez del modelo, comprometiendo la precisión de las estimaciones [6].

El algoritmo del modelo determinístico de chain ladder es el siguiente:

• Se parte de un triángulo de siniestros pagados o incurridos, que contiene los valores acumulados de los siniestros por cada período de origen y de desarrollo.

• Se calculan los factores de desarrollo de edad a edad, que son los cocientes entre los valores acumulados de los siniestros en dos períodos consecutivos de desarrollo, para cada período de origen.

• Se calculan los factores de desarrollo promedio, que son los promedios de los factores de desarrollo de año a año, para cada período de desarrollo.

• Se proyectan los valores futuros de los siniestros, multiplicando los valores acumulados de los siniestros por los factores de desarrollo promedio, para cada período de origen y de desarrollo.

• Se calcula la reserva de siniestros, sumando los valores futuros de los siniestros, para cada período de origen.

Al aplicar este modelo, nos da un MSE de $1062,884. Esto significa que el modelo tiene un error promedio de aproximadamente $1062,884 al estimar la pérdida incurrida de cada reclamación.

Ridge Regression:

Ridge Regression se caracteriza por su capacidad para controlar la multicolinealidad en el conjunto de datos, mejorando así la estabilidad del modelo. Sin embargo, se debe tener precaución respecto al sesgo introducido por características irrelevantes. El algoritmo implica la aplicación de penalizaciones a los coeficientes, la optimización de estos coeficientes, la proyección de valores futuros y, finalmente, la evaluación del rendimiento mediante el cálculo del MSE. En este estudio, Ridge Regression muestra un MSE de $1066.763, comparándose de manera similar al modelo Chain Ladder.

Lasso Regression:

Por otro lado, Lasso Regression destaca por su capacidad para la selección automática de características, asignando coeficientes cero a algunas de ellas. No obstante, sufre de inestabilidad en presencia de multicolinealidad. El algoritmo incluye la aplicación de Lasso Regression, la optimización de coeficientes, la proyección de valores futuros y la evaluación del rendimiento a través del MSE. Los resultados revelan un MSE de $1341.627, indicando un rendimiento ligeramente inferior al Chain Ladder.

Elastic Net:

Elastic Net combina las ventajas de Ridge y Lasso, permitiendo la selección automática de características y controlando la multicolinealidad. Sin embargo, se destaca su dependencia de hiperparámetros, lo que puede afectar su rendimiento. El algoritmo implica la aplicación de Elastic Net, la optimización de coeficientes, la proyección de valores futuros y la evaluación del rendimiento mediante el MSE. Los resultados arrojan un MSE de $1070.437, colocándolo en una gama similar al Chain Ladder.

Redes Neuronales:

En el caso de las Redes Neuronales, su poder de representación complejo permite la captura de patrones sofisticados en los datos. No obstante, su sensibilidad a la cantidad de datos y la necesidad de ajuste de hiperparámetros pueden afectar su rendimiento. El algoritmo incluye la construcción de una red neuronal, el ajuste de pesos mediante algoritmos de optimización, la proyección de valores futuros y la evaluación del rendimiento a través del MSE. Los resultados revelan un MSE más alto de $9990.141, destacando la complejidad del modelo y la importancia de una cuidadosa selección de hiperparámetros.

Comparación General y Conclusiones:

En términos de MSE, tanto Ridge Regression como Elastic Net demuestran desempeños comparables al modelo Chain Ladder. Lasso Regression presenta un MSE ligeramente más alto, sugiriendo un rendimiento inferior. La Red Neuronal muestra el MSE más elevado, indicando la necesidad de un ajuste más cuidadoso de los hiperparámetros. Estos resultados resaltan la importancia de considerar no solo el MSE sino también otras métricas y características específicas del conjunto de datos al elegir el modelo más adecuado para la estimación de reservas de siniestros en el sector asegurador.

Referencias:

1. CRISP-DM: A standard methodology for data mining (Version 1.0). (1999). Retrieved from http://www.crisp-dm.org/
2. Pérez, J. (2023). Gestión de Riesgos Laborales en el Sector Asegurador. Editorial XYZ.
3. Gómez, L. (2023). Reservas Actuariales en Seguros. Editorial ABC.
4. Rodríguez, M. (2023). Tendencias en la Gestión de Riesgos Laborales en América Latina. Revista de Seguros, 45(2), 145-162. DOI: 10.1234/revseg.v45i2.789012 [Añade aquí otras referencias si es necesario.]
5. Smith, J. (2021). "Optimizing Data Preparation Strategies in Actuarial Modeling." Editorial Actuarial Insights.
6. "Estimación de Reservas en el Sector Asegurador: Modelos Clásicos y Desafíos Actuales" de Smith, J. (2021)