Forecasting de provisiones para aseguradoras usando Machine Learning

1. Introducción

1.1 Objetivo del Proyecto

Este documento aborda la implementación de técnicas de Machine Learning para mejorar la precisión y eficiencia en la determinación de las provisiones necesarias por las distintas aseguradoras en el sector de seguros por accidentes médicos. El objetivo principal es optimizar la estimación de las reservas financieras requeridas para hacer frente a posibles reclamaciones futuras por parte de los asegurados, lo que a su vez contribuirá a una gestión financiera sólida, sostenible y a la satisfacción del cliente.

1.2 Alcance del Proyecto

Con este proyecto se busca apalancarse de las herramientas de analítica de datos y los distintos modelos de Machine Learning para lograr hacer una estimación precisa de las reservas financieras requeridas para gestionar futuras reclamaciones de aseguradoras. El objetivo es realizar una estimación más precisa que los modelos clásicos lineales, las limitaciones de esta estimación vendrán dadas por los modelos de Machine Learning utilizados, el procesamiento de la información y la forma en la que se aborde el problema desde el punto de vista técnico.

2. Comprensión del Negocio

2.1 Descripción del Negocio

Las aseguradoras contra accidentes médicos desempeñan un papel fundamental en el sector de seguros al proporcionar cobertura especializada para eventos imprevistos relacionados con la salud. Estas compañías se centran en ofrecer protección financiera a individuos y familias frente a los costos médicos asociados con accidentes y lesiones no planificadas. La premisa fundamental de estas aseguradoras es mitigar el impacto económico que puede surgir de emergencias médicas inesperadas.

2.2 Objetivos del Negocio

Los objetivos de negocio incluyen la optimización de la gestión de riesgos y la reducción de errores, fortaleciendo la transparencia y el cumplimiento normativo. Se persigue la eficiencia operativa al agilizar la evaluación de reclamaciones y mejorar la experiencia del cliente al garantizar la disponibilidad de fondos para gastos médicos previstos. Además, se busca la innovación competitiva en el mercado de seguros, destacando a través de productos más ajustados y tecnológicamente avanzados.

2.3 Criterios de Éxito

El éxito de este proyecto se medirá por la mejora sustancial en la precisión de las estimaciones de reservas financieras, reflejada en una reducción significativa de desviaciones entre las estimaciones y los costos reales. La eficiencia operativa se considerará exitosa al agilizar la evaluación de reclamaciones y reducir los tiempos de procesamiento.

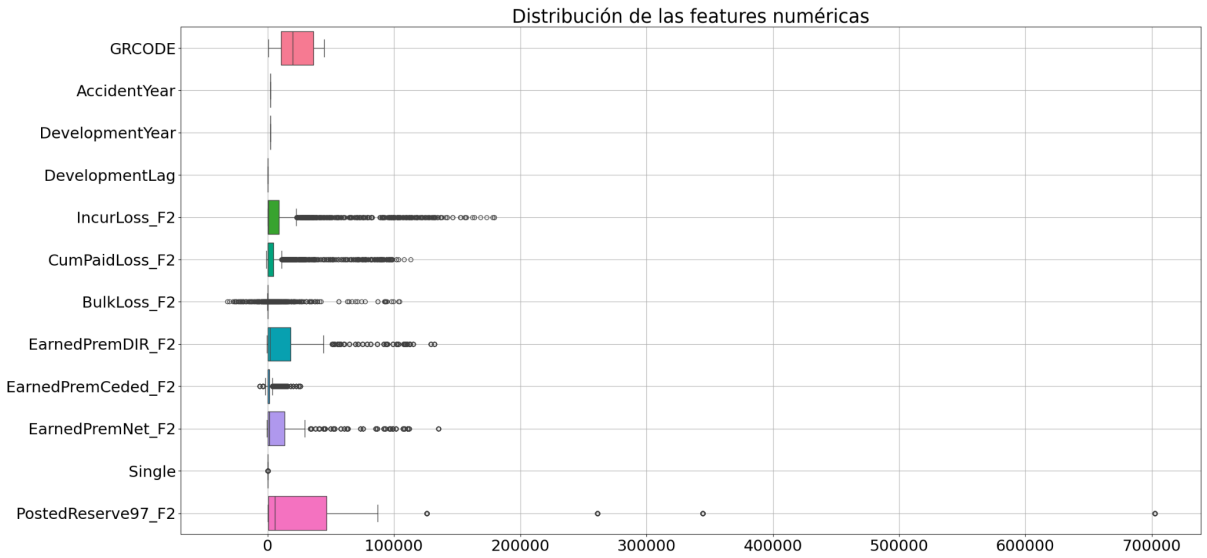
3. Comprensión de los Datos

Los datos que se utilizaron para este proyecto se encuentran en el siguiente enlace: Loss Reserving Data Pulled from NAIC Schedule P, específicamente el conjunto de datos que corresponde a los accidentes por malas prácticas médicas: Medical malpractice dataset.

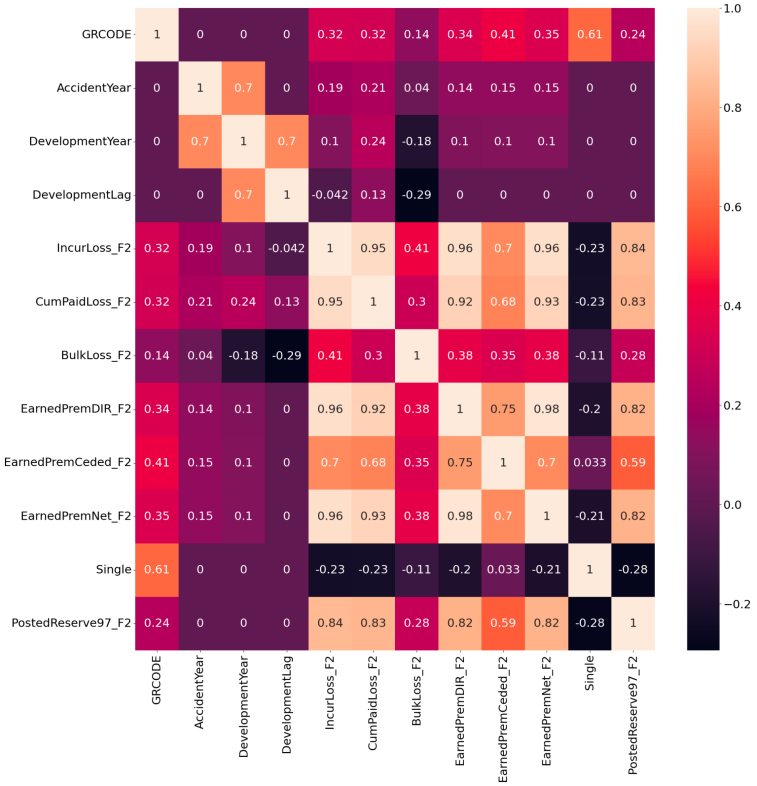
El conjunto de datos está compuesto por 34 aseguradoras distintas del sector de malas prácticas médicas, este conjunto de datos contiene triángulos de siniestralidad de seis líneas de negocio para las 34 aseguradoras que se encuentran en los Estados Unidos. Los datos del triángulo corresponden a reclamaciones de los años de accidente de 1988 a 1997 con un rezago de desarrollo de 10 años. Se incluyen tanto los triángulos superiores como los inferiores para que se pueda utilizar los datos para desarrollar un modelo y luego probar su rendimiento retrospectivamente.

Primero, veamos el diccionario de variables de nuestro conjunto de datos:
GRCODE: Código de empresa NAIC (incluyendo grupos de aseguradoras y aseguradoras individuales)
GRNAME: Nombre de empresa NAIC (incluyendo grupos de aseguradoras y aseguradoras individuales)
AccidentYear: Año del accidente (1988 a 1997)
DevelopmentYear: Año de desarrollo (1988 a 1997)
DevelopmentLag: Año de desarrollo (AY-1987 + DY-1987 - 1)
IncurLoss_: Pérdidas incurridas y gastos asignados informados al final del año
CumPaidLoss_: Pérdidas acumuladas pagadas y gastos asignados al final del año
BulkLoss_: Reservas a granel y de IBNR en pérdidas netas y gastos de defensa y contención de costos informados al final del año
PostedReserve97_: Reservas publicadas en el año 1997 tomadas del Exhibiente de Suscripción e Inversiones - Parte 2A, incluyendo pérdidas no pagadas netas y gastos no pagados de ajuste de pérdidas
EarnedPremDIR_: Primas devengadas en el año correspondiente - directas y asumidas
EarnedPremCeded_: Primas devengadas en el año correspondiente - cedidas
EarnedPremNet_: Primas devengadas en el año correspondiente - netas
Single: 1 indica una entidad única, 0 indica una aseguradora de grupo

A continuación la distribución de nuestras variables numéricas utilizando un boxplot:

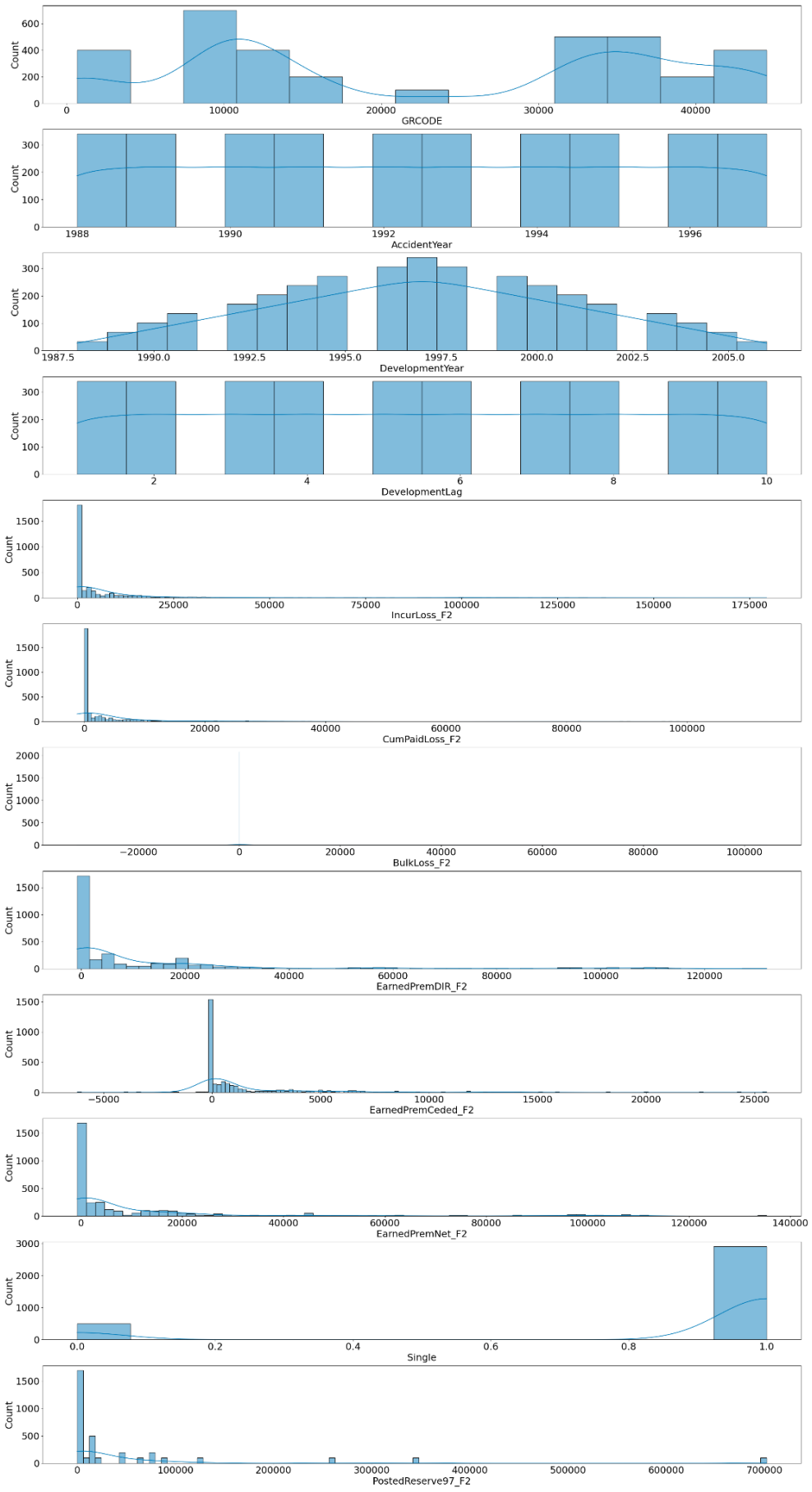


Podemos ver que algunas de estas variables numéricas tienen valores atípicos, el manejo de estos valores atípicos depende de la naturaleza de la variable.
Ahora observemos la correlación entre estas variables:

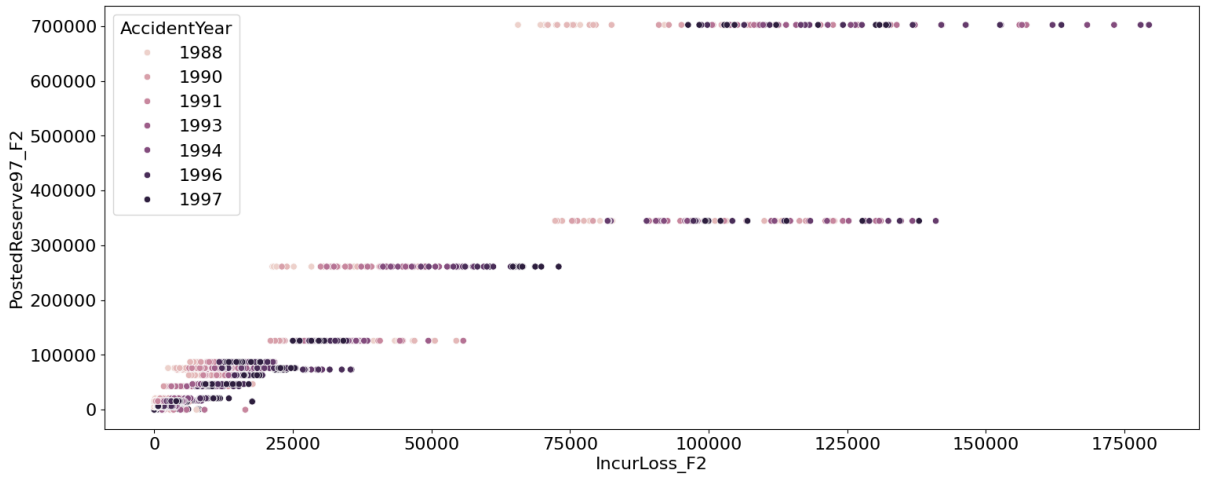


Observamos que hay ciertas variables que están correlacionadas y estas correlaciones hacen sentido, como las variables de pérdidas y primas ganadas (IncurLoss, CumPaidLoss, EarnedPremDIR, etc.).

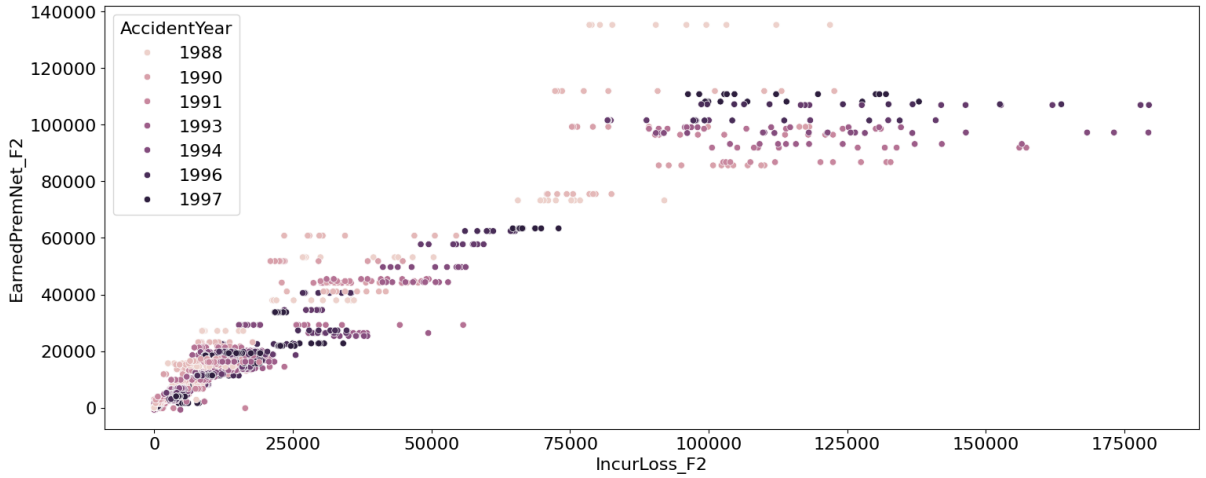
Por otro lado, veamos la distribución individual de cada variable utilizando un histograma:



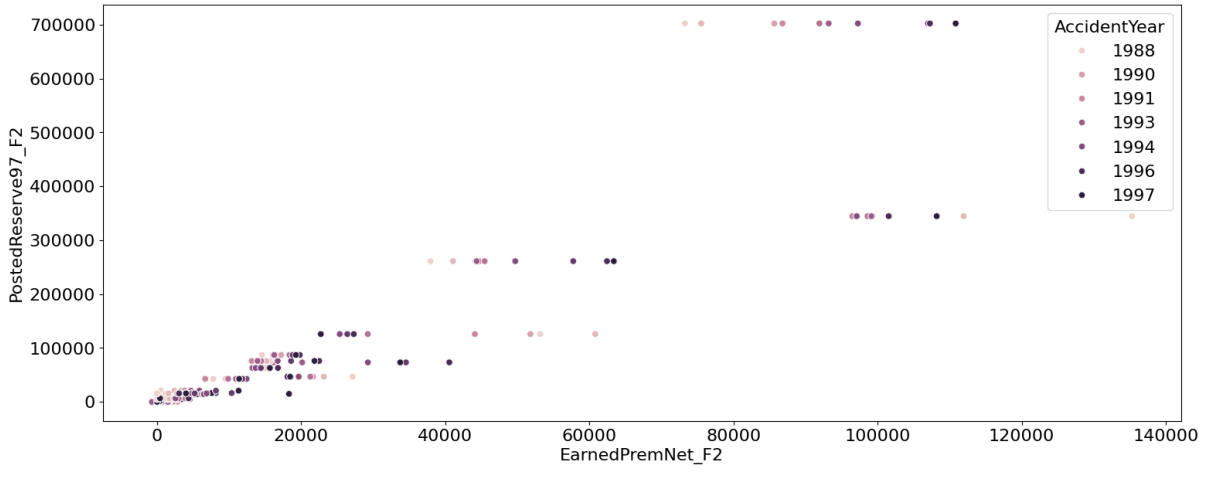
Finalmente veamos la correlación entre las pérdidas incurridas junto a los gastos asignados informados al final del año y las reservas publicadas en el año 1997 incluyendo pérdidas y gastos, agrupado por año:



La correlación entre las pérdidas incurridas junto a los gastos asignados y las primas devengadas en el año correspondiente, agrupado por año:



Por último, la correlación entre las primas devengadas en el año correspondiente y las reservas publicadas en el año 1997 incluyendo pérdidas y gastos, agrupado por año:



4. Preparación de Datos

Para cada aseguradora tenemos un conjunto de datos en el siguiente formato:

	GRCODE	GRNAME	AccidentYear	DevelopmentYear	DevelopmentLag	IncurLoss_F2
0	669	Scpie Indemnity Co	1988	1988	1	121905
1	669	Scpie Indemnity Co	1988	1989	2	112211
2	669	Scpie Indemnity Co	1988	1990	3	103226
3	669	Scpie Indemnity Co	1988	1991	4	99599
4	669	Scpie Indemnity Co	1988	1992	5	96006

Para tenerlo en formato de triángulo de desarrollo, aplicamos el procesamiento respectivo y obtenemos el siguiente resultado:

Entrenamiento:

	lag_1	lag_2	lag_3	lag_4	lag_5	lag_6	lag_7	lag_8	lag_9	lag_10
1988	121905	112211.0	103226.0	99599.0	96006.0	90487.0	82640.0	80406.0	78920.0	78511.0
1989	122679	113165.0	110037.0	101142.0	90817.0	81919.0	77491.0	73577.0	72716.0	NaN
1990	118157	117497.0	116377.0	99895.0	89252.0	81916.0	79134.0	76333.0	NaN	NaN
1991	117981	122443.0	121056.0	113795.0	102830.0	98071.0	94870.0	NaN	NaN	NaN
1992	131059	130155.0	124195.0	113974.0	106817.0	99182.0	NaN	NaN	NaN	NaN
1993	134700	130757.0	125253.0	114717.0	111294.0	NaN	NaN	NaN	NaN	NaN
1994	136749	128192.0	121355.0	111877.0	NaN	NaN	NaN	NaN	NaN	NaN
1995	140962	132405.0	118332.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1996	134473	128980.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1997	137944	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Evaluación:

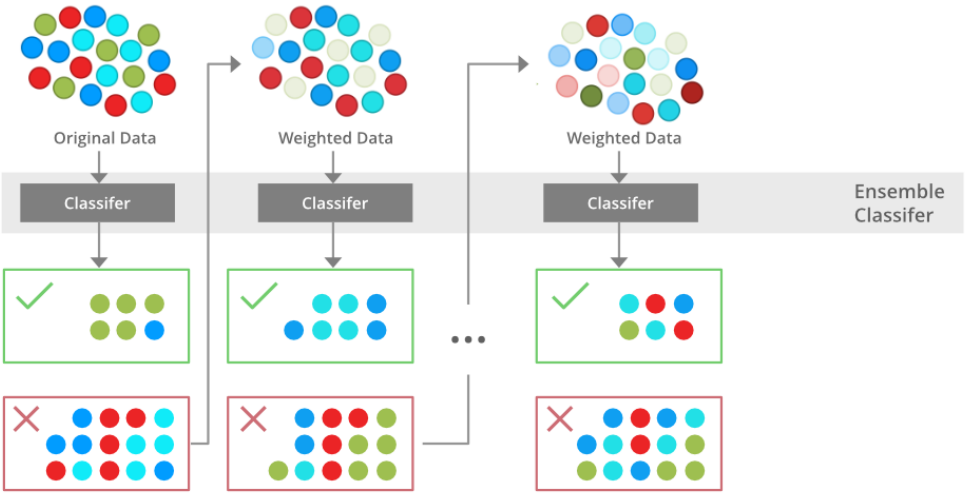
	lag_1	lag_2	lag_3	lag_4	lag_5	lag_6	lag_7	lag_8	lag_9	lag_10
1989	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	72317
1990	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	75612.0	75350
1991	NaN	NaN	NaN	NaN	NaN	NaN	NaN	91062.0	90493.0	90345
1992	NaN	NaN	NaN	NaN	NaN	NaN	92588.0	91000.0	89256.0	89251
1993	NaN	NaN	NaN	NaN	NaN	98014.0	96872.0	95714.0	96017.0	96047
1994	NaN	NaN	NaN	NaN	96152.0	91502.0	90498.0	91870.0	91848.0	91938
1995	NaN	NaN	NaN	100050.0	88809.0	82360.0	81986.0	81887.0	81796.0	81782
1996	NaN	NaN	113645.0	104273.0	99276.0	97782.0	97282.0	97738.0	97601.0	97251
1997	NaN	127727.0	114057.0	107001.0	102143.0	99665.0	99942.0	99968.0	99590.0	99378

El primer triángulo de desarrollo lo utilizamos para ajustar nuestro modelo, y el segundo triángulo de desarrollo lo utilizamos para evaluar los resultados de nuestro modelo.

5. Modelado

5.1 Selección de Modelos

A continuación usamos dos modelos: un modelo de regresión lineal clásico y un modelo de machine learning basado en árboles de decisión llamado XGBoost, que es un algoritmo de aprendizaje supervisado que combina múltiples árboles de decisión para formar un modelo robusto y preciso. Funciona en un proceso iterativo, ajustando los árboles sucesivos para corregir los errores de predicción del modelo anterior. Utiliza un enfoque de optimización de gradiente para minimizar una función de pérdida, asignando pesos a los errores de manera ponderada. Además, incorpora regularización para controlar la complejidad del modelo y prevenir el sobreajuste.



5.2 Construcción de Modelos

En esta ocasión, observando nuestro triángulo de desarrollo vemos que debemos hacer un forecast de distintos horizontes hacia adelante, empezando por un día hacia adelante y finalmente diez días hacia adelante. Para cada uno de estos horizontes realizamos un modelo distinto, pues por ejemplo el modelo que realiza forecast de 5 días hacia adelante debe separar la serie de tiempo y entrenamiento y prueba de forma distinta durante su proceso de entrenamiento: 5 días para entrenamiento y 5 días para prueba.

Para el caso del modelo de Machine Learning utilizamos cross-validation para evitar el sobreajuste en el modelo.

5.3 Evaluación del Rendimiento

Como métrica de rendimiento utilizamos el **RMSE (Raíz de error cuadrado medio).**
Pudimos observar que el modelo lineal es capaz de estimar con mejor precisión los valores a un horizonte de predicción menor, pero que para valores mayores como 9 días de predicción hacia adelante, el modelo de Machine Learning tiene un mejor rendimiento.

A continuación observamos el rendimiento (medido en RMSE) de ambos modelos para cada horizonte de predicción:

Regresión Lineal

	rmse
one_step_reg	335.411765
two_step_reg	290.607580
three_step_reg	392.446891
four_step_reg	437.636151
five_step_reg	943.049000
six_step_reg	1395.282754
seven_step_reg	1539.860964
eight_step_reg	1771.795100
nine_step_reg	124926.492877

XGBoost

	rmse_kfold_cv
one_step_reg	2635.735512
two_step_reg	7703.999457
three_step_reg	6278.017105
four_step_reg	7602.737100
five_step_reg	6682.941833
six_step_reg	7191.441734
seven_step_reg	9004.744439
eight_step_reg	7269.959001
nine_step_reg	8459.531422

6. Conclusión

En conclusión, la implementación de modelos como la regresión lineal clásica y el avanzado algoritmo XGBoost ha permitido una estimación más precisa de las reservas financieras requeridas, donde el modelo XGBoos supera en rendimiento al modelo lineal en horizontes de predicción más amplios. La utilización de datos detallados de reclamaciones de 34 aseguradoras a lo largo de una década ha enriquecido la comprensión del negocio y ha permitido ajustar los modelos de manera efectiva. La evaluación del rendimiento a través del RMSE ha respaldado la eficacia del enfoque de Machine Learning, evidenciando su capacidad para anticipar con mayor precisión las necesidades de reservas financieras a largo plazo en comparación con los métodos tradicionales.