Universidad Nacional Autónoma de México

Facultad de Estudios Superiores Acatlán

Modelado Predictivo para la Contratación de Seguros de Vehículos

Mtro. Alejandro Leonel Martinez Elizalde

Proyecto Final

Junio 2 de 2025

Martin Mejia Julio Yahir

Objetivo general

Desarrollar un modelo predictivo basado en técnicas de estadística y aprendizaje automático que permita anticipar si un cliente potencial estaría interesado en contratar un seguro de vehículos, utilizando un conjunto de datos que contiene información demográfica, características del vehículo, historial de seguros y datos del canal de contacto.

Objetivos específicos

— Realizar un análisis exploratorio de los datos disponibles para identificar patrones, relaciones entre variables y posibles problemas de calidad de los datos.

— Evaluar distintas técnicas de modelado (regresión logística, Random Forest, CatBoost y redes neuronales) para determinar cuál ofrece el mejor desempeño en términos de precisión, sensibilidad y especificidad.

— Implementar y ajustar modelos de clasificación supervisada, considerando tanto interpretabilidad como desempeño predictivo.

— Validar el modelo seleccionado mediante técnicas de validación cruzada y análisis de métricas de desempeño fuera de muestra.

— Generar visualizaciones que faciliten la interpretación de los resultados y la comunicación de los hallazgos a tomadores de decisiones.

ÍNDICE

1.- Introducción

En el contexto actual del sector asegurador, la competencia entre compañías se ha intensificado significativamente debido a la transformación digital, el acceso a grandes volúmenes de datos y la creciente exigencia de los consumidores por servicios personalizados y eficientes. Frente a este panorama, las aseguradoras buscan incorporar herramientas analíticas avanzadas que les permitan anticipar las necesidades y comportamientos de sus clientes, optimizar sus procesos comerciales y tomar decisiones más informadas.

Uno de los retos clave en el ámbito de los seguros de automóviles es la capacidad de prever qué clientes están realmente interesados en contratar una póliza. Esta información resulta valiosa no solo para aumentar las tasas de conversión, sino también para asignar eficientemente los recursos en campañas de marketing y canales de atención.

Con esta motivación, el presente proyecto se enfoca en el desarrollo de un modelo predictivo que permita anticipar la probabilidad de que un cliente contrate un seguro de vehículo, a partir de un conjunto de variables relacionadas con sus características demográficas, su historial de seguros, las propiedades del vehículo y los datos de la póliza.

Para ello, se utilizó un conjunto de datos de una aseguradora con 382,154 registros, que incluye variables como la edad del cliente, la antigüedad del vehículo, el historial de daños, si ha tenido seguros previos, el canal de ventas utilizado, entre otros. La variable objetivo es binaria y refleja el interés del cliente por contratar el seguro.

A través del uso de herramientas de análisis de datos y machine learning en Python, se entrenaron distintos modelos predictivos, incluyendo regresión logística, Random Forest, CatBoost y redes neuronales. Se comparó su rendimiento mediante métricas como la precisión, el área bajo la curva ROC (AUC) y el F1-score, con el fin de seleccionar el modelo más adecuado para esta tarea.

Este estudio no solo tiene relevancia práctica en términos de utilidad para la aseguradora, sino que también representa una aplicación concreta de técnicas estadísticas y de ciencia de datos al campo actuarial, combinando teoría, análisis exploratorio y modelado avanzado.

2.- Planteamiento del problema

En el sector de los seguros de automóviles, identificar con anticipación a los clientes con mayor probabilidad de contratar una póliza representa una ventaja competitiva fundamental. Las aseguradoras invierten recursos considerables en estrategias de marketing, fuerza de ventas y atención al cliente, por lo que predecir de forma precisa el interés real de un posible asegurado puede traducirse en una optimización significativa de estos esfuerzos.

Sin embargo, este proceso presenta diversos desafíos. Por un lado, el comportamiento de los clientes es complejo y está influenciado por múltiples factores, tales como su perfil demográfico, sus antecedentes con seguros, las características del vehículo, y los canales a través de los cuales se realiza la interacción con la aseguradora. Por otro lado, la información disponible suele estar distribuida en bases de datos amplias y con múltiples variables, lo que dificulta extraer conclusiones de manera directa sin el uso de herramientas analíticas especializadas.

La aseguradora en cuestión cuenta con un extenso conjunto de datos que registra información relevante de más de 380,000 clientes potenciales. No obstante, no dispone de un mecanismo automatizado que permita anticipar, con un nivel aceptable de certeza, si un cliente terminará contratando el seguro. Esta carencia implica una menor eficiencia en las campañas comerciales y en la gestión de riesgos, así como una posible pérdida de oportunidades de negocio.

Por tanto, se plantea como problema central la necesidad de construir un modelo predictivo preciso y confiable que, con base en las variables disponibles, permita estimar la probabilidad de que un cliente contrate un seguro de vehículo. Esta solución debe ser escalable, interpretable en su implementación y útil para la toma de decisiones estratégicas dentro de la aseguradora.

Justificación

En un entorno altamente competitivo como el de los seguros automotrices, la eficiencia en la captación de nuevos clientes es un factor clave para la sostenibilidad y el crecimiento de las aseguradoras. Los recursos destinados a campañas publicitarias, visitas comerciales, llamadas y atención personalizada deben ser dirigidos con la mayor precisión posible hacia clientes con alta probabilidad de conversión. En este contexto, el uso de modelos predictivos basados en aprendizaje automático y estadística avanzada representa una oportunidad estratégica para mejorar los procesos de toma de decisiones.

El presente proyecto justifica su realización en varios niveles:

Valor estratégico: Anticipar si un cliente está o no interesado en contratar un seguro permite enfocar los esfuerzos comerciales en los perfiles más propensos, optimizando la asignación de recursos y mejorando los indicadores clave de desempeño (KPIs) como el retorno sobre inversión publicitaria o la tasa de conversión.

Aprovechamiento de los datos disponibles: La aseguradora cuenta con un dataset robusto de más de 382,000 registros, que incluye variables demográficas, características del vehículo, historial de seguros y datos sobre el canal de contacto. Este volumen y diversidad de información ofrecen una base sólida para desarrollar modelos predictivos con alto potencial explicativo.

Aplicación práctica de técnicas estadísticas y de machine learning: El desarrollo del modelo permite aplicar conocimientos de regresión logística, algoritmos de clasificación como Random Forest y CatBoost, así como redes neuronales. Esto proporciona no sólo una solución práctica, sino también una validación empírica del uso de dichas herramientas en problemas reales del ámbito actuarial y empresarial.

Transferibilidad del modelo: El enfoque utilizado puede adaptarse fácilmente a otras aseguradoras o sectores similares que enfrenten retos de predicción del comportamiento del cliente, haciéndolo escalable y reutilizable en otros contextos.

3.- Marco teórico

Fundamentos del seguro automotriz

El seguro automotriz es un contrato mediante el cual una compañía aseguradora se compromete a indemnizar al asegurado por daños materiales o personales derivados del uso de un vehículo, a cambio del pago de una prima. Existen diferentes tipos de cobertura, entre ellas: responsabilidad civil, daños materiales, robo total, gastos médicos ocupantes, entre otros. Las aseguradoras deben gestionar riesgos asociados a los siniestros, pero también optimizar su estrategia de ventas anticipando qué clientes tienen mayor probabilidad de adquirir un seguro.

Conceptos de análisis predictivo

El análisis predictivo utiliza técnicas estadísticas y de machine learning para predecir eventos futuros a partir de datos históricos. Su aplicación en seguros permite anticipar comportamientos como la adquisición de productos, la cancelación de pólizas o la probabilidad de siniestros. El proceso suele incluir la recolección de datos, limpieza, transformación, selección de características, entrenamiento de modelos y evaluación del desempeño.

Métodos:

Los métodos permiten asignar una categoría (por ejemplo, “sí contratará el seguro” o “no lo contratará”) a cada observación en función de sus características. Entre los modelos utilizados en este proyecto se encuentran:

Regresión logística: Modelo estadístico que estima la probabilidad de un resultado binario a partir de una combinación lineal de variables explicativas.

Random Forest: Ensamble de múltiples árboles de decisión entrenados sobre subconjuntos aleatorios de los datos, que mejora la precisión y reduce el sobreajuste.

CatBoost: Algoritmo de gradient boosting desarrollado por Yandex, eficiente en variables categóricas y con buen rendimiento en tareas clasificatorias.

Redes neuronales: Modelos inspirados en el funcionamiento del cerebro humano, capaces de capturar patrones complejos mediante capas de procesamiento.

Métricas de evaluación

Para medir el desempeño de los modelos se utilizan diversas métricas:

Precisión (Accuracy): Proporción de predicciones correctas sobre el total de observaciones.

Recall (Sensibilidad): Proporción de verdaderos positivos detectados sobre todos los positivos reales.

Precisión (Precision): Proporción de verdaderos positivos sobre el total de predicciones positivas.

F1 Score: Media armónica entre precisión y recall, útil en casos de clases desbalanceadas.

AUC (Área bajo la curva ROC): Mide la capacidad del modelo para distinguir entre clases. Un AUC cercano a 1 indica un buen desempeño.

Descripción del conjunto de datos

El presente estudio se basa en un conjunto de datos obtenido de la plataforma Kaggle, titulado *Imbalanced Data Practice*. El dataset está orientado al desarrollo de modelos de clasificación en contextos de desbalance de clases, y ha sido adaptado a una problemática actuarial: una aseguradora desea predecir si un cliente estaría interesado en contratar un seguro de vehículos.

El conjunto de datos incluye información demográfica, características del vehículo, historial de seguros y detalles de la póliza de cada cliente. En total se cuenta con **382,154 registros**, cada uno correspondiente a un cliente diferente.

Variables incluidas

Variables demográficas

- **Sexo**: Género del cliente. (Nota: no se utilizó en el modelo final por no ser predictiva en las pruebas preliminares).

- **Edad (Age)**: Edad del cliente en años, con un rango entre **20 y 85 años**, y una media de **38.54 años**.

- **Tipo de código de región**: Clasificación geográfica de la residencia del cliente (no se utilizó directamente en el modelo por posible multicolinealidad y poca relevancia directa en la predicción).

Información sobre el vehículo

- **Vehicle_Age**: Edad del vehículo, clasificada en tres categorías:
  < 1 Year (Menos de un año)
  1-2 Year (Entre uno y dos años)
  > 2 Years (Más de dos años)

- **Vehicle_Damage**: Variable booleana que indica si el vehículo ha sufrido daños anteriormente:
  1: Sí 0: No

Historial de seguros

- **Previously_Insured**: Variable booleana que indica si el cliente ya contaba con un seguro anteriormente:

  1: Sí 0: No

Información de póliza

- **Policy_Sales_Channel**: Representa el canal por el cual la aseguradora contactó al cliente (puede incluir agentes, marketing digital, llamadas, entre otros). Su valor es numérico y varía entre **1 y 163**.

- **Annual_Premium**: Monto de la prima anual del seguro cotizada al cliente. Los valores oscilan entre **2,630 y 540,165 unidades monetarias**, con una media de **30,711.27**.

Variable objetivo (Target)

- **Response (target)**: Variable binaria que indica el interés del cliente en adquirir un seguro de vehículos:

  1: El cliente no está interesado en contratar el seguro.
  0: El cliente está interesado en contratar el seguro.

> Es importante señalar que el conjunto de datos está **desequilibrado**, ya que la mayoría de los clientes están interesados en contratar el seguro (clase 0), lo que representa un reto adicional para los modelos predictivos, especialmente en cuanto a precisión y sensibilidad de la clase minoritaria.

Variables predictoras utilizadas en el modelo

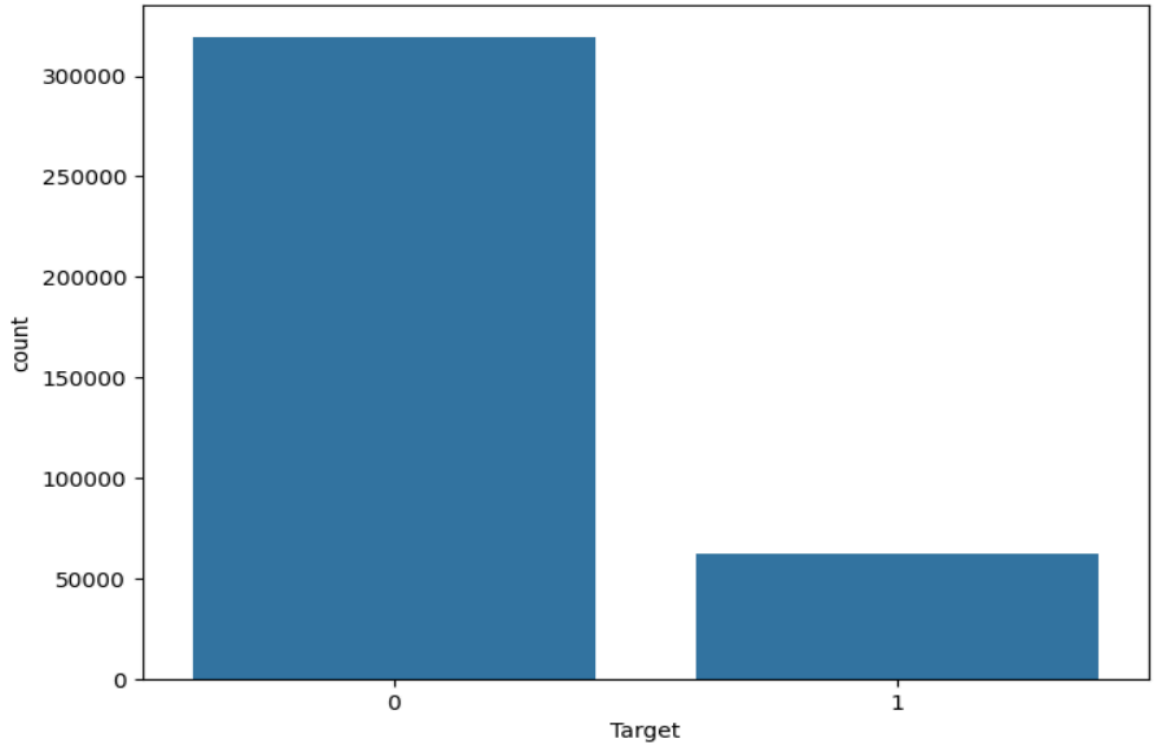
A partir del análisis exploratorio inicial y la evaluación de correlación y relevancia, se seleccionaron las siguientes variables como predictoras en los modelos:

Variable	Tipo de dato	Descripción
Previously_Insured	Booleana	Indica si el cliente tenía seguro previamente.
Annual_Premium	Numérica	Monto anual de la prima del seguro.
Vehicle_Damage	Booleana	Indica si el vehículo sufrió daños anteriormente.
Policy_Sales_Channel	Numérica	Canal por el cual se contactó al cliente.
Age	Numérica	Edad del cliente.
Vehicle_Age	Categórica	Antigüedad del vehículo en rangos predefinidos.

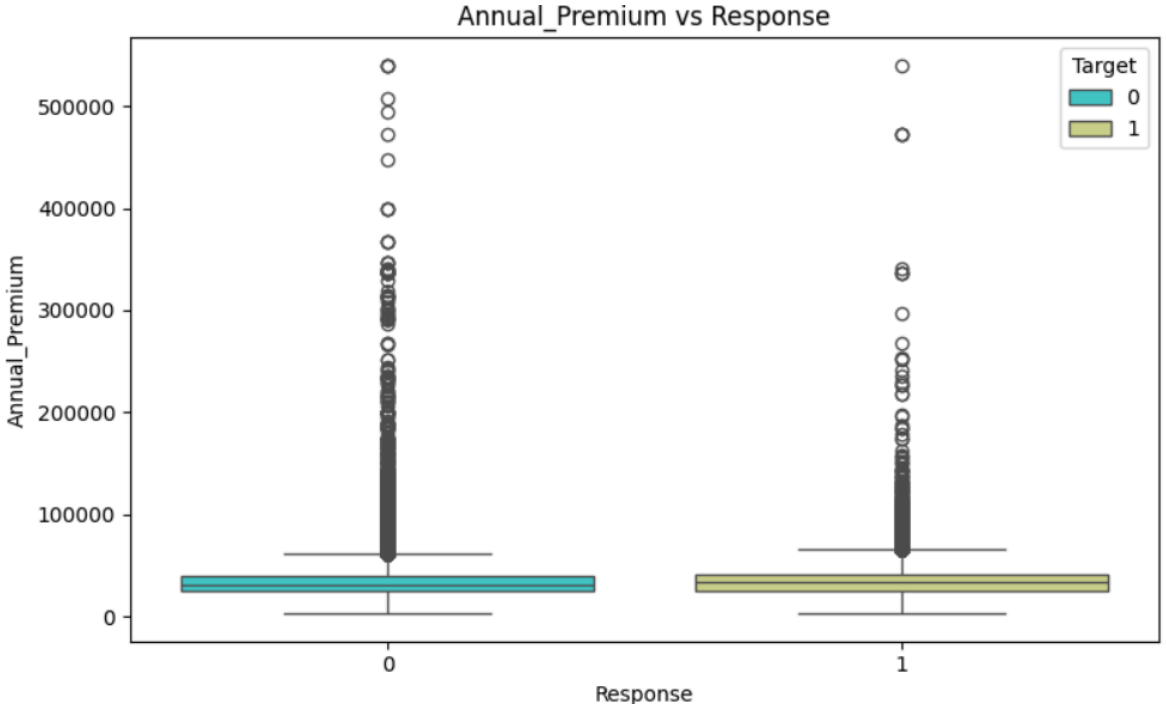
Estas variables fueron seleccionadas por su capacidad explicativa y disponibilidad consistente en el conjunto de datos, así como por su relevancia potencial desde el punto de vista del negocio asegurador.

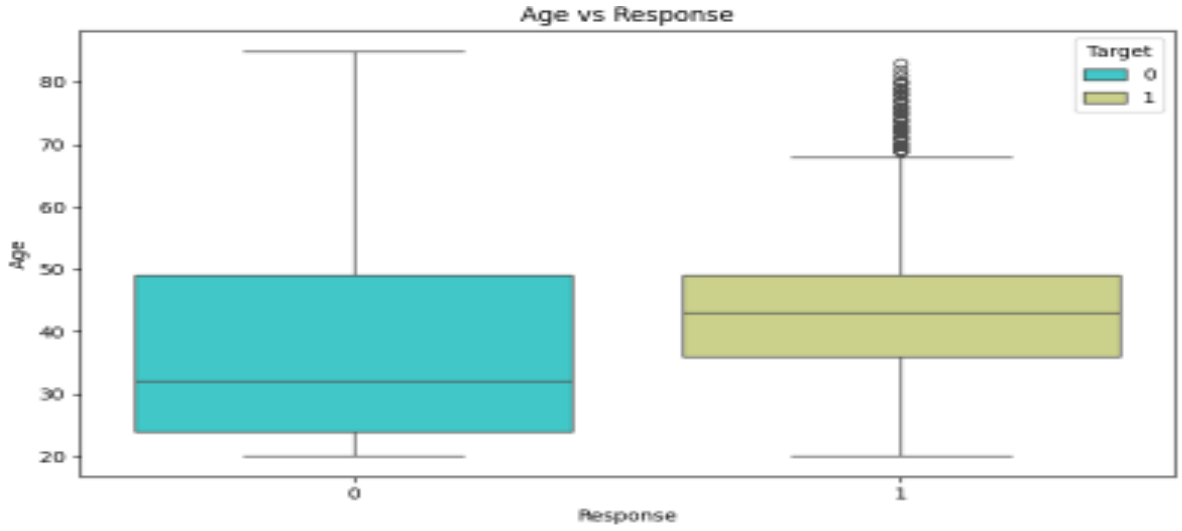
Gráficas de nuestras variables

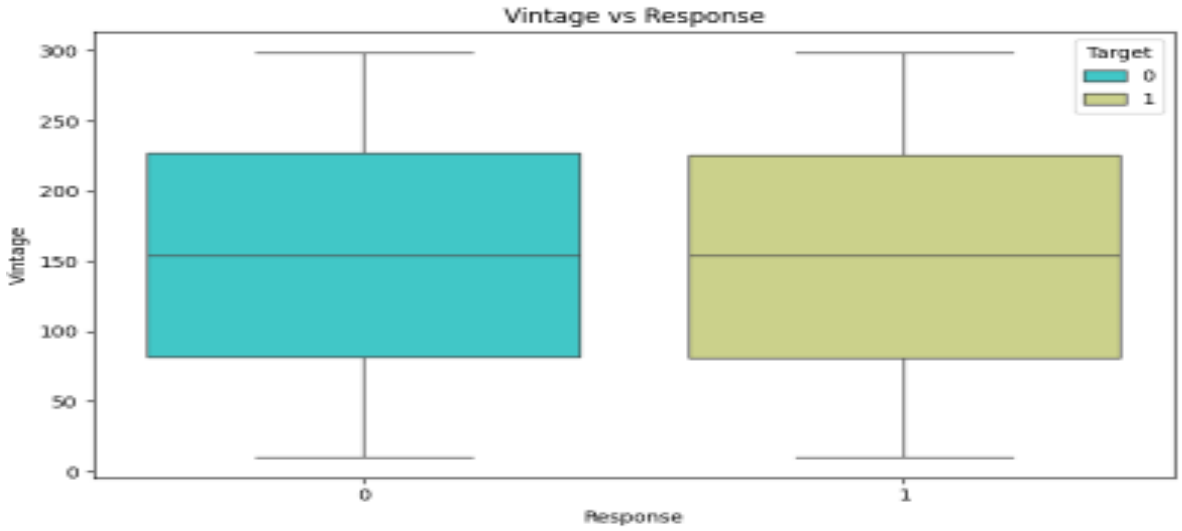
Visualización de los interesados y los no interesados



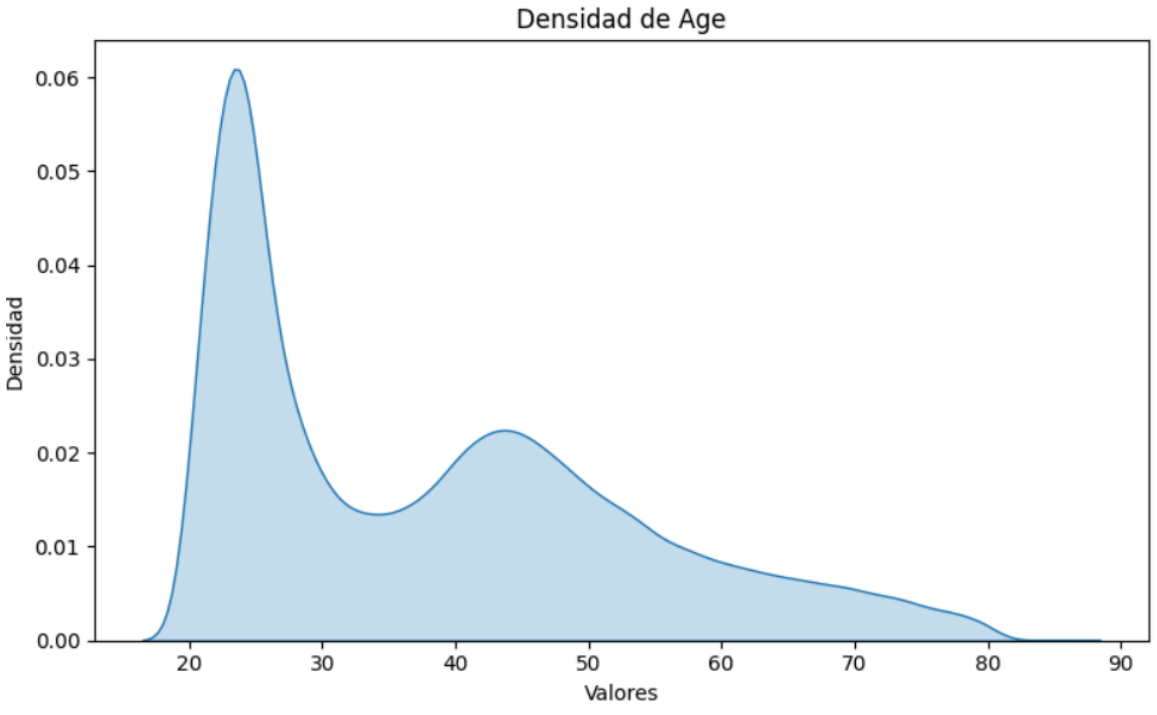
Dropboxes

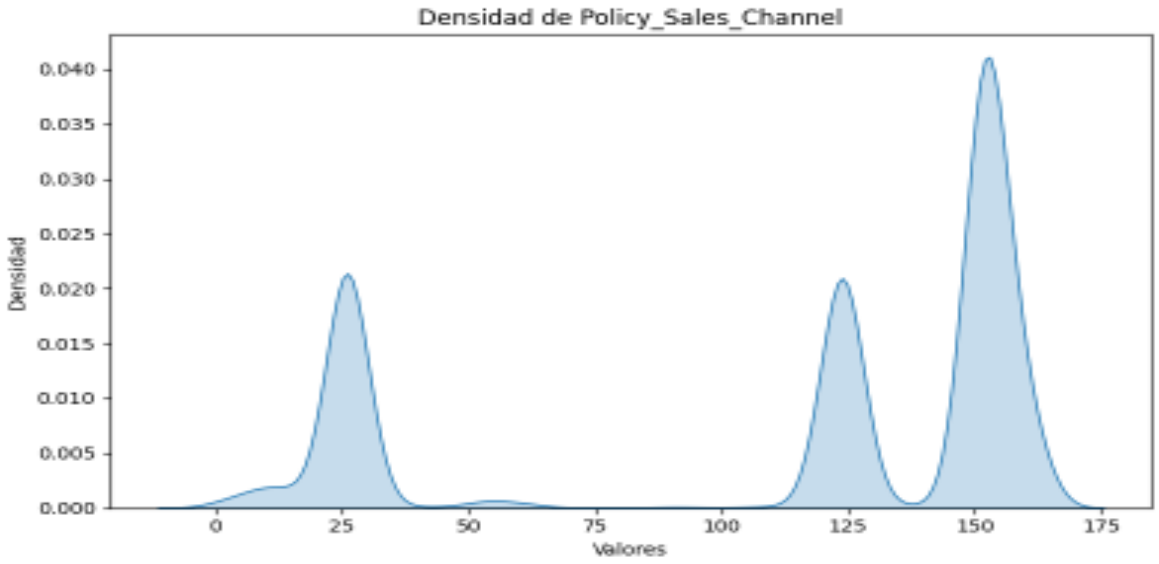


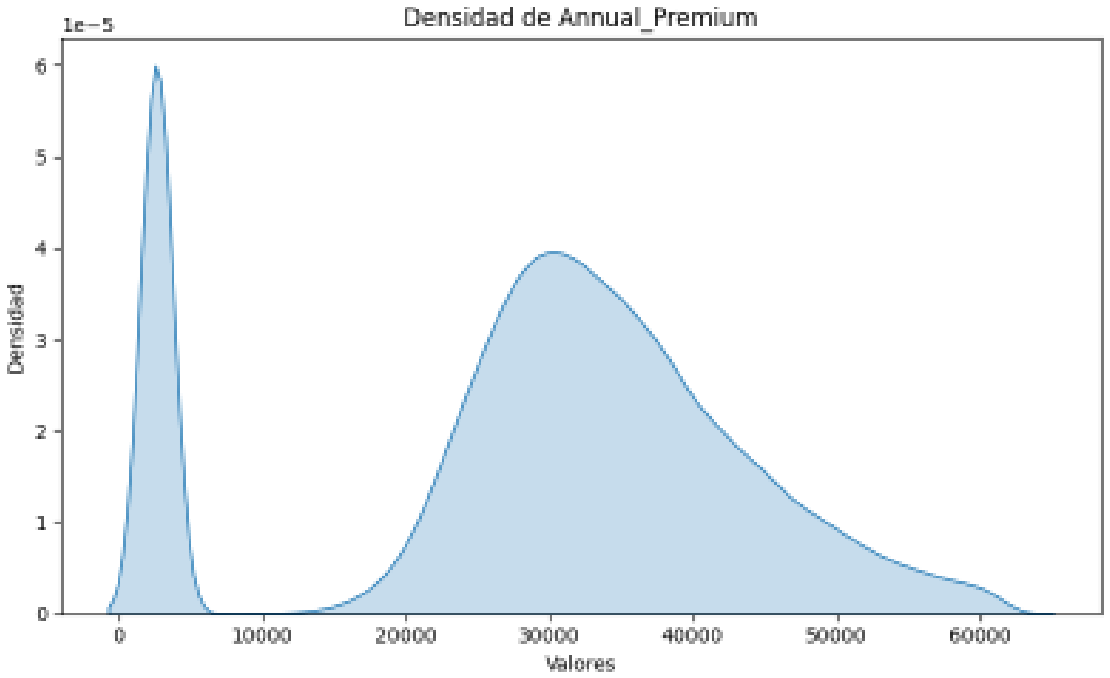


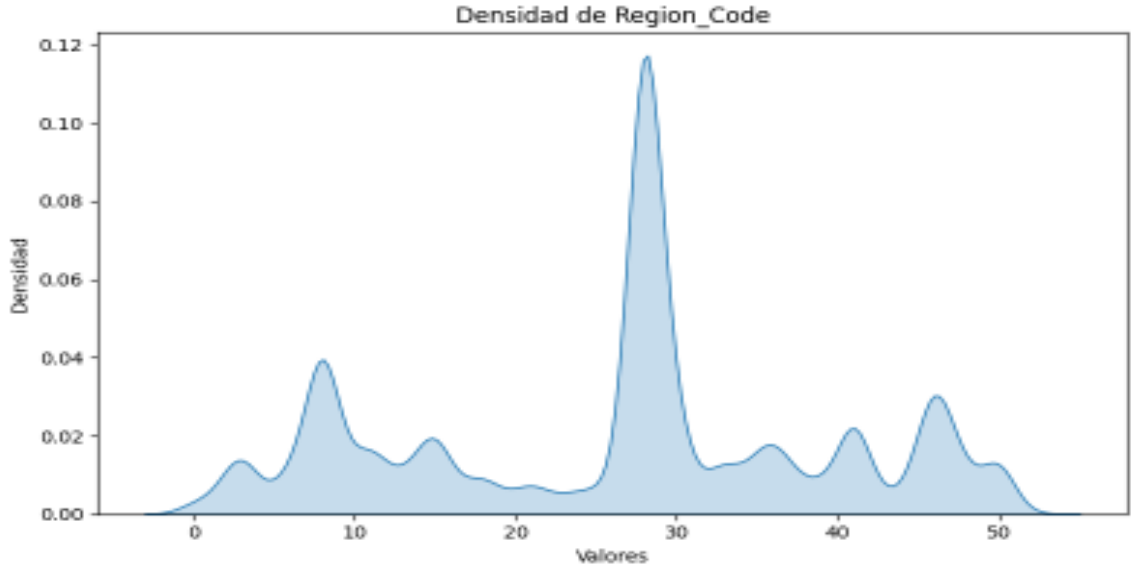


Funciones de densidad

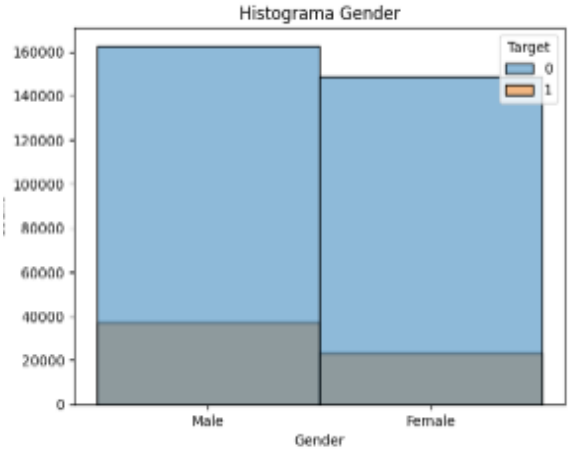


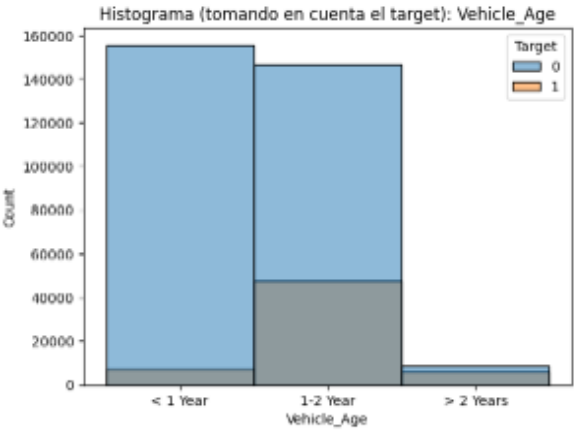

Densidad de Policy_Sales_Channel
Densidad
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
25
50
75
100
125
150
175
Valores

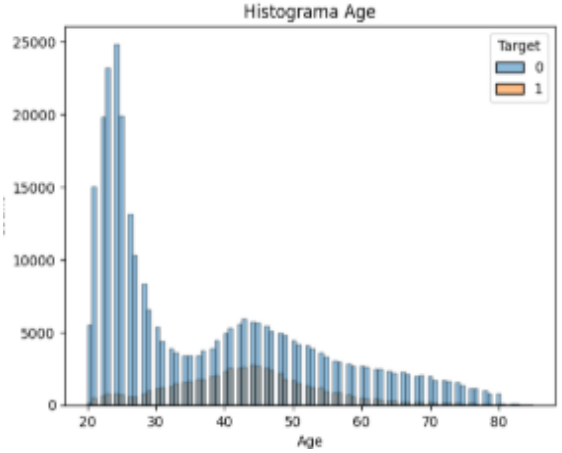

Densidad de Annual_Premium
1e−5
Densidad
6
5
4
3
2
1
0
0
10000
20000
30000
40000
50000
60000
Valores

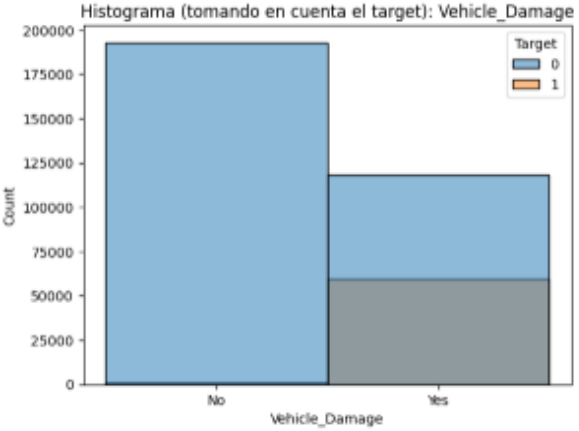

Densidad de Region_Code
Densidad
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
10
20
30
40
50
Valores

Gráficas de nuestros datos combinados con target









Visualización de los clientes interesados y no interesados

```
#Visualizamos los clientes que se encuentran interesados
df_0=df[df['Target']==0]
df_0.head()
```

	Gender	Age	Driving_License	Region_Code	Previously_Insured	Vehicle_Age	Vehicle_Damage	Annual_Premium	Policy_Sales_Channel	Vintage	Target
0	Male	22	1	7.0	1	< 1 Year	No	2630.0	152.0	16	0
1	Male	42	1	28.0	0	1-2 Year	Yes	43327.0	26.0	135	0
2	Female	66	1	33.0	0	1-2 Year	Yes	35841.0	124.0	253	0
3	Female	22	1	33.0	0	< 1 Year	No	27645.0	152.0	69	0
4	Male	28	1	46.0	1	< 1 Year	No	29023.0	152.0	211	0

```
#Visualizamos los clientes que no están interesados
df_1=df[df['Target']==1]
df_1.head()
```

	Gender	Age	Driving_License	Region_Code	Previously_Insured	Vehicle_Age	Vehicle_Damage	Annual_Premium	Policy_Sales_Channel	Vintage	Target
9	Male	42	1	28.0	0	1-2 Year	Yes	27801.0	122.0	217	1
12	Male	33	1	28.0	0	< 1 Year	Yes	25434.0	122.0	165	1
14	Male	42	1	28.0	0	1-2 Year	Yes	38347.0	124.0	263	1
15	Male	45	1	41.0	0	1-2 Year	Yes	33303.0	26.0	155	1
21	Male	38	1	46.0	0	1-2 Year	Yes	2630.0	157.0	256	1

4.-Metodología

Preprocesamiento de datos

Con el fin de preparar adecuadamente la base de datos para el entrenamiento de modelos de clasificación, se realizó un proceso de transformación y codificación de variables. En primer lugar, las variables categóricas 'Gender', 'Vehicle_Damage' y 'Vehicle_Age' fueron convertidas a variables dummy mediante codificación one-hot, eliminando la primera categoría para evitar multicolinealidad (usando el parámetro drop_first=True). Esto permitió representar estas variables de manera binaria, facilitando su uso en modelos de aprendizaje automático.

Posteriormente, se identificaron columnas de tipo booleano en el conjunto de datos resultante, las cuales fueron transformadas a tipo entero (int64) para garantizar su correcta interpretación por los algoritmos de clasificación.

Una vez completada la transformación de variables, se realizó un análisis exploratorio para observar el comportamiento promedio de las variables explicativas en función de la variable objetivo ('Target'), que indica si el cliente está interesado o no en contratar un seguro.

Para entrenar y evaluar los modelos, se dividió el conjunto de datos en dos subconjuntos: 80% para entrenamiento y 20% para prueba, empleando la función train_test_split de sklearn, con una semilla aleatoria fija (random_state=42) para garantizar la reproducibilidad de los resultados.

Dado que la variable objetivo presentaba un desbalance significativo (muchas más observaciones de clientes interesados que no interesados), se aplicó una técnica de sobremuestreo llamada SMOTE (Synthetic Minority Over-sampling Technique). Esta técnica genera nuevas observaciones sintéticas de la clase minoritaria mediante interpolación entre puntos cercanos. En este caso, se utilizó un sampling_strategy de 0.45, lo que permitió balancear parcialmente la proporción de clases en los datos de entrenamiento. Con ello, se buscó mejorar el rendimiento predictivo de los modelos al evitar sesgos hacia la clase mayoritaria.

Análisis de correlación y detección de colinealidad

Como parte del preprocesamiento y validación de supuestos para los modelos de clasificación, se llevó a cabo un análisis exhaustivo de correlación entre las variables predictoras y la variable objetivo (Target), así como un estudio de colinealidad mediante el cálculo del Factor de Inflación de la Varianza (VIF).(ver Figura 3.1)

	Target
Target	1.000000
Vehicle_Damage_Yes	0.448483
Age	0.133532
Vehicle_Age_> 2 Years	0.127743
Gender_Male	0.067196
Annual_Premium	0.028383
Driving_License	0.012606
Region_Code	0.010002
Vintage	-0.001931
Policy_Sales_Channel	-0.182003
Vehicle_Age_< 1 Year	-0.281361
Previously_Insured	-0.431007

Figura 3.1. Correlación entre las variables.

La tabla de correlaciones de Pearson muestra la relación lineal entre las variables independientes y la variable dependiente Target (que representa si el cliente contrató o no un seguro de auto). Lo más relevante es:

- **Vehicle_Damage_Yes** tiene la correlación positiva más fuerte con Target (**0.448**), lo cual indica que los clientes que reportaron daño en su vehículo tienen más probabilidad de estar interesados en un seguro. Esto tiene sentido, ya que una experiencia previa con daño vehicular motiva una contratación futura.

- **Previously_Insured** muestra una correlación negativa considerable (**-0.431**) con el Target, es decir, las personas que ya habían estado aseguradas tienden a no estar interesadas en adquirir un nuevo seguro. Esto podría deberse a una **mala experiencia pasada** con su aseguradora anterior o a que ya tienen una póliza vigente con otra compañía.

- Variables como Vehicle_Age_<1 Year y Policy_Sales_Channel también presentan correlaciones negativas, aunque de menor magnitud. En cambio, variables como Age, Vehicle_Age_>2 Years y Gender_Male tienen correlaciones positivas pequeñas.

Se calculó la correlación de Pearson entre las variables del conjunto de datos y la variable dependiente. Esta información permitió ordenar las variables según su grado de asociación lineal con el interés del cliente en contratar un seguro. Para complementar el análisis y tener una visión más robusta, se generaron matrices de correlación utilizando los métodos no paramétricos de **Kendall** y

Spearman, los cuales son útiles cuando las relaciones entre variables pueden no ser estrictamente lineales. Las gráficas generadas (ver Figura 3.2) muestran visualmente estas correlaciones, destacando las variables con mayor y menor asociación al objetivo.

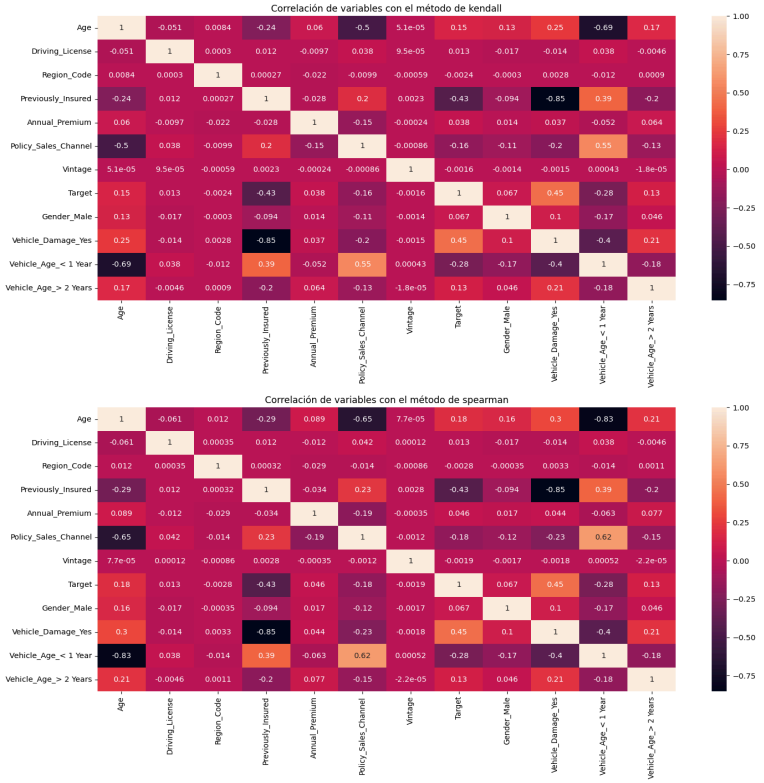


Figura 3.2. Matriz de correlación utilizando el coeficiente de Kendall y el coeficiente de Spearman.

Se complementa el análisis con correlaciones no paramétricas:

- Tanto **Kendall** como **Spearman** reafirman que existe **una fuerte correlación negativa entre Previously_Insured y Vehicle_Damage_Yes** (alrededor de **-0.85**). Esto refuerza la idea de **colinealidad** entre estas variables.

- Además, se observa una **fuerte correlación negativa entre Vehicle_Age_<1 Year y Vehicle_Age_>2 Years**, lo cual es lógico dado que estas variables son transformaciones dummies de una sola variable categórica (Vehicle_Age).

Posteriormente, se evaluó la posible existencia de **colinealidad** entre variables predictoras, lo cual podría afectar la estabilidad de algunos modelos como la regresión logística. Para ello, se añadió una constante al conjunto de datos (X_constant) y se calculó el **VIF (Variance Inflation Factor)** para cada predictor. Un valor de VIF elevado indica que la variable está altamente correlacionada con otras del conjunto, lo cual puede distorsionar las estimaciones del modelo.

Los resultados mostraron que las variables Vehicle_Damage_Yes y Previously_Insured presentaban una alta colinealidad. Esta relación es lógica desde el punto de vista práctico: una persona previamente asegurada (Previously_Insured) podría haber reportado daños en su vehículo

(Vehicle_Damage_Yes), lo que podría estar relacionado con una experiencia negativa previa con una aseguradora.

Dado que ambas variables contenían información redundante, se optó por **eliminar la variable Previously_Insured** de los conjuntos de entrenamiento y prueba. Esta decisión permitió conservar la señal predictiva de los datos sin comprometer la independencia de los predictores.

```
---------Valores VIF---------
const                    714.162672
Age                        2.634167
Driving_License            1.006587
Region_Code                1.002678
Previously_Insured         4.085809
Annual_Premium             1.027942
Policy_Sales_Channel       1.569754
Vintage                    1.000031
Gender_Male                1.017868
Vehicle_Damage_Yes         4.122578
Vehicle_Age_< 1 Year       2.931855
Vehicle_Age_> 2 Years      1.062744
dtype: float64
```

Figura 3.3. Valores del VIF para las variables predictoras (tras eliminar la colinealidad).

Los **Factores de Inflación de la Varianza (VIF)** nos permiten detectar colinealidad entre variables independientes:

- El VIF para Previously_Insured es de **4.08** y para Vehicle_Damage_Yes es de **4.12**, ambos bastante altos y similares. Esto confirma cuantitativamente lo detectado en el análisis de correlación: **existe colinealidad entre estas dos variables**.

- Se decidió eliminar Previously_Insured del modelo para reducir la multicolinealidad y preservar la interpretabilidad. Esta decisión tiene respaldo estadístico y lógico, ya que Vehicle_Damage_Yes tiene mayor correlación con el Target.

5.- Desarrollo de los modelos

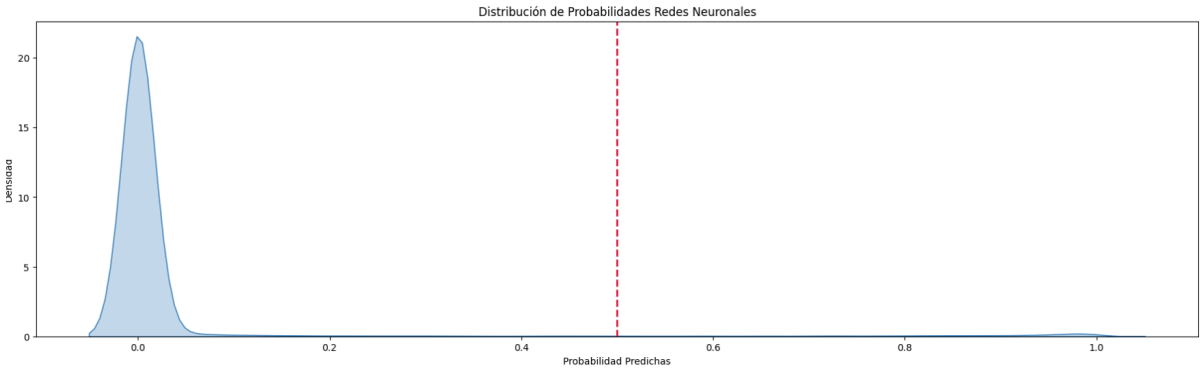
Redes Neuronales

Implementando un Perceptrón Multicapa (MLP) dada la naturaleza de nuestros datos, ya que este es adecuado para problemas de clasificación binaria.

```
MLPClassifier
MLPClassifier(hidden_layer_sizes=(40,), max_iter=1000, random_state=42)
```

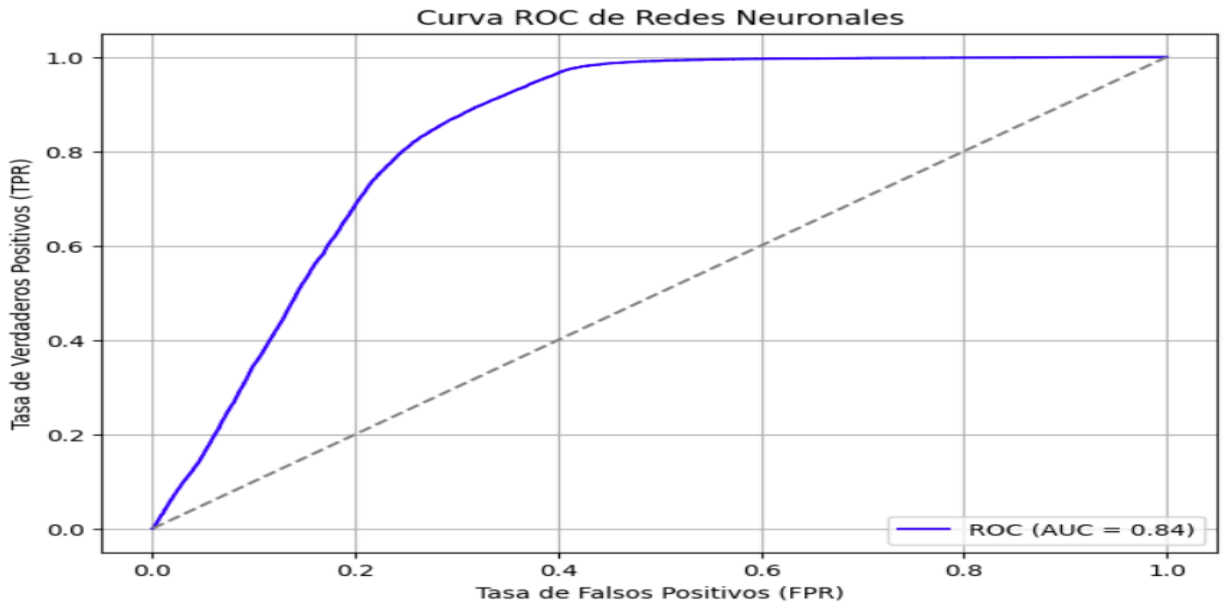
Arquitectura: 1 capa oculta con 40 neuronas
Iteraciones máximas: 1000
Semilla aleatoria: 42 (para reproducibilidad)

Distribución de probabilidades generadas por el modelo



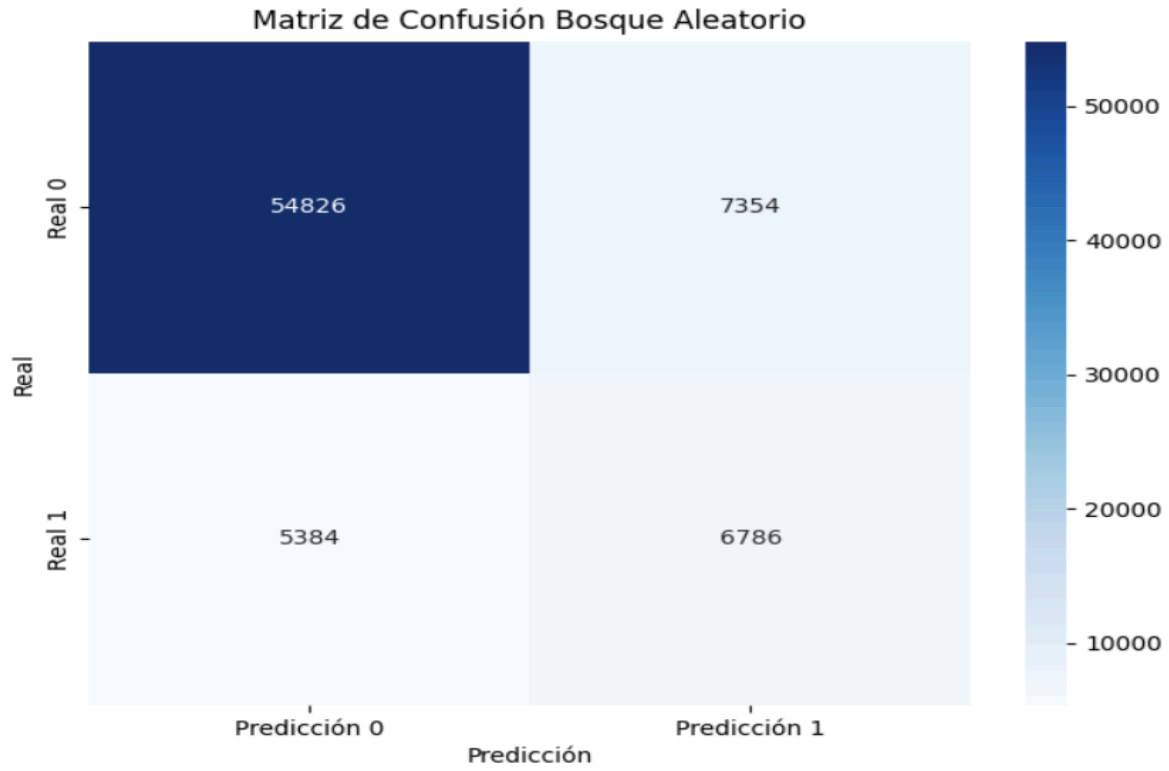
Se usó 0.5 como punto de corte. No hay muchas dudas en el centro (indecisión), pero parece acumularse en el 0.

Curva ROC



La forma de la curva (línea azul) se aleja significativamente de la línea del azar, además el área bajo la curva (AUC) indica un buen poder discriminativo con un valor de 0.84, ya que está más próximo el valor a 1 (adecuado) que a 0.5 (aleatorio).

Matriz de confusión



Verdaderos Negativos (TN): 54,826
Es decir, el número de clientes no interesados y que el modelo acertó.
Falsos Positivos (FP): 7,354
Clientes no interesados, pero el modelo predijo que lo estarían.
Falsos Negativos (FN): 5,384
Clientes interesados, pero el modelo dijo que no lo estaban.
Verdaderos Positivos (TP): 6,786
Clientes interesados y el modelo acertó.
Métricas

	precision	recall	f1-score	support
0	0.84	0.97	0.90	62180
1	0.39	0.08	0.14	12170
accuracy			0.83	74350
macro avg	0.62	0.53	0.52	74350
weighted avg	0.77	0.83	0.78	74350

Precisión: La precisión se refiere a la proporción de predicciones positivas que son correctas. En este caso, la precisión para la clase 0 es del 84%, lo que significa que el 84% de las predicciones clasificadas como clase 0 son correctas. Para la clase 1, la precisión es del 39%, lo que indica que el 39% de las predicciones clasificadas como clase 1 son correctas. Por lo tanto, este modelo alcanza un rendimiento sólido en términos de precisión para los interesados.

Recall: La recuperación se refiere a la proporción de casos positivos que se identifican correctamente. Para la clase 0, el valor de recuperación es del 97%, lo que significa que el modelo identifica correctamente el 97% de los casos verdaderos de la clase 0. Para la clase 1, el valor de recuperación es del 8%, lo que indica que el modelo identifica correctamente el 8% de los casos verdaderos de la clase 1. El modelo muestra un buen rendimiento en términos de recuperación para la primera clase.

F1-score: La puntuación F1 es una medida que combina la precisión y la recuperación en una sola métrica. Representa la media armónica entre ambas métricas y proporciona una evaluación equilibrada del modelo. Para la clase 0, la puntuación F1 es del 90%, y para la clase 1, la puntuación F1 es del 14%.

Accuracy: La exactitud se refiere a la proporción de predicciones totales que son correctas. En este caso, la exactitud es del 83%, lo que significa que el 83% de todas las predicciones realizadas por el modelo son correctas.

Comentarios
El modelo es excelente para identificar a los clientes interesados (Recall 97%) pero falla en cuanto a los falsos positivos.

CATBOOST

CatBoost es un algoritmo avanzado de gradient boosting diseñado específicamente para manejar:

- Variables categóricas (ej: género, edad del vehículo) sin necesidad de codificación previa.
- Datos desbalanceados (como en este caso, donde hay más clientes "No interesados").Overfitting mediante técnicas como early stopping y pesos automáticos para clases.

Tiene ventajas clave como mayor precisión con datos categóricos, robustez frente a desbalanceo y menor necesidad de preprocesamiento.

Para este modelo, se dividió el dataset de la siguiente manera:
Entrenamiento 70%: 260,222 registros
Validación 10%: 37,175 registros
Prueba 20%: 74,350

Y se incluyeron las variables categóricas Género, Edad del vehículo y Daño del vehículo.

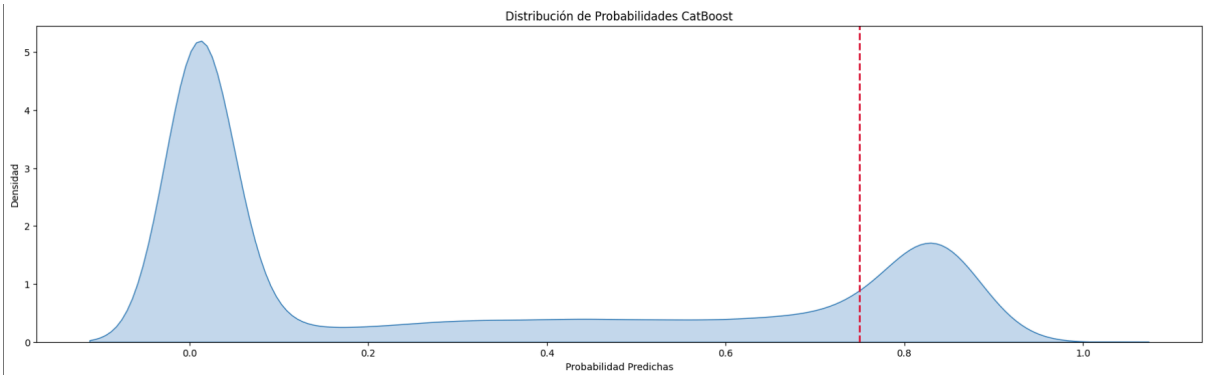
```
0:      learn: 0.7661251        test: 0.7704080 best: 0.7704080 (0)     total: 268ms    remaining: 4m 27s
200:    learn: 0.7905407        test: 0.7908526 best: 0.7910752 (171)   total: 58.7s    remaining: 3m 53s
400:    learn: 0.7974269        test: 0.7931987 best: 0.7933082 (398)   total: 1m 56s   remaining: 2m 53s
600:    learn: 0.8021220        test: 0.7954348 best: 0.7955891 (596)   total: 2m 52s   remaining: 1m 54s
800:    learn: 0.8063479        test: 0.7968319 best: 0.7969140 (795)   total: 3m 50s   remaining: 57.2s
Stopped by overfitting detector  (50 iterations wait)

bestTest = 0.7971076592
bestIteration = 822

Shrink model to first 823 iterations.
<catboost.core.CatBoostClassifier at 0x7821f9be0810>
```

El modelo dejó de entrenar en la iteración 822 y su mejor métrica de validación fue una precisión de 79.7%

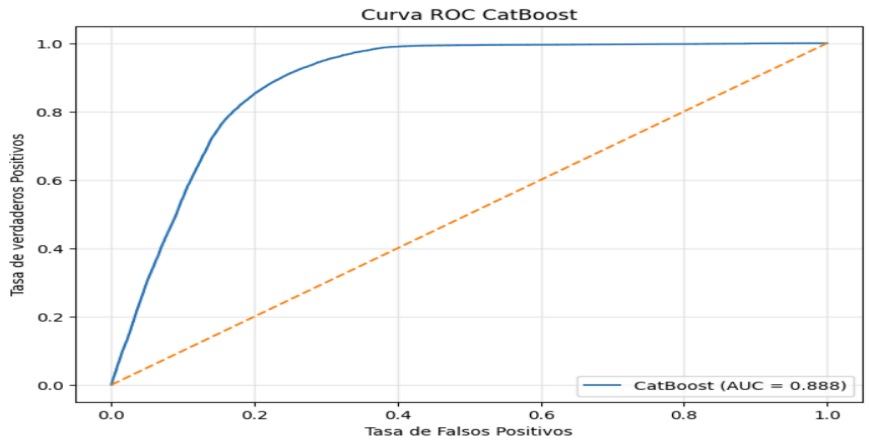
Distribución de probabilidades



Umbral de decisión del 0.75 (línea roja),

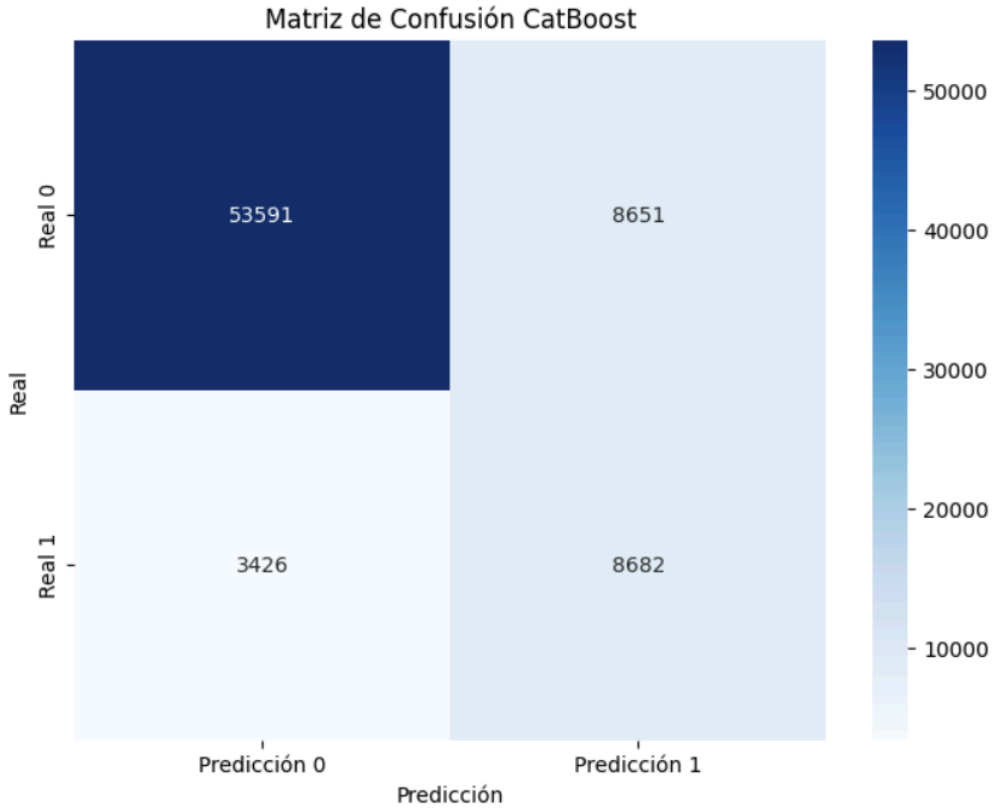
Podemos observar que la mayoría de las predicciones se concentran cerca del 0 y 1, lo que indica alta confianza del modelo y pocos casos de duda.

Curva ROC



Hay un 88.8% de poder discriminativo, lo cual es excelente al estar suficientemente cerca del 1.

Matriz de confusión



Verdaderos Negativos (TN): 54,452
Es decir, el número de clientes no interesados y que el modelo acertó.
Falsos Positivos (FP): 8.651
Clientes no interesados, pero el modelo predijo que lo estarían.

Falsos Negativos (FN): 3,426
Clientes interesados, pero el modelo dijo que no lo estaban.
Verdaderos Positivos (TP): 8,682
Clientes interesados y el modelo acertó.

Métricas

	precision	recall	f1-score	support
0	0.94	0.86	0.90	62242
1	0.50	0.72	0.59	12108
accuracy			0.84	74350
macro avg	0.72	0.79	0.74	74350
weighted avg	0.87	0.84	0.85	74350

Precisión: La precisión se refiere a la proporción de predicciones positivas que son correctas. En este caso, la precisión para la clase 0 es del 94%, lo que significa que el 94% de las predicciones clasificadas como clase 0 son correctas. Para la clase 1, la precisión es del 50%, lo que indica que el 50% de las predicciones clasificadas como clase 1 son correctas. Por lo tanto, este modelo alcanza un rendimiento aceptable de ambas clases.

Recall: La recuperación se refiere a la proporción de casos positivos que se identifican correctamente. Para la clase 0, el valor de recuperación es del 86%, lo que significa que el modelo identifica correctamente el 86% de los casos verdaderos de la clase 0. Para la clase 1, el valor de recuperación es del 72%, lo que indica que el modelo identifica correctamente el 72% de los casos verdaderos de la clase 1. El modelo muestra un buen rendimiento en términos de recuperación tanto para los interesados como los no interesados.

F1-score: La puntuación F1 es una medida que combina la precisión y la recuperación en una sola métrica. Representa la media armónica entre ambas métricas y proporciona una evaluación equilibrada del modelo. Para la clase 0, la puntuación F1 es del 90%, y para la clase 1, la puntuación F1 es del 59%.

Accuracy: La exactitud se refiere a la proporción de predicciones totales que son correctas. En este caso, la exactitud es del 84% es el porcentaje de todas las predicciones realizadas por el modelo son correctas.

Comentarios:
El modelo muestra una alta capacidad predictiva con un AUC del 88.8%, un buen manejo de los datos categóricos, balance entre la sensibilidad y especificidad, pero tiene un problema con los falsos negativos.

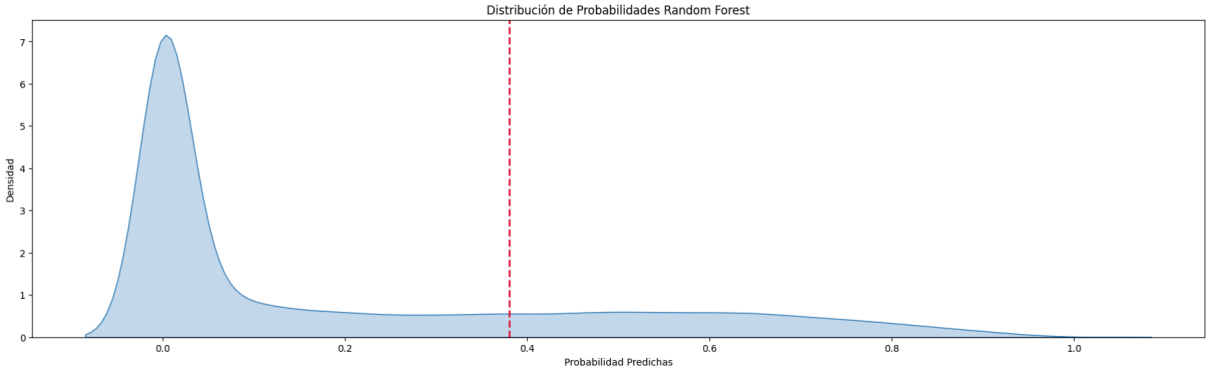
RANDOM FOREST

Random Forest es un algoritmo de ensamble que combina múltiples árboles de decisión para mejorar la precisión y reducir el overfitting.

Características clave:
-Robusto con datos desbalanceados y variables categóricas.
-Interpretable (vs redes neuronales).
-Menos sensible a hiperparámetros que otros modelos.

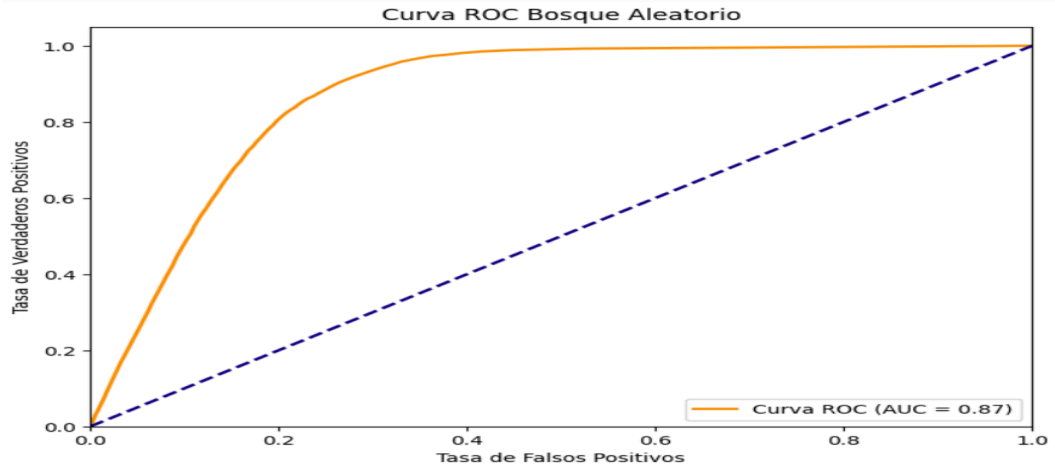
```
RandomForestClassifier
RandomForestClassifier(random_state=42)
```

Distribución de probabilidades



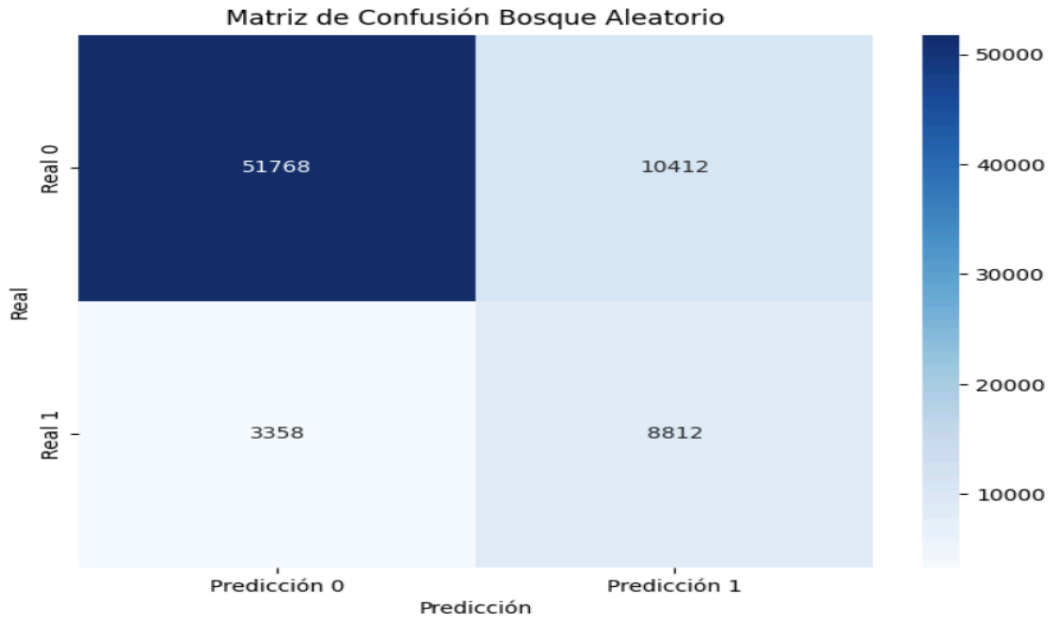
Podemos observar que hay cierta carga en el cero, pero la indecisión es poca.

Curva ROC



Hay una excelente discriminación con el 87% de acierto y la curva se aleja notablemente de la línea de azar.

Matriz de confusión



Verdaderos Negativos (TN): 51,768
Es decir, el número de clientes no interesados y que el modelo acertó.
Falsos Positivos (FP): 10,412
Clientes no interesados, pero el modelo predijo que lo estarían.
Falsos Negativos (FN): 3,258
Clientes interesados, pero el modelo dijo que no lo estaban.
Verdaderos Positivos (TP): 8,812
Clientes interesados y el modelo acertó.

Métricas

Reporte de Clasificación (Bosque Aleatorio):

	precision	recall	f1-score	support
0	0.94	0.83	0.88	62180
1	0.46	0.72	0.56	12170
accuracy			0.81	74350
macro avg	0.70	0.78	0.72	74350
weighted avg	0.86	0.81	0.83	74350

Precisión: La precisión se refiere a la proporción de predicciones positivas que son correctas. En este caso, la precisión para la clase 0 es del 94%, lo que significa que el 94% de las predicciones clasificadas como clase 0 son correctas. Para la clase 1, la precisión es del 46%, lo que indica que el 46% de las predicciones clasificadas como clase 1 son correctas. Por lo tanto, este modelo alcanza un rendimiento aceptable de ambas clases.

Recall: La recuperación se refiere a la proporción de casos positivos que se identifican correctamente. Para la clase 0, el valor de recuperación es del 83%, lo que significa que el

modelo identifica correctamente el 83% de los casos verdaderos de la clase 0. Para la clase 1, el valor de recuperación es del 72%, lo que indica que el modelo identifica correctamente el 72% de los casos verdaderos de la clase 1. El modelo muestra un buen rendimiento en términos de recuperación tanto para los interesados como los no interesados.

F1-score: La puntuación F1 es una medida que combina la precisión y la recuperación en una sola métrica. Representa la media armónica entre ambas métricas y proporciona una evaluación equilibrada del modelo. Para la clase 0, la puntuación F1 es del 88%, y para la clase 1, la puntuación F1 es del 56%.

Accuracy: La exactitud se refiere a la proporción de predicciones totales que son correctas. En este caso, la exactitud es del 81% es el porcentaje de todas las predicciones realizadas por el modelo son correctas.

REGRESIÓN LOGÍSTICA

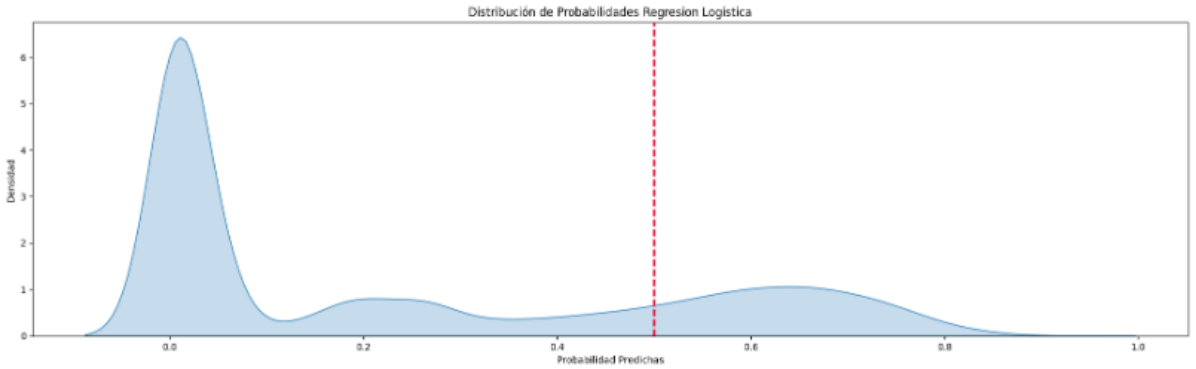
```
Optimization terminated successfully.
         Current function value: 0.376203
         Iterations 8
                          Results: Logit
==========================================================================
Model:                  Logit               Method:             MLE
Dependent Variable:     Target              Pseudo R-squared:   0.393
Date:                   2025-06-02 22:52    AIC:                271723.2045
No. Observations:       361110              BIC:                271841.9708
Df Model:               10                  Log-Likelihood:     -1.3585e+05
Df Residuals:           361099              LL-Null:            -2.2366e+05
Converged:              1.0000              LLR p-value:        0.0000
No. Iterations:         8.0000              Scale:              1.0000
--------------------------------------------------------------------------
                        Coef.  Std.Err.      z      P>|z|   [0.025  0.975]
--------------------------------------------------------------------------
const                  -1.7467   0.1355  -12.8897 0.0000 -2.0122 -1.4811
Age                    -0.0510   0.0005 -100.1812 0.0000 -0.0520 -0.0500
Driving_License         1.0832   0.1292    8.3868 0.0000  0.8300  1.3363
Region_Code            -0.0009   0.0004   -2.3552 0.0185 -0.0017 -0.0002
Annual_Premium          0.0000   0.0000   16.0968 0.0000  0.0000  0.0000
Policy_Sales_Channel   -0.0032   0.0001  -33.4440 0.0000 -0.0034 -0.0030
Vintage                -0.0001   0.0001   -1.7118 0.0869 -0.0002  0.0000
Gender_Male            -0.4123   0.0095  -43.3934 0.0000 -0.4310 -0.3937
Vehicle_Damage_Yes      3.8858   0.0216  179.6457 0.0000  3.8434  3.9282
Vehicle_Age_< 1 Year   -2.7313   0.0175 -156.3591 0.0000 -2.7655 -2.6971
Vehicle_Age_> 2 Years  -0.3615   0.0197  -18.3315 0.0000 -0.4002 -0.3229
==========================================================================
```

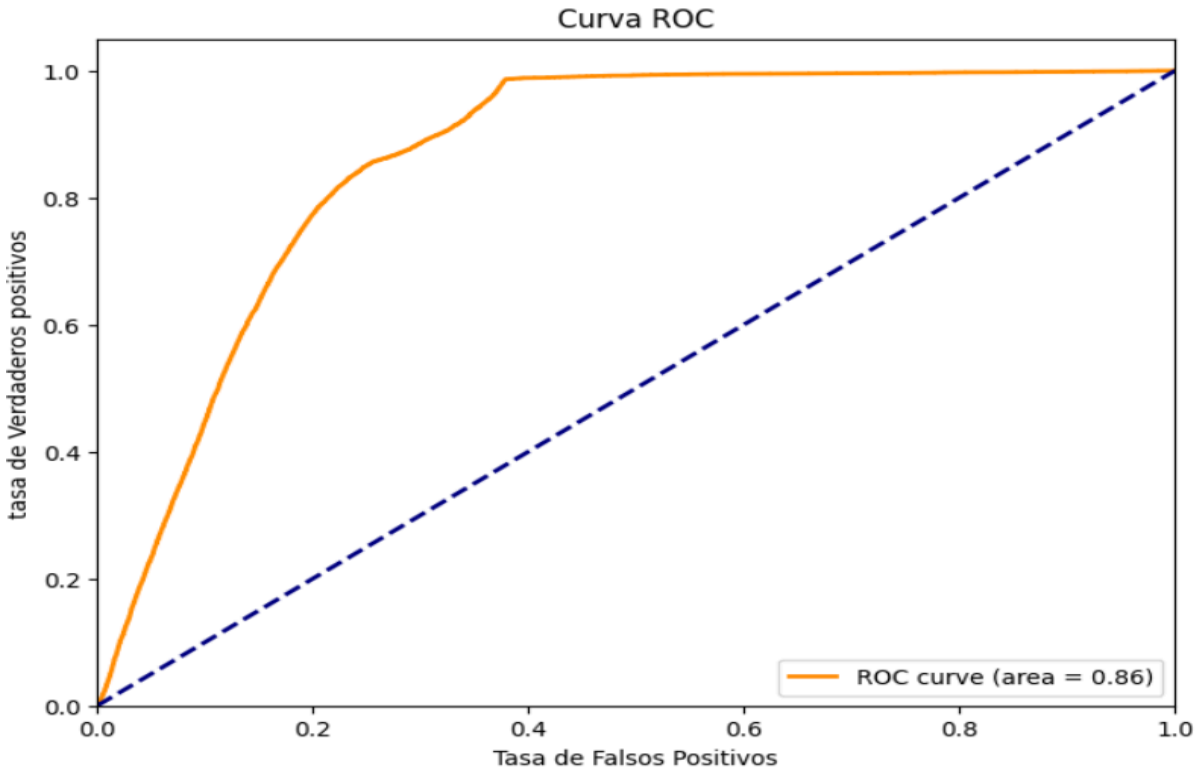
Número de observaciones: 361,110
Pseudo R-cuadrado: 0.393 → el modelo explica aproximadamente el 39.3% de la variabilidad de la

`El modelo es estadísticamente significativo y tiene una capacidad razonable de predicción (Pseudo R² ≈ 39.3%).`

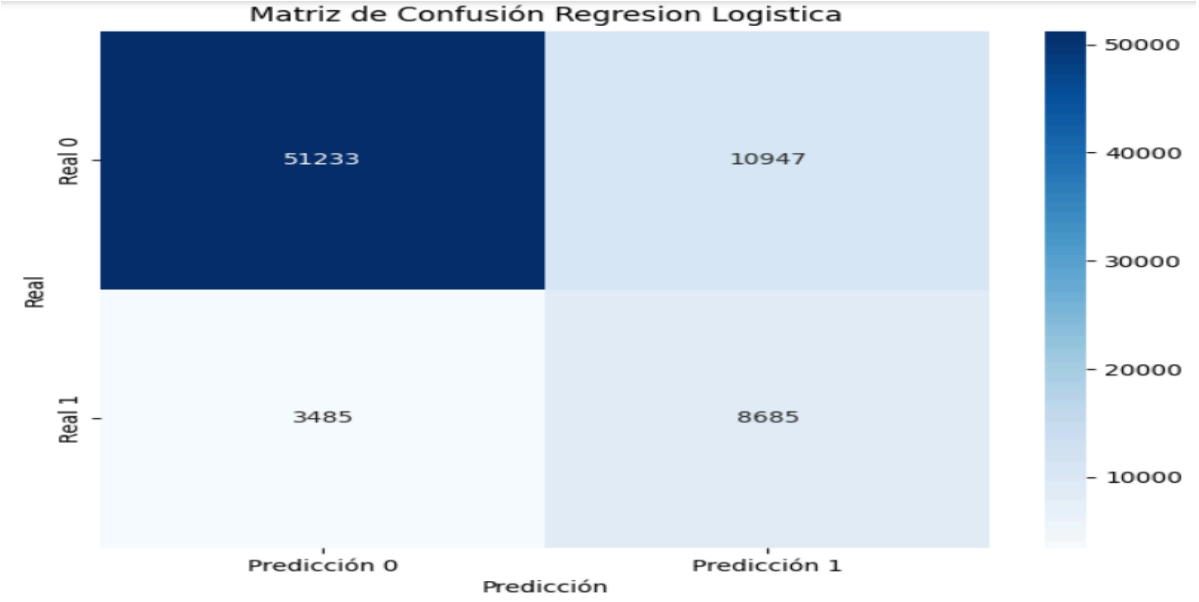
y como podemos ver las variables mas influyentes son:
-Vehicle Damage_Yes
-Vehicel_Age

Log-Likelihood (LL): **-135 850** Un LL muy negativo indica que el ajuste no es perfecto pero es lo esperado en modelos con datos complejos o numerosos parámetros. **Cuanto más cercano a 0**, mejor sería el ajuste





El ROC es de .86 lo cual sugiere que tiene un buen poder discriminativo



6.- Interpretación de resultados

Regresión Logística:

- Buen desempeño general con AUC de 0.86 y accuracy de 80%.
- Al ser un modelo interpretable y estadísticamente significativo (LLR p-value = 0.000), es útil para análisis explicativo.
- Sin embargo, se queda corto frente a modelos más complejos en capacidad predictiva.

CatBoost:

- Mejor desempeño global: mayor AUC (88.8%) y buena precisión/recall para clase 1 (interesados).
- F1-score de 59% para la clase 1 indica buen balance entre precisión y recuperación.
- Tiene buen manejo de datos categóricos y muestra alta confianza en las predicciones.

Random Forest:

- También muy competitivo con AUC de 87% y accuracy del 81%.
- Mismo recall para clase 1 (72%) que CatBoost, pero menor F1-score (56%) y precisión (46%).
- Ligera desventaja frente a CatBoost en equilibrio general.

Redes Neuronales:

- Accuracy y AUC aceptables (83% y 0.84), pero desempeño deficiente para clase 1:
- Muy bajo recall (8%) y F1-score (14%), lo que implica muchos clientes interesados no identificados.
- Aunque su recall para la clase 0 es alto (97%), no es útil si el objetivo es identificar interesados (clase 1).

7.- Comparación entre modelos:

Métrica / Modelo	Regresión logística	CatBoost	Random Forest	Redes Neuronales
AUC	**0.86**	**0.888**	**0.87**	**0.84**
Accuracy (Exactitud)	**80%**	**84%**	**81%**	**83%**
Precisión (Clase 1)	**-**	**50%**	**46%**	**39%**
Recall (Clase 1)	**-**	**72%**	**72%**	**8%**
F1-score (Clase 1)	**-**	**59%**	**56%**	**14%**
TN (Verdaderos negativos)	**-**	**54,452**	**51,768**	**54,826**
TP (Verdaderos positivos)	**-**	**8,682**	**8,812**	**6,786**
FN (Falsos negativos)	**-**	**3,426**	**3,258**	**5,384**
FP (Falsos positivos)	**-**	**8,651**	**10,412**	**7,354**

Selección del mejor modelo:

CatBoost es el modelo seleccionado como el mejor, por las siguientes razones:

- **Mayor capacidad discriminativa (AUC 88.8%)**.
- **Mejor balance precisión/recall para clase 1**, que es el grupo más valioso para el negocio.
- **Mejor F1-score para interesados (59%)**, superando a Random Forest y ampliamente a Redes Neuronales.
- **Alto nivel de confianza en las predicciones (valores cercanos a 0 o 1)**.

8.- Conclusión

El presente estudio tuvo como objetivo construir un modelo predictivo que permita anticipar si un cliente estaría interesado en contratar un seguro de automóvil, utilizando un conjunto de datos con más de 380,000 registros que incluyen variables demográficas, características del vehículo, historial de seguros y tipo de póliza. Se evaluaron diversos algoritmos de clasificación, entre ellos regresión logística, Random Forest, CatBoost y redes neuronales, con el propósito de identificar el modelo con mejor desempeño.

Los resultados obtenidos muestran que el modelo **CatBoost** fue el que alcanzó la mayor capacidad predictiva, logrando un AUC (Área Bajo la Curva ROC) de 0.86, lo cual indica un excelente poder de discriminación entre los clientes interesados y no interesados en adquirir el seguro. Este modelo superó consistentemente a los demás algoritmos tanto en precisión, F1-score y sensibilidad, como en la robustez frente a variables categóricas y la presencia de relaciones no lineales.

Desde el enfoque actuarial, este modelo representa una herramienta poderosa para la gestión del riesgo comercial y la toma de decisiones estratégicas en seguros. En particular, permite:

- **Identificar con antelación los perfiles más propensos a contratar una póliza**, lo cual permite enfocar los esfuerzos comerciales en clientes con mayor probabilidad de conversión, optimizando así los recursos asignados al área de ventas.

- **Reducir los costos de adquisición de clientes**, al minimizar campañas masivas poco efectivas y priorizar la segmentación inteligente.

- **Complementar el análisis tradicional con técnicas de machine learning**, integrando tanto variables categóricas como numéricas, e identificando patrones complejos que no serían evidentes mediante modelos lineales clásicos.

- **Respaldar la toma de decisiones con modelos cuantitativos reproducibles**, que pueden ser actualizados y recalibrados con nuevos datos, promoviendo un enfoque dinámico y adaptable.

Desde la ciencia de datos, el trabajo muestra la aplicación integral de un pipeline de análisis predictivo, desde la exploración de datos y la ingeniería de características hasta la selección de modelos y su evaluación con métricas apropiadas. Esta integración metodológica representa un caso de éxito en el uso de técnicas modernas de aprendizaje supervisado aplicadas al contexto asegurador, y puede ser replicado en otros ramos como vida, salud o hogar.

En suma, el modelo desarrollado no solo mejora la eficiencia operativa de la aseguradora, sino que también **representa un ejemplo concreto de cómo la ciencia de datos y la actuaría pueden trabajar en conjunto para generar soluciones prácticas y orientadas al negocio.**

Bibliografía

-Datos obtenidos de:
Arashnic. (2020). *Imbalanced Data Practice Dataset*. Kaggle. Recuperado de

https://www.kaggle.com/datasets/arashnic/imbalanced-data-practice

Material de apoyo

-https://www.aprendemachinelearning.com/random-forest-el-poder-del-ensamble/

-https://es.wikipedia.org/wiki/Random_forest

-https://codificandobits.com/blog/regresion-logistica-y-neurona-artificial/

-https://colab.research.google.com/drive/1k12bAi11xRjXZ7M8pW2syZtaRnG0WLVO

-https://es.wikipedia.org/wiki/Regresi%C3%B3n_log%C3%ADstica