ENTREGA CONTRATACIÓN SEGURO MÓVIL

UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE CIENCIAS MATEMÁTICAS
MÁSTER EN INGENIERÍA MATEMÁTICA

ASIGNATURA: Estadística Aplicada y Minería de Datos
PROFESOR: Daniel Velez Serrano

ALUMNAS: Siria Catherine Íñiguez Brito,
Ana Marta Oliveira dos Santos

CURSO: 2025-2026

Madrid, Enero de 2026

Índice

1. INTRODUCCIÓN

Este trabajo aborda un problema de clasificación binaria en el contexto del sector de telecomunicaciones. Una compañía *TELCO* desea optimizar la comercialización de un seguro para dispositivos móviles dirigido a su base de clientes. Para ello, se dispone de un conjunto de datos históricos que recogen el comportamiento de los clientes frente a la oferta del servicio (`ContrataSeguroMovil`: 1 si contrata, 0 si no), junto con diversas variables explicativas.

El objetivo principal es desarrollar un modelo capaz de identificar, con la mayor precisión posible, al **10 % de clientes con mayor probabilidad de contratar el seguro**. Para lograrlo, se ha seguido una metodología de *Data Mining*, ejecutada principalmente en **SAS Enterprise Miner**, que abarca desde el diagnóstico y preparación de los datos hasta la construcción, evaluación y comparación de múltiples modelos.

El desarrollo del estudio se estructura en las siguientes etapas: primero, se realiza un análisis exploratorio y un diagnóstico de calidad de los datos. Segundo, se divide el conjunto en muestras estratificadas para entrenamiento, validación y test. Tercero, se construyen y ajustan diferentes tipos de modelos: Árboles de Decisión (con y sin *undersampling*), Regresión Logística (incluyendo un modelo con transformaciones WOE) y dos modelos ensamblados *Gradient Boosting* y *Random Forest*. Finalmente, se comparan los modelos en función de su capacidad de *lift* en el primer decil, determinando cuál es el más adecuado para la selección del público objetivo óptimo.

Esta memoria documenta cada paso del proceso, las configuraciones específicas en SAS Miner, la interpretación de los resultados y la justificación de la elección del modelo final.

2. DATOS

2.1. Descripción del conjunto de datos

El conjunto de datos disponible recoge, a nivel de `ID_CLIENTE`, la respuesta de los clientes a una campaña comercial, junto con diversas variables explicativas registradas en un momento anterior al lanzamiento de dicha campaña.

	ID_CLIENTE	EDAD	TARIFA_ACTUAL	TARIFA_PREVIA	CONSUMO	ARPU_MEDIO	REGION	ContrataSeguroMovil
1	2	32	3	3	1.13	34.64	2	0
2	3	36	1	3	1.15	11.55	4	0
3	4	19	2	3	1.04	21.08	2	0
4	6	54	2	2	1.08	22.59	2	0
5	7	34	3	3	1.08	32.64	3	0
6	8	32	2	2	1.14	23.31	2	0
7	9	27	1	5	1.01	10.83	3	0
8	11	54	2	2	1.12	22.53	5	0
9	12	39	4	4	1.03	41.32	3	0
10	13	42	1	1	1.18	12.27	5	0
11	14	51	3	3	1.02	30.78	2	0
12	15	63	2	3	1.18	24.43	4	0
13	16	54	1	1	1.02	10.93	2	0
14	17	53	4	4	1.12	45.63	4	0
15	18	35	2	3	1.15	23.15	5	0

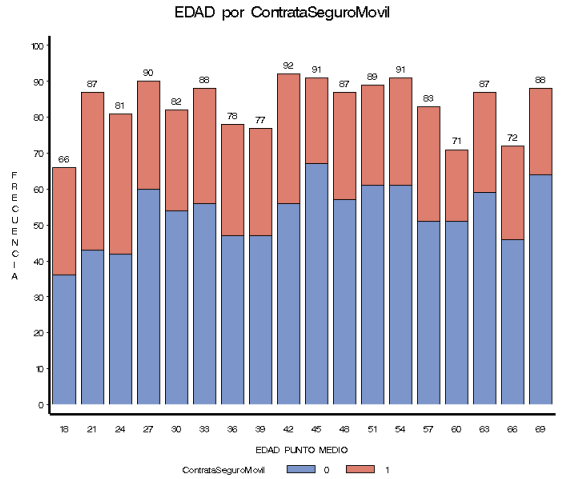
Figura 1: Vista de la tabla del conjunto de datos

El conjunto de datos está compuesto por las siguientes variables:

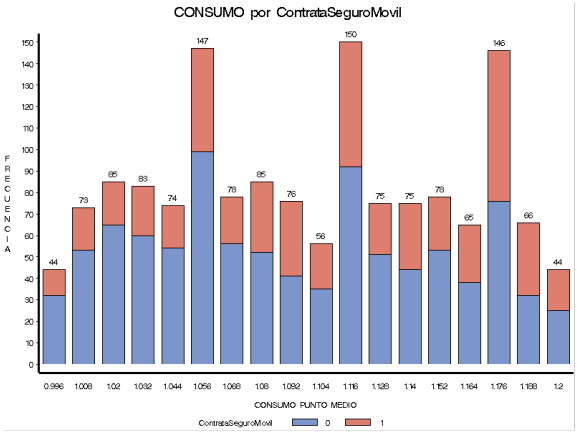
- **`EDAD`**: edad del cliente.
- **`TARIFA_ACTUAL`**: tarifa contratada por el cliente en el momento inmediatamente anterior al contacto comercial.
- **`TARIFA_PREVIA`**: tarifa contratada por el cliente el año anterior a la campaña.
- **`CONSUMO`**: nivel de consumo de servicios de telefonía.
- **`ARPU_MEDIO`** (*Average Revenue Per User*): ingreso medio generado por el cliente en el momento de la campaña.
- **`REGION`**: región geográfica de residencia del cliente.
- **`ContratacionSeguroMovil`**: variable objetivo la cual recoge el resultado de la campaña.

La variable **ContratacionSeguroMovil** definida como la variable *target* del estudio es de naturaleza binaria, tomando el valor 1 si el cliente contrata el seguro para móviles y 0 en caso contrario. En cuanto a la tipología de las variables explicativas, se estableció la siguiente clasificación:

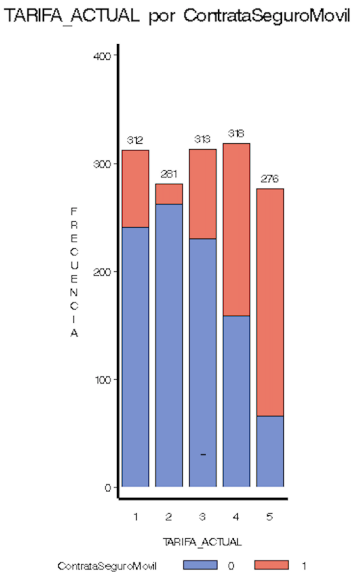
- **Variables intervalares**: `EDAD`, `CONSUMO` y `ARPU_MEDIO`.
- **Variables ordinales**: `TARIFA_ACTUAL` y `TARIFA_PREVIA`.
- **Variable nominal**: `REGION`.



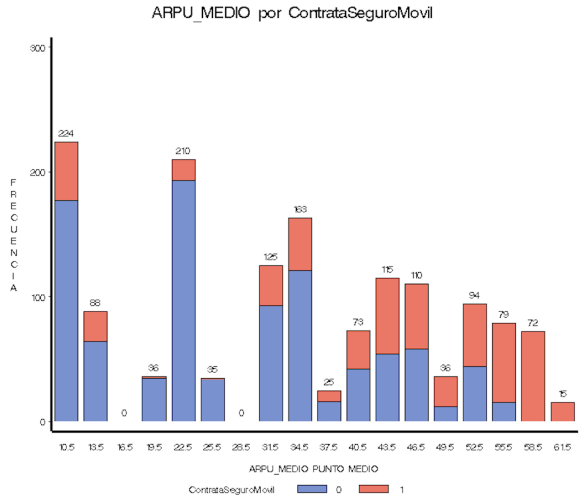
(a) Distribución de edades



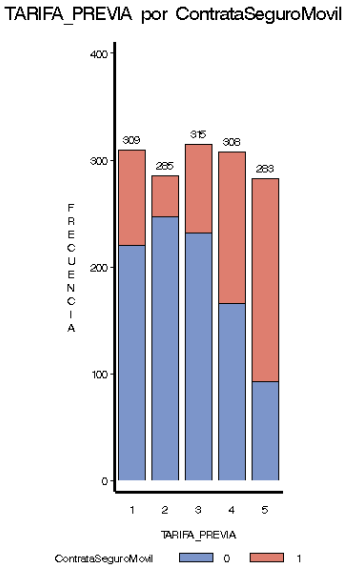
(b) Consumo mensual



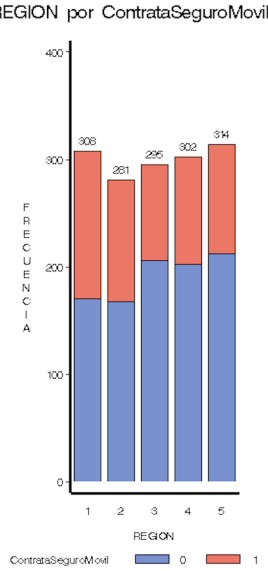
(c) Tarifa actual



(d) ARPU medio



(e) Tarifa previa



(f) Región

Figura 2: Análisis exploratorio de las variables

Se realizó un análisis exploratorio mediante histogramas segmentados por la variable objetivo como se aprecia en la Figura 2, cuya primera impresión señala a las variables de `TARIFA` y `ARPU_MEDIO` como candidatas prometedoras para el análisis posterior.

En la Figura 2c se presenta la distribución de clientes según la variable categórica `TARIFA_ACTUAL`. Se observa que, para las tarifas más bajas, predomina claramente la no contratación del seguro. A medida que aumenta el nivel de la tarifa, la proporción de clientes que contratan el seguro se incrementa, siendo especialmente notable en la tarifa más alta, donde la clase contratante llega a ser mayoritaria. Este patrón se repite en el caso de la `TARIFA_PREVIA` (Figura 2e), lo que sugiere que estas variables podrían ser indicadores significativos a la hora de predecir la contratación del seguro.

Por otra parte, la Figura 2d muestra la distribución del `ARPU_MEDIO` agrupado en intervalos. En los niveles bajos de `ARPU_MEDIO` predomina la no contratación del seguro, mientras que en los intervalos de `ARPU_MEDIO` medio y alto se aprecia un aumento de la proporción de clientes contratantes.

2.2. Diagnóstico de Calidad de Datos

Se realizó un diagnóstico para la detección valores faltantes (*missing values*) y valores atípicos (*outliers*) del conjunto de datos, compuesto por 1.500 observaciones, cuyos resultados se resumen en la Figura 3. Los hallazgos principales revelan un alto nivel de calidad y limpieza en los datos:

Variables de clase

Obs	NAME	LEVEL	CODE	FREQUENCY	TYPE	CRAW	NRAW	FREQPERCENT	NMISSPERCENT
1	ContrataSeguroMovil	1	1	542	N		1	36.1333	36.1333
2	ContrataSeguroMovil	0	0	958	N		0	63.8667	63.8667
3	REGION	1	4	308	N		1	20.5333	20.5333
4	REGION	2	0	281	N		2	18.7333	18.7333
5	REGION	3	2	295	N		3	19.6667	19.6667
6	REGION	4	1	302	N		4	20.1333	20.1333
7	REGION	5	3	314	N		5	20.9333	20.9333
8	TARIFA_ACTUAL	1	1	312	N		1	20.8000	20.8000
9	TARIFA_ACTUAL	2	2	281	N		2	18.7333	18.7333
10	TARIFA_ACTUAL	3	0	313	N		3	20.8667	20.8667
11	TARIFA_ACTUAL	4	3	318	N		4	21.2000	21.2000
12	TARIFA_ACTUAL	5	4	276	N		5	18.4000	18.4000
13	TARIFA_PREVIA	1	4	309	N		1	20.6000	20.6000
14	TARIFA_PREVIA	2	1	285	N		2	19.0000	19.0000
15	TARIFA_PREVIA	3	0	315	N		3	21.0000	21.0000
16	TARIFA_PREVIA	4	3	308	N		4	20.5333	20.5333
17	TARIFA_PREVIA	5	2	283	N		5	18.8667	18.8667

Variables de intervalo

Obs	NAME	NMISS	N	MIN	MAX	MEAN	STD	SKEWNESS	KURTOSIS
1	ARPU_MEDIO	0	1500	10.05	60.99	33.1515	15.4390	0.017102	-1.22431
2	CONSUMO	0	1500	1.00	1.20	1.0982	0.0585	0.043231	-1.19991
3	EDAD	0	1500	18.00	70.00	43.6673	15.2911	0.014438	-1.17576

Figura 3: Diagnóstico de calidad de los datos.

- **Valores Faltantes:** El análisis confirma que el conjunto de datos está completo.
- **Distribución de Variables Categóricas:** La inspección de las frecuencias relativas (columna `FREOPERCENT`) muestra distribuciones balanceadas. Para `REGION`, las proporciones oscilan entre 18.73 % y 20.93 %; para `TARIFA_ACTUAL` y `TARIFA_PREVIA`, entre 18.40 % y 21.20 %. No se identifican categorías con frecuencias anómalamente bajas (inferiores al 1-2 %) que pudieran requerir agrupación.
- **Valores Atípicos en Variables Numéricas:** Los coeficientes de asimetría (`Skeumess`), todos próximos a 0 (entre 0.014 y 0.043), indican distribuciones simétricas. Los rangos reportados—`ARPU_MEDIO` (`MIN`=10.05, `MAX`=60.99), `CONSUMO` (1.00-1.20) y `EDAD` (18-70)—son coherentes y dentro del contexto del negocio. No se detectan valores extremos.

En conclusión, la ausencia total de valores faltantes y la no detección de *outliers* evidentes en esta fase exploratoria permiten avanzar a las etapas de análisis y modelado sin necesidad de aplicar técnicas de imputación o corrección de valores extremos.

2.3. Partición de las muestras

El conjunto de datos original, denominado `contratacionseguromovil`, fue dividido en dos subconjuntos principales siguiendo una proporción de **80 %** para entrenamiento-validación y **20 %** para test. Esta partición dio lugar a las tablas `data_csm_trainval` y `data_csm_test`, respectivamente. La división se realizó utilizando **SAS Base**, aplicando un muestreo estratificado con el objetivo de preservar la proporción original de la variable objetivo en ambas muestras. El código empleado para llevar a cabo esta partición se encuentra detallado en `Tablas_EV-T`. Además, en **SAS Miner** siempre que se ha utilizado un nodo 'Partición de datos' su configuración ha sido 70 % para entrenamiento y 30 % para validación.

2.4. Estrategia de modelización

Para la construcción y comparación de modelos predictivos se siguieron dos estrategias diferenciadas:

1. En los modelos de **Árbol de Decisión** y **Regresión Logística**, se entrenaron versiones con y sin aplicación de **bajo muestreo (undersampling) 50-50**. Este procedimiento tiene como finalidad balancear la distribución entre la clase mayoritaria y la clase minoritaria en el conjunto de entrenamiento.
2. En el caso del modelos ensamblados para el **Gradient Boosting** y **Random Forest** los modelos fueron entrenados sin aplicar técnicas de bajo muestreo.

3. MODELOS

3.1. Árboles de Decisión

En lo que respecta al desarrollo de modelos basados en árboles de decisión, tras un análisis comparativo, se optó como mejor candidato el `ARBOL_MODIF_GINI` cuyo modelo incorpora una técnica de bajomuestreo. Con dicho modelo se identificó un mejor desempeño para identificar el 10 % de la población con mayor disposición de contratación.

Para la construcción de este árbol, se utilizaron las siguientes opciones técnicas:

- **Ajuste de probabilidades a priori:** Se configuraron probabilidades a priori del 36 % para la clase 1 y 64 % para la clase 0. Este ajuste, realizado en el nodo de decisión, permite alinear las salidas del modelo con la distribución real de la población.
- Utilizar decisiones sí, utilizar probabilidades a priori sí.
- **Criterio de división:** Criterio objetivo nominal: Índice Gini, para maximizar la pureza de los nodos.
- **Tamaño de la hoja:** 20 en los nodos terminales.
- **Método de poda:** Evaluación, utilizando como medida evaluación: mejora con fracción de evaluación 0.1.
- **Tamaño del árbol:** El modelo resultante presenta una estructura con un total de 29 nodos.

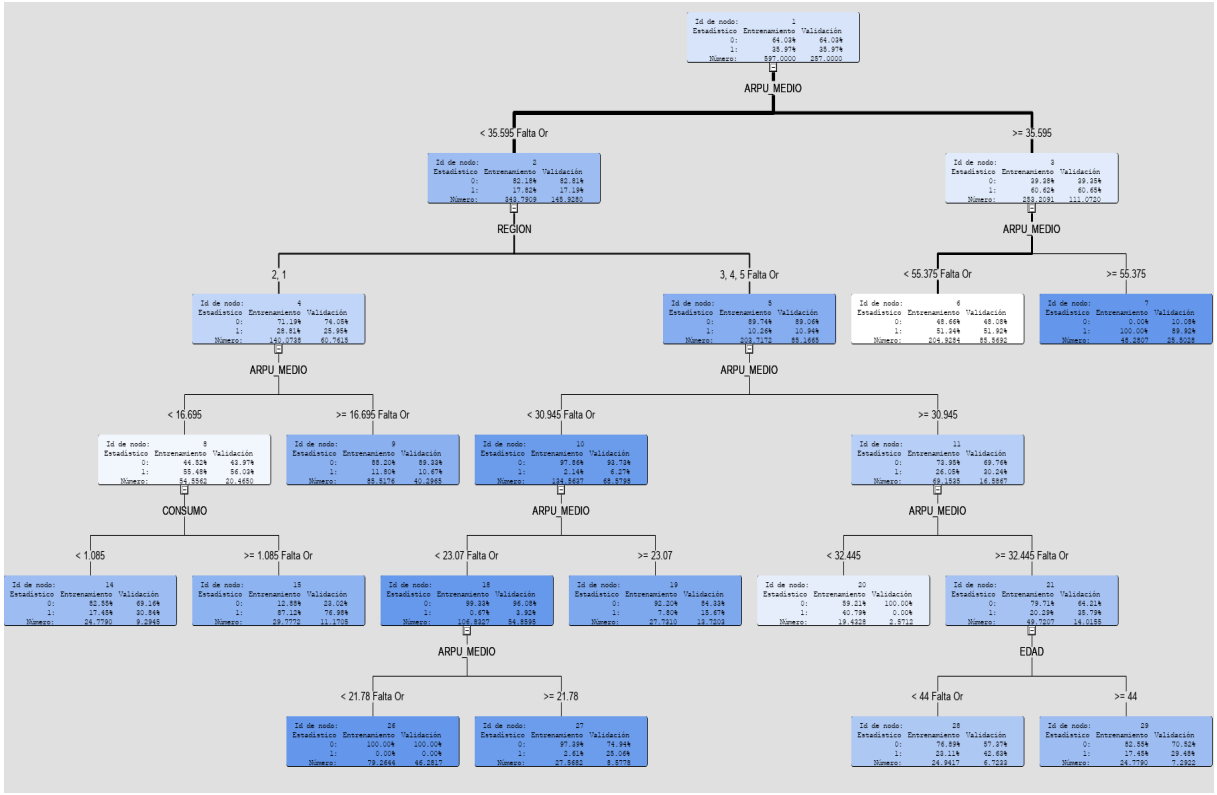


Figura 4: Estructura del árbol de decisión con bajo muestreo (`ARBOL_MODIF_GINI`)

El análisis de la estructura inicial del árbol de decisión (Figura 4) identifica a la variable ARPU_MEDIO como el principal factor discriminativo en la contratación del seguro móvil. El primer nodo de decisión establece un umbral en ARPU_MEDIO = 35.595, separando a los clientes en dos segmentos con comportamientos claramente diferenciados en términos de contratación.

En la rama izquierda, correspondiente a clientes con ARPU_MEDIO inferior a 35.595, se observa una baja tasa de contratación del seguro. En este nodo, la proporción de clientes que no contratan el seguro es claramente mayoritaria, lo que indica que los clientes con menor ingreso medio presentan una respuesta limitada a la campaña. Dentro de este segmento, el árbol introduce variables adicionales como REGION, CONSUMO y EDAD, con el objetivo de identificar subgrupos con una ligera mejora en la probabilidad de contratación. No obstante, incluso en estos nodos más específicos, la clase cero sigue siendo predominante, lo que evidencia que la mayoría de los clientes de bajo ARPU_MEDIO no contratan el seguro. Esto sugiere que las acciones comerciales dirigidas a estos perfiles tendrían una efectividad limitada.

Por otro lado, la rama derecha del árbol, correspondiente a clientes con ARPU_MEDIO mayor o igual que 35.595, presenta una tasa de contratación significativamente superior. En este segmento, el porcentaje de clientes que contratan el seguro aumenta de forma notable respecto a la rama de bajo ARPU_MEDIO. Además, el árbol introduce un segundo punto de corte relevante en ARPU_MEDIO = 55.375, que permite identificar un subsegmento con una tendencia mayor a la contratación. En algunos nodos terminales de esta rama, la proporción de clientes que contratan el seguro es claramente dominante, alcanzando valores cercanos al 90 % en la muestra de validación, lo que pone de manifiesto una elevada afinidad de estos perfiles con el producto ofertado.

Dentro de este segmento de alto ARPU_MEDIO, el árbol refina la segmentación incorporando variables como REGION, CONSUMO y EDAD. Estas variables permiten discriminar perfiles concretos con diferentes niveles de propensión, observándose que determinados rangos de edad y niveles de consumo concentran una mayor proporción de contrataciones. Esta evidencia sugiere que, además del nivel de gasto del cliente, existen otros factores que influyen en la decisión de contratar el seguro móvil.

Desde una perspectiva de negocio, la estructura del árbol respalda una estrategia comercial orientada a priorizar a los clientes con ARPU_MEDIO elevado, especialmente aquellos con valores superiores a 55.375, donde la probabilidad de contratación es sustancialmente mayor. En estos segmentos, una asignación preferente de recursos comerciales, como campañas personalizadas o acciones proactivas de venta, resulta justificable. Por el contrario, para los clientes con ARPU_MEDIO bajo, donde el árbol muestra de forma consistente una baja tasa de contratación incluso tras una segmentación adicional, una estrategia basada en acciones de bajo coste o campañas masivas parece más adecuada.

En conclusión, el árbol de decisión proporciona un modelo predictivo eficaz y, además, ofrece una interpretación clara y respaldada por los datos observados en los distintos nodos, alineada con la lógica comercial y útil para la toma de decisiones estratégicas orientadas a maximizar la eficiencia de la campaña de seguros móviles.

Los estadísticos de ajuste del árbol (Figura 5) muestran consistencia entre entrenamiento y validación, con tasas de clasificación errónea similares (21.8 % vs 23.0 %) y errores cuadráticos medios cercanos (0.146 vs 0.164). Esta estabilidad sugiere que el modelo generaliza adecuadamente a nuevos datos sin presentar sobreajuste significativo.

Target	Etiqueta target	Estadísticos de ajuste	Etiqueta de estadísticos	Entrenamiento	Validación
ContrataSeguroMovil		_NOBS_	Suma de frecuencias	597	257
ContrataSeguroMovil		_MISC_	Tasa de clasificación ...	0.217755	0.229572
ContrataSeguroMovil		_MAX_	Error absoluto máximo	0.954545	1
ContrataSeguroMovil		_SSE_	Suma de errores cuad...	174.763	84.08581
ContrataSeguroMovil		_ASE_	Error cuadrático medio	0.146368	0.163591
ContrataSeguroMovil		_RASE_	Raíz del error cuadráti...	0.38258	0.404464
ContrataSeguroMovil		_DIV_	Divisor para ASE	1194	514
ContrataSeguroMovil		_DFT_	Grados de libertad tot...	597	.
ContrataSeguroMovil		_APROF_	Beneficio medio para ...	0.767396	0.755783
ContrataSeguroMovil		_PROF_	Beneficio total para C...	458.1351	194.2362
ContrataSeguroMovil		_PASE_	Error cuadrático medi...	0.156355	0.176247
ContrataSeguroMovil		_PMISC_	Tasa de clasificación ...	0.232604	0.244217

Figura 5: Estadísticos de ajuste del árbol de decisión con bajo muestreo

La capacidad del modelo para cumplir el objetivo de identificar al 10 % de clientes con mayor propensión a contratar el seguro se evalúa mediante la curva de mejora acumulada (*lift*). En la Figura 6 se comparan las versiones del árbol de decisión con y sin aplicación de bajo muestreo. Los resultados muestran que el modelo entrenado con bajo muestreo y posterior ajuste de probabilidades a priori presenta un **lift superior en el primer decil**, lo que indica una mayor concentración de clientes que finalmente contratan el seguro.

Este comportamiento es especialmente relevante desde una perspectiva de negocio, ya que la campaña está orientada a focalizar los esfuerzos comerciales en un subconjunto reducido de clientes. Por tanto, el mejor desempeño del modelo con bajo muestreo en los primeros deciles justifica su selección como el modelo de árbol de decisión de referencia, que será comparado con el resto de modelos propuestos en fases posteriores del análisis.

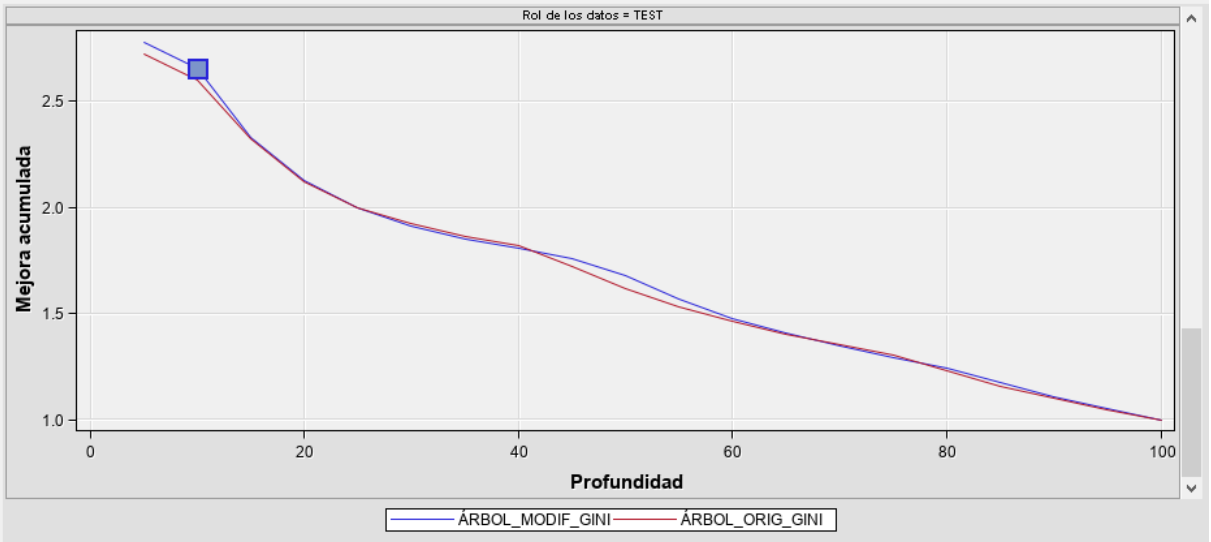


Figura 6: Lift para los árboles de decisión

3.2. Regresión Logística

Para el ajuste del modelo de regresión logística se llevaron a cabo diversas pruebas con el objetivo de identificar el modelo más adecuado. En una primera fase, se ajustaron modelos variando el método de selección de variables, comparando la utilización de todas las variables disponibles, con los métodos *Stepwise* y *Backward*. Tras la comparación de los resultados obtenidos, se concluyó que el modelo ajustado mediante el uso de todas las variables proporcionaba un mejor ajuste a los datos.

Posteriormente, se exploraron dos enfoques adicionales: un modelo de regresión logística con bajomuestreo (50-50) y otro con transformaciones WOE (*Weight of Evidence*). Como resultado del análisis se seleccionó el modelo con transformaciones WOE, el cual transforma las variables en variables contínuas basadas en la capacidad de discriminación de sus intervalos respecto a la variable objetivo. Este enfoque permite utilizar variables que capturan relaciones no lineales entre las variables explicativas y el target.

En la figura 7 se muestra la curva de mejora acumulada (*cumulative lift*), donde se comparan los tres modelos analizados: el modelo con utilización de todas las variables, el con bajomuestreo y el con transformaciones WOE. Del análisis de dicha curva se observa que el modelo WOE presenta el mayor Lift en el primer decil. Dado que el objetivo de estudio es identificar el 10 % de los clientes con mayor probabilidad de respuesta positiva, se selecciona este modelo.

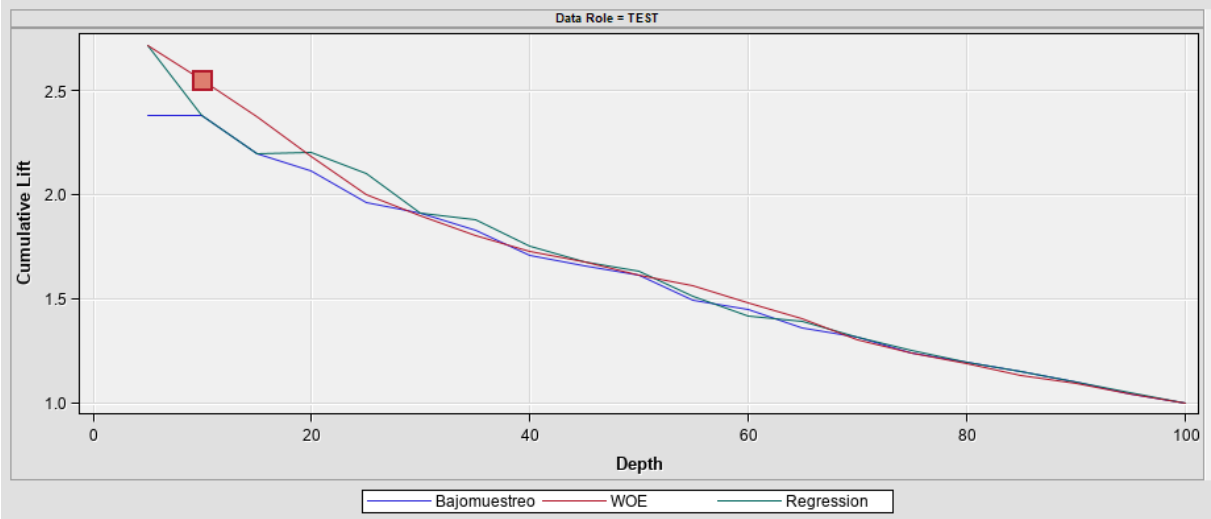


Figura 7: Lift acumulado para los distintos modelos

Además, analizando la tabla de estadísticos de ajuste del modelo WOE, se observa que este presenta un coeficiente Gini de 0.64 en la muestra de test y uno de 0.66 en la muestra de entrenamiento. Este comportamiento sugiere que el modelo no presenta sobreajuste y que las agrupaciones realizadas mediante el nodo Agrupación Interactiva son sólidas, capturando patrones de comportamiento que se mantienen en la población general.

Modelo de Regresión Logística con Transformaciones WOE
Para la construcción del modelo de regresión logística con transformaciones WOE se optó por realizar una modelización basada en la discretización de las variables de entrada, para lo cual se utilizó un nodo de agrupación interactiva.

En primer lugar, se definieron las siguientes opciones para la transformación y discretización de las variables:

- **Transformación binária:** Método *Weighting*
- **Método Binning:** Quantil
- **Número de intervalos:** 20

En el apartado de opciones de agrupación se establecieron los siguientes parámetros:

- **Método de agrupación:** Criterio óptimo, lo que permite realizar la discretización mediante árboles de decisión.
- **Número máximo de grupos a considerar:** 5
- Se ha activado la opción "Ajustar WOE", con un factor de ajuste de 0.5.

El método de selección de variables se dejó configurado como "Ninguno", con el objetivo de favorecer el proceso de discretización de las variables.

Una vez finalizado este proceso, se aplicó un nodo Metadatos para rechazar el rol de las variables no transformadas, garantizando así que únicamente las variables transformadas mediante WOE fueran utilizadas en el modelo.

Posteriormente, en el nodo de regresión logística (denominado WOE), se fijaron las siguientes opciones:

- **Tipo de modelo:** regresión logística con función de *linkaje* logit.
- **Método de selección de variables:** paso a paso.

Se eligió el método paso a paso dado que, al encontrarse las variables previamente alineadas con el *target* tras la transformación WOE, se facilita la identificación de variables con capacidad informativa relevante, al situarlas en una escala común y favorecer así la estabilidad del proceso de selección.

En la Figura 8 se observa que las variables `ARPU_MEDIO` (IV=1.548) y `TARIFA_ACTUAL` (IV=1.198) presentan una capacidad de discriminación extrema. Desde una prespectiva de negocio, esto indica que el ingreso y el tipo de contrato actual son los principales predictores de respuesta a la campaña. Asimismo, el Gini elevado de `ARPU_MEDIO` (55.4) sugiere que la ordenación de los clientes de mayor a menor propensión de contratación puede realizarse de forma eficaz.

El análisis de los gráficos del *Event Rate* presentados en la Figura 9, permite profundizar en la identificación de los perfiles más receptivos a la campaña. En el caso de `ARPU_MEDIO`, se observa un pico de respuesta en el Grupo 4 (`ARPU_MEDIO` entre 36.69 y 55.2), lo que sugiere que estos clientes son los más propensos a la contratación. En relación a la variable `CONSUMO`, se identifica una relación no lineal, destacando los Grupos 3 y 5 como los más activos, lo que justifica el uso de transformaciones WOE. Por su parte, en la variable `TARIFA_ACTUAL` se aprecia una tendencia clara: a medida que aumenta el nivel de tarifa, hacia el grupo 5, la tasa de eventos se dispara, indicando que los clientes con tarifas más altas tienen mucha más predisposición de contratar el seguro.

Adicionalmente, el análisis muestra que en la variable `REGION` no existe un grupo claramente dominante. En cuanto a la variable EDAD, se concluye que el grupo 3 (edades entre 33 y 50) es el que presenta una mayor propensión a responder positivamente a la campaña.

Variable	Gini Statistic	Information Value
ARPU_MEDIO	55.433	1.548
TARIFA_ACTUAL	54.685	1.198
TARIFA_PREVIA	44.185	0.704
CONSUMO	22.336	0.163
EDAD	13.095	0.068
REGION	9.264	0.027

Figura 8: Event Rate

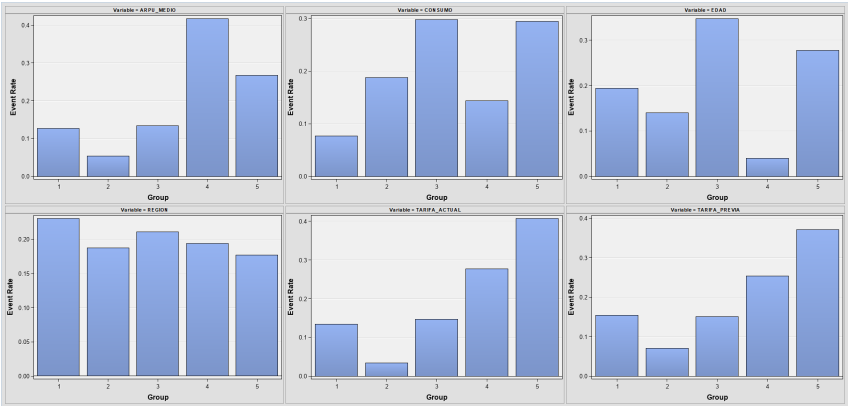


Figura 9: IV

Si se analiza la columna de Estimación de la Figura 10, se observa el efecto estimado de cada factor sobre la probabilidad de contratación. Todas las estimaciones resultan negativas, lo que indica una relación directa entre la evidencia de evento capturada por las transformaciones WOE y la probabilidad final de contratación. En términos de negocio, esta estructura confirma que el modelo respecta las tendencias observadas en el análisis univariante: por ejemplo, en el caso de la variable `TARIFA_ACTUAL`, a valores más elevados corresponde una mayor probabilidad de contratación. Esta lógica es consistente en todas las variables incluidas en el modelo. Además, dado que todos los parámetros presentan un p-valor inferior a 0.05, se concluye que todas las variables aportan información estadísticamente significativa al modelo conjunto.
El análisis de los *ODDs RATIO* permite interpretar como varía la "ventaja" de contratación por cada unidad de incremento del WOE. La variable `WOE_REGION` es la que tiene el valor más alejado del 1 (0.148). Esto indica que es la variable que genera cambios más drásticos en la probabilidad al pasar de una región a otra. Por su parte, `WOE_EDAD` (0.321) y `WOE_CONSUMO` (0.324) muestran un impacto moderado y muy similar entre sí. Por último, `WOE_ARPU_MEDIO` (0.564) y `WOE_TARIFA_ACTUAL` (0.575) presentan un efecto más suave y estable dentro del modelo multivariante.

Análisis de estimadores de verosimilitud máxima

Parámetro	DF	Estimación	Error estándar	Chi-cuadrado de Wald	Pr > ChiSq	Estimador estandarizado	Esp(Est)
Intercept	1	-0.5859	0.0883	44.04	<.0001		0.557
WOE_ARPU_MEDIO	1	-0.5729	0.1614	12.59	0.0004	-0.4568	0.564
WOE_CONSUMO	1	-1.1265	0.2340	23.17	<.0001	-0.2538	0.324
WOE_EDAD	1	-1.1361	0.3298	11.87	0.0006	-0.1622	0.321
WOE_REGION	1	-1.9082	0.5402	12.48	0.0004	-0.1735	0.148
WOE_TARIFA_ACTUAL	1	-0.5542	0.1710	10.50	0.0012	-0.3680	0.575

Figura 10: Estimaciones

Estimadores de ratio de probabilidad

Efecto	Estimador de punto
WOE_ARPU_MEDIO	0.564
WOE_CONSUMO	0.324
WOE_EDAD	0.321
WOE_REGION	0.148
WOE_TARIFA_ACTUAL	0.575

Figura 11: ODDS obtenidos

Si bien el análisis univariante identificó a `ARPU_MEDIO` y `TARIFA_ACTUAL` como los predictores con mayor capacidad discriminante individual, la regresión logística pone `REGION` y `EDAD` como los factores que ejercen un mayor peso relativo para ajustar la probabilidad final de contratación cuando se consideran de forma conjunta con el resto de variables.

3.3. Modelo Ensamblado

3.3.1. Gradient Boosting

Para la construcción del primer modelo ensamblado, se implementó un algoritmo de *Gradient Boosting*, un método que combina múltiples árboles de decisión de forma secuencial. Cada nuevo árbol corrige los errores del anterior, mejorando progresivamente la precisión del modelo.

La configuración específica del Gradient Boosting se estableció de la siguiente manera:

- **Iteraciones y aprendizaje:** Se entrenaron 250 árboles con una tasa de aprendizaje de 0.05 con una proporción de entrenamiento del 67 % .
- **Complejidad de los árboles base:** Cada árbol individual se limitó a una profundidad máxima de 2 niveles y máximo 2 ramas por nodo.
- **Tamaño mínimo de hoja:** Se configuró una fracción de hoja de 0.05, lo que significa que cada nodo terminal (hoja) debe contener al menos el 5 % de las observaciones.
- **Selección de variables:** Se activó la selección automática de predictores, permitiendo que el algoritmo identifique y utilice principalmente las variables con mayor poder predictivo para la contratación del seguro móvil.

El análisis de importancia de variables del modelo Gradient Boosting (Figura 12) confirma el rol determinante de `ARPU_MEDIO` (importancia 1.0 en ambos conjuntos). Un hallazgo relevante es `TARIFA_ACTUAL`, que con un ratio de validación de 1.58 muestra una mayor capacidad predictiva en datos no vistos respecto al entrenamiento. En contraste, `CONSUMO` presenta un ratio de 0.28, indicando menor consistencia entre conjuntos.

Nombre de la variable	Etiqueta	Número de reglas de división ▼	Importancia	Importancia de validación	Ratio de validación para la importancia de entrenamiento
ARPU_MEDIO		255	1	1	1
REGION		100	0.456233	0.444379	0.974017
CONSUMO		78	0.327985	0.090615	0.276278
EDAD		77	0.323416	0.129775	0.401263
TARIFA_PREVIA		22	0.16083	0.159664	0.992745
TARIFA_ACTUAL		5	0.081441	0.129459	1.589611

Figura 12: Importancia de variables en el modelo Gradient Boosting.

Aunque los estadísticos de ajuste globales (Figura 13) muestran consistencia del modelo (ASE: 0.127 → 0.153, MISC: 16.6 % → 22.2 %), estas métricas evalúan el rendimiento global. Para el objetivo específico de capturar al 10 % de clientes más propensos, la métrica clave es el lift en el primer decil, que se analiza en la siguiente sección comparativa.

Target	Etiqueta target	Estadísticos de ajuste	Etiqueta de estadísticos	Entrenamiento	Validación
ContrataSeguroMovil		_NOBS_	Suma de frecuencias	831	356
ContrataSeguroMovil		_SUMW_	Suma de casos pond...	1662	712
ContrataSeguroMovil		_MISC_	Tasa de clasificación ...	0.166065	0.22191
ContrataSeguroMovil		_MAX_	Error absoluto máximo	0.960701	0.956266
ContrataSeguroMovil		_SSE_	Suma de errores cuad...	210.9002	108.9855
ContrataSeguroMovil		_ASE_	Error cuadrático medio	0.126895	0.15307
ContrataSeguroMovil		_RASE_	Raíz del error cuadráti...	0.356224	0.391241
ContrataSeguroMovil		_DIV_	Divisor para ASE	1662	712
ContrataSeguroMovil		_DFT_	Grados de libertad tot...	831	.

Figura 13: Estadísticos de ajuste globales del modelo Gradient Boosting.

3.3.2. Random Forest

Como segundo modelo ensamblado, se propone el ajuste de un modelo de *Random Forest*, con el objetivo de capturar posibles relaciones complejas y no lineales entre las variables explicativas y la variable objetivo.

Para la configuración de los parámetros del modelo de la parte de entrenamiento se definieron las siguientes opciones.

- **Árbol:** Se fijó el número máximo de árboles en 50, y se utilizó un muestreo aleatorio con una proporción del 67 % del conjunto de datos original para la construcción de cada árbol.
- **Reglas de segmentación:** se ha fijado la profundidad máxima de los arboles en 50 niveles, el número de variables considerados en la búsqueda de *splits* en 5 variables y el número máximo de categorías en la búsqueda de *splits* en 30.
- **Nodos:** Se ha establecido 1 como el número mínimo de observaciones por nodo.

En relación a la parte de *Scoring*, se han considerado 3 variables.

Los estadísticos de ajuste global del modelo evidencian su consistencia (Figura 14). En particular, el ASE presenta un ligero incremento de 0.130 a 0.139, mientras que el MISC aumenta de 18.9 % a 21.6 % al pasar del conjunto de entrenamiento al de validación. Estas variaciones moderadas indican la ausencia de sobreajuste y confirman el adecuado desempeño del modelo en datos no vistos.

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
ContrataSeguroMovil		ASE	Average Squared Er...	0.130178	0.139008
ContrataSeguroMovil		DIV	Divisor for ASE	1662	712
ContrataSeguroMovil		MAX	Maximum Absolute ...	0.941707	0.944033
ContrataSeguroMovil		NOBS	Sum of Frequencies	831	356
ContrataSeguroMovil		RASE	Root Average Squar...	0.360801	0.372838
ContrataSeguroMovil		SSE	Sum of Squared Err...	216.3552	98.97397
ContrataSeguroMovil		DISF	Frequency of Classi...	831	356
ContrataSeguroMovil		MISC	Misclassification Rate	0.188929	0.216292
ContrataSeguroMovil		WRONG	Number of Wrong C...	157	77

Figura 14: Estadísticos de ajuste globales del modelo Random Forest

Dado que este tipo de modelos presenta una menor interpretabilidad directa, se recurre al análisis de gráficos de importancia de variables, que permiten evaluar la contribución relativa de cada predictor dentro del conjunto ensamblado (Figura 15).

En el gráfico se observa un dominio claro de las variables relacionadas con el gasto, destacando `ARPU_MEDIO` y `TARIFA_ACTUAL`, que aparecen situadas en el extremo derecho. Este resultado indica que estas variables actúan como los predictores más relevantes del modelo, con una contribución notablemente superior al resto.

Asimismo, se aprecia un salto significativo en la importancia tras la variable `CONSUMO`, lo que sugiere que el modelo basa fundamentalmente su capacidad predictiva en las primeras tres variables.

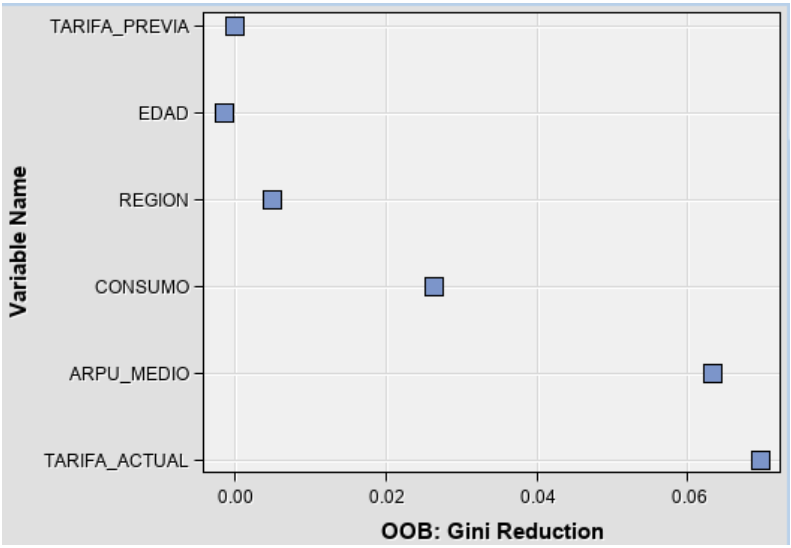


Figura 15: Importancia de las Variables

El análisis de importancia de variables indica que el comportamiento que se pretende predecir está principalmente asociado a características relacionadas con el gasto y tarifa contractada actualmente, mientras que variables demográficas como la edad o la localización geográfica presentan una contribución relativa menor dentro del modelo.

3.3.3. Elección

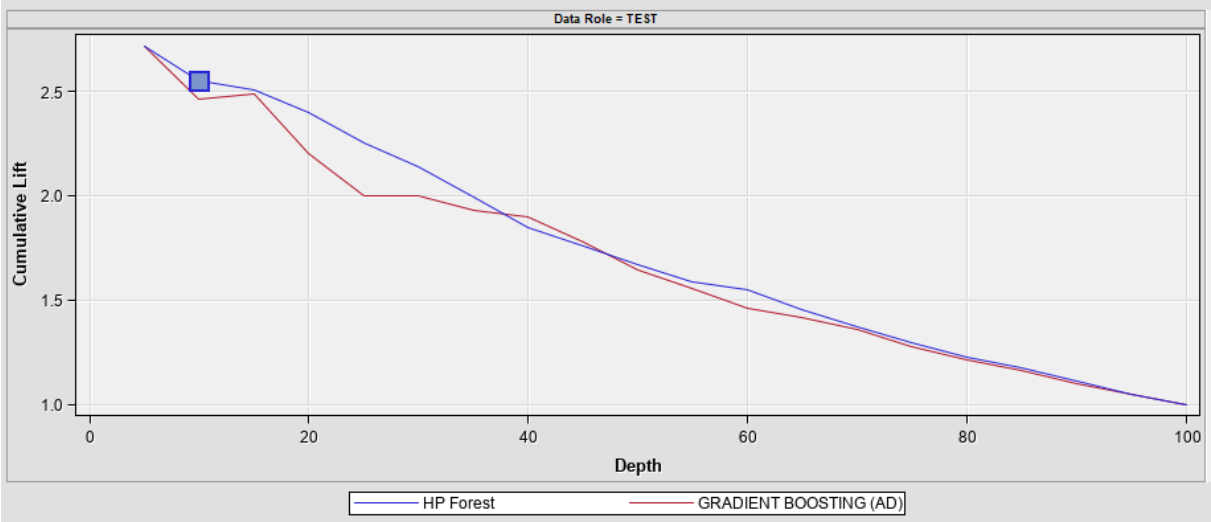


Figura 16: Lift acumulado modelos ensamblados

Comparando ambos modelos ensamblados, *Random Forest* y *Gradient Boosting*, y utilizando nuevamente como criterio de selección el Lift acumulado en el 10 % de la población, en coherencia con el objetivo principal de la campaña, se selecciona el modelo *Random Forest* como el modelo ensamblado ajustado. (Figura 16).

Este modelo se utilizará, por tanto, en la comparación final con el resto de modelos propuestos, con el fin de determinar la alternativa más adecuada para la selección del público objetivo de la campaña.

3.4. Comparaciones

Con el fin de determinar el modelo más apropriado para el objetivo del estudio, se realizó una comparación entre los tres modelos previamente seleccionados:

- Árbol de decisión con bajomuestreo
- Regresión logística con transformaciones WOE
- Modelo ensamblado *Random Forest*

Dado que el objetivo principal es definir el público óptimo que se seleccionaría en la muestra de test de una campaña digital al 10 % de los clientes con mayor probabilidad de responder positivamente, el criterio de selección del modelo se centró en en *Lift* acumulado en el primer decil, *Depth*=10. Esto resulta especialmente adecuado en campañas dirigidas, ya que evalúa la capacidad del modelo para ordenar correctamente a los clientes según su propensión a respuesta en el segmento de mayor interés.

Cabe destacar que, desde un punto de vista global de calidad predictiva, el modelo Random Forest presenta valores superiores en métricas como el ROC Index y el índice de Gini. No obstante, estas métricas evalúan el rendimiento del modelo a lo largo de toda la distribución y no están específicamente orientadas a la optimización del segmento objetivo de la campaña.

A partir de la comparación visual de las curvas de Lift acumulado presentadas en la Figura 17, se observa que el modelo del árbol de decisión con bajomuestreo es el que alcanza el mayor Lift acumulado en el depth de 10 llegando a un valor de 2.65, superando a la regresión logística con transformaciones WOE y al modelo *Random Forest* en el segmento objetivo.

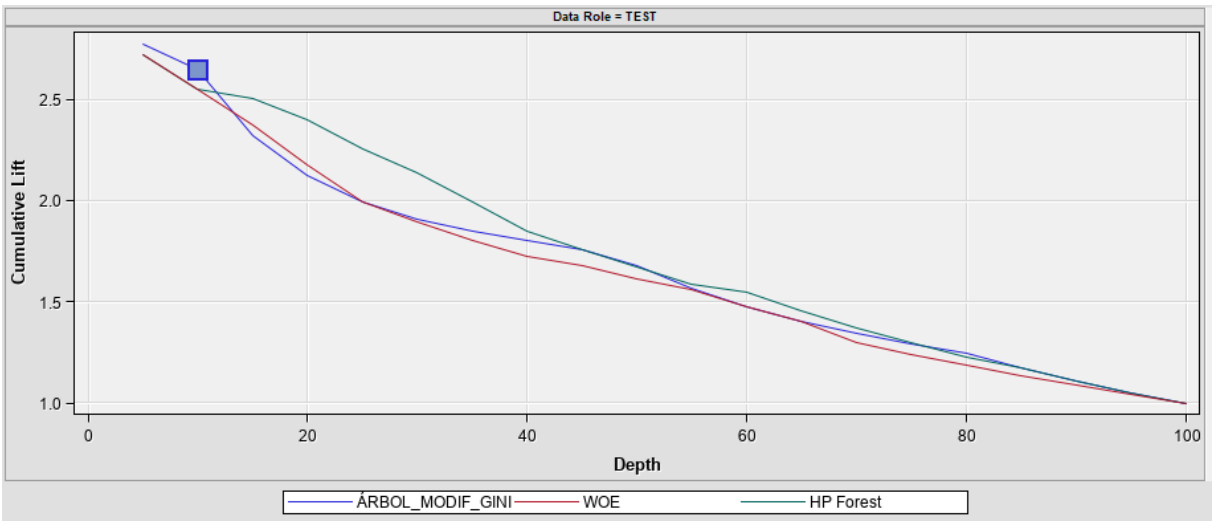


Figura 17: Comparativa Lift

En consecuencia, y aun cuando el Random Forest presenta mejores valores de ROC Index y Gini, se selecciona el árbol de decisión con bajomuestreo como el modelo más adecuado para esta campaña. Esta decisión responde a que, en un escenario de recursos

limitados donde solo se contactará al 10 % de la base de clientes, la prioridad no es la precisión global del modelo, sino su capacidad de discriminación en el tramo superior de la distribución.

Más allá de su superioridad en el target específico, la elección del árbol aporta una ventaja operativa: la accionabilidad. Mientras que el Random Forest opera como una 'caja negra', la estructura jerárquica del árbol permite traducir los hallazgos en reglas de negocio inmediatas. Por ejemplo, la relevancia del ARPU_MEDIO identificada por el modelo facilita la segmentación de estrategias: permite priorizar acciones comerciales de alto valor en clientes con consumos superiores a los umbrales críticos detectados, mientras se reservan canales automáticos o de bajo coste para los segmentos de menor propensión.

En conclusión, parece que este modelo es el mejor para el caso de uso, ya que maximiza el éxito en el objetivo de la campaña al obtener el mayor Lift en el primer decil. Además, su interpretabilidad proporciona una segmentación clara para el equipo de ventas, permitiendo traducir los datos en reglas de negocio directas y fáciles de ejecutar.

4. CONLCUSIONES

En el presente trabajo se ha abordado el problema de la selección del público objetivo para una campaña de contratación de seguro móvil, aplicando distintas técnicas de modelización supervisada, como son los árboles de decisión, regresiones logísticas y modelos ensamblados. Tras la comparación de los modelos realizada a partir del Lift acumulado en el primer decil pues el objetivo es identificar al 10 % de clientes más porpensos a realizar la contratación del seguro móvil, se ha seleccionado el árbol de decisión con bajomuestreo como el modelo más adecuado para el objetivo planteado.

En conclusión, el estudio evidencia la importancia de alinear los criterios de evaluación del modelo con el objetivo final del negocio, así como la necesidad de no basar la selección del modelo únicamente en métricas globales de ajuste. La metodología seguida permite justificar de forma sólida la elección del modelo final y proporciona una base consistente para su aplicación en campañas dirigidas reales.