Árboles de Decisión

Siguiendo con los modelos predictivos, en este apartado se analizará el algoritmo de los Árboles de Decisión, CART en adelante, con el mismo propósito específico: la clasificación de clientes en categorías de riesgo crediticio. En particular, nos enfocaremos en discendir entre aquellos clientes que puedan tener dificultades de pago y aquellos que son financieramente solventes.

El algoritmo de Árboles de Decisión se revela como una herramienta particularmente poderosa en este contexto, ya que su capacidad para modelar relaciones complejas entre variables puede proporcionar insights para la toma de decisiones financieras. Exploraremos cómo el algoritmo selecciona de manera inteligente las variables más influyentes para segmentar eficientemente el conjunto de datos, permitiendo la identificación de patrones que podrían indicar riesgos financieros.

Algoritmo

En este contexto, la estructura de un Árbol de Decisión se modela de forma análoga a un proceso de decisiones estratégicas:

- Cada nodo interno del árbol representa una evaluación crítica sobre un atributo financiero específico. Estas evaluaciones sirven como puntos clave para discernir las distintas condiciones financieras de los clientes.
- Las ramas que se desprenden de cada nodo interno representan las diferentes trayectorias que un cliente puede seguir según el resultado de la evaluación realizada en ese nodo.
- Las hojas del árbol en el contexto financiero contienen la información crucial: la etiqueta o el valor predicho relacionado con la capacidad del cliente para afrontar compromisos financieros. Esto puede manifestarse como una clasificación de riesgo, como “solvente” o “en riesgo”, proporcionando una guía clara para las decisiones crediticias.

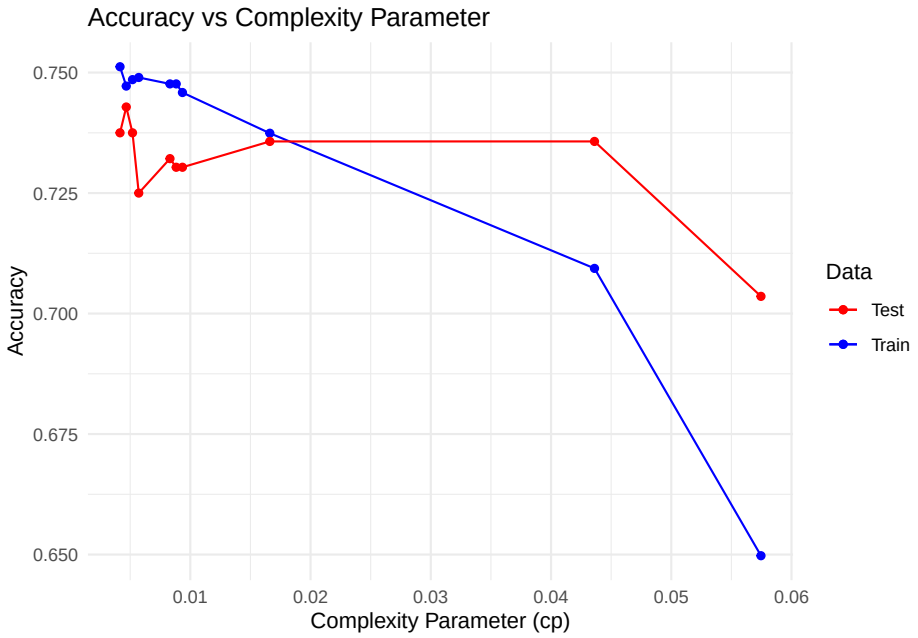
Así pues, a continuación se procede a realizar dicho análisis predictivo.

Desarrollo del CART

Para inciar el desarrollo del modelo, el primer paso es encontrar el valor óptimo del complexity parameter, o parámetro de complejidad, que controla la cantidad de ramificaciones y nodos terminales en el árbol. Este parámetro juega un papel importante en la regularización del árbol, evitando que éste se vuelva demasiado complejo y se adapte demasiado a los datos de entrenamiento, lo que podría resultar en un sobreajuste del modelo.

Entonces, para encontrar este valor óptimo del parámetro de complejidad, se entrena el modelo con los datos balanceados de Train y se realiza un proceso de crosvalidación con 10 folds. Entonces calculamos el accuracy para cada 10 valores posibles del complexity parameter tanto para los datos train como test.

Figura 149: Evolución de la precisión (obtenida mediante validación cruzada) dependiendo del parámetro de complejidad



En el gráfico se observa como el primer valor del complexity parameter es el que reporta un mayor accuracy para el conjunto de datos de entrenamiento. No obstante, no únicamente buscamos el hiperparámetro que nos aporte un mayor accuracy, sino que también nos interesa encontrar un cp con el que además de maximizar el accuracy, evitemos overffiting (sobreajuste del modelo). Así pues, observamos como el segundo valor del complexity parameter nos da un valor que no se aleja mucho del accuracy óptimo y, además, nos evita en una gran manera un overfitting. Por lo tanto, concluimos que el cp óptimo para nuestro árbol de decisión final es 0.0046729.

Validación del modelo

Una vez ejecutado el modelo CART, con el objetivo de validar el modelo, se muestra en una tabla la matriz de confusión y se calcula la precisión con la que el algoritmo ha predicho la variable Target tanto en la población del Train como en la del Test, para observar si ha habido un sobreajuste o no.

Cuadro 73: Matriz de confusión del conjunto de validación

	Realidad	
	No moroso	Potencial moroso
Predicción		
No moroso	295	120
Potencial moroso	25	120

Cuadro 74: Medidas de Validación para el modelo CART

	Train	Test
Accuracy	0.7894971	0.7410714
Sensitivity	0.5711319	0.5000000
Specificity	0.9532710	0.9218750
Recall	0.5711319	0.5000000
F1	0.6993007	0.6233766
Precision	0.9016393	0.8275862

La anterior salida nos muestra la matriz de confusión junto con diversos estadísticos que tratan de explicar como de bien o mal ha predicho el algoritmo de CART. Así pues, como se observa, las medidas de validación son aproximadamente las mismas tanto para Train como para Test, por lo tanto, reafirmamos que no hay un sobreajuste en el modelo (como ya se había dicho anteriormente).

En este caso, la precisión ha sido del 0.7411 %, lo que indica que el algoritmo ha predicho correctamente el 74.1071 % de los individuos de Test. Esto indica que el modelo es capaz de clasificar correctamente a la mayoría de los clientes en categorías de riesgo crediticio.

La sensibilidad del modelo, que mide la capacidad de identificar clientes potencialmente morosos, es del 50 % en el conjunto de prueba. Esto sugiere que hay margen para mejorar en la identificación de clientes con dificultades de pago.

Por otro lado, el modelo muestra una alta especificidad, del 92.1875 %, indicando su habilidad para identificar clientes no morosos con precisión.

Si observamos otras métricas disponibles, apreciaremos como la precisión, que mide la exactitud de las predicciones positivas, es del 82.7586 % en el conjunto de prueba. Esto significa que cuando el modelo predice que un cliente es potencialmente moroso, es correcto en aproximadamente el 82.76 % de las veces. Por último, podemos apreciar como la puntuación F1, que equilibra precisión y recuperación, es del 62.3 % en el conjunto de prueba, indicandonos que el modelo logra un buen equilibrio entre la precisión de las predicciones positivas y la capacidad para recuperar casos positivos.

En resumen, el modelo muestra un buen rendimiento general, especialmente en términos de especificidad, pero hay margen para mejorar en la identificación de clientes potencialmente morosos, como lo sugiere la sensibilidad y la puntuación F1 en ambos conjuntos.

Prueba ácida

Cuadro 75: Matriz de confusión del conjunto de validación desbalanceado

	Realidad	
	No moroso	Potencial moroso
Predicción		
No moroso	856	68
Potencial moroso	63	12

Cuadro 76: Medidas de Validación con el conjunto test desbalanceado para el modelo CART

	Train	Test_desbalanceado
Accuracy	0.7894971	0.8688689
Sensitivity	0.5711319	0.1500000
Specificity	0.9532710	0.9314472
Recall	0.5711319	0.1500000
F1	0.6993007	0.1548387
Precision	0.9016393	0.1600000

El accuracy del modelo en el conjunto de prueba desbalanceado ha sido del 0.8689 %, lo que indica que el 0.8689 % de las predicciones fueron buenas. Sin embargo, la exactitud puede ser engañosa en conjuntos de datos desbalanceados, donde la mayoría de las observaciones pertenecen a una clase particular. Por otra parte, la sensibilidad en el conjunto de prueba desbalanceado es bastante baja, solo del 15 %. Esto significa que el modelo tiene dificultades para identificar correctamente a los clientes morosos. La sensibilidad es especialmente crucial en situaciones financieras, ya que representa la capacidad del modelo para capturar la totalidad de los casos positivos (morosos) reales. En este caso, el bajo valor de sensibilidad indica que el modelo está dejando pasar un número significativo de clientes morosos sin detectarlos.

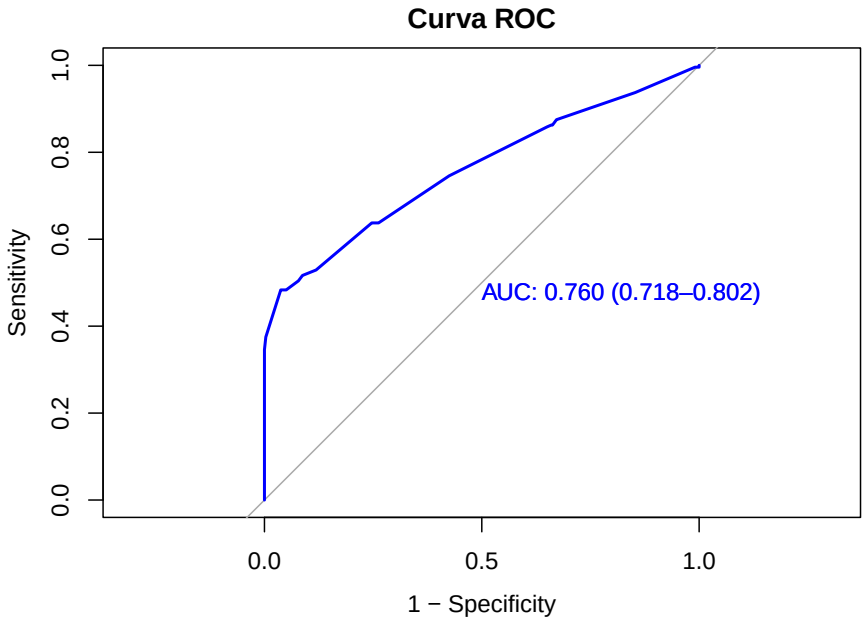
Así pues, la especificidad que mide la capacidad del modelo para identificar correctamente los casos negativos (clientes no morosos), es del 93.14 %. Esto sugiere que el modelo tiene un buen rendimiento al identificar a los clientes que no son morosos. Sin embargo, es importante destacar que la alta especificidad podría deberse al desbalance en los datos, ya que hay más clientes no morosos en el conjunto de prueba.

Finalmente, el valor de F1 es del 15.48 %, lo que refleja un equilibrio entre precisión y sensibilidad. Este valor relativamente bajo sugiere que hay margen de mejora en la capacidad del modelo para identificar clientes morosos sin comprometer demasiado la precisión. De la misma manera, en cuanto a la precisión (Precision), es del 16 %, lo que significa que de las instancias que el modelo predice como morosas, solo el 16 % son realmente morosas. Este valor puede ser bajo, indicando que el modelo podría estar generando demasiados falsos positivos.

Curva ROC

Para un análisis más profundo sobre la calidad de predicción del modelo, se representa la curva ROC y se interpreta su área bajo la curva (AUC).

Figura 150: Curva ROC

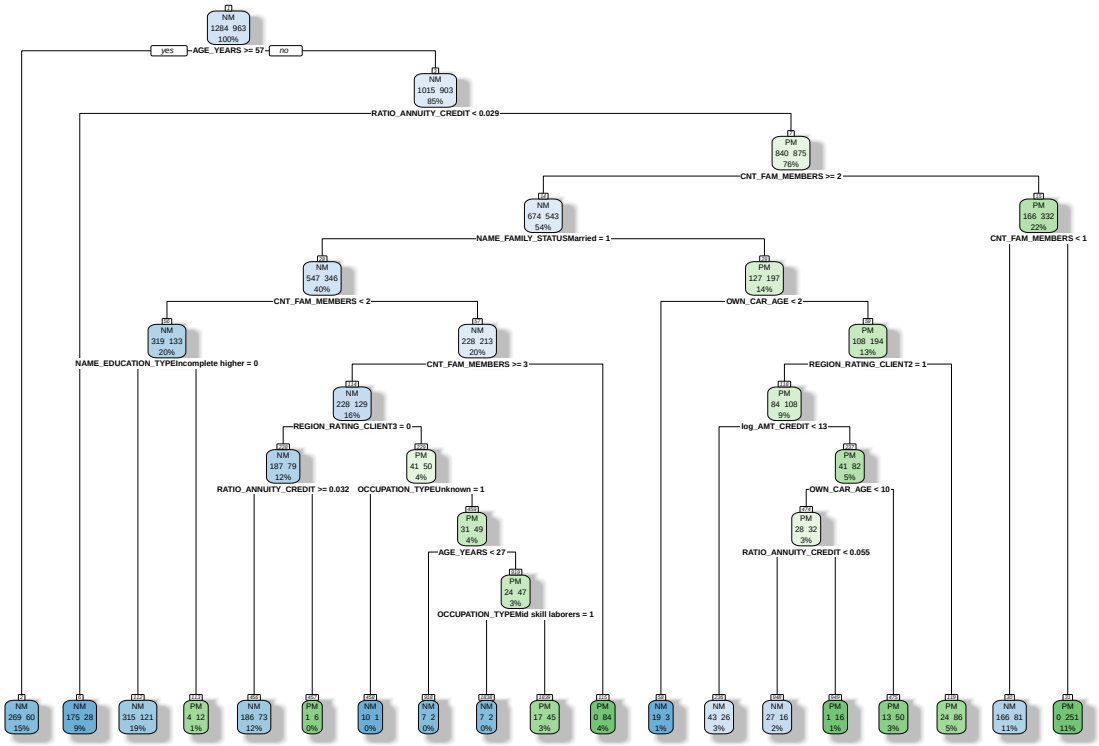


El AUC (Área Bajo la Curva) de 0.760 en la curva ROC sugiere que el modelo tiene una capacidad moderadamente buena para distinguir en la clasificación binaria entre morosos y no morosos. En otras palabras, el modelo es mejor que una clasificación aleatoria, es prometedor y sugiere que el modelo tiene un rendimiento decente en términos de discriminación. Sin embargo, hay un pequeño margen para mejorar.

Árbol de decisión

A continuación, se presenta el árbol de decisión final con el parámetro de complejidad óptimo elegido.

Figura 151: Árbol de clasificación de la variable TARGET, obtenido con la complejidad 'óptima'



El árbol de decisión generado se inicia evaluando la edad del solicitante. Si la edad es mayor o igual a 57 años, el modelo tiende a clasificar al individuo como "No Moroso" con un accuracy del 85 %. Este primer nivel de decisión sugiere que la edad es un factor determinante en la predicción de la no morosidad.

Por otro lado, dentro de la categoria de clientes más jovenes, el árbol se ramifica según el ratio anualidad/crédito (RATIO_ANNITY_CREDIT). Aquellos con un ratio inferior a 0.0295, se los clasifica como no morosos con un accuracy del 91 %, sugiriendo que clientes con cargas de anualidad más bajas en comparación con su crédito son menos propensos a tener dificultades de pago.

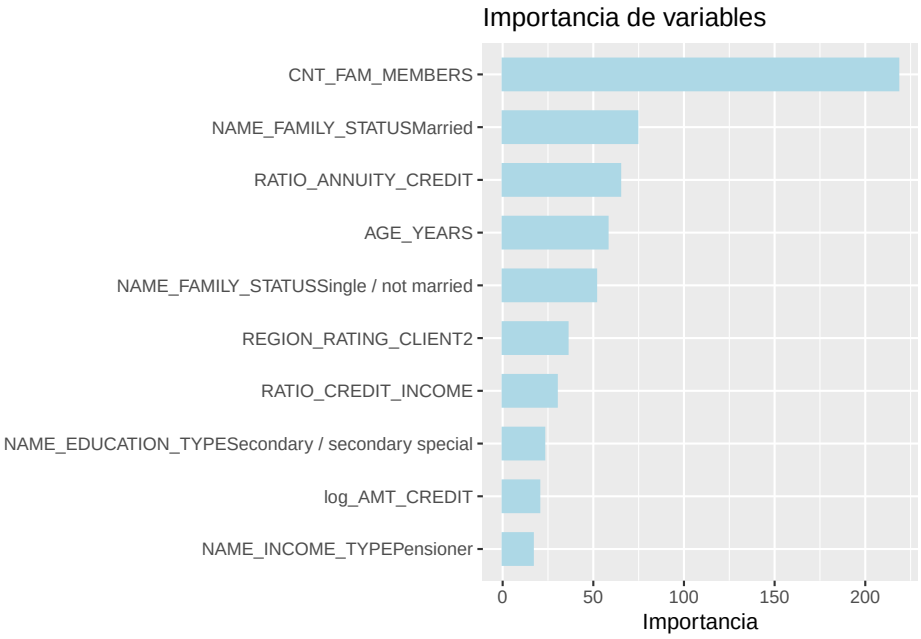
En contraste, para clientes con un ratio mayor o igual a 0.0295, factores adicionales como el estado civil, número de miembros familiares y educación influyen en la clasificación.

Clientes casados y con más de 2 miembros familiares tienden a tener una probabilidad de ser clasificados correctamente de no ser morosos (accuracy) de un 80 %. En casos específicos, como aquellos con menos de 2 miembros familiares y educación incompleta, la probabilidad ser clasificados en clientes no morosos alcanza el 88 %.

La segmentación se profundiza aún más considerando variables como la región del cliente, la ocupación y la relación anualidad/crédito. En situaciones particulares, como ocupaciones desconocidas y ratios anualidad/crédito superiores a 0.03193, la probabilidad de morosidad se incrementa significativamente (71.81 %).

Este orden de variables en el árbol está determinado por la importancia relativa de cada variable en la tarea de clasificación. Las variables que ofrecen una mayor separación entre las clases son utilizadas en los niveles iniciales del árbol.

Figura 152: Importancia de las variables en CART



La variable "CNT_FAM_MEMBERS" (Número de miembros de la familia) es la característica más influyente en la clasificación del riesgo crediticio, con una importancia relativa del 218.19 %. Esto sugiere que la composición familiar tiene un impacto significativo en la capacidad de pago.

Por otro lado, el estado civil "Married" (Casado) y la relación entre la anualidad y el crédito ("RATIO_ANNUITY_CREDIT") son también factores cruciales, con importancias del 74.26 % y 64.80 %, respectivamente. Estos indican que el estado civil y la relación entre la anualidad y el crédito desempeñan un papel fundamental en la toma de decisiones crediticias.

Además, la "Edad" ("AGE_YEARS") del solicitante es otra variable clave, con una importancia del 57.85 %. Esto refuerza la conclusión de que la edad es un factor importante en la predicción de la no morosidad.

En resumen, las variables más influyentes, como el número de miembros de la familia, estado civil, relación anualidad/crédito y edad, resaltan la importancia de aspectos fundamentales en la evaluación del riesgo. Además, factores sociodemográficos como la ubicación geográfica y la educación juegan un papel crucial en la toma de decisiones crediticias.

Conclusiones

Como la sensibilidad obtenida ha sido mucho más baja que la especificidad, concluimos que el modelo tiene más dificultades para identificar los casos positivos reales (morosos) en comparación con su habilidad para identificar correctamente los casos negativos reales (no morosos). Este resultado no nos es beneficioso en la clasificación, ya que en este contexto quizás sea mejor detectar adecuadamente casos posotivos (morosos), para así reducir el número de clientes morosos.

Así pues concluimos que el modelo CART proporciona una herramienta valiosa para la clasificación de riesgo crediticio, destacando la importancia de variables clave como los miembros de la familia, la edad y la relación entre la anualidad y el crédito. Para mejorar aún más, es recomendable explorar ajustes en la sensibilidad y considerar otras técnicas de modelado que puedan aportar mejoras específicas para el objetivo del problema.