Random Forest

El Random Forest, es un método que destaca por su capacidad para realizar predicciones precisas y robustas de conjuntos de datos. Utiliza un conjunto de árboles de decisión, cada uno entrenado en subconjuntos aleatorios de datos y utilizando subconjuntos aleatorios de características. Al combinar las predicciones de múltiples árboles, el Random Forest reduce el sobreajuste y mejora la generalización del modelo. Este enfoque de "ensamblado" hace que el algoritmo sea resistente al ruido y capaz de manejar conjuntos de datos grandes y complejos. Además, el Random Forest proporciona información sobre la importancia relativa de las características, lo que facilita la identificación de variables influyentes en el proceso de toma de decisiones.

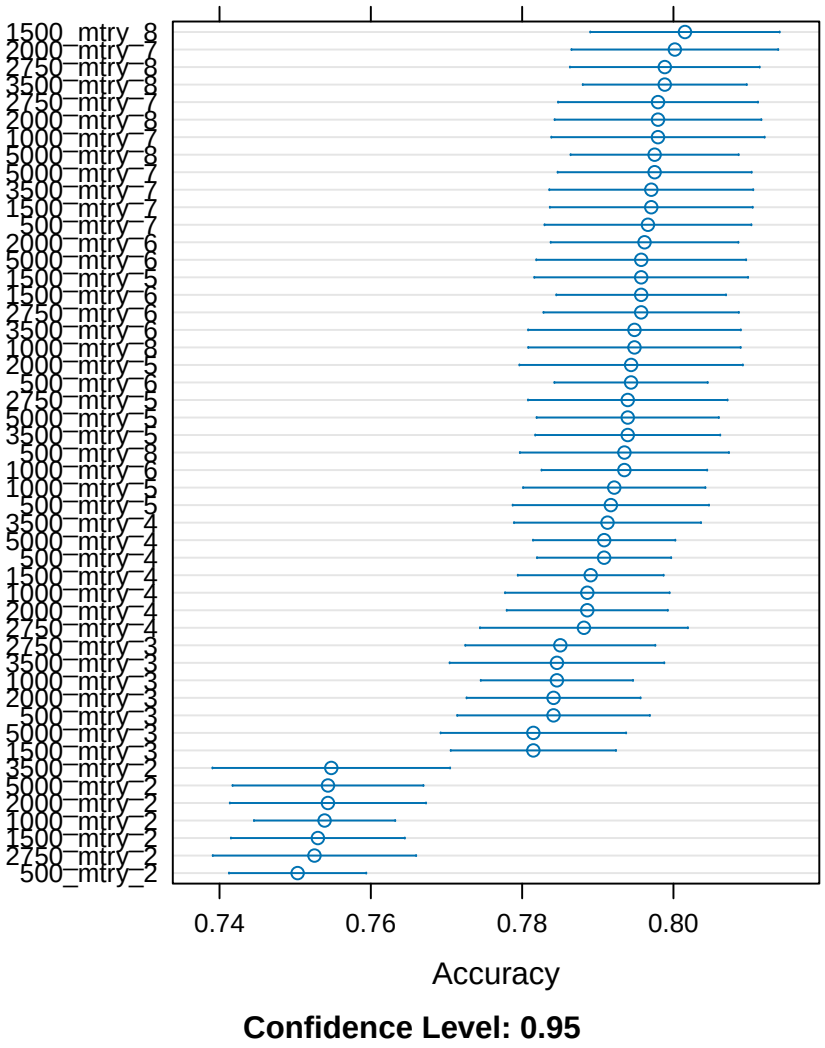
Selección de parámetros

Para llevar a cabo el Random Forest, se escogen sus dos parámetros principales:

1. **Número de árboles (`ntree`):** Este parámetro especifica la cantidad de árboles de decisión que se construirán en el bosque. Un mayor número de árboles generalmente mejora la estabilidad y la precisión del modelo, pero también aumenta el costo computacional. Sin embargo, hay un punto de rendimiento óptimo después del cual agregar más árboles puede no aportar mejoras significativas y puede conducir a un sobreajuste.
2. **Número de características seleccionadas en cada nodo (`mtry`):** Este parámetro indica cuántas características (variables) se deben considerar al hacer una división en cada nodo de un árbol. Una elección adecuada de `mtry` es crucial para el rendimiento del modelo. Valores bajos pueden llevar a la construcción de árboles más decorados y correlacionados, aumentando el riesgo de sobreajuste. Valores altos pueden hacer que los árboles sean más similares y reducir la diversidad del bosque.

Para encontrar los valores óptimos de ambos parámetros, se realizará un search grid para evaluar el accuracy del modelo para cada par de valores de los parámetros. Del parámetro ntree se probarán los valores 1000, 1500, 2000, 2750, 3500 y 5000. No se probará un número mayor por el alto coste computacional que supone. Del mtry, se prueban los valores del 2 al 8, ya que el valor propuesto por la literatura sería 4. Para entrenar el modelo para cada par de variables se realiza un 10-fold cross validation.

Figura 153: Random Forest para pares de valores de parámetros Ntree y Mtry



Se prueban los pares de variables propuestos y en este gráfico, ordenado de mayor a menor accuracy, se observa que el mayor accuracy lo presenta modelo de Random Forest con ntree=1500 y mtry=8, con valores alrededor del 80 %.

Desarrollo y validación del modelo

A partir de la 10-folds cross validation del modelo escogido, en los datos de entrenamiento se obtiene un Recall de 'r sensitivity_train' y una Especificidad de 'r specificity_train'.

A continuación se presenta la matriz de confusión del conjunto de validación en el modelo de Random Forest entrenado.

Cuadro 77: Matriz de confusión del conjunto de validación

	Realidad	
	No moroso	Potencial moroso
Predicción		
No moroso	284	84
Potencial moroso	36	156

Y a partir de las matrices de confusión de ambos conjuntos de datos train y test, se calculan las diferentes métricas de validación del modelo:

Cuadro 78: Medidas de Validación para el modelo Random Forest con conjunto test balanceado

	Train	Test
Accuracy	0.8067033	0.7857143
Sensitivity	0.8404118	0.6500000
Specificity	0.7891156	0.8875000
F1	0.8139564	0.7222222
Precision	0.6780893	0.8125000

El modelo Random Forest es más o menos exacto en las predicciones de la variable respuesta del conjunto de datos de validación, con un accuracy del ‘r MC$overall[“Accuracy”]’ indica que el modelo acierta en casi el 79 % de las predicciones. Además, se observa que el modelo no presenta excesivo sobrajuste porque el Accuracy es aproximadamente igual para ambos conjuntos de datos train y test.

Analizando los valores de la Sensibilidad y Especificidad, con un valor de Sensibilidad del 65 % en el conjunto de prueba, el modelo tiene una capacidad moderada para identificar a los clientes morosos, lo que significa que el modelo está perdiendo algunas instancias de morosidad, mientras que con una Especificidad de casi el 90 % tiene una capacidad significativamente más alta para identificar los clientes no morosos, este valor minimiza los falsos positivos y ayuda a evitar que clientes no morosos sean clasificados incorrectamente como morosos.

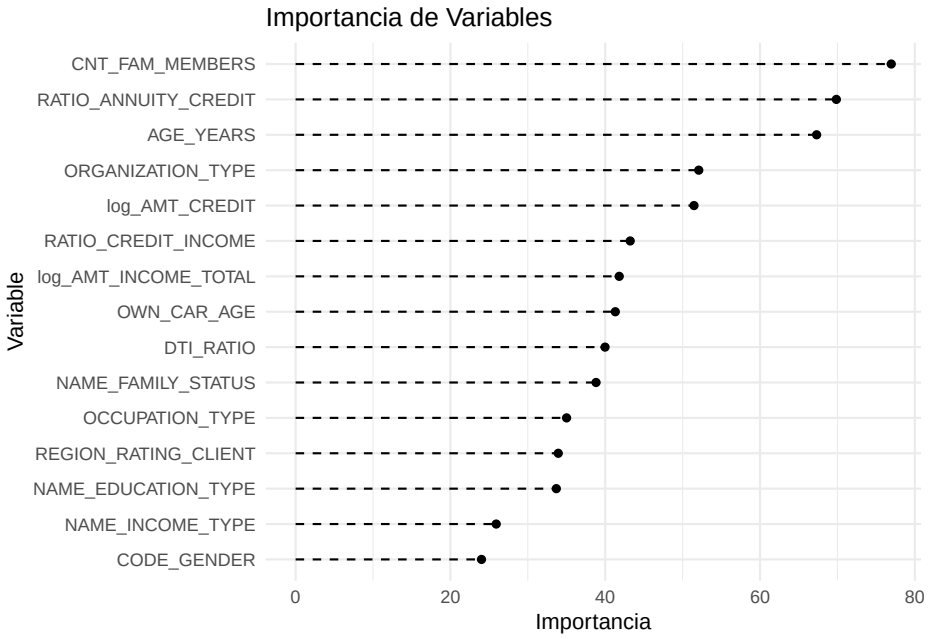
Sobre el F1, con un valor del 72.56 % en el conjunto de prueba, el F1 Score indica un equilibrio razonable entre precisión y sensibilidad. En problemas de morosidad, donde las consecuencias pueden ser significativas, es importante buscar un equilibrio entre identificar correctamente a los morosos y evitar clasificar incorrectamente a los no morosos.

Por último, analizando la Precisión, con un valor del 82.11 % en el conjunto de prueba indica que, cuando el modelo predice que un cliente es moroso, tiene una alta probabilidad de que sea correcto. Sin embargo, se debe considerar en conjunto con la sensibilidad para no pasar por alto clientes morosos.

Variables importantes

La importancia de las variables en Random Forest se basa en cuánto contribuyen al aumento de la homogeneidad de las clases cuando se utilizan para hacer divisiones en los árboles de decisión del bosque.

Figura 154: Importancia de las variables en Random Forest



Se consideraran las variables más importantes las primeras cuatro que se presentan en el gráfico. Se observa que la Ratio entre la anuidad del préstamo y el crédito total solicitado, el DTI (Debt-to-income) ratio (que mide la capacidad del cliente para pagar la annuity de su préstamo en relación con sus ingresos), la Ratio entre el crédito pedido y el salario anual del prestatario (también se puede contar como el número de años que se tarda en devolver el crédito) y el tipo de organización en la que trabajan, son las cuatro variables más importantes, y que por tanto influyen más en la predicción de morosidad de un cliente en el modelo de Random Forest.

Prueba ácida

Cuadro 79: Matriz de confusión del conjunto de validación desbalanceado

	Realidad	
	No moroso	Potencial moroso
Predicción		
No moroso	819	13
Potencial moroso	100	67

Cuadro 80: Medidas de Validación con el conjunto test desbalanceado para el modelo Random Forest

	Train	Test
Accuracy	0.8067033	0.8868869
Sensitivity	0.8404118	0.8375000
Specificity	0.7891156	0.8911861
F1	0.8139564	0.5425101
Precision	0.6780893	0.4011976

En el conjunto de entrenamiento, el modelo acierta alrededor del 80.67 % de las predicciones, mientras que en el conjunto de prueba, alcanza un 88.69 %. El accuracy es una métrica general que puede ser engañosa en conjuntos de datos desbalanceados, como en el caso de este conjunto de datos de validación, ya que puede estar dominada por la clase mayoritaria. La sensibilidad mide la proporción de casos positivos que el modelo identifica correctamente. En el conjunto de entrenamiento, el modelo identifica correctamente alrededor del 84.04 % de los casos positivos, mientras que en el conjunto de prueba, la sensibilidad es del 83.75 %, lo que indica que el modelo está capturando bien los casos positivos. El modelo también presenta una alta Especificidad para el conjunto de validación desbalanceado, con una tasa del 89,11 %, indicando una buena capacidad para predecir casos negativos.

En el conjunto de prueba, el F1 Score es del 54.25 %. Aunque esta puntuación es más baja que en el conjunto de entrenamiento, sigue siendo una métrica relevante para evaluar el rendimiento del modelo en un conjunto de datos desbalanceado. Un F1 Score más bajo en el conjunto de prueba sugiere que el modelo tiene dificultades para equilibrar precisión y recall en un entorno desbalanceado, y puede requerir ajustes para mejorar su rendimiento en la clasificación de la clase minoritaria. La precisión en el conjunto de prueba es del 40.12 %, lo que significa que el modelo tiene una proporción más baja de verdaderos positivos entre todas las instancias clasificadas como positivas, situación esperada en conjuntos desbalanceados.

En conclusión, el modelo Random Forest exhibe una sensibilidad excepcionalmente alta, indicando su habilidad sobresaliente para identificar correctamente los casos positivos. Esta capacidad para minimizar los falsos negativos, es decir, la tendencia del modelo a perder muy pocos casos positivos reales, es extremadamente valiosa, especialmente en el contexto financiero, donde la identificación precisa de la clase positiva es esencial. La baja incidencia de falsos negativos indica que nuestro modelo tiende a errar en el lado de la precaución, asegurándose de no perder casos positivos importantes. Sin embargo, es importante tener en cuenta que la alta sensibilidad va acompañada de un aumento en los falsos positivos, es decir, este modelo tiende a clasificar más instancias como positivas de lo necesario.