Universidad ORT Uruguay

Diploma de Especialización en Analítica de Negocios

Obligatorio

Entregado como requisito para la aprobación del curso de Machine Learning Supervisado.

Andrés Bähre - 261046

Kary Francia - 269671

Guillermo Vázquez – 268211

Profesor: Guillermo Magnou

07/12/2022

ÍNDICE

Obligatorio: Proyecto Corporativo

1. Descripción del problema de negocio:

El problema de negocio se centrará en una empresa que tiene una cartera de créditos para ofrecer y que le preocupa minimizar el riesgo de no otorgar un crédito a un buen cliente, para poder ser una financiera más rentable y crecer.

2. Objetivos:

Objetivo analítico: Encontrar un modelo de machine learning supervisado, comparando los algoritmos de clasificación en base a su bondad de ajuste, para predecir si los clientes de una cartera de créditos incurrirán en Default en base a su historial crediticio y otros parámetros.

Objetivo de negocio: Conocer las condiciones para otorgar un crédito a un cliente y mitigar el riesgo de no considerar prestatarios (clientes) que sean buenos pagadores.

3. Descripción de la base de datos:

La base de datos cuenta con información de 20,185 clientes. Esta información se distribuye en 19 variables (1 variable de respuesta categórica binomial y 18 variables independientes o predictores).

De la variable de respuesta, se tiene que 7,211 observaciones (35.7%) son Default y 12,974 observaciones (65.3%) son No Default. El objetivo del análisis es clasificar a los clientes de acuerdo a su probabilidad de incumplimiento (Default). Esta variable al ser categórica y binomial tomará los valores de 0 si es No Default (el cliente pagó la parte correspondiente de su crédito) y 1 si es el cliente incurrió en Default (el cliente no pagó la parte correspondiente de su crédito).

De las variables independientes, se tiene que 12 son variables categóricas y 8 son variables numéricas.

Luego, se observaron descriptivamente y cualitativamente las observaciones para detectar presencia de valores perdidos, existencia de outliers e inconsistencia entre el tipo de variable (cuantitativas o cualitativas). En este caso, no se detectaron valores perdidos, se detectó la presencia de outliers y también se corroboró que las variables se condigan con su naturaleza.

CNT_CHILDREN	AMT_INCOME_TOTAL	YEAR_BIRTH	YEAR_EMPLOYED	CNT_FAM_MEMBERS	ANT_CLI_MONTH
Min. : 0.000	Min. : 27000	Min. :22.00	Min. : 0.000	Min. : 1.000	Min. : 0.00
1st Qu.: 0.000	1st Qu.: 121500	1st Qu.:35.00	1st Qu.: 2.000	1st Qu.: 2.000	1st Qu.: 9.00
Median : 0.000	Median : 157500	Median :44.00	Median : 5.000	Median : 2.000	Median :19.00
Mean : 0.416	Mean : 186120	Mean :44.84	Mean : 6.567	Mean : 2.183	Mean :22.66
3rd Qu.: 1.000	3rd Qu.: 225000	3rd Qu.:54.00	3rd Qu.: 9.000	3rd Qu.: 3.000	3rd Qu.:35.00
Max. :19.000	Max. :1575000	Max. :70.00	Max. :44.000	Max. :20.000	Max. :60.00

De acuerdo al análisis exploratorio de los datos, se observa que no hay valores perdidos. Sin embargo, la variable OCCUPATION_TYPE tiene datos nulos que representan el 31.1% de los datos totales.

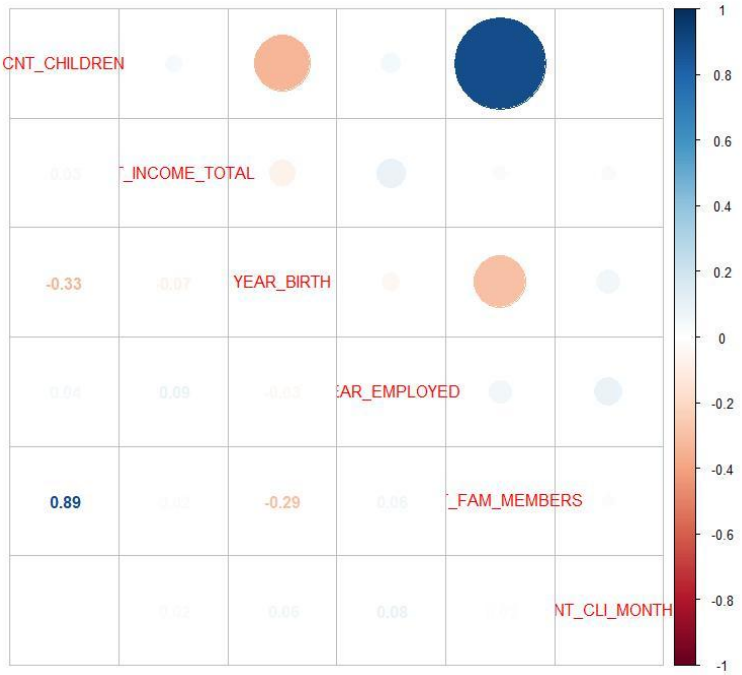
Por otro lado, las siguientes variables no presentan una adecuada distribución de sus datos:

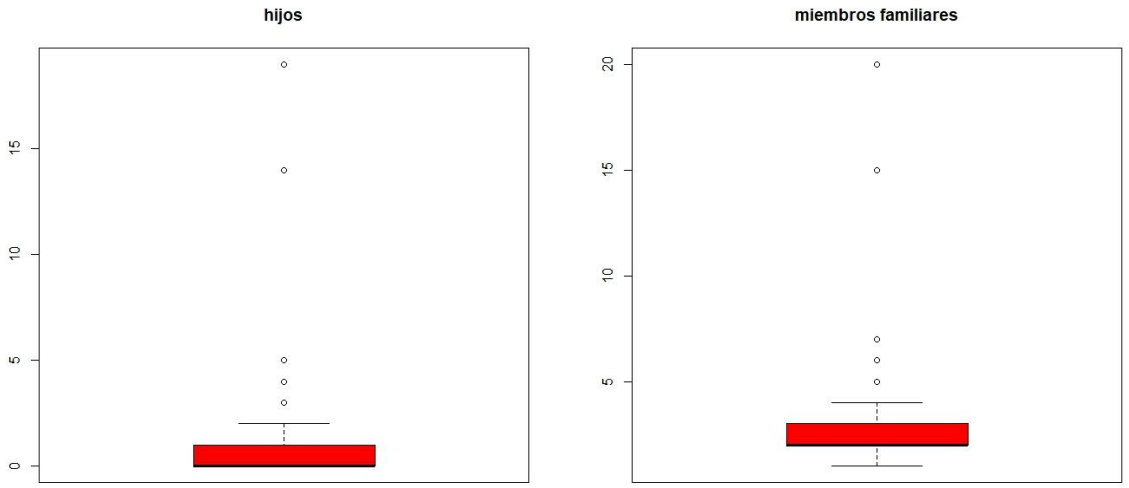
- NAME_INCOME_TYPE: student = 9 observaciones
- NAME_EDUCATION_TYPE: Academic degree = 19 observaciones
- NAME_HOUSING_TYPE: co-op apartment =77 observaciones
- OCCUPATION_TYPE: SD=6280, HR staff=46, IT staff=41, Realty agents=39, secretaries=78, waiters/barme staff =70.

Cabe resaltar que, la variable tipo de ocupación no afecta al análisis porque presenta una gran cantidad de datos nulos. Asimismo, no es relevante para el análisis si el cliente tiene correo electrónico, teléfono propio o en el trabajo. Por lo que se procederá a retirar las siguientes variables para el análisis:

- FLAG_WORK_PHONE
- FLAG_PHONE
- FLAG_EMAIL
- OCCUPATION_TYPE

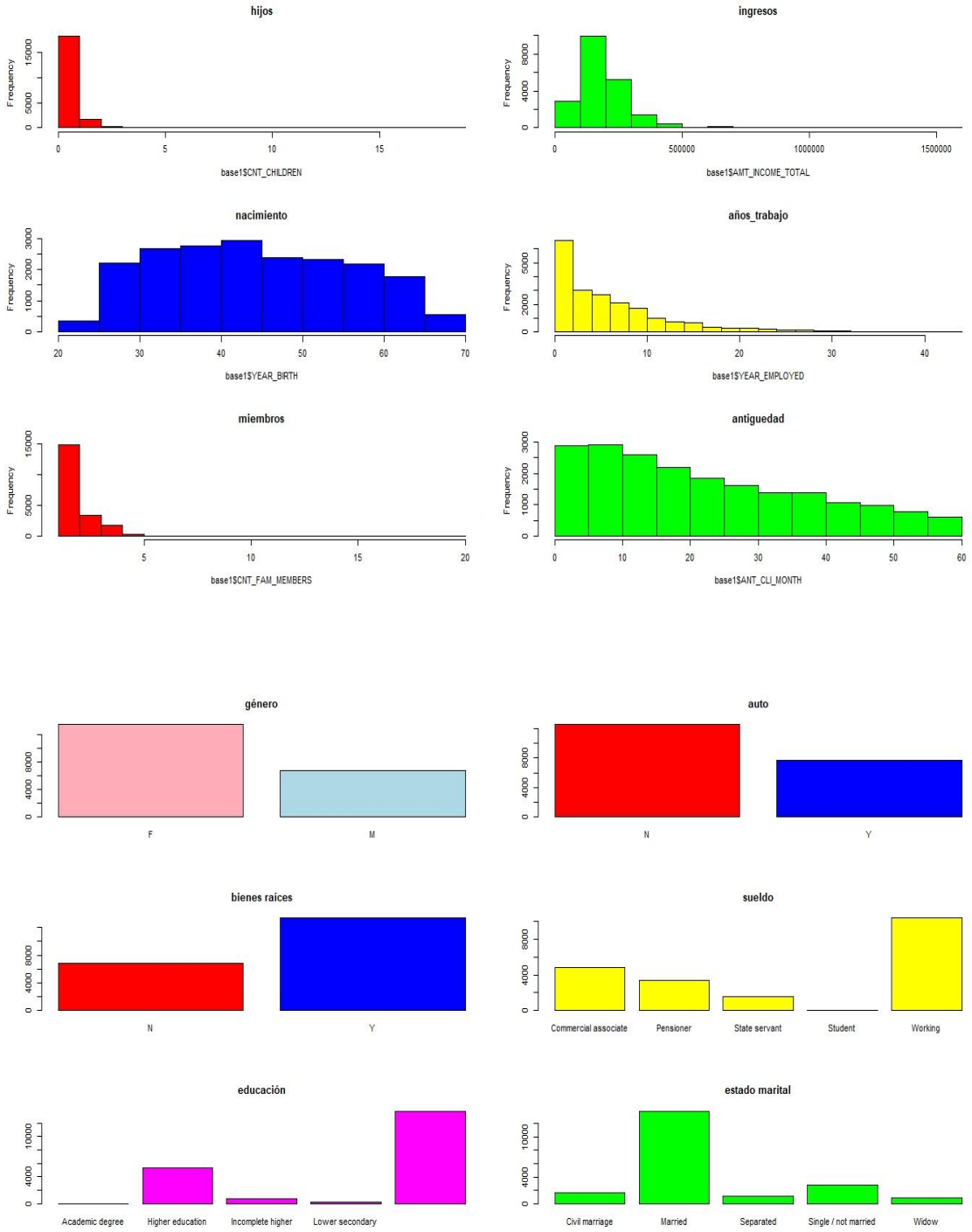
Por último, se va a utilizar como primer modelo el algoritmo de regresión logística, por ello se analiza la matriz de correlación entre las variables independientes, pudiendo observarse que presentan poca o nula correlación, por ende no hay multicolinealidad entre las variables. A excepción de las variables CNT_CHILDREN y CNT_FAM_MEMBERS, que se entienden están directamente correlacionadas, debido a que la cantidad de hijos está incluida de alguna manera en la cantidad de miembros de la familia, es decir a mayor cantidad de hijos también mayor cantidad de miembros de la familia.





Sin embargo, a pesar de lo antes mencionado, se decidió utilizar todas las variables cuantitativas para construir los modelos.

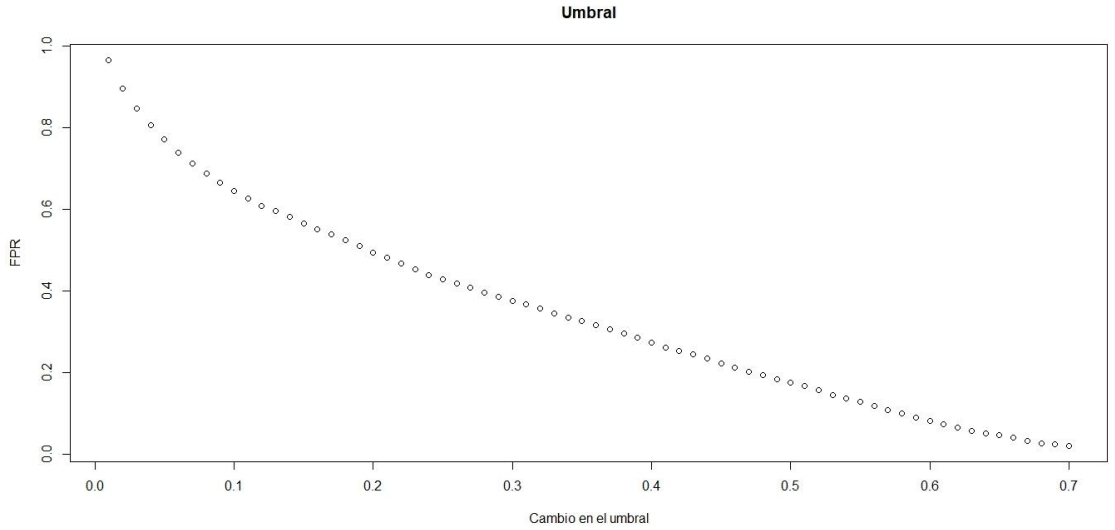
Resumiendo las variables en un histograma se obtiene lo siguiente:



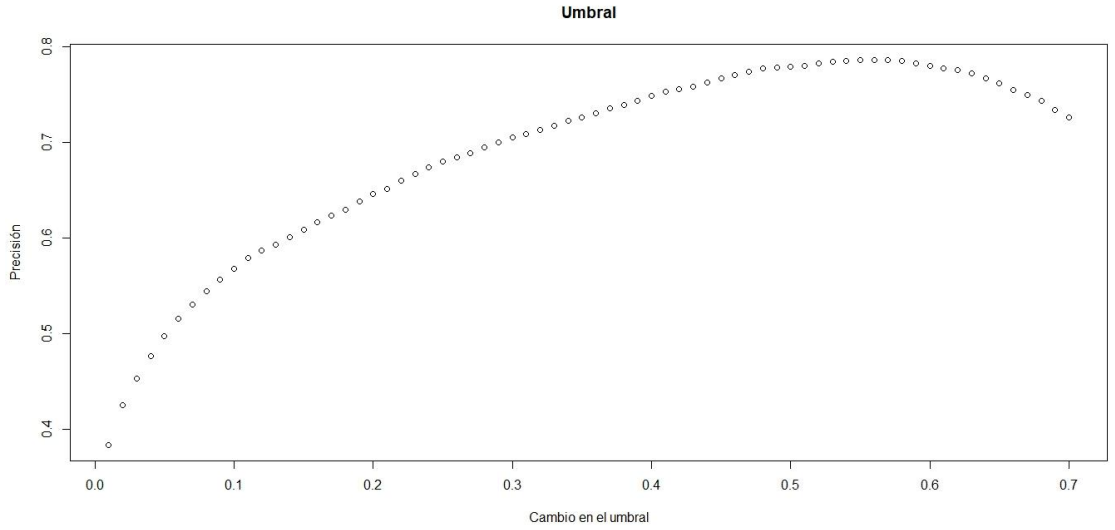
Luego de que se realizó el análisis de los datos y seleccionar las variables para construir los modelos de predicción, se dividió la base en train (70%) y test (30%) - Cross validation simple, a fin de entrenar y validar la calidad de clasificación de cada modelo.

4. Resultados con regresión logística:

Para determinar la matriz de confusión en base a la regresión logística, es importante determinar el umbral que se le va a asignar al modelo para determinar si es default o no default. Evaluando de 0.1 a 0.7 como umbral para la regresión, se tiene el siguiente gráfico para FPR:



Por otro lado, al evaluar la precisión en función del umbral, se obtiene:



Para lograr una precisión óptima se llega con 0,55 de punto de corte. Por otro lado, es importante considerar una proporción de clientes Default y no Default similar a toda la base en la matriz de confusión de train, por lo que se evalúa la distribución en el conjunto de entrenamiento:

Modelo Logit	No Default	Default
Base (20.185)	12,974 (64%)	7,211 (36%)
Test (Punto corte 0,5)	3,927 (65%)	2,129 (35%)

Test (Punto corte 0,54)	4,160 (69%)	1,896 (31%)
Test (Punto corte 0,55)	4,224 (70%)	1,832 (30%)
Test (Punto corte 0,56)	4,297 (71%)	1,759 (29%)

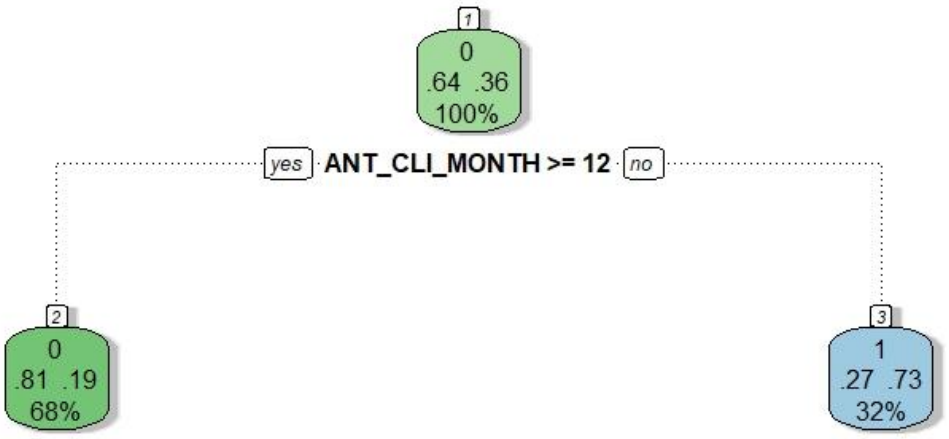
La matriz de confusión de train y test para el punto de corte de 0.55 quedan las clasificaciones para cada una de las bases de la siguiente manera:

Modelo Logit: Train	Real: No Default	Real: Default
Modelo: No Default	7,873	1,882
Modelo: Default	1,146	3,228

Modelo Logit: Test	Real: No Default	Real: Default
Modelo: No Default	3,454	770
Modelo: Default	501	1,331

5. Resultados con CART:

Tanto el árbol original como el segundo modelo de árbol, condicionado para que tenga una partición diferente, generado de la ganancia de información del índice 'information' (también se probó con 'gini'), ambos dan el mismo resultado: El 68% de las observaciones presenta un ratio utilizado de la antigüedad de los clientes mayor o igual a 12 meses, siendo clasificados como buenos clientes (No Default), y el 32% corresponde a los malos clientes (Default). Por lo tanto, se puede ir observando la precisión de estos modelos con esta predicción.



Con respecto a la evaluación de la importancia de las variables, ambos árboles tienen en consideración como variable importante a la antigüedad del cliente con una gran relevancia. El resto de las variables están muy por debajo en cuanto a relevancia. Si bien aparecen valores diferentes entre ambos árboles, al ser dicha variable tan relevante, opaca las demás y concluyen en modelos similares.

La matriz de confusión de train y test quedan de la siguiente manera:

Árbol: Train	Real: No Default	Real: Default
Modelo: No Default	7,831	1,801
Modelo: Default	1,223	3,274

Árbol: Test	Real: No Default	Real: Default
Modelo: No Default	3,416	775
Modelo: Default	504	1,361

6. Análisis con Random Forest:

Este modelo se hizo aplicando 5,000 árboles de decisión y 4 nodos en cada árbol como resultado de la optimización de los hiperparámetros. Además, mtry=4 coincide con la raíz cuadrada de las 14 variables explicativas.

Debido a la desventaja del modelo, no se puede realizar una clara interpretación de las reglas de decisión ya que se combinan los resultados de múltiples árboles paralelamente. Sin embargo, se observa que la variable más importante sigue siendo la antigüedad del cliente en este modelo.

La matriz de confusión de train y test quedan de la siguiente manera:

RF: Train	Real: No Default	Real: Default
Modelo: No Default	8,974	139
Modelo: Default	80	4,936

Modelo RF: Test	Real: No Default	Real: Default
Modelo: No Default	3,490	746
Modelo: Default	430	1,390

Al evaluar indicadores tan altos de sensibilidad y especificidad, se considera la posibilidad de encontrar overfitting en el modelo,en parte por la complejidad del modelo y la variable de entrada de antigüedad de clientes la cual tiene un gran peso sobre las otras, pudiendo influir a

que se esté generando el sobreajuste, incluso se vio que en CART no daba la posibilidad de realizar particiones diferentes, por lo que el algoritmo podría estar trabajando con muchos árboles similares entre sí, demostrando la inestabilidad del modelo, el cual se considera podría ser descartado como un modelo para resolver el problema del negocio.

7. Análisis con Boosting:

La variable más importante también se mantiene como la antigüedad del cliente en este modelo. Para este modelo se utilizó una tasa de aprendizaje de 0.1, con 6 nodos de profundidad y 100 interacciones. Al igual que con el random forest se pierde interpretación de las reglas.

La matriz de confusión de train y test quedan de la siguiente manera:

Boosting: Train	Real: No Default	Real: Default
Modelo: No Default	7,996	1,669
Modelo: Default	1,058	3,406

Boosting: Test	Real: No Default	Real: Default
Modelo: No Default	3,451	773
Modelo: Default	469	1,363

8. Comparación de los modelos y selección del modelo predictivo:

El enfoque para comparar los modelos se ajustará en poder identificar correctamente a los clientes Default, considerando como la métrica principal a la Tasa de Falsos Positivos ,debido a que este describe la probabilidad de clasificar a un cliente como Default siendo este un buen cliente. A su vez, se cuidará de que la proporción final de la clasificación de los datos (Default y No Default) se mantenga lo más cercano posible a la proporción de los datos de la base original, es decir se observará que el modelo rechace cerca de un 35.7% a los clientes que serían Default, tomando en cuenta también que la precisión del modelo sea el mejor.

Entonces, con la finalidad de realizar la comparación de los modelos y para mejorar la especificidad se decidió subir el umbral por encima del 0.50. A continuación, se presentan las siguientes tablas con la información obtenida de las matrices de confusión tanto con el punto de corte 0.56 como con 0.55:

- *Punto de corte 0.56:*

Se observa que para la base de entrenamiento el modelo que presenta un mejor rendimiento corresponde a Random Forest, debido a que la Tasa de Falsos Positivos es la menor con 0.007 y que el modelo con mayor error es el de los árboles de decisión. Sin embargo, para evaluar el correcto desempeño de los modelos se analizarán las métricas obtenidas en test y su variación con respecto al conjunto de train.

En Train

Modelo	Sensitividad	Especificidad	Precision	AUC	F1	FPR
Reg. Logística completa	0.615	0.883	0.786	0.749	0.725	0.117
Reg. Logística simplificada	0.614	0.884	0.786	0.749	0.724	0.116
Árboles de decisión	0.645	0.865	0.785	0.755	0.739	0.135
Random Forest	0.979	0.993	0.988	0.986	0.986	0.007
Boosting	0.670	0.883	0.806	0.777	0.762	0.117

En Test

Modelo	Sensitividad	Especificidad	Precision	AUC	F1	FPR
Reg. Logística completa	0.614	0.881	0.789	0.748	0.724	0.119
Reg. Logística simplificada	0.612	0.882	0.788	0.748	0.723	0.118
Árboles de decisión	0.637	0.871	0.788	0.754	0.736	0.129
Random Forest	0.647	0.888	0.803	0.768	0.749	0.112
Boosting	0.635	0.883	0.796	0.759	0.739	0.117

En la tabla de test, se observa que el error mayor se obtiene en los árboles de decisión a pesar de que disminuye en -0.006 con respecto al train. Si bien el error en Random Forest es el menor, la variación con train es grande para todas las medidas de bondad de ajuste pudiendo incurrir el modelo en overfitting. En cuanto a los modelos de regresión logística completa, simplificada y boosting, presentan métricas similares en su bondad de ajuste, por lo que se definirá al mejor modelo entre los tres aquel que tenga una variación que mejore la precisión del modelo, manteniendo al mismo tiempo una buena especificidad, para este caso es el modelo boosting. Sin embargo, se entiende que este es un modelo secuencial y tiende al overfitting, teniendo en cuenta ello, se decidió analizar a los modelos con el punto de corte 0.55 y ver si mejoraba la predicción antes de tomar una decisión.

Variación porcentual

Modelo	Sensitividad	Especificidad	Precision	AUC	F1	FPR	Proporción Default
Reg. Logística completa	-0.001	-0.002	**0.003**	-0.001	-0.001	0.002	**29.0%**
Reg. Logística simplificada	-0.002	-0.002	0.002	-0.001	-0.001	0.002	**29.0%**
Árboles de decisión	-0.008	0.006	0.003	-0.001	-0.003	-0.006	30.8%
Random Forest	-0.332	-0.105	-0.185	-0.218	-0.237	0.105	30.1%
Boosting	-0.035	**0.000**	**-0.010**	-0.018	-0.023	**0.000**	**29.9%**

- *Punto de corte 0.55:*

En Train

Modelo	Sensitividad	Especificidad	Precision	AUC	F1	FPR
Reg. Logística completa	0.632	0.873	0.786	0.752	0.733	0.127
Reg. Logística simplificada	0.631	0.875	0.787	0.753	0.733	0.125
Árboles de decisión	0.645	0.865	0.785	0.755	0.739	0.135
Random Forest	0.979	0.993	0.988	0.986	0.986	0.007
Boosting	0.670	0.883	0.806	0.777	0.762	0.117

En Test

Modelo	Sensitividad	Especificidad	Precision	AUC	F1	FPR
Reg. Logística completa	0.633	0.873	0.790	0.753	0.734	0.127
Reg. Logística simplificada	0.632	0.872	0.789	0.752	0.733	0.128
Árboles de decisión	0.637	0.871	0.788	0.754	0.736	0.129
Random Forest	0.647	0.887	0.803	0.767	0.748	0.113
Boosting	0.635	0.883	0.796	0.759	0.739	0.117

Se observa que para el punto de corte 0.55, el análisis se centra en comparar los modelos de regresión logística con el Boosting. Los modelos logísticos tienen un mejor desempeño en comparación al punto de corte 0.56, siendo un mejor modelo que boosting debido a que discrimina mejor a los Default con una proporción de 30.3%. Con respecto a la variación de las métricas de bondad de ajuste, se destaca que el modelo logístico completo no aumenta el error en test a diferencia del simplificado que aumenta el error en 0.003. Además, la precisión aumenta en 0.004 y el área bajo la curva en 0.001, mejorando así el modelo su capacidad de discriminación, a diferencia del boosting que la disminuye. Asimismo, el modelo logístico completo mantiene una buena especificidad por lo que se seleccionará este modelo para predecir a los Default.

Variación porcentual

Modelo	Sensitividad	Especificidad	Precision	AUC	F1	FPR	Proporción Default
Reg. Logística completa	0.001	**0.000**	**0.004**	**0.001**	0.001	**0.000**	**30.3%**
Reg. Logística simplificada	0.001	-0.003	**0.002**	-0.001	**0.000**	0.003	30.3%
Árboles de decisión	-0.008	0.006	0.003	-0.001	-0.003	-0.006	30.8%
Random Forest	-0.332	-0.106	-0.185	-0.219	-0.238	0.106	30.1%
Boosting	-0.035	**0.000**	-0.01	-0.018	-0.023	**0.000**	29.9%

9. Conclusiones y recomendaciones:

Luego del análisis realizado se concluye que los mejores modelos para predecir los clientes Default son Boosting y la regresión logística. La elección del mejor modelo se enfocó desde la perspectiva del negocio, considerando que la financiera podría incurrir en 2 posibles errores: darle un crédito a clientes que luego no pagarán el crédito, o bien no dar un crédito a clientes que sí pagarían.

Se decidió ir por la Tasa de Falsos Positivos , debido a que el caso de estudio presenta un 65.3% de clientes No Default y pocas observaciones de clientes Default, siendo esta base poco balanceada al igual que en la realidad de muchas empresas inmersas en el otorgamiento de créditos. Asimismo, suponemos que no convenía que la financiera deje de prestarle a alguien que fuese clasificado en el modelo como Default siendo que podría ser un buen cliente, por que a nivel de negocio puede existir una pérdida asociada al volumen de crédito entregado y una pérdida de ganancia por los intereses respectivamente. En esta situación que se planteó este equipo de trabajo, se concluyó que para el negocio es más crítico no entregar créditos a clientes que sí pagarán.

En este sentido se recomienda que el modelo mejore la especificidad para reducir la cantidad de clientes clasificados como falsos positivos, qué es lo mismo que buscar un FPR mínimo, sin dejar de lado una precisión óptima del modelo.

Por lo tanto, evaluando la especificidad, y considerando un punto de corte de 0,55, se obtiene que el modelo logístico es el que mejor se ajusta al objetivo del negocio, logrando un nivel de precisión casi del 79% y logrando un FPR de 11,9%. En líneas generales, el modelo no presentó un overfitting en el conjunto de test, siendo la variación de sus métricas de bondad de ajuste mejor que el promedio obtenido por los otros modelos. Además, la ventaja que tiene este modelo es que es conocido y utilizado frecuentemente por la industria financiera por lo que se cree será fácil de implementar.