



Clasificación de deudores con datos del BCRA

Estadística Computacional - Maestría en Estadística Aplicada

Nicolás Gottig

Junio - 2024

Motivación

El Banco Central de la República Argentina (BCRA) publica mensualmente un informe consolidado de deudas actuales e históricas (24 meses), denominado ‘Central de Deudores del Sistema Financiero’ elaborado en función de los datos recibidos de distintos tipos de entidades financieras. Cada deuda informada al BCRA es acompañada de su situación que es una aproximación a la cantidad de días de atraso en el cumplimiento de pago (situación 1, 2, 3, etc..). Siendo sit. 1 y 2 situación normal y riesgo bajo, 3 riesgo medio, 4 alto, 5 irrecuperable y 6 irrecuperable por disposición técnica.

Se cuenta con una muestra aleatoria de cuits de personas físicas que tenían al menos una de deuda en el sistema financiero en junio de 2019 y sus características crediticias entre dic-2018 y jun-2019, su estado de deuda en jun-2019 y si el cuit está en situación 3 o superior en todas las deudas del cuit entre Jul-2019 y Jun-2020, en ese caso se lo clasifica como caso de default.

En el informe se describen las características generales del conjunto de datos, además se ajustan y comparan algoritmos basados en árboles CART y CHAID, luego se los compara con regresión logística stepwise como método de clasificación tradicional, y con modelos basados en Bagging y Boosting, concretamente el algoritmo Random Forest, el algoritmo XGBoost y el algoritmo LightGBM, itinerando a través de sus parámetros.

Por último, se describen las medidas de clasificación de los modelos, seleccionando uno y optimizando el punto de clasificación a través de una función de costo.

Análisis descriptivo

Algunas consideraciones:

- Se reagruparon los cuits que empiezan con “23” o “24” en una única categoría “global” para facilitar la interpretación en el análisis descriptivo y porque la categoría “24” tiene sólo 90 casos.

- Sólo hay valores nulos en la maxima situación de deudas garantizadas y no garantizadas. En las garantizadas hay muchos más valores nulos porque sólo el 1.9 % del total de cuits tiene garantías. Se decide eliminar la variable ‘max_sit_mes_con_garantía’ que tiene el 97.3 % de los casos nulos y la variable ‘max_sit_mes_sin_garantía’ dado que algunos modelos tuvieron problemas de convergencia, y es preferible a eliminar los casos pudiendo sesgar los parámetros. Los modelos incluyeron el ajuste con y sin la variable ‘max_sit_mes_sin_garantía’, siendo su aporte nada notable en las métricas de interés.
- Hay 2 personas extranjeras que no se consideran en el análisis (DNI 60 millones) por su escasa participación relativa en el conjunto de datos.
- Se eliminan los registros con igual fecha que la variable respuesta: *peor_situacion_respuesta* y *mora_mayor_30_dias*.
- Si se elimina la columna ‘id_individuo’ hay 13 casos repetidos que son eliminados del análisis.
- Se crean variables auxiliares para identificar si el cuit tuvo alguna vez default histórico (dic-2018 a jun-2019): *tiene_default_histórico* que toma valor 1 si lo tuvo y 0 si no lo tuvo. También se crea la variable *‘tiene_garantia_histórica’* que toma valor 1 si la persona tuvo una proporción mayor a 0 de deuda garantizada histórica y 0 si no lo tuvo.
- Se considerarán las variables expresadas en pesos para el análisis descriptivo pero se eliminarán en los modelos, conservando solo la proporción media en nivel 1, 2, etc... sobre el total. Dado que los montos en pesos son demasiado variantes en el tiempo, por lo que no son buenos predictores para futuros casos de mora.
- Para el ajuste de los modelos también se eliminan las variables registradas el mes inmediato anterior al default, con el objetivo de tener información histórica y que el estado inmediato anterior a la observación de default no condicione la regla de decisión. Las variables eliminadas son *n_deudas_actual*, *situacion_mes_actual*, *tiene_garantia_actual*, *mora_30_dias_mes_actual*

Características generales de los CUIT

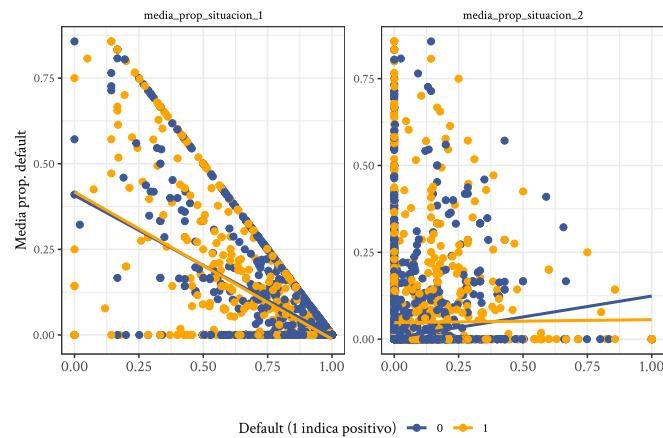


Figura 1: Relación entre default, prop. de default, situación 1 y 2.

Del total de 18006 casos, el 8.7 % estuvo situación crediticia más grave mayor o igual 3 en todas las deudas (default). Sólo el 1.9 % tuvo deuda garantizada a jun-2019 y el 4.2 % tuvo alguna proporción de deuda más grave mayor o igual a 3 entre Dic-2018 y Jun-2019.

Se observan correlaciones esperadas entre las variables. Por ejemplo, para los casos que en algún momento estuvieron en default, la proporción de deuda en default está fuertemente correlacionada (linealmente) de forma inversa con

la proporción de deuda en etapa 1, aunque está muy débilmente correlacionada de forma directa con la proporción de deuda en etapa 2 (figura 1).

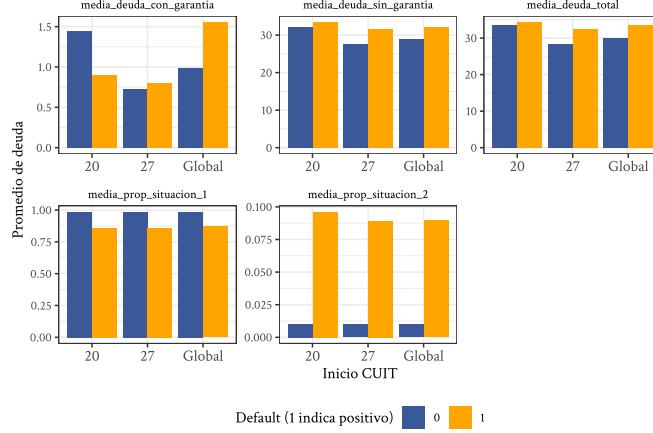


Figura 2: Casos de default, montos de deuda y situaciones generales

Las relaciones entre los casos de default y los créditos seguros son distintas respecto a los montos solicitados y la participación relativa en cada situación. Aunque no se observan, a priori, importantes clasificadores en las características observadas en cada tipo de cuit (excepto la media de la proporción de la situación 2 sobre el total de créditos). Los promedios de deuda desagregados por tipo de persona y situación de falta (figura 2), indican que los casos que cayeron en default tienden a endeudarse por montos mayores.

La deuda con garantía en general es menor, representando en promedio alrededor del 5 % del total de la deuda. Sin embargo, la tendencia es igual; excepto en los cuit iniciados en 20, cuya media garantizada es notablemente mayor en aquellos casos que no cayeron en situación de falta. En otras palabras: los cuit iniciados en 20 están asociados a mayores montos bajo garantía.

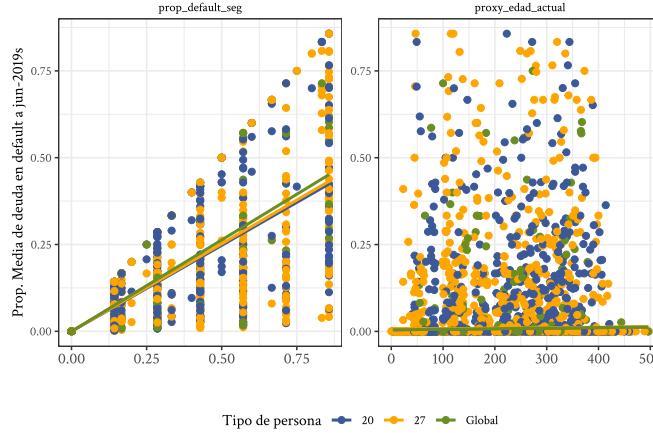


Figura 3: Relación entre características de default y tipo de persona

Por otro lado, no se observan tendencias notables respecto a las características etarias o de género respecto a la proporción de deuda en default histórico, o el default en el periodo de interés. Si es más notable la relación entre la proporción de deuda en default histórico (a junio del 2019) y la proporción de meses en los cuales el cuit estuvo en default. Sugiriendo, intuitivamente, que los malos historiales crediticios se sostienen en el tiempo (figura 3).

pecto a las características de los CUIT que cayeron en default, se observa en la figura 4 que el algoritmo CART identifica a la proporción de deuda en situación 1 como el mejor clasificador.

Si la proporción de la media de deuda en la situación uno es mayor a 0.97, entonces se clasifica al cuit como cumplidor, representando el 89 % de los casos y una probabilidad de default del 0.05. Ahora bien, si no se cumple esta condición, la situación del mes actual es otro clasificador de falta. Si la situación fuese distinta a uno, la probabilidad de default es 0.62 (¡12 veces más grande!).

Si la situación en el mes actual efectivamente fuese situación uno, entonces dependería de la media de deuda en default. Si ésta es mayor a 8.2 mil, entonces la probabilidad de default es 0.55, 11 veces más grandes que si la proporción de la media de deuda en situación 1 es mayor a 0.97 y un 11 % menos que si la situación del mes actual fuese distinta a uno.

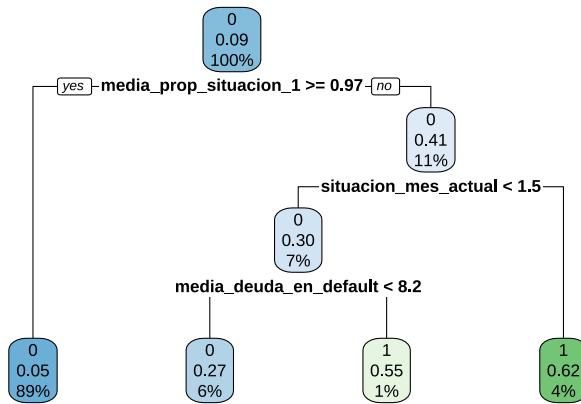


Figura 4: Criterios de clasificación para casos de default

Comparación de modelos CART y CHAID

Se eliminan todas las variables medidas en miles de pesos (dadas las características económicas del país y su rápido efecto en la medida de los créditos obtenidos en el tiempo) y las variables observadas el mes anterior al default. Se itera sobre los parámetros cp que determina la complejidad del modelo en el caso del algoritmo CART y sobre valores de $alpha$ para el algoritmo CHAID (previamente se categorizan todas las variables continuas en deciles).

Además, se divide el conjunto de datos en un conjunto de entrenamiento (70 %) uno de prueba (20 %) con el que se buscarán los mejores parámetros y uno de validación (10 %) con el que se expondrán los resultados finales (tablas).

En el algoritmo *CART* se itineró por distintos valores del parámetro de complejidad o cp y del número mínimo de observaciones requeridas en un nodo para que se realice una división o $minsplit$. Los resultados se validaron con el conjunto de prueba. El modelo final seleccionado es aquel con parámetros $cp=0.00$, $minsplit = 20$ con un $auc = 0.705$ en prueba y $auc = 0.712$ en validación.

En el algoritmo *CHAID* se itineró por distintos valores de los parámetros $alpha2$, $alpha3$ y $alpha4$ (valores críticos para la separación o unión de categorías en la prueba $Ji-2$). Los resultados se validaron con el conjunto de prueba. El modelo final seleccionado es aquel con parámetros $alpha2 = 0.02$, $alpha3 = 0.01$ y $alpha4 = 0.05$ con un $auc = 0.740$ en prueba y $auc = 0.727$ en validación. El resultado final de la calidad de clasificación, sobre el conjunto de validación, se encuentra en la tabla a continuación:

Modelo	RMSE	AUC	Lift_10	Accuracy	Recall	Precision	Specificity	F1_Score
CART	0.272	0.712	4.641	0.893	0.451	0.388	0.934	0.417
CHAID	0.259	0.728	4.510	0.889	0.503	0.383	0.925	0.435

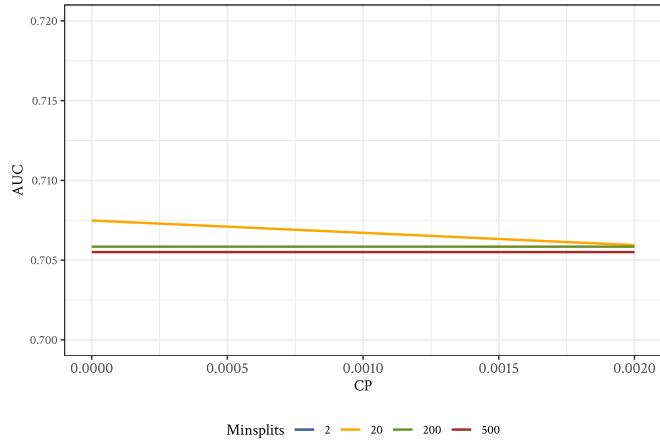


Figura 5: AUC para distintos valores de Cp y Minsplit

Árboles combinados

Para evitar el sobreajuste y el sesgo de los predictores más influyentes se hiperparametrizarán y ajustarán 3 modelos de ensamble: uno basado en el algoritmo Random Forest, uno a través del método de Gradient Boosting con el algoritmo XGBoost y el algoritmo LightGBM. Se compararán ademas con una regresión logística stepwise sin optimizar otros parámetros (como la regularización).

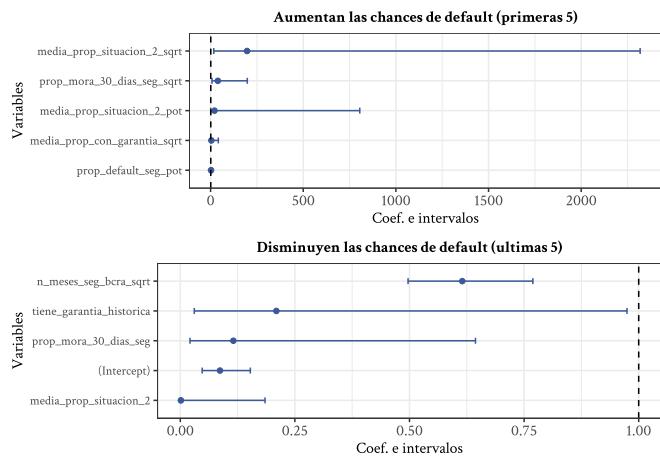


Figura 6: Efecto de las variables sobre las chances de default

Regresión logística Se ajustó un modelo de regresión logística maximal y uno sólo con intercepto. A través de regresiones stepwise mediante la función step del paquete stats, que compara las variaciones en el AIC incluyendo y excluyendo variables, se detectó que el mejor modelo binomial es aquel que incluye un conjunto de variables originales y transformadas ordenadas por magnitud del efecto (figura 6).

variables relevantes

*media_prop_situacion_2_sqrt, prop_mora_30_dias_seg_sqrt, media_prop_situacion_2_pot,
 media_prop_con_garantia_sqrt, prop_default_seg_pot, max_situacion_mes, proxy_edad_actual_pot,
 n_meses_seg_bcra_sqrt, tiene_garantia_historica, prop_mora_30_dias_seg, (Intercept),
 media_prop_situacion_2*

Como descripción general, la raíz cuadrada de la proporción de deuda en situación 2 en el semestre previo es un gran determinante del default, habiendo 196 veces más chances de no pagar. En segundo lugar, el cuadrado de la proporción de meses con default en el semestre pasado aumenta en 38 veces las chances de no pagar. El efecto de la edad, aunque es significativamente distinto de 1 en términos probabilísticos, tiene un efecto muy poco notable en las chances de no pagar, mientras que haber tenido garantía en el semestre previo reduce las chances de caer en default en un 79 %.

Random forest A continuación se buscan los mejores hiperparámetros para el algoritmo Random Forest, variando en *Mtry*: La cantidad de regresores aleatorios a utilizar en cada arbol. *Trees*: La cantidad de árboles en cada iteración y *Min_n*: La cantidad mínima de datos que requiere cada nodo para dividirse. Se añaden además dos variables uniformes aleatorias que varían entre [0,1] y [2,5], para compararlas con el efecto del resto de las variables. Dada la relevancia de estas variables en las primeras iteraciones, se modificó el parámetro *maxnodes* (cantidad de nodos terminales) para evitar el sobreajuste.

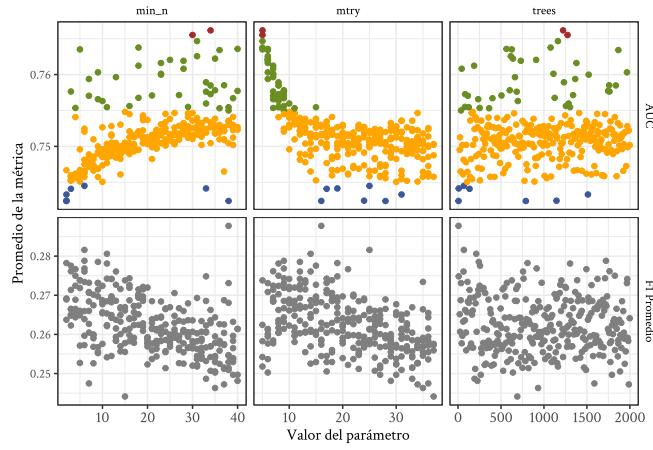
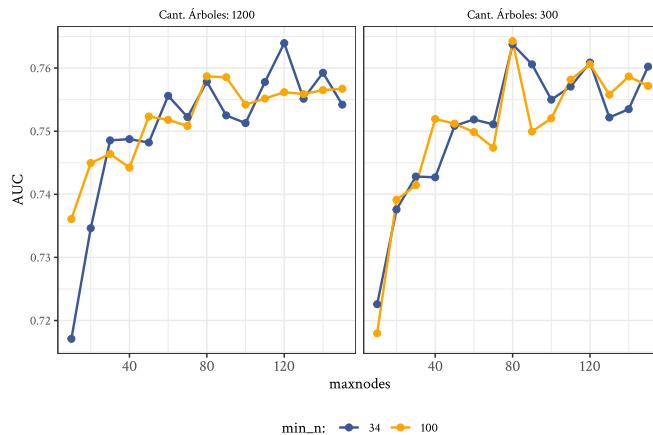


Figura 7: Distribución de las métricas en los distintos parámetros

En la figura 7 (los colores representan intervalos del AUC) se exponen 300 modelos con distintos hiperparámetros y sus resultados en FC y AUC. Es notable que menos variables mejoran la capacidad predictiva general (AUC), siendo el mejor modelo aquel con 5 variables. Los incrementos en la cantidad de datos necesarios para dividir un nodo también mejoran el AUC, mientras que la relación con la cantidad de árboles es menos distinguible. Se realiza una segunda iteración manteniendo fija la cantidad de variables:



Entonces se elige como modelo final el Random Forest con $mtry = 5$, $trees = 300$, $max_nodes = 80$ y $min_n = 100$ aunque con $trees = 1200$ el modelo parece más estable. Su $AUC = 0.764$ en el conjunto de entrenamiento y $AUC = 0.741$ en el conjunto de validación. Se puede observar además, que las variables randomizadas parecen tener una gran importancia en el modelo, esto puede ser producto del sobreajuste.

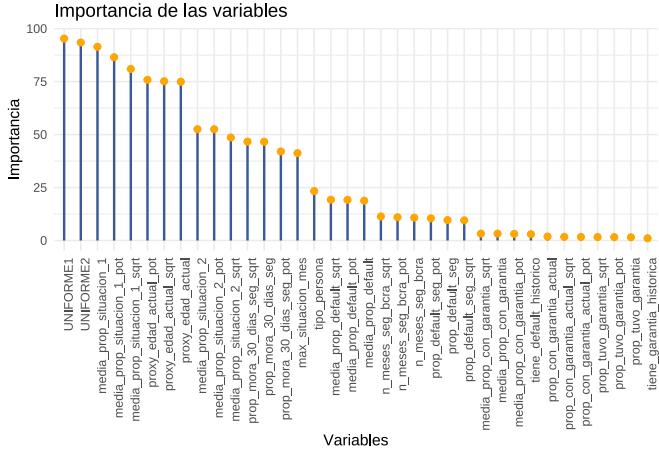


Figura 8: Comparación de la relevancia de variables

Extreme Gradient Boosting (XGBoost) Este método construye modelos con buena capacidad predictiva a partir de modelos más débiles, reconociendo los errores de las iteraciones previas. En primer lugar se itineró sobre el *learning rate (eta)* y se obtuvo que el mejor es de 0.01. A partir de allí, se optimizaron los parámetros *nrounds* (cantidad de iteraciones), *max_depth* (La profundidad máxima de cada árbol), *subsample* (cantidad de observaciones incluidas en cada árbol) y *colsample_bytree* (La fracción de características que se utilizan para entrenar cada árbol).

Mientras que valores altos de algunos parámetros (nrounds o max_depth) pueden generar sobreajuste, parámetros como gamma o min_child_weight pueden reducirlo.

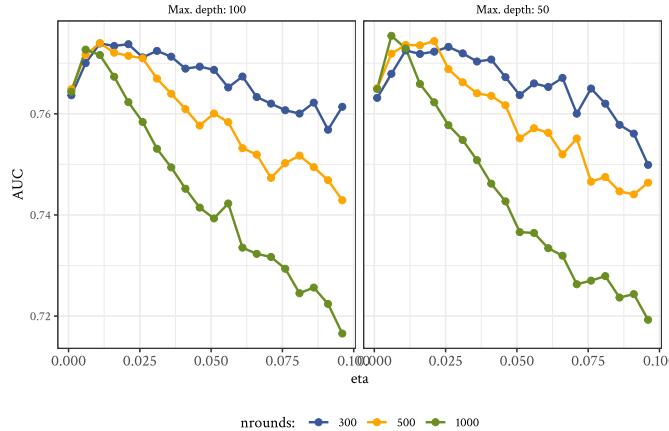


Figura 9: AUC Para hiperparámetros de XGBoost

Se elige finalmente el XGBoost con parámetros $eta = 0.006$, $nrounds = 1000$, $max_depth = 50$, $subsample = 0.8$, $colsample_bytree = 0.05$, con un $AUC = 0.750$ en el conjunto de prueba y $AUC = 0.765$ en el conjunto de validación.

Light GBM Es una variante de Gradient Boosting, donde el árbol crece hoja por hoja en lugar de hacerlo por profundidad o nivel. En general, tiene una velocidad de entrenamiento más rápida y más precisión (al dividir por hojas en lugar de por niveles). Los parámetros a optimizar serán en primer lugar el *learning rate*, luego se buscarán los mejores valores para *feature_fraction* y *num leaves*

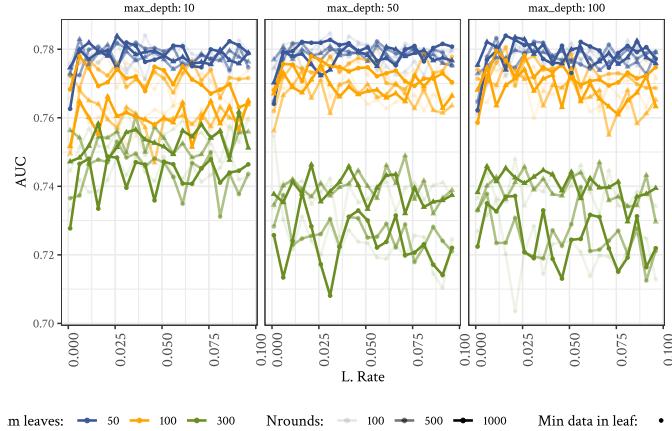


Figura 10: AUC para hiperparámetros de LGBM

El mejor modelo hallado es aquel con $learning_rate = 0.021$, $num_leaves = 50$ y $feature_fraction = 0.5$. Su AUC es de 0.781 en el conjunto de validación y 0.784 en el conjunto de validación. El resto de los parámetros del modelo son $bagging_fraction = 0.8$, $max_depth = 10$ y $min_data_in_leaf = 1$ que mejoren el AUC.

Modelo	RMSE	AUC	Lift_10	Accuracy	Recall	Precision	Specificity	F1_Score
CART	0.27	0.71	4.64	0.89	0.45	0.39	0.93	0.42
CHAID	0.26	0.73	4.51	0.89	0.50	0.38	0.92	0.44
REG.LOG.	0.26	0.77	4.64	0.88	0.54	0.36	0.91	0.43
RAND.FOR.	0.25	0.74	4.54	0.87	0.53	0.29	0.90	0.38
XGBOOST	0.24	0.77	5.00	0.87	0.56	0.30	0.90	0.39
LightGBM	0.24	0.78	4.92	0.85	0.61	0.26	0.86	0.36

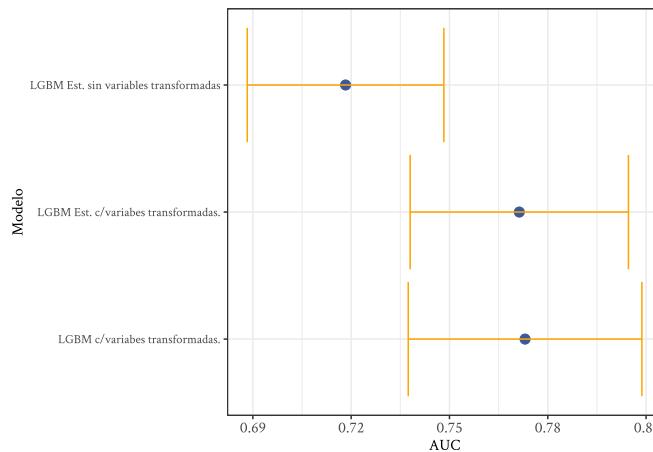


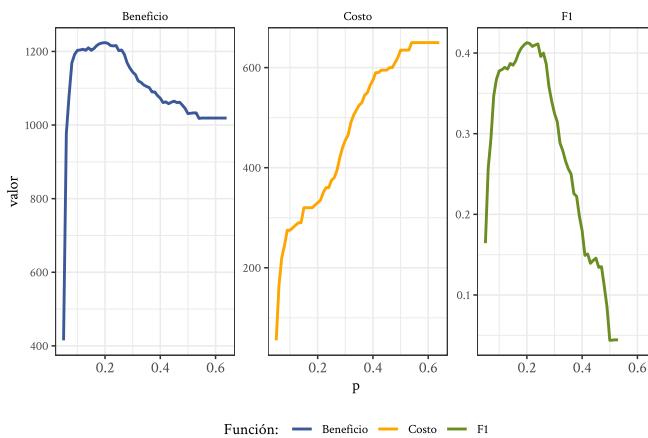
Figura 11: Cómparación de algoritmo Light GBM sobre distintos conjuntos de datos

Selección del modelo final y elección del punto de corte

Considerando que el costo de entregar un crédito y que el cuit caiga en default es de 5 créditos otorgados correctamente, se pueden plantear las siguientes funciones, siendo $FN(p)$: *Cant. de falsos negativos* y $VN(p)$: *Cant. de verdaderos negativos* y p : *punto de corte*, y C es *Costo*, I es *Ingreso* y B *Beneficio*: $C(p) = 5FN(p)$, $I(p) = VP(p)$, $B(p) = VN(p) - 5 * FN(p)$.

En primer lugar se elegirá el mejor modelo entre los ajustados. También se analizaron distintos algoritmos sobre el conjuntos de datos sin variables transformadas, estandarizado, y con variables agrupadas con métodos multivariados. Sin embargo, la mejor capacidad predictiva se obtuvo en el conjunto de datos que incluye las transformaciones no lineales. En este conjunto, el mejor algoritmo es el LightGBM (figura 11).

Considerando el mejor modelo (Light GBM) las funciones asociadas a cada valor de P se pueden observar en la figura 12, encontrando una probabilidad de corte de 0.20 como mejor clasificador si el objetivo es maximizar los beneficios con la función planteada anteriormente.



Conclusiones y uso del modelo para futuros préstamos

Se analizaron 18.006 observaciones de 15 variables y sus transformaciones a los efectos de encontrar la mejor regla de clasificación entre deudores que cayeron en default y los que no, en julio del 2019. El conjunto de datos, en general, se caracterizó por tener una baja participación de deudores en default y de deudores con garantía.

Asimismo, la proporción de la media en situación 1 entre diciembre del 2018 y junio de 2019 es una variable relevante, dado que modifica las chances la impagabilidad. La garantía de la deuda es otra variable relevante, que reduce las chances de impagabilidad.

Se evaluaron 6 algoritmos, siendo el algoritmo Light GBM el mejor hallado con un AUC de 0.78. Sin embargo, esto es menos del 1 % mejor que la regresión logística.

Luego, se optimizó el criterio de clasificación, identificando que para $p = 0.20$ se maximiza los beneficios, mientras que $p = 0.05$ minimiza el costo.

Dado que el modelo incluye información sobre el comportamiento histórico reciente de los cuit y las medidas de clasificación son las mejores encontradas con los datos disponibles, no debería ser utilizado para evaluar futuros créditos, dado que se debería contar con características *más estructurales*. Es decir, con atributos de los propietarios más invariantes en el tiempo, como ser historiales crediticios de los últimos años, ingresos anuales, patrimonio, etc. Por último, el mejor modelo basado en el algoritmo Light GBM sin incorporar las variables transformadas arroja un AUC máximo de 0.77.