Reporte Modelo de Riesgo para independientes/informales

Índice

Visión general

La población de informales/independientes es aquella compuesta por cuentahabientes del Banco que, a pesar de contar con movimiento en sus productos pasivos, no tienen una fuente comprobable de ingresos y/o trabajan por su cuenta.

Dentro de esta población se hallan no solo personas residentes del Banco sino también clientes internacionales cuya cuenta principal no es necesariamente la del Banco en cuestion.

En cualquier caso, el objetivo es medir el nivel de riesgo de estos cuentahabientes para desarrollar productos a la medida, que muchas veces serán el primer producto crediticio de estas personas y que, por ello, son difícilmente puntuables por las metodologías tradicionales y/o APC.

La solución además permitirá desarrollar un esquema de prospección para los nuevos cuentahabientes que aperturen con Banco en sus productos de Digital Banking.

Muestra de desarrollo: análisis descriptivo de la data

Se recibieron por parte del banco las siguientes tablas.

Data	Tipo	Num Variables	Num Registros	Datos Nulos/Nan	%
Base_Automovi	.xlsx	7	5,060	1,769	4.99%
Base_Clientes_Sin_Experiencia	.xlsx	46	143,607	974,969	14.76%
Base_Consultas_APC	.xlsx	22	225,753	367,923	7.41%
Base_Independientes	.xlsx	46	143,607	974,969	14.76%
Compras TDC	.txt	16	1,529,607	163,512	0.67%
Pagos TDC	.txt	16	6,102,754	325,487	0.33%

Se realizó una breve exploración de los datos, de la cual resaltamos los siguiente:

i. Las Datas identificadas como Compras y Pagos TD presentan estructuras idénticas
ii. Las Datas Independientes y Clientes sin Experiencia tienen la misma estructura, cantidad de registros y datos nulos.
iii. Todas las Datas son de tipo transaccional, es decir, el Primary Key (Pk) que identifica inequívocamente a cada "Cliente" aparece más de una vez.

Resumen de las bases:

Pagos TDC (txt): conjunto de datos tipo transaccional con 6.102.754 registros, contentivos de campos tipo fecha, números y categóricos, de estos últimos podemos decir algunos (TIPO_PRODUCTO, CATEGORIA) poseen más de 30 Clases, sobre las cuales se realiza una reclasificación.

Compras TDC (txt): conjunto de datos tipo transaccional con 1.622.927 registros, contentivos de campos tipo fecha, números y categóricos, de estos últimos podemos decir algunos (CATEGORIA) poseen más de 30 Clases, sobre las cuales se realiza una reclasificación.

Base Independientes (xlsx): conjunto de datos tipo transaccional con 143.607 registros, contentivos de campos tipo fecha, números y categóricos, de estos últimos podemos decir algunos (Cargo, DescripcionIndustria, Nacionalidad, Provincia, Ciudad) poseen más de 30 Clases, sobre las cuales se realiza una reclasificación.

Base Consultas APC (xlsx): conjunto de datos tipo transaccional con 225.753 registros, contentivos de campos tipo fecha, números y categóricos, de estos últimos podemos decir algunos (dc_nom_asoc, dc_descr_obs_corta,) poseen más de 30 Clases, sobre las cuales se realiza una reclasificación.

Base Clientes SinExp (xlsx): conjunto de datos tipo transaccional con 143.607 registros, contentivos de campos tipo fecha, números y categóricos, de estos últimos podemos decir algunos (Cargo, DescripcionIndustria, Nacionalidad, provincia, ciudad) poseen más de 30 Clases, sobre las cuales se realiza una reclasificación.

Base Automóvil (xlsx): conjunto de datos tipo transaccional con 5.060 registros, contentivos de campos tipo fecha, números y categóricos, de estos últimos podemos decir algunos (Color, DescripMarcaAuto, DescripModeloCarro) poseen más de 30 Clases, sobre las cuales se realiza una reclasificación.

Notas:

- Los campos de tipo transversal entre las Bases de Datos, es decir, aquellos que indican la misma información en las diferentes tablas, fueron homologados siguiendo las instrucciones proporcionadas por Banesco en el diccionario de datos.
- Los valores nulos o atípicos serán tratados según el dato al que hagan referencia. Ejemplo, si un registro tiene valores nulos en el histórico de TDC, es un indicativo de que no cuenta con dicho producto, lo que podría ser de valor para la predicción. Para mayor detalle, consultar el tratamiento de variables.
- Las reclasificaciones realizadas sobre los campos categóricos mencionados en el listado anterior serán evidenciadas en el diccionario de datos correspondiente a la versión final del modelo.
- Para ver el análisis descriptivo completo, revisar el documento Reporte de Calidad de Datos.

Variable objetivo I: definición de la PD y Riesgo en el tiempo

Considerando que el objetivo del modelo es evaluar a clientes sin experiencia crediticia dentro del banco, el análisis del riesgo se realizó principalmente sobre la base con el mismo nombre.

El primer paso fue revisar los tipos de productos que habían disponibles para el segmento: TDC, CONSUMO PERSONAL. Se unificaron ambas bases y se realizó la limpieza correspondiente sobre registros duplicados, lo que nos dejó la siguiente cantidad de observaciones:

- 67.674 registros de cierres de TDC
- 75.933 DE CONSUMO PERSONAL

Transformaciones y tratamientos de la data

Se calcularon las morosidades a 30 días, a 60 días y a 90 días. Es decir, se identificó en qué cierres las personas cayeron en morosidad mayor a 30, 60 y 90 días respectivamente.

Posteriormente se calculó la madurez en meses que tiene el crédito en cada una de las fechas de corte. Dicho de otra manera, se calcula el tiempo transcurrido entre la fecha de cierre de mes y la fecha del inicio del crédito.

Finalmente, con estos cálculos y unas serie de transformaciones, se analizan las morosidades por cada tipo de segmento, con el objetivo de comprender un poco la cartera en este aspecto y ver si existen diferencias y similitudes que nos ayuden a agrupar estos productos y escoger la variable objetivo del modelo.

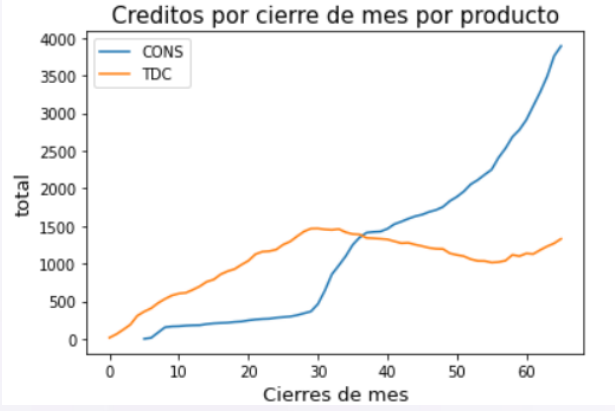


Gráfico 1

- Los cierres de mes van desde febrero de 2017 hasta julio de 2022 para TDC y desde julio de 2017 hasta julio de 2022 para consumo
- Crecimiento de producto de TDC en Abril 2018- enero 2019 y en consumo a partir de marzo de 2021

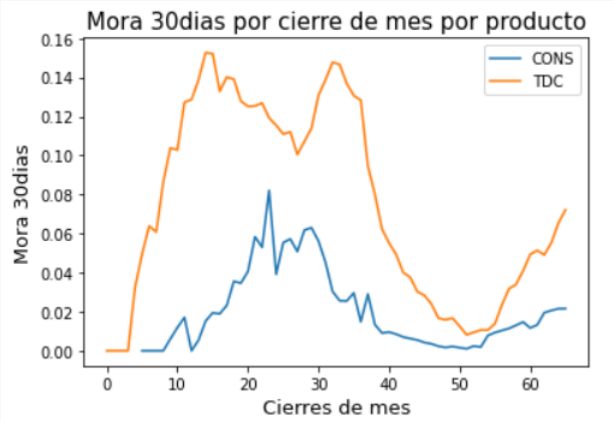


Gráfico 2

- Incremento en la morosidad de TDC marzo de 2018 hasta finales de 2019
- Incremento en la morosidad de CONSUMO desde enero hasta julio de 2019

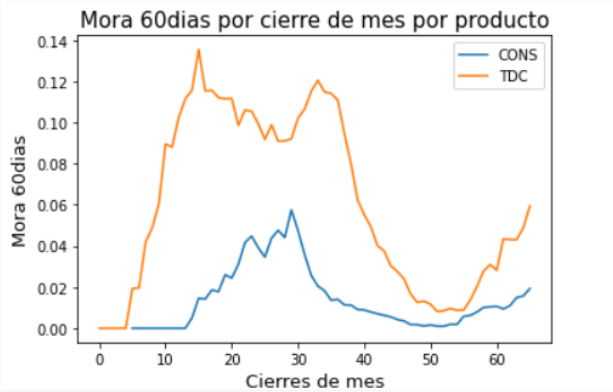
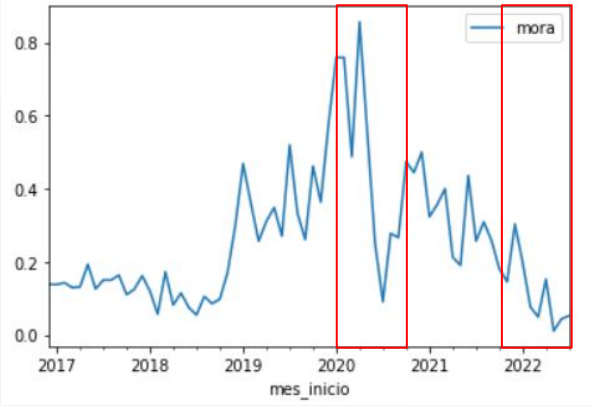


Gráfico 3

- El producto TDC el que tiene una mayor tasa de morosidad superior a 60 días.
- Muy similar al comportamiento de las curvas de la mora mayor a 30 días, pero con porcentajes menores.

Gráfico 4 - Vintage Analysis 30+

- Se ilustra la morosidad mensual mayor a 30 días. Se observan anomalías en los meses correspondientes a la pandemia (2020) y en las cosechas más recientes (2022), Ambos se descartan para mitigar la inestabilidad de la mora.

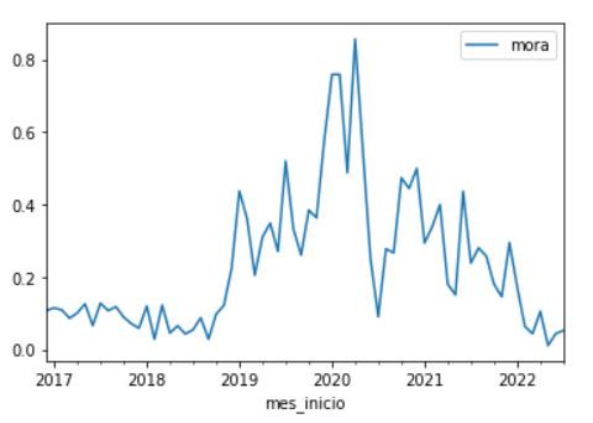


mes_inicio	mean	sum	count	mora
2016-12-01	0.861538	56	65	0.138462
2017-01-01	0.862069	75	87	0.137931
2017-02-01	0.857143	102	119	0.142857
2017-03-01	0.870504	121	139	0.129496
2017-04-01	0.868687	86	99	0.131313
2017-05-01	0.806723	96	119	0.193277
2017-06-01	0.874172	132	151	0.125828
2017-07-01	0.849624	113	133	0.150376
2017-08-01	0.850000	119	140	0.150000
2017-09-01	0.836364	92	110	0.163636
2017-10-01	0.889655	129	145	0.110345

Por cuestiones de espacio, la tabla completa se envía por separado.

Gráfico 5 - Vintage Analysis 60+

- Se ilustra la morosidad mensual mayor a 60 días. Se observa un comportamiento similar al encontrado en el Gráfico 5 con las cosechas de 2020 y 2022.



Dados estos resultados, el siguiente paso es determinar el deterioro (clientes que escalan de mora 30+ a mora 60+ en el siguiente mes) para identificar la definición ideal de mora de cara a la predicción (variable objetivo). En la base hay un total de 3.751 evaluaciones seleccionadas a partir de los resultados de los Gráficos 4 y 5, de las cuales:

MOROSIDAD	TOTAL	EMPEORAN AL SIG MES	% DE DESMEJORA
MOROSOS 30 DÍAS	1115	716	64.2%
MOROSOS 60 DÍAS	890	638	71.7%

Como parte de nuestras prácticas estándar, seleccionamos una definición de mora que supere el 50% de deterioro con respecto a su predecesor. De acuerdo con los resultados, tanto los clientes que tienen 30 o más días de mora, como los que tienen 60 días o más, presentan un porcentaje superior al 50%.

Ya con la definición más acotada se observa un (64.2%), lo que nos indica que con la definición de 30 días o más, podríamos anticiparnos con eficiencia al deterioro y prevenir que más clientes se retrasen por períodos superiores a un mes calendario.

El siguiente paso es la selección de la ventana de tiempo en la cual se observará el atraso en cada cosecha.

ventana30	Total_Clientes	Total_Clientes_Acum	Acum_porcentaje	Acum_porc_Data_Total
1.0	23	23	2.062780	0.613170
2.0	2	25	2.242152	0.666489
3.0	8	33	2.959641	0.879765
4.0	32	65	5.829596	1.732871
5.0	58	123	11.031390	3.279126
6.0	38	161	14.439462	4.292189
7.0	57	218	19.551570	5.811784
8.0	41	259	23.228700	6.904825
9.0	33	292	26.188341	7.784591
10.0	53	345	30.941704	9.197547
11.0	45	390	34.977578	10.397227
12.0	49	439	39.372197	11.703546
13.0	46	485	43.497758	12.929885
14.0	37	522	46.816143	13.916289
15.0	49	571	51.210762	15.222607
16.0	53	624	55.964126	16.635564
17.0	33	657	58.923767	17.515329
18.0	27	684	61.345291	18.235137
19.0	33	717	64.304933	19.114903

Nuevamente y de acuerdo con nuestro estándar metodológico, nuestro objetivo con la tabla anterior es encontrar el umbral de deterioro para la definición anterior en relación a los meses transcurridos. Por defecto, utilizamos una ventana de 12 meses de maduración, a menos que el porcentaje de deterioro en un mes anterior al 12vo supere el 70%.

Para ello, hemos de observar la columna “Acum_porcentaje”, que hace referencia al porcentaje acumulado de clientes morosos que incurrieron en un atraso para el mes en cuestión (véase columna “ventana_30”).

Tal y como se denota, en el mes 12 contamos ya con un 39.37% de acumulación de clientes morosos y, dado que en meses previos no existe una acumulación mayor al 70%, confirmamos la selección de dicho horizonte temporal para la variable objetivo.

NOTA: Si bien se observa un incremento en el porcentaje después del mes 12, si utilizamos ventanas más amplias, tendríamos que descartar registros que no cumplan con dicha maduración, lo que empeoraría significativamente la calidad de la predicción, dado que solo contamos con menos de 4.000 observaciones aplicables.

Variable objetivo II: Uso de las Reestructuraciones

Se realizó el siguiente tratamiento sobre los créditos reestructurados:

1. Se dividió la data de independientes, filtrada por activos, en 2 bases: una con fecha menor o igual a 30-04-2020 (base menor) y otra mayor a dicha fecha (base mayor)

2. En la base mayor se filtró por “categoría_cambio” igual a 2, y por cada 'evaluación' (u otorgamiento de crédito), se halló el mes de cierre mínimo y se categorizó cómo “mínima fecha de reestructurado”.

3. En la base menor se calculó, a partir de la variable “fecha_reestructurado”, la mínima fecha de reestructuración de cada prestatario.

4. Se unieron las dos bases obtenidas a partir de los puntos 2 y 3, para generar una sola base con la fecha de reestructuración por cada evaluación.

5. Finalmente, en esta base, al haber un margen de error en cuanto al límite de las bases de datos y su categorización o no para la fecha 2020-04-30, se volvió a calcular la min fecha para la variable mes_reestructurado calculada en los puntos 2 y 3.

6. La construcción de esta base nos permitió identificar cuáles clientes fueron reestructurados durante los 12 primeros meses de haber sido otorgado el crédito, reflejándose así en una marca de bueno o malo por motivo de reestructuración.

Variable objetivo III: Definición final de default

La definición de mal pagador que utilizamos tras aplicar los tratamientos explicados en las páginas anteriores, fue la siguiente:

Todo cliente cuyo atraso sea igual o mayor a 30 días en los primeros 12 meses del crédito y/o que hayan sido reestructurados en el mismo periodo, será categorizado como “malo”, asignándole el valor “0” dentro del formato binario de la variable “Default”

Debajo se ejemplifica esto último con los diferentes buckets de mora, incluyendo reestructuraciones. La tabla completa se comparte en un documento aparte por cuestiones de espacio.

mes_inicio	CLIENTE	Default30	Default60	Default90	Default_R
2019-03-01	14052	1	1	1	0
2018-05-01	200111260	1	1	1	0
2018-01-01	200194137	1	1	1	0
2017-06-01	200202823	1	1	1	0
2018-06-01	201421604	1	1	1	0
2020-06-01	201442214	1	1	1	0
2019-12-01	201476823	1	1	1	0
2018-10-01	201604443	1	1	1	0
2017-01-01	201894919	1	1	1	0
2018-01-01	201955981	1	1	1	0
2021-10-01	202268555	1	1	1	0

La distribución final dentro de la muestra con la cual se construirán las diversas iteraciones del modelo, es la siguiente:

TOTAL APROBADOS	BUEN PAGADOR	MAL PAGADOR
3,341	2,662	679
100%	79.7%	20.3%

Tratamiento de datos: Selección preliminar de variables explicativas

Resumen del tratamiento de variables explicativas: Para esta etapa se tomaron las variables sociodemográficas como nivel educativo, estado civil, ciudad, género, ingreso mensual, entre otras. También se calcularon otras como la antigüedad del cliente en su último empleo referente a la fecha de inicio, así como la edad. Adicionalmente se hicieron unas series de transformaciones para las variables de saldos activos y saldos pasivos referidas antes del inicio del nuevo crédito. De la misma manera se construyeron variables de morosidad anteriores a la fecha de inicio.

CLIENTE	mes_inicio	SALDO_PAS_UM	SALDO_TDC_UM	SALDO_CONS_UM	SALDO_HIPO_UM	SALDO_AUTO_UM
14052	2019-03-01	NaN	1806.87	NaN	NaN	NaN
200111260	2018-05-01	5.176583e+05	4173.95	130027.61	212845.43	NaN
200194137	2018-01-01	NaN	9101.25	NaN	NaN	NaN
200202823	2017-06-01	6.415601e+04	103.21	NaN	NaN	NaN
201421604	2018-06-01	2.736780e+05	12111.35	NaN	951855.85	NaN
201442214	2020-06-01	2.818773e+04	11614.25	NaN	NaN	NaN
201476823	2019-12-01	2.178684e+05	NaN	48733.01	164153.45	NaN
201604443	2018-10-01	3.175138e+06	9845.44	NaN	NaN	NaN
201894919	2017-01-01	NaN	NaN	NaN	NaN	NaN
201955981	2018-01-01	3.517331e+04	5179.30	NaN	279182.61	25923.03
202268555	2021-10-01	1.076789e+06	NaN	NaN	NaN	NaN

CLIENTE	mes_inicio	DIAS_MORA_UM	DIAS_MORA_U3M
14052	2019-03-01	0.0	0.0
200111260	2018-05-01	0.0	0.0
200194137	2018-01-01	0.0	0.0
200202823	2017-06-01	0.0	0.0
201421604	2018-06-01	0.0	0.0
201442214	2020-06-01	0.0	0.0
201476823	2019-12-01	0.0	0.0
201604443	2018-10-01	0.0	0.0
201894919	2017-01-01	NaN	NaN
201955981	2018-01-01	0.0	0.0
202268555	2021-10-01	NaN	NaN

Tratamiento de datos: Information Value (IV). Selección definitiva de variables explicativas

Se realizó un análisis bivariado, con el objetivo de medir la capacidad predictiva que tenía cada variable explicativa en relación la marca de default.

La siguiente tabla es el resultado del peso de las variables de entrada.

Éste es un paso previo a la construcción del modelo, por lo que no refleja un peso o ponderación asignado a las variables, únicamente refleja si la información tiene un valor para la discriminación de buenos y malos pagadores, y qué tan alto es dicho valor. Véase la tabla de abajo para mayor detalle en la interpretabilidad de estos resultados.

Variables	Importancia
score	0.053765655
ratio_i_spas	0.04146124
provincia__OTROS	0.03556504
DescripcionIndustria__CONTABILIDAD	0.029156975
DescripcionIndustria__ALIMENTOS	0.026578683
Cargo__ABOGADO	0.026028201
Cargo__COMERCIANTE	0.025808146
Descripcion_Nivel_Educacion__Maestria	0.024557855
Cargo__OTROS	0.023448162
DescripcionIndustria__ASESORIAYASISTENCIA	0.023133313
DescripcionIndustria__ROPACALZADOYACCESORIOS	0.022371791
antiguedad_empleo	0.022271644
DescripcionIndustria__ARQUITECTURAYURBANISMO	0.02226568
ciudad__OTROS	0.021929083
max_dias_atraso	0.020695766
DescripcionIndustria__ACUICULTURAPESCADERIA	0.020646263
ciudad__SANMIGUELITO	0.019731674
edad	0.019051995
IngresoMensual	0.01890237
Cargo__GERENTESUBGERENTE	0.018816212
Estadocivil__MARRIED	0.018373983
risk.score	0.018087408

TiempoRelacionBanco	0.018061649
Descripcion_Nivel_Educacion__Secundaria	0.018056886
DescripcionIndustria__OTROS	0.017656524
DescripcionIndustria__MERCANCIA	0.017577183
ratio_i_sprom	0.017410684
ratio_i_cuotas	0.01650267
provincia__CHIRIQUI	0.015985651
carrier.name__TelefonicaMoviles	0.015962064
DescripcionIndustria__ADMINISTRACION	0.015876802
Genero__FEMALE	0.015468751
cantidad_Dependientes	0.01530837
DescripcionIndustria__CONSTRUCCION	0.015262432
carrier.name__DIGICEL	0.015048769
DescripcionIndustria__LEGAL	0.014889843
ciudad__ARRAIJAN	0.014622675
carrier.name__Cable&WirelessPanama	0.014496556
ciudad__PANAMA	0.014447735
provincia__PANAMA	0.014215466
Descripcion_Nivel_Educacion__Universitario	0.014162857
carrier.name__CLARO	0.013746467
Descripcion_Nivel_Educacion__SinEscolaridad	0.013175295
Cargo__DIRECTOR	0.012282703
DescripcionIndustria__AGRICULTURAAGRONOMIA	0.012235917
Cargo__ADMINISTRADOR	0.012133883
Estadocivil__SINGLE	0.012126152
ciudad__LACHORRERA	0.012034421
provincia__PANAMAOESTE	0.011203136
Descripcion_Nivel_Educacion__OTROS	0.011132495
provincia__VERAGUAS	0.010934826
DescripcionIndustria__BELLEZAYESTETICA	0.009883337
ciudad__DAVID	0.009664206
provincia__COCLE	0.005786362

Algoritmo: Descripción y aplicación

Se ha desarrollado un algoritmo de machine learning no supervisado basado en modelos logarítmicos y árboles de decisión. El algoritmo puede ser empleado utilizando metodologías de tratamiento de datos estándar para la industria financiera, por lo que es una herramienta flexible y confiable para la construcción de modelos predictivos, principalmente orientados a originación (predicción de default).

Para la aplicación de este algoritmo y una vez aplicados todos los tratamientos pertinentes a la data (vistos en la secciones previas de este reporte), se debe realizar un proceso de "dumificación" de las variables categóricas, lo que consiste en convertir los posibles valores de una variable, en variables binarias, por ejemplo: "género" pasa a ser "género_M", cuyos posibles valores son Si (1) o No (2). Este proceso mejora la eficacia general del algoritmo, que, al ser un modelo logarítmico, funciona mejor con información numérica y binaria.

Una vez inyectada la información al algoritmo y tras realizar procesos de parametrización que son parte de nuestra IP, el algoritmo realiza micro segmentaciones sobre la población y sus variables. Dichas agrupaciones son las "ramas" del árbol de decisión, y suelen alcanzar el orden de los miles.

Dentro de estas segmentaciones se construyen a su vez variables de segundo nivel de forma automática, que pueden partir de combinaciones tan sencillas como la relación "Salario-Cargas Familiares", a variables más complejas como índices que agrupan el consumo en un periodo de tiempo dado. El resultado final luce similar al siguiente ejemplo:

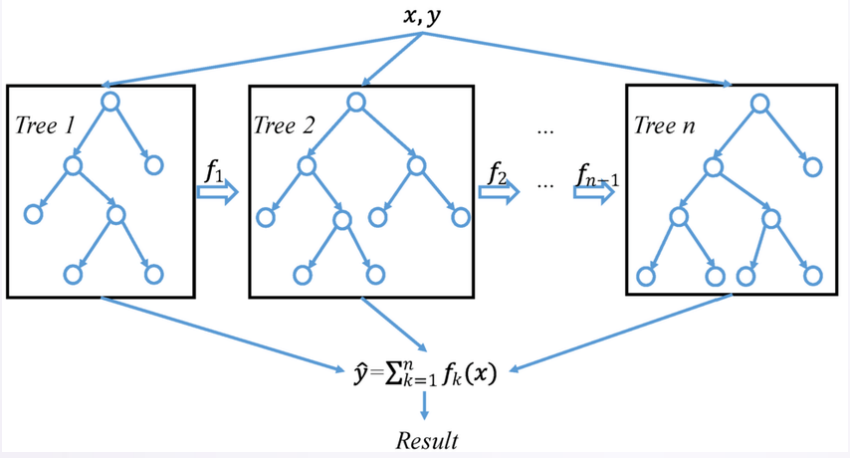


Imagen ilustrativa, no representa el resultado real del presente reporte.

Resultados: Entrenamiento y validación (Benchmarking)

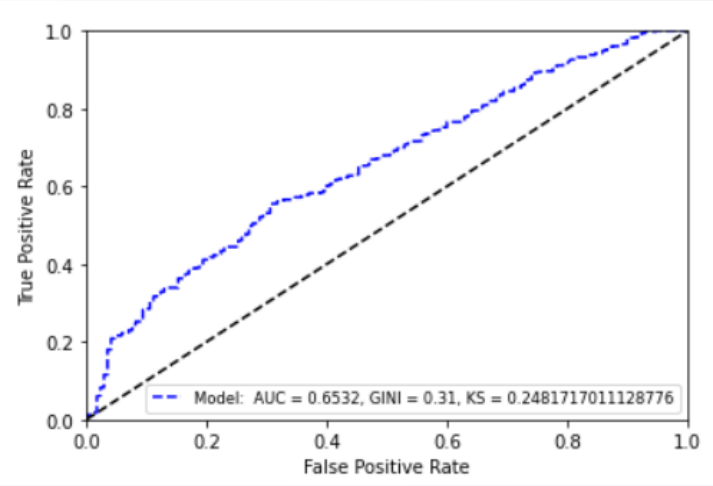
Se realizaron diferentes iteraciones (versiones) del modelo construido para el Banco, con el objetivo de comparar el impacto de a) priorizar el uso de la mínima cantidad posible de información, sacrificando performance o b) priorizar el máximo performance aprovechando el máximo nivel de información disponible. Los resultados fueron los siguientes:

V1

AUC: 0.65

GINI: 0.31

KS: 0.24



Esta versión se concibió bajo la hipótesis de no utilizar variables de comportamiento que, por su naturaleza, solo están disponibles en clientes con historial crediticio. Las variables removidas fueron:

- DIAS_MORA_U3M
- DIAS_MORA_UM
- SALDO_PAS_UM
- SALDO_TDC_UM
- SALDO_CONS_UM
- SALDO_HIPO_UM
- SALDO_AUTO_UM

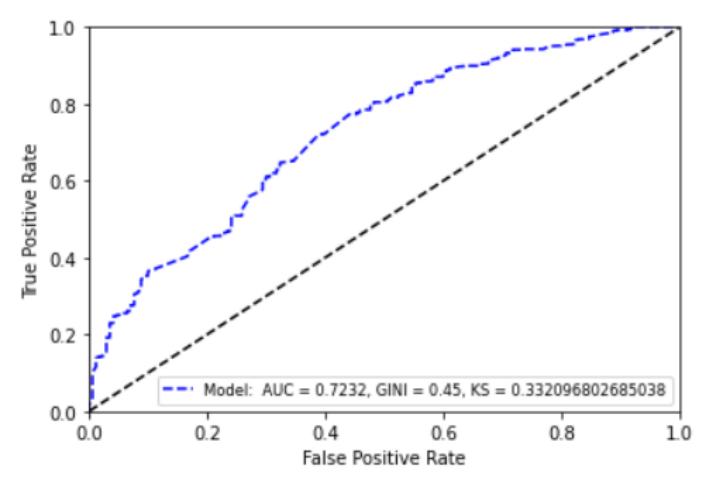
La versión fue descartada dado que los resultados en la base de test tuvieron métricas no satisfactorias (AUC menor al 0.70 y KS menor al 0.30).

V2

AUC: 0.72

GINI: 0.45

KS: 0.33



Esta versión se concibió bajo la hipótesis de evaluar solo los saldos de productos activos de los clientes, descartando el atraso para ser más flexible en la entrada de nuevos clientes sin experiencia. Las variables removidas fueron:

- DIAS_MORA_U3M
- DIAS_MORA_UM

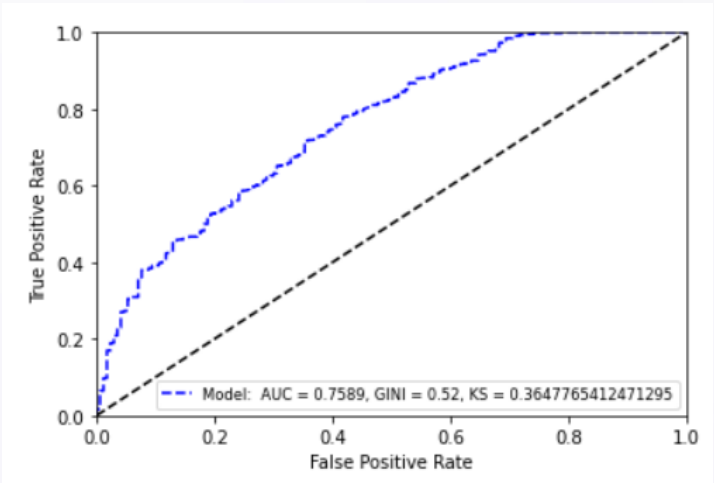
Si bien los resultados a nivel de métricas se encuentran por encima del mínimo estándar requerido (AUC igual o mayor a 0.70 y KS igual o mayor al .30), se observaron mejores resultados en otras versiones.

V3

AUC: 0.76

GINI: 0.52

KS: 0.36



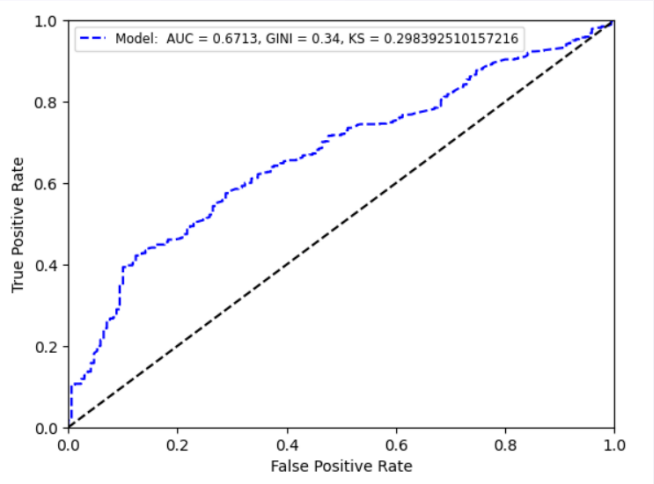
Esta versión se concibió bajo la hipótesis de emplear toda la información disponible y considerando que el modelo sería aplicado únicamente a cuentahabientes Banesco. Esta versión arrojó los mejores resultados en la base de test.

V4

AUC: 0.67

GINI: 0.34

KS: 0.30



En esta versión se descarta la data de comportamiento interno, se añaden las variables disponibles de APC, se mantiene la información de pasivos y se realizó un tratamiento sobre las variables de ingreso (unificación) y de nacionalidad (transformación a variable dicotómica -Local, Extranjero-.

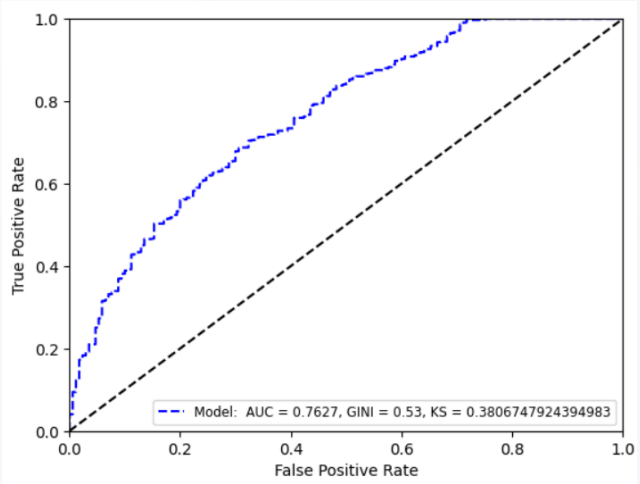
Las métricas mejoran a la V1 en test, especialmente en el KS, aunque siguen por debajo del umbral óptimo en AUC (70).

V5

AUC: 0.76

GINI: 0.53

KS: 0.38



Esta versión utiliza el máximo de información disponible, incluyendo información de APC, para la cual solo existe un nivel de match para poco menos del 5% de los registros totales de la muestra de desarrollo.

Las métricas mejoran en comparación a la V3 e incluye el tratamiento explicado en la V4.

(Ver variables empleadas en la próxima página).

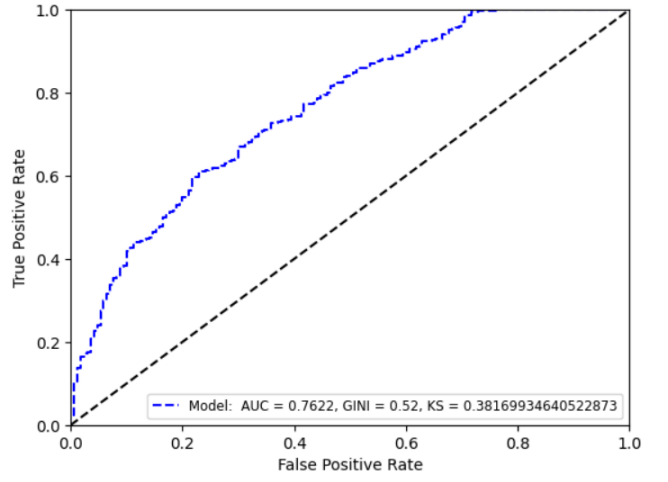
score	max_dias_atraso	prom_saldo_actual	CLIENTE	mes_consulta
732	14	18211.824000	600109452	2018-11
635	0	52381.108000	600087896	2019-02
690	0	875.720000	600033336	2019-02
586	30	13926.093333	600008786	2019-02
631	606	23067.228333	600018848	2019-02
754	0	1973.310000	600157305	2019-03
751	0	52444.338333	600087896	2019-05
773	0	1833.828333	600157305	2019-05
644	0	10797.360000	600119958	2019-05
527	60	7911.940000	600041378	2019-05

V6

AUC: 0.76

GINI: 0.52

KS: 0.38



Esta versión remueve las variables de APC pero mantiene los tratamientos de la V4.

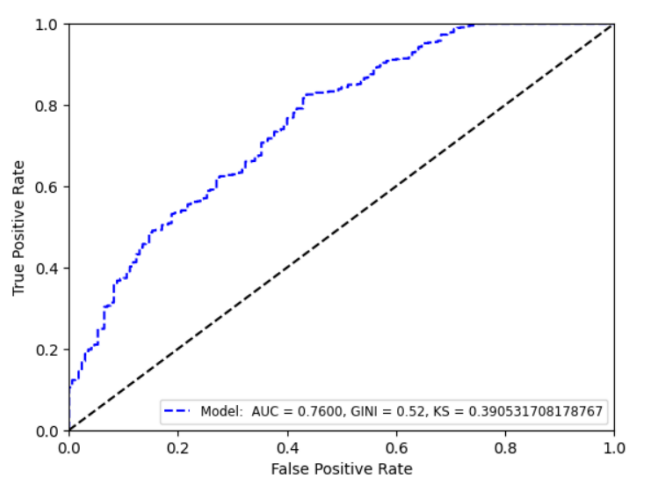
Las métricas mejoran sustancialmente en comparación a la V3 gracias al tratamiento sobre las variables de ingreso y nacionalidad.

V7

AUC: 0.76

GINI: 0.52

KS: 0.39



En esta versión utiliza toda la información disponible y los tratamientos de versiones anteriores. Se reemplazaron las variables de Saldos por un ratio que combina éstas últimas con el ingreso mensual.

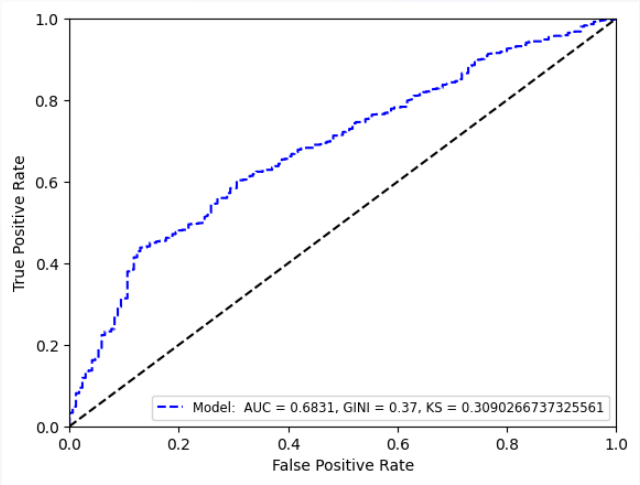
El KS mejora (el más alto obtenido hasta ahora), el AUC y GINI disminuyen ligeramente en comparación a otras versiones.

V8

AUC: 0.68

GINI: 0.37

KS: 0.31



Esta versión solo utiliza información de pasivos y APC, mantiene las transformaciones a las variables de versiones anteriores e incorpora los ratios de saldos con ingreso mensual.

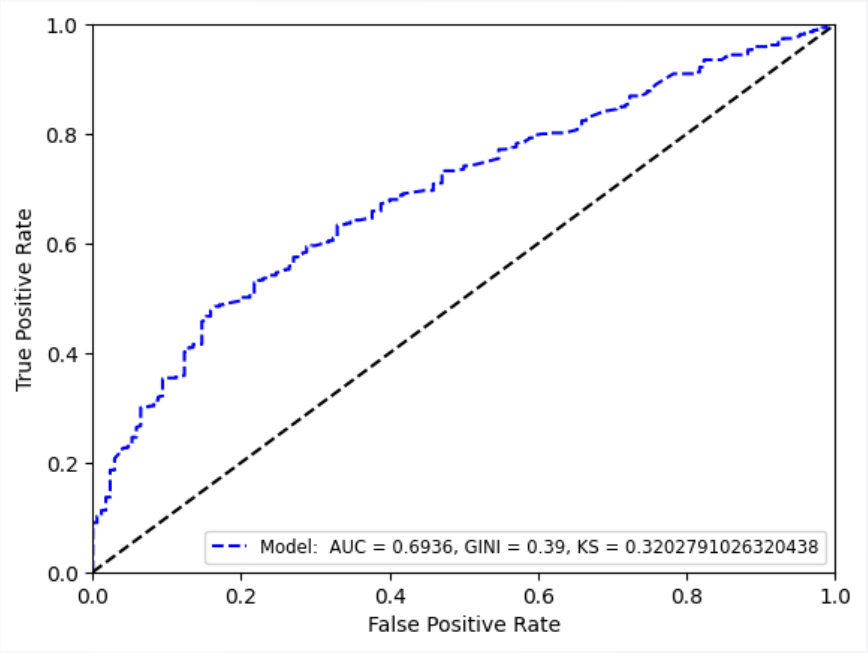
En comparación al resto de versiones sin comportamiento interno, el KS y el GINI presentan un incremento sustancial. El AUC aumenta ligeramente y se acerca al 70%. Mejores métricas hasta ahora para este tipo de modelo.

V9

AUC: 0.69

GINI: 0.39

KS: 0.32



Esta versión solo utiliza información de pasivos y APC, mantiene las transformaciones a las variables de versiones anteriores e incorpora el porcentaje de endeudamiento aproximado como monto_cuotas/ingreso_mensual.

En comparación al resto de versiones sin comportamiento interno, el KS y el GINI presentan un incremento sustancial. El AUC aumenta ligeramente y se acerca al 70%. Mejores métricas hasta ahora para este tipo de modelo.

Aspectos relevantes:

- En esta versión se recibió el cruce del apc de parte del cliente, lo que aumentó el porcentaje de match de un 5% a un 29% de las bases de datos lo que incidió en el aumento de las métricas.

VERSIÓN SELECCIONADA PARA EL PASO A PRODUCCIÓN

Modelo con Data Telco

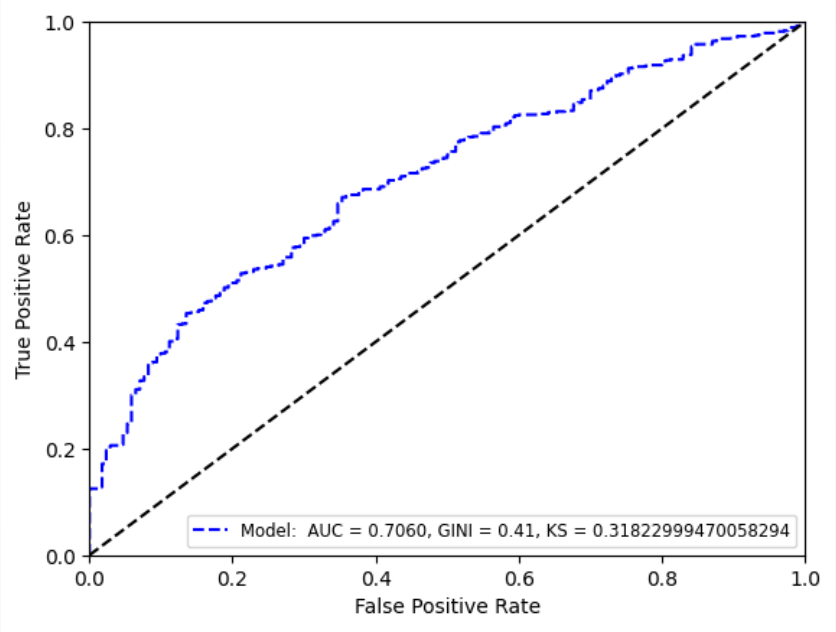
Esta versión utiliza información de pasivos y APC, mantiene las transformaciones a las variables de versiones anteriores e incorpora variables de data telco.

V10

AUC: 0.71

GINI: 0.41

KS: 0.32



En comparación con la versión V9 sin ninguna variable alternativa, el AUC aumenta ligeramente a un 70.6%, teniendo un incremento porcentual del 1.79% respecto a la versión anterior.

Version Modelo	AUC	Gini	KS
V9	0.6936	0.39	0.32
V10	0.7060	0.41	0.32
Mejora	**+1.79%**	**+5.13%**	-

Modelo final: Tabla de performance V9 (Deciles)

En el siguiente análisis se presenta como se distribuyen los 3,341 registros de los créditos aprobados por el banco de acuerdo al Score que les asigna el modelo en cuantiles con igual cantidad de casos (10% c/u).

Este análisis es fundamental para la definición de estrategias y políticas de otorgamiento de créditos.

Grupo	QUASH Score		Total	PD	Odds	PD acumulada	Odds acumulado
	Min	Max					
1	0.969	1.000	334	2.84%	34.2	2.84%	34.2
2	0.944	0.969	334	5.41%	17.5	4.13%	23.2
3	0.915	0.944	334	8.10%	11.3	5.45%	17.3
4	0.852	0.915	334	12.34%	7.1	7.17%	12.9
5	0.797	0.852	334	18.48%	4.4	9.43%	9.6
6	0.740	0.797	334	23.35%	3.3	11.75%	7.5
7	0.702	0.740	334	27.86%	2.6	14.05%	6.1
8	0.670	0.702	333	30.64%	2.3	16.13%	5.2
9	0.629	0.670	335	33.72%	2.0	18.08%	4.5
10	0.000	0.628	335	41.30%	1.4	20.40%	3.9

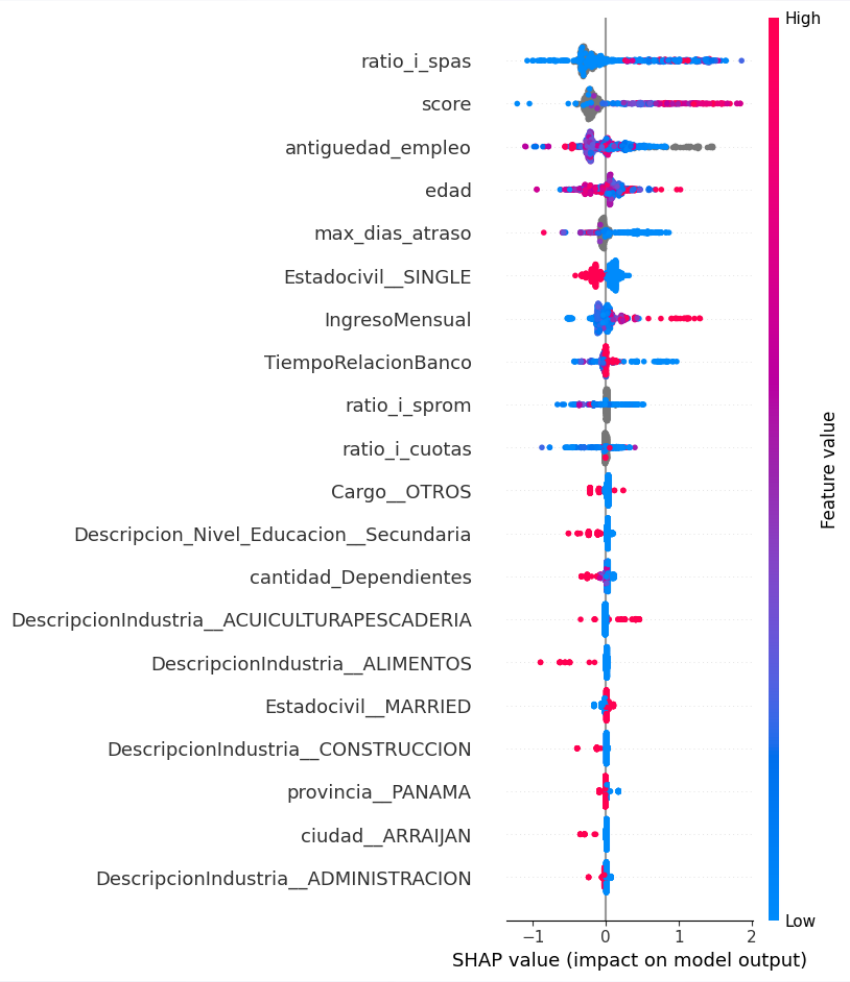
- Grupo: Identificador del cuantil
- Score: Puntaje asignado por el modelo.
- Total: Cantidad de casos en el grupo
- PD: Porcentaje de casos con la marca de “mal pagador” en el grupo.
- Odds: Cantidad de clientes “buenos pagadores” por cada “mal pagador”.
- PD Acumulada: Porcentaje acumulado de casos con la marca de “mal pagador”.
- Odds Acumulados: Cantidad acumulada de “buenos pagadores” por cada “mal pagador”.

Interpretabilidad del modelo: Análisis Shapley (V9)

El Shapley es un análisis que resulta de la librería con el mismo nombre. Básicamente, permite interpretar los resultados de un modelo de machine learning, en el cual no se ponderan manualmente las variables, sino que el propio modelo hace dicha ponderación.

Para interpretar correctamente el gráfico, se lee de la siguiente manera:

- Eje Y (Izquierda): Ordenado por importancia de la variable
- Eje Y (Derecha): Tonalidades rojas representan valores numéricos altos. Tonalidades azules, valores numéricos bajos.
- Eje X: Score Shap (Ponderación del modelo). Rangos entre -5 y 5. Valores negativos implican mayor PD y peor QUASH Score, valores positivos menor PD y mejor QUASH Score.
- Círculos: Observaciones, registros únicos.



Categorías más riesgosas

Análisis descriptivo de las categorías más riesgosas con relación a la probabilidad de default (datos reales) versus score del modelo (datos de predicción).

DescripcionIndustria	n	score_q	buenos	malos	PD
ALIMENTOS	81	0.7226	58	23	0.2840
LEGAL	94	0.7437	64	30	0.3191
ARQUITECTURAYURBANISMO	73	0.7480	54	19	0.2603
TRANSPORTETERRESTRE	75	0.7527	55	20	0.2667
CONTABILIDAD	48	0.7533	33	15	0.3125
CONSTRUCCION	188	0.7726	144	44	0.2340
ADMINISTRACION	1088	0.7808	842	246	0.2261
MERCANCIA	97	0.7817	73	24	0.2474
BELLEZAYESTETICA	44	0.7825	37	7	0.1591
ASESORIAYASISTENCIA	156	0.7874	131	25	0.1603
ROPACALZADOYACCESORIOS	82	0.7889	63	19	0.2317
SERVICIODECONSULTORIA	55	0.8035	45	10	0.1818
OTROS	733	0.8122	601	132	0.1801
AGRICULTURAAGRONOMIA	59	0.8224	45	14	0.2373
PUBLICIDADYMERCADEO	43	0.8240	36	7	0.1628
ELECTRICIDADYELECTRONICA	44	0.8278	35	9	0.2045
TECNOLOGIADEINFORMACION	73	0.8305	62	11	0.1507
RESTAURANTESYSERVICIOS	76	0.8494	66	10	0.1316
MEDICINA	41	0.8816	39	2	0.0488
ACUICULTURAPESCADERIA	191	0.9307	179	12	0.0628

Categorías más riesgosas

Análisis descriptivo de las categorías más riesgosas con relación a la probabilidad de default (datos reales) versus score del modelo (datos de predicción).

Cargo	n	score_q	buenos	malos	PD
COMERCIANTE	89	0.7564	68	21	0.2360
ABOGADO	55	0.7957	40	15	0.2727
OTROS	578	0.8190	465	113	0.1955
AGENTECOMERCIAL	45	0.8669	42	3	0.0667
GERENTESUBGERENTE	333	0.8920	299	34	0.1021
ADMINISTRADOR	376	0.8927	341	35	0.0931
DIRECTOR	58	0.8947	56	2	0.0345

provincia	n	score_q	buenos	malos	PD
PANAMAOESTE	378	0.7815	294	84	0.2222
PANAMA	2191	0.7942	1724	467	0.2131
OTROS	51	0.8008	39	12	0.2353
CHIRIQUI	338	0.8071	276	62	0.1834
HERRERA	92	0.8202	78	14	0.1522
COCLE	122	0.8286	103	19	0.1557
COLON	84	0.8337	73	11	0.1310
VERAGUAS	85	0.8424	75	10	0.1176

Categorías más riesgosas

Análisis descriptivo de las categorías más riesgosas con relación a la probabilidad de default (datos reales) versus score del modelo (datos de predicción).

Estadocivil	n	score_q	buenos	malos	PD
SINGLE	1281	0.7786	1007	274	0.2139
MARRIED	2009	0.8089	1611	398	0.1981
OTHER	51	0.8795	44	7	0.1373

Descripcion_Nivel_Educacion	n	score_q	buenos	malos	PD
OTROS	15	0.7358	8	7	0.4667
Secundaria	379	0.7709	286	93	0.2454
Universitario	2604	0.7984	2074	530	0.2035
SinEscolaridad	53	0.8077	46	7	0.1321
Maestria	290	0.8352	248	42	0.1448

Genero	n	score_q	buenos	malos	PD
FEMALE	1051	0.7886	839	212	0.2017
MALE	2290	0.8028	1823	467	0.2039