



Universidad
Carlos III de Madrid

Máster Universitario en Ciencias Actariales y Financieras (2015-2016)

Trabajo de Fin de Máster

**"Desarrollo y Validación de Modelo de Scoring de Admisión
para Tarjetas de Crédito con metodología de Inferencia de
Denegados"**

David Martínez Zapata

Tutores

José Miguel Rodríguez-Pardo del Castillo

Jesús Ramón Simón del Potro

Universidad Carlos III de Madrid - Puerta de Toledo

Esta tesis es propiedad del autor. No está permitida la reproducción total o parcial de este documento sin mencionar su fuente. El contenido de este documento es de exclusiva responsabilidad del autor, quien declara que no ha incurrido en plagio y que la totalidad de referencias a otros autores han sido expresadas en el texto.

Resumen

Este trabajo presenta los modelos de Scoring, introduce los distintos tipos de modelos existentes, así como las metodologías más utilizadas. Utilizando datos de solicitudes de tarjeta de crédito realizadas por clientes extranjeros de una entidad del sector bancario, se construye un modelo de scoring a partir de la metodología logit. Como variables independientes, se incluyen los WOE's de variables que recogen características del cliente, del crédito, del historial bancario e información laboral. Posteriormente, se evalúa el poder predictivo y la estabilidad del modelo a través del estadístico KS y el índice de Gini. Adicionalmente, se dispone de una muestra de clientes denegados, cuyo comportamiento será inferido individualmente en función de sus características. El modelo final incluye esta población denegada y es comparado con el modelo inicial.

Los datos cedidos por la entidad financiera no han sido mostrados en ningún momento. El presente trabajo no compromete la confidencialidad de la entidad, así como la de sus clientes.

Abstract

This working paper introduces Scoring models and the existing types of them, as well as the different methodologies that can be carried out. Using credit card applications made by foreign clients of an entity belonging to the banking sector, a Scoring model is constructed with the logit methodology. As independent variables, the WOE's of variables containing personal, occupational, credit and delinquency information are included. Later, predictive power and stability are assessed through the KS and Gini statistics. Additionally, a rejected population sample is provided, whose credit performance is inferred based on their characteristics. The final model, including this rejected population, is compared with the initial model.

At no time have the data provided by the financial entity been exhibited. The current working paper does not jeopardize the privacy of the entity, as well as for its clients.

Índice

1	Introducción	5
1.1	Modelos de Scoring.....	6
1.1.1	Definición de Scoring.....	6
1.1.2	Tipos de Scoring	7
1.1.3	Beneficios del modelo de Scoring	8
1.1.4	Limitaciones del modelo de Scoring.....	9
1.1.5	Metodologías de Scoring.....	11
1.2	Resultados del trabajo.....	12
1.3	Contribuciones del trabajo.....	13
2	Revisión de la literatura	13
3	Descripción de la muestra e hipótesis de partida.....	18
3.1	Descripción de la muestra.....	18
3.2	Definición de default	19
3.3	Definición de tasa de mora	20
4	Metodología y validación del modelo.....	20
4.1	Análisis univariante	20
4.1.1	Prueba Chi-cuadrado, WOE e Information Value.....	21
4.1.2	Variables numéricas	23
4.1.3	Variables categóricas.....	33
4.2	Análisis multivariante.....	39
4.3	Regresión logística.....	41
4.4	Obtención de los scores y Scorecard.....	44
4.5	Poder predictivo del modelo inicial	46
4.5.1	Test de Kolmogorov-Smirnov	46
4.5.2	Índice de GINI	47
4.5.3	Poder predictivo en la muestra de Desarrollo	48

4.5.4	Poder predictivo en la muestra Hold-Out	49
4.5.5	Poder predictivo en la muestra Out of Time.....	52
4.6	Estabilidad del modelo inicial.....	54
4.7	Inferencia de denegados y reestimación del modelo	56
4.7.1	Análisis univariante	58
4.7.2	Análisis multivariante.....	59
4.7.3	Regresión logística.....	61
4.7.4	Scorecard.....	62
4.7.5	Poder predictivo del modelo.....	64
4.7.6	Estabilidad del modelo	69
4.8	Validación Bootstrapping	71
5	Resultados y Conclusiones	74
6	Agradecimientos	78
7	Referencias Bibliográficas	79
8	Anexos.....	82
8.1	Anexo 1: Nacionalidades que componen cada grupo en el modelo inicial	82
8.2	Anexo 2: Profesiones que conforman cada grupo en el modelo inicial.....	83
8.3	Anexo 3: Provincias que conforman cada grupo en el modelo inicial	83
8.4	Anexo 4: Tipos de estado civil que componen cada grupo en el modelo inicial	84
8.5	Variables no incluidas en el modelo inicial	84
8.6	Índice de tablas y gráficos	85
8.6.1	Índice de tablas	85
8.6.2	Índice de gráficos	87
8.6.3	Índice de ilustraciones.....	89

1 Introducción

La actividad financiera se ha incrementado enormemente en las últimas décadas, debido a diversos factores como la globalización, el creciente uso de internet o la creación de distintos tipos de productos financieros. Sin embargo, toda rentabilidad trae consigo un riesgo del que las entidades financieras deben protegerse, ya sea convirtiendo ese riesgo en riesgo de otro tipo (Por ejemplo, convertir riesgo de tipo de interés en riesgo de contraparte a través de un swap), o sacrificando rentabilidad (Ventas de cartera de créditos).

A medida que los bancos están cada vez más interconectados entre sí (puesto que una entidad financiera puede ser cliente de otra, que a su vez es cliente de una tercera), y crece su expansión geográfica, se hace más necesario reforzar la regulación de su actividad. Esta regulación es muy estricta, debido a que las entidades bancarias tienen cada vez más influencia en la sociedad, y su objetivo es mantener una estabilidad monetaria y financiera, así como la protección del cliente.

Es por ello que una entidad financiera debe identificar los riesgos a los que está expuesta, debe saber medirlos y ser capaz de minimizarlos o protegerse de ellos en caso de que sea necesario. Entendiendo como riesgo la probabilidad de ocurrencia de un hecho que tuviera consecuencias adversas para la entidad. Los riesgos a los que se enfrenta se agrupan en cuatro grandes bloques: riesgo de mercado, riesgo de crédito, riesgo de liquidez y riesgo operacional. Todos ellos afectan en mayor o menor medida a una entidad bancaria.

Se puede definir el riesgo de mercado como la probabilidad de que ocurra una pérdida debido a movimientos adversos en los precios de instrumentos financieros. Dentro de este bloque se enmarcan el riesgo de tipo de interés, el riesgo de tipo de cambio, riesgo de precio de las acciones, etc. Una entidad puede protegerse de este riesgo por medio de una gran variedad de derivados financieros.

El riesgo de liquidez se define como la imposibilidad de una entidad de hacer frente a sus pasivos. Para mitigar este riesgo, los bancos están obligados a depositar un porcentaje de sus pasivos en Bancos Centrales. La importancia que la regulación le ha dado a este aspecto se debe a la gravedad de sus consecuencias. La imposibilidad por parte de los clientes de recuperar sus ahorros depositados en entidades bancarias crea severos problemas de desconfianza en el sector.

Se conoce como riesgo operacional a aquel que puede ocasionar pérdidas debido a errores humanos, fallos del sistema, errores informáticos, etc. Es un riesgo del que es difícil protegerse debido a que son eventos cuya ocurrencia y consecuencias son muy complicadas de modelizar.

El riesgo de crédito es en el que se va a centrar principalmente este trabajo. Se refiere a pérdidas ocasionadas por la imposibilidad del deudor de hacer frente al pago de la deuda que ha contraído con la entidad financiera. Las pérdidas que puede ocasionar este tipo de crédito son muy importantes para un banco, puesto que la concesión de créditos es una de sus fuentes principales de rentabilidad. Para hacer frente a este riesgo, la regulación impone una serie de parámetros que los bancos deben calcular periódicamente, y en función a éstos, almacenar una cantidad de capital, (un colchón) que permita mitigar las posibles pérdidas. Estos parámetros son:

- PD (Probability of Default): Probabilidad de que un cliente determinado sea incapaz de devolver el préstamo que se le ha concedido. Para ello, es necesario establecer una definición de default¹.
- LGD (Loss Given Default): Una vez que un cliente determinado ha entrado en default, la LGD es el porcentaje del préstamo que el banco no recupera finalmente.
- EAD (Exposure at Default): Es la cantidad que está expuesta al riesgo en el momento del impago. Suele coincidir con el saldo de la operación.

Sin embargo, una de las formas que tiene un banco de protegerse del riesgo de crédito consiste en establecer una política de concesión de préstamos. Un modo de llevar a cabo esta medida es a través de modelos de Scoring, que permiten evaluar al cliente previamente, y tomar una decisión basada en resultados estadísticos.

1.1 Modelos de Scoring

1.1.1 Definición de Scoring

Un modelo de scoring (Mester, 1997) es un método de evaluar el riesgo de crédito de solicitudes de préstamos (scoring de admisión) o de préstamos ya concedidos anteriormente (scoring de comportamiento). El objetivo es aislar el efecto de una serie de características personales o propias del producto en la probabilidad de impago del cliente, utilizando datos históricos y técnicas estadísticas. El modelo da como resultado una puntuación o "score" que el banco puede utilizar para calificar al cliente y tomar una decisión respecto a la concesión o

¹ Según Basilea II, un cliente entra en estado de default cuando tiene un retraso de más de 90 de días en el pago de una cuota en alguno de los 12 meses siguientes al mes en que es observado.

no. Para ello se utiliza información histórica de la entidad de créditos concedidos que han resultado en impago, créditos concedidos que han acabado satisfactoriamente y créditos no concedidos a los que se les estimará un comportamiento (técnica conocida como "reject inference").

Un buen modelo de scoring es aquel que identifica como buenos a los clientes que no causaron incidencias en el pago del crédito, e identifica como malos a aquellos que no fueron capaces de pagar. Es asumible que el modelo cometa errores y califique como buenos a clientes que han entrado en default. Por ello es importante realizar análisis adecuados que permitan minimizar dichos errores.

1.1.2 Tipos de Scoring

El manual de Gestión del Riesgo para Actividades relativas a Tarjetas de Crédito (FDIC², 2007) distingue distintos tipos de Scorings que utilizan los bancos actualmente:

Bureaus de Crédito

A pesar de que los modelos desarrollados internamente es lo más habitual últimamente, los bancos siguen utilizando los servicios de bureaus de crédito, sobre todo cuando no disponen de una muestra consistente de defaults, debido a que el modelo de scoring que desean realizar está dirigido a un nuevo producto o segmento.

Estos scorings están realizados por empresas especializadas en gestión de riesgo, toma de decisiones o segmentación de marketing. Las tres empresas más importantes de este sector son Experian, TransUnion y Equifax. FICO (Fair Isaac and Company) fue pionero en desarrollar modelos matemáticos para predecir default basados en información crediticia de estas agencias. Por ellos estos modelos toman el nombre de FICO scores.

Estos modelos tienen la ventaja de ser rápidos y baratos de implementar. Sin embargo, las empresas que los distribuyen son reacias a mostrar su metodología, sus supuestos y sus fuentes de información, por lo que a menudo, son "cajas negras".

Scoring de admisión

Modelos utilizados para asignar una puntuación a un solicitante de un crédito utilizados para tomar un decisión previa a la aprobación o denegación del mismo.

² FDIC: Federal Deposit Insurance Corporation, es una agencia independiente creada por el Congreso de los EEUU como consecuencia de la Gran Depresión de 1929.

Suelen utilizar como variables explicativas características propias del cliente, como la antigüedad laboral, antigüedad en la residencia actual, tipo de vivienda, etc., o propias del préstamo, como el importe solicitado. Los scorings de decisión pueden servir, a su vez, para tomar decisiones de los términos y condiciones del préstamo.

Attrition scoring

Se utilizan para detectar a aquellos clientes propensos a cancelar sus cuentas, a ser inactivos o reducir considerablemente su posición.

Scoring de comportamiento

Los scoring de comportamiento se utilizan para evaluar el riesgo de una cartera de clientes que ya cuentan con productos del banco. En contraposición a los modelos de admisión, que evalúan solicitudes de productos ya sea por parte de clientes o no clientes.

Se utilizan para tomar decisiones de gestión, y llevar un control de la exposición del banco a determinados riesgos.

Scoring para la recuperación de créditos

Califica a los clientes por la probabilidad de recuperar pagos retrasados. Permiten gestionar mejor a los clientes irregulares (aquellos que tienen un retraso en algún pago inferior a 90 días), así como tomar decisiones de recuperación en función del grado de morosidad o irregularidad.

Scoring para detección del fraude

Se utiliza para detectar aquellas cuentas con potenciales actividades fraudulentas. Identifican comportamientos sospechosos en cuentas, los cuales se puntúan con un score.

Estos son algunos de los distintos scorings que se utilizan. Los modelos de scoring más importante, sin embargo, son los FICO scores, los scoring de admisión y los scoring de comportamiento.

1.1.3 Beneficios del modelo de Scoring

El uso de modelos de scoring para evaluar clientes y operaciones está justificado por las ventajas que estos presentan. Entre ellas, Mester (1997) destaca el ahorro de tiempo y recursos y garantizar que se está aplicando la misma política de aceptación para todos los clientes, independientemente de sus características personales.

En cuanto al ahorro de tiempo, un modelo de scoring eficiente puede reducir considerablemente el tiempo requerido para tomar una decisión sobre la admisión o no de un préstamo. El tiempo que ahorre el modelo dependerá de cuánto confíe el banco en él, y qué grado de atención quiera emplear en aquellos clientes que se acercan al punto de corte. Incluso cuando el banco decide no confiar totalmente en el modelo, éste permite concentrar los recursos en operaciones dudosas, ya que los clientes que obtengan una buena puntuación será debido a que cumplen los criterios establecidos a la hora de construir el modelo.

Respecto a la objetividad, de cara a la regulación, es importante justificar que no se establecen criterios de aceptación en base a características personales de índole racial, familiar, de género o cualquier aspecto con una sensibilidad especial. Un modelo de scoring permite demostrar a la entidad bancaria, en caso de usar un factor que perjudique en mayor medida a una población determinada, que dicho factor se está utilizando por motivos de negocio. Schreiner (2002) defiende que un modelo de scoring trata de la misma forma a dos clientes con idénticas cualidades.

Además, Schreiner (2002) indica que la puntuación que se proporciona a un cliente se puede demostrar fácilmente a través de la tarjeta de puntuación o "scorecard", con lo cual, es más sencillo que demostrar una decisión subjetiva tomada por un analista. La tarjeta de puntuación permite mostrar de forma sencilla los resultados que hay detrás de la compleja metodología del scoring.

1.1.4 Limitaciones del modelo de Scoring

Uno de los aspectos más importantes a la hora de desarrollar un modelo es conocer sus limitaciones y conocer hasta qué punto se puede confiar en él.

La precisión de sus predicciones dependerá del cuidado con el que se haya construido el modelo, según Mester (1997). Si el modelo no se ha calibrado con precisión, las consecuencias de su uso puede ser muy perjudiciales. La muestra sobre la que se debe desarrollar debe ser rica en buenos y malos (préstamos que han pagado exitosamente y préstamos fallidos, respectivamente). Los datos deben ser lo más actuales posibles, y el modelo debe ser revisado periódicamente para analizar si las relaciones entre variables son similares a las existentes en el momento inicial.

Hay que tener en cuenta, además, que la muestra de desarrollo no debe contener solamente préstamos concedidos, ya que en ese caso, el modelo estaría sesgado hacia una población concreta. Por ello, se deben incluir los préstamos que fueron denegados y estimar su

comportamiento. Para comprobar el desempeño, habitualmente la muestra de partida se divide en dos grupos: muestra de desarrollo y muestra de testeo. Esta última tendrá las mismas características que la primera y se utilizará para probar el poder predictivo del modelo.

Igualmente, el modelo de scoring no determinará adecuadamente la probabilidad de fallido de un cliente cuyas características difieran de las de los individuos incluidos en la muestra de desarrollo. Si la población con la que fue calibrada el modelo no es suficientemente diversa, el poder predictivo del modelo será muy pobre para determinadas subpoblaciones. Es decir, es importante tener en cuenta para qué individuos conviene utilizar un modelo u otro.

Schreiner (2002) añade que una de las limitaciones del scoring es que asume que gran parte del riesgo está relacionado con características observables o cuantificables, como la edad del cliente, tipo de empleo, etc. En línea con este argumento, según la metodología que se utilice para calibrar el modelo, se asume que las relaciones entre variables son lineales. Poder capturar relaciones no lineales entre variables es costoso de llevar a cabo, y el resultado no es tan comprensible como una tarjeta de puntuación.

Otra limitación es que la ventana temporal utilizada para desarrollar el modelo no es una decisión trivial. Una scorecard calibrada en momentos de estabilidad económica puede no predecir bien la insolvencia de clientes con iguales características que los de la muestra en tiempos de crisis. Por ello, destaca la importancia de revisar periódicamente el modelo y comprobar la estabilidad de la población, es decir, que la distribución de las variables no ha cambiado significativamente, y que las relaciones entre ellas permanece inalterada.

Además, existe el riesgo de que cuando un modelo de scoring se ha implantado y está listo para su uso, ya se considere anticuado. Por su naturaleza, estos modelos necesitan gran cantidad de información y lleva tiempo desarrollarlos. Puede ser que alguna de las condiciones que se han asumido en el modelo haya cambiado en el transcurso de su desarrollo.

Por último, no todos los factores observables pueden ser incluidos en un modelo de scoring. Hay determinadas características que no se pueden utilizar para discriminar a los clientes. Un ejemplo es el Equal Credit Opportunity Act (ECOA) de 1974, que regula la prohibición de discriminar la calidad crediticia de un cliente en base a la raza, el sexo, el estado civil, religión o nacionalidad en Estados Unidos. Si alguna de estas características influyera de forma significativa en la capacidad crediticia, no incluirlas supondría un coste en términos de precisión del modelo.

Conocer las limitaciones de un modelo de scoring puede ayudar a evitar pérdidas ocasionadas por una dependencia excesiva de modelos matemáticos. El exceso de confianza en un modelo ha subrayado la importancia de su validación. En el presente trabajo se llevarán a cabo diversas pruebas de validación del modelo.

1.1.5 Metodologías de Scoring

Hay diversas metodologías que se utilizan para desarrollar modelos de scoring. Desde modelos lineales de probabilidad hasta modelos de valoración de opciones y redes neuronales.

Dentro de los modelos lineales de probabilidad se encuentran los logit y los probit. Son de las técnicas más comunes para modelizar probabilidades de default en base a características históricas del cliente y datos históricos de crédito. Ambas técnicas asumen una relación lineal entre el default y las variables independientes. La metodología logit³ asume que la probabilidad de default sigue una distribución logística. Por el contrario, la metodología probit asume que la probabilidad de default tiene una distribución normal.

Otra de las metodologías para el modelo de scoring de crédito es el análisis discriminante, que en lugar de devolver una probabilidad de default, divide la muestra entre clientes de alto riesgo y clientes de bajo riesgo. Se trata de encontrar factores explicativos que permitan distinguir a un grupo de otro.

Estas tres tipologías mencionadas anteriormente son las metodologías clásicas de scoring. En los últimos años han ido apareciendo métodos más sofisticados.

Uno de estos métodos es el basado en precios de opciones, y se puede aplicar a empresas. Este método consiste en comparar los pasivos del prestatario con una opción put suscrita sobre sus activos. Cuando los activos caen por debajo del nivel de los pasivos, la compañía entra en default. El modelo, basado en el modelo estructural de Merton (1974), infiere la probabilidad de default de una empresa en función de la volatilidad de sus activos, que normalmente depende de la volatilidad de las acciones.

Una de las metodologías más destacadas son las redes neuronales, algoritmos de inteligencia artificial que permiten, a través de la experiencia, el aprendizaje automático para capturar de la manera más adecuada la relación entre los factores explicativos y la probabilidad de fallido y a la vez determinar qué características son las más importantes para explicar esta probabilidad. Esta metodología es más flexible que las anteriores ya que no requiere

³ La función logit tiene la siguiente forma: $logit(x) = \ln\left(\frac{p}{1-p}\right) = \ln(p) - \ln(1-p)$

establecer ningún supuesto sobre la distribución de las variables o sobre la forma en que las variables están relacionadas. El aprendizaje automático, o *machine learning*, consiste en que el propio ordenador aprende de la información ya procesada y genera nuevos comportamientos. Este método es muy complejo, y más complicado de entender que los métodos estándar, por ello que éstos se sigan utilizando habitualmente en el mundo de la banca.

En el presente trabajo se aplicará un modelo logit cuyos resultados serán reescalados para proporcionar una puntuación. Las variables se agruparán en buckets a partir de los cuales se calcularán los Woe's (*Weight of evidence*), que se utilizarán como variables independientes del modelo. Los Woes determinan el poder predictivo de una variable independiente en relación a una variable dependiente. Su cálculo se explicará más detenidamente en los próximos apartados, no obstante, la fórmula es la siguiente:

$$WOE_i = \ln \left(\frac{\% \text{ Buenos}_i}{\% \text{ Malos}_i} \right)$$

Donde %Buenos_i es el porcentaje de buenos del grupo i sobre el total de buenos, y % Malos es el porcentaje de malos en el grupo i sobre el total de malos. Los Woes se utilizan habitualmente para discretizar variables continuas, agrupándolas según la distribución de buenos y malos a lo largo de la variable independiente. En este caso, se considera bueno a aquel individuo que devuelve satisfactoriamente el préstamo, y malo a aquel individuo que ha entrado en default.

1.2 Resultados del trabajo

Se han obtenido dos modelos de scoring diferentes, uno con una muestra que únicamente contiene solicitudes de clientes a los cuales se les concedió el crédito, y otro modelo, construido a partir del primero (mediante la metodología de inferencia de denegados) con una muestra que contiene a clientes admitidos y clientes denegados. El modelo con denegados presenta resultados ligeramente mejores en cuanto a poder predictivo. Sin embargo, al ser aplicado en muestras diferentes, la discriminación de clientes buenos y malos es algo inferior. En cuanto a la estabilidad del modelo, es decir, cómo de similar es la clasificación de clientes en muestras posteriores a la de desarrollo; el modelo realizado a partir de denegados, muestra resultados mejores que el modelo inicial. Por estas razones, el modelo seleccionado finalmente para la gestión del riesgo de crédito y la gestión de admisión de solicitudes de tarjetas será el modelo que incluye población de denegados.

1.3 Contribuciones del trabajo

En este trabajo se desarrolla un modelo de admisión para tarjetas de crédito de clientes no nacionales de una entidad del sector bancario. Permite observar la relación entre un conjunto de características del cliente y del crédito, y el evento de default. Además, se llevará a cabo análisis univariante y multiariante utilizando técnicas y habilidades adquiridas en el transcurso del Máster. El modelo se ha realizado enteramente con el software SAS, a través de la versión gratuita disponible en su página web⁴, con código desarrollado expresamente con este fin.

El trabajo se estructura de la siguiente forma, en el apartado 2 se hace un breve resumen de la literatura existente sobre modelos de Scoring y técnicas utilizadas en estos; el apartado 3 describe la muestras utilizadas y se da una definición de cliente "bueno" y cliente "malo"; el apartado 4 explica detalladamente la metodología seguida para la obtención del modelo final y se divide en: análisis univariante, análisis multivariante, regresión logística, poder predictivo del modelo, estabilidad del modelo e inferencia de denegados; y por último, el apartado 5 resume los resultados obtenidos e introduce las conclusiones del trabajo.

2 Revisión de la literatura

Existe una extensa literatura que trata las distintas metodologías de scoring, las técnicas más habituales de medir el poder de discriminación, o la elección de un punto de corte óptimo desde un punto de vista estadístico.

Leung et al. (2008) exponen brevemente el procedimiento a seguir en el desarrollo de un modelo de scoring. Adicionalmente, utilizan una muestra obtenida de un importante banco australiano, compuesta por 38.766 observaciones y 138 variables. Los pasos que identifican son los siguientes: limpieza de los datos, que consiste en eliminar variables que no puedan observarse en el proceso de admisión, eliminar variables con un alto porcentaje de valores perdidos y eliminar observaciones con datos atípicos. A continuación, categorización de variables continuas, se simplifican las variables continuas y se obtienen sus WOES. La selección de variables, realizada a través del método de regresión stepwise, que añade y elimina variables teniendo en cuenta si añaden información o no. El siguiente paso es la creación de la muestra de desarrollo y la muestra de validación, que consiste en dividir la muestra aleatoriamente y desarrollar el modelo con el 80% de las observaciones aproximadamente. El 20% restante se utilizará para comprobar su robustez. Por último, la validación del modelo,

⁴ http://www.sas.com/en_my/industry/higher-education/on-demand-for-academics.html

consistente en evaluar su poder de discriminación y su estabilidad a través de los estadísticos Kolmogórov-Smirnov (KS) y el índice Gini.

Mester (1997) aporta una visión general de los modelos de scoring. Incluye una breve introducción de la creación de estos modelos en los años 50 para poder calificar préstamos al consumo. Define el scoring crediticio como un método de evaluación del riesgo de crédito a partir de datos históricos y características personales. Continúa su exposición describiendo los usos que se le puede dar a los modelos de scoring, las distintas metodologías utilizadas para su desarrollo, los beneficios del scoring, algunas de sus limitaciones y sus implicaciones para el sector bancario. Mester concluye con la reflexión de que la técnica del scoring será determinante para la Banca Personal y de Pequeñas Empresas en el transcurso de los próximos años.

Obrová (2012) repasa brevemente la historia del credit scoring, mencionando a Fair Isaac and Company como los primeros en desarrollar modelos matemáticos para este fin. En su repaso general a los modelos de scoring, revisa la metodología de construcción del modelo a través de la regresión logística binaria utilizando dummies como variables explicativas. Además, presenta el coeficiente de Gini y la curva de Lorenz como estadísticos útiles para evaluar la habilidad de un modelo de separar a los clientes buenos de los malos.

El manual de gestión del riesgo para actividades relacionadas con tarjetas de crédito (2007), en su octavo capítulo, trata de forma muy exhaustiva todo lo relativo a los modelos de scoring, desde los propósitos para los que construye (controlar selección de riesgos, gestionar el riesgo a partir de una política de precios, evaluar nuevos tipos de préstamo, reducir tiempo de decisión para la aceptación de un préstamo, etc.). Explica los distintos tipos de modelos scoring que se utilizan habitualmente, mencionados anteriormente. Se detallan las distintas etapas de desarrollo de un modelo, ya sea la selección de una muestra apropiada, suficientemente grande para captar la variabilidad de la población a la que esté dirigido, selección de variables explicativas y categorización de éstas, obtención de la tarjeta de puntuación y posterior validación del modelo. Al igual que múltiples artículos, este capítulo expone las limitaciones del scoring: necesidad de una muestra robusta con gran cantidad de buenos y malos, sensibilidad a cambios en la distribución de las variables y a cambios en la situación macroeconómica. Como novedad, se presenta el proceso de validación que debe llevar a cabo un banco antes de implementar un modelo interno: réplica independiente del modelo, documentación del proceso, justificación de los supuestos y back-testing.

Altman et al. (1998) revisan la literatura existente sobre credit scoring de los 20 años anteriores. Comentan los trabajos de otros autores en materia de modelos de scoring realizados utilizando regresión logística, análisis discriminante, modelos estructurales y redes neuronales.

Thomas (2000) explica la historia del credit scoring. Aunque el primer modelo matemático fue desarrollado por Fair Isaac and Company, la idea de diferenciar entre grupos de buenos y malos fue introducida en estadística por R. A. Fisher en su artículo de 1936. Además, explican ampliamente las distintas metodologías que utilizan los modelos de scoring: análisis discriminante, regresiones utilizando variables dummies, regresiones agrupando las variables en buckets, regresión logística, regresión probit, árboles de decisión, redes neuronales, algoritmos genéticos o el algoritmo de vecinos cercanos. También comentan las posibilidades que ofrece el scoring de comportamiento, permite a los prestamistas tomar decisiones de gestión en base al futuro desempeño de sus clientes.

Avery et al. (1996) introducen el scoring para hipotecas y los factores que determinan el impago de los préstamos hipotecarios, tanto teóricamente (como el capital del prestatario, este aspecto se relaciona con las teorías basadas en el precio de opciones, si el capital del prestatario cae por debajo del pasivo, tiene la opción de vender la propiedad; otro factor teórico son los llamados *triggering events*, como son un desempleo repentino o una enfermedad grave) como empíricamente (un factor empírico, según los autores, es el loan-to-value, la proporción de garantía que está financiada por el préstamo). Además, los autores estudian la distribución del TMS (The Mortgage Score, desarrollado por Equifax) a través de la población, para ello, utilizan la historia crediticia, situación de la hipoteca y código postal de una muestra de hogares proporcionada por Equifax. Encuentran que los hogares que están sujetos a una hipoteca tienden a tener score más bajo que las personas que viven individualmente o los hogares sin hipoteca.

Majer (2006) utiliza 500 solicitudes de préstamos entre agosto de 2004 y mayo de 2005 de un banco que opera en Polonia. La muestra consiste en 21 variables además de la fecha de la solicitud. A partir de esta muestra desarrolla dos modelos logit, uno utilizando WOES y otro con variables continuas. Una vez estimados ambos modelos procede a su validación por medio del KS y el Gini. Los resultados que obtiene son que el modelo estimado a partir de WOES tiene un porcentaje mayor de aciertos (malos clasificados como malos y buenos clasificados como buenos) que el modelo logit estándar, además, el modelo con WOES tiene valores más elevados de KS y Gini, y mejor desempeño de la curva de Lorenz.

García et al. (2005) realizan un modelo de scoring utilizando dos técnicas distintas: regresión logística y análisis discriminante. Para ello utilizan una muestra de clientes de entidades financieras de Castilla y León que hubieran formalizado una operación de crédito o préstamo. El método de selección de variables en ambos modelos es *stepwise* o pasos sucesivos. Como resultados obtienen que los dos modelos desarrollados tienen la misma capacidad predictiva, un 97,1%. Se diferencian en que el análisis discriminante predice el 100% de los buenos clientes y el 88,9% de los malos; y la regresión logística predice el 98,1% y 94,4% respectivamente.

Boj et al. (2009) proponen un método de selección de modelos de scoring basado en distancias. El criterio de selección se establece en función del coste del error, que tiene en cuenta las probabilidades de mala selección de clientes buenos y malos. Para poder utilizar este criterio es necesario asignar unos valores del coste de conceder un crédito a un cliente malo, o de denegarlo a un cliente bueno, para ello, acude a bibliografía. Este criterio lo aplica a datos obtenidos gratuitamente del repositorio Statlog. Primero sobre una muestra de 1000 individuos alemanes, y posteriormente sobre una muestra de 690 individuos australianos. Concluyen que no existe un modelo óptimo para todas las carteras, sino que hay que estudiar cada entorno por separado.

Boj et al. (2012) estudian diferentes criterios para seleccionar un punto de corte adecuado en el modelo de regresión logística. Estos criterios son el coeficiente Kolmogorov-Smirnov (KS) y el índice de Gini, además de la curva ROC. Nuevamente, estudia probabilidades de mala clasificación y funciones de coste del error, y prueba estos criterios sobre una muestra de datos alemanes y otra de datos australianos. La curva ROC, utilizada para localizar el punto de corte donde el KS se maximiza, y el índice de Gini son presentadas como medidas de calidad de ajuste del modelo. Sin embargo, como conclusiones se puede extraer que el punto de corte es diferente si lo que se pretende es maximizar el KS, maximizar el Gini o minimizar la probabilidad de mala clasificación.

Laruccia et al. (2000) estiman un modelo de regresión logística para predecir el evento de default en una muestra de 8230 pequeñas y medianas empresas mayoritariamente del norte de Italia. Las variables que utilizan son variables con información contable de cada empresa, como el ratio patrimonio neto sobre el total de pasivo o el ratio de deuda a corto plazo sobre efectivo. Su objetivo es clasificar la muestra en distintas clases de riesgo en función del score resultante del modelo, es decir, en función de su probabilidad de default.

Rezác et al. (2011) indican que, dado que los modelos de scoring son ampliamente utilizados para predecir la probabilidad de impago de un cliente, es muy importante que el modelo esté bien ajustado. Introducen medidas para comprobar la calidad de un modelo de scoring, basadas en la función de distribución, como el KS, la curva de Lorenz o el índice de Gini, y medidas basadas en la función de densidad, como la diferencia de medias entre los grupos de buenos y malos o el Information Value (IV). Utilizan datos de una entidad financiera de Europa del Este correspondientes a 176.878 préstamos concedidos entre Enero de 2004 y Diciembre de 2005, con un 10,5% de tasa de mora. Obtienen un KS de 0,3394, un índice de Gini de 0,4623 y un Information Value de 0,712.

Huang et al. (2007) destacan la importancia de los modelos de scoring y su labor como instrumento principal de decisión. Debido a su importancia, no deben ajustarse únicamente a la muestra de desarrollo, por ello, normalmente, parte de la muestra se mantiene aparte (muestra *holdout*). Esta proporción suele ser un 70% muestra de desarrollo, 30% muestra *holdout*. Además de esta muestra, es recomendable incluir la población de denegados a la hora de desarrollar el modelo, para evitar un sesgo hacia los aprobados. Una vez que se ha calibrado el modelo y se ha puesto en práctica con la muestra *holdout*, se comprueba la estabilidad y poder predictivo con una muestra posterior, llamada *out of time (oot)*. El problema reside en que, parte del poder predictivo del modelo se pierde al utilizar la muestra *oot* para la población aceptada. Los autores buscan posibles razones que expliquen este hecho, una de ellas es el efecto de la inferencia de denegados y los cambios en la distribución de la población entre la muestra de desarrollo y la muestra *out of time*.

En cuanto a la metodología de la regresión logística aplicada a otros campos, Duarte et al. (2007) realizan un modelo de regresión para determinar la probabilidad de éxito en el primer semestre de la carrera de ingeniería en una muestra de 217 alumnos. Como variables independientes utiliza una medida de la expectativa del alumno, valoración subjetiva del ratio entre el esfuerzo realizado y el resultado obtenido, promedio de la nota obtenida en estudios anteriores, nota estandarizada en matemáticas, nota estandarizada en lengua y nota estandarizada en ciencias. Para comprobar la validez el modelo utiliza la matriz de confusión para obtener la probabilidad de error en la predicción, e intenta minimizar la sensibilidad (probabilidad de calificar como exitoso a un alumno no exitoso) y especificidad (probabilidad de calificar como no exitoso a un alumno exitoso).

Respecto a la metodología a utilizar para categorizar variables continuas, Fuentes (2013) introduce diversas técnicas como la representación gráfica de la variable en cuestión, técnicas

basadas en el rango intercuantílico o los resultados obtenidos al realizar la prueba de asociación de χ^2 de Pearson. Esta última medida de asociación será utilizada en el presente trabajo.

3 Descripción de la muestra e hipótesis de partida

3.1 Descripción de la muestra

En este apartado se introducirá la muestra con la que se desarrollará el modelo de scoring, y las muestras con las que se validará.

La muestra completa se compone de 28.650 solicitudes de tarjeta de crédito de clientes extranjeros de una entidad del sector bancario en los años 2007 y 2008. Se dispone de información personal del cliente (edad, estado civil, nivel de estudios, años viviendo en la residencia actual, etc.), información laboral (tipo de profesión, tipo de contrato, años de antigüedad laboral, ingresos netos justificados, etc.), información bancaria (año de antigüedad siendo cliente, activos y pasivos en el banco, provincia en la que se solicita el préstamo, importe solicitado, etc.) e información crediticia (puntuación de scoring TRIAD, número de incidencias anteriores). De cada cliente se observa además, en caso de que se le haya concedido el préstamo, si lo ha pagado satisfactoriamente o no. En las 25.505 solicitudes cuyo desempeño puede ser observado, hay un total de 21.510 clientes buenos, y 3.995 clientes malos.

La muestra principal se ha dividido en tres submuestras:

Muestra de desarrollo: Compuesta por 24.247 solicitudes del año 2007. Esta muestra a su vez, se ha dividido en dos muestras mediante un muestreo aleatorio, por lo que no hay diferencias significativas en la distribución de las variables de ambas submuestras.

Muestra de desarrollo del modelo: Es el 70% de la muestra anterior. Se utilizará para el desarrollo del modelo. Se compone de 16.973 observaciones, de las cuales 14.392 corresponden a clientes buenos y 2.581 a clientes malos (tasa de mora de 15,21%).

Muestra Hold-Out: Compuesta por 7.274 observaciones, el 30% de la muestra anterior. Hay un total de 6.120 clientes buenos y 1.154 clientes malos (tasa de mora de 15,86%).

Muestra Out of Time: Se compone de 1.258 solicitudes de préstamo posteriores a la de la muestra de desarrollo, en este caso, son solicitudes realizadas a lo largo del año 2008. Se

utilizará para comprobar la estabilidad del modelo y su poder predictivo. Tiene un total de 998 clientes buenos y 260 clientes malos (tasa de mora ligeramente superior, 20,67%).

Muestra de denegados: Se compone de 3.145 solicitudes realizadas en 2007 de clientes a los que no se les concedió el préstamo. Para estos clientes no se dispone de información de desempeño, no es posible saber si son buenos o malos. Sin embargo, sí se dispone del resto de información, por lo que, a través del modelo construido con la muestra de desarrollo, se podrá inferir su comportamiento y su probabilidad de impago. Este tratamiento resulta en un total de 2.273 clientes buenos y 872 clientes malos (tasa de mora 27,73%, más alta debido a que son clientes potencialmente malos).

Muestra	Observaciones	Buenos	Malos	Tasa de Mora
Desarrollo	16.973	14.392	2.581	15,21%
Hold-out	7.274	6.120	1.154	15,86%
Out of Time	1.258	998	260	20,67%
Denegados	3.145	2.273	872	27,73%
Total	28.650	23.783	4.867	16,99%

Tabla 1: Muestras utilizadas

3.2 Definición de default

Para poder realizar un modelo de scoring, cuyo objetivo es identificar a clientes malos y separarlos de los buenos, es necesario establecer una definición de qué es bueno y qué es malo. En esta definición, entran en juego dos factores: los días de atraso en un pago para considerar a un cliente como moroso, y el desempeño, que es la ventana temporal en la que se observará la peor situación de dicho cliente. Los días de atraso determinan si un cliente es moroso o no. La marca de mora determina la situación de default durante el periodo de desempeño.

Un cliente malo es aquel que presenta un atraso en un pago, de 30 días o más, en alguno de los 12 meses siguientes a la fecha en que se observa su situación.

Es decir, un cliente que en febrero de 2016 se ha atrasado 30 días en el pago de una cuota, será considerado como malo a partir de febrero 2015, como mínimo los 12 meses siguientes, hasta que se ponga al día de los pagos, momento en que pasará al estado de cura. El estado de cura es el estado posterior al default. Para que un cliente vuelva al estado normal desde default, debe pasar 12 meses en cura. Sin embargo, este concepto no es relevante para el modelo de scoring.

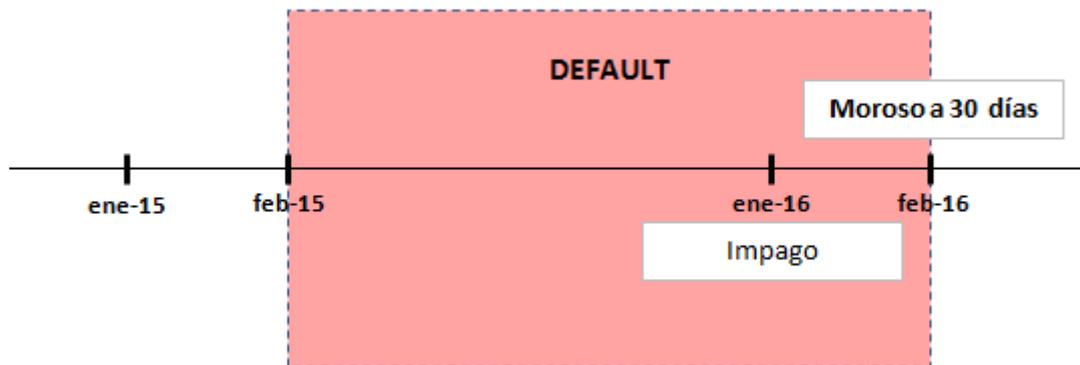


Ilustración 1: Explicación de la situación de default

3.3 Definición de tasa de mora

La variable que el modelo pretende explicar es el evento de default, explicado anteriormente. Esta variable se codifica de forma que es igual a 1 para clientes malos, y 0 para clientes buenos. Cada variable se intenta categorizar de forma que cada tramo tenga una tasa de mora diferenciada y un número suficiente de individuos. La tasa de mora del tramo i , viene dada por la siguiente expresión:

$$Tasa\ de\ mora_i = \frac{Malos_i}{Malos_i + Buenos_i}$$

Es decir, mide la proporción de clientes malos sobre el total de clientes en dicho tramo.

4 Metodología y validación del modelo

Esta sección explica detalladamente los pasos seguidos para la obtención de la scorecard final, desarrollada con una muestra que incluye inferencia de denegados, y su posterior validación evaluando el poder predictivo y la estabilidad de los scores.

4.1 Análisis univariante

En este apartado se realiza el análisis univariante de las variables más explicativas para el evento de default. Se ha realizado de forma separada para las variables numéricas y categóricas.

Primero, se introducirán los conceptos de Chi cuadrado, Weight of Evidence (WOE) e Information Value (IV).

Para las variables numéricas se ha estudiado su distribución e histograma, así como sus estadísticos descriptivos: media, mediana y percentiles. Posteriormente se ha agrupado cada variable en buckets y obtenido los WOES de manera que se maximice el Information Value.

Para las variables categóricas, se ha estudiado la frecuencia y la tasa de mora de cada clase. Posteriormente, se han agrupado las clases tratando de formar grupos lo más homogéneos posibles y con suficiente población.

4.1.1 Prueba Chi-cuadrado, WOE e Information Value

Prueba Chi-cuadrado

Introducida por Karl Pearson en 1900, esta prueba sirve para decidir si dos criterios de clasificación sobre el mismo conjunto de individuos son independientes o no. En este caso, esta prueba se ha utilizado para seleccionar puntos de corte que mejor separen la muestra en base a la distribución de buenos y malos. Esta metodología es también la utilizada por Fuentes (2013) y Nieto (2010).

Para realizar esta prueba es necesario una tabla de contingencia 2x2, en la que se separe la muestra estableciendo un punto de corte en una variable.

X/Y	Y=0	Y=1	n_y
X < C	n_{11}	n_{12}	n_{1*}
X > C	n_{21}	n_{22}	n_{2*}
n_x	n_{*1}	n_{*2}	n

El estadístico se calcula de acuerdo a la siguiente fórmula:

$$Q = \sum_{i=1}^2 \sum_{j=1}^2 \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

donde:

$$e_{ij} = \frac{n_{i*} * n_{*j}}{n}$$

e_{ij} es la esperanza que se esperaría encontrar si los criterios de clasificación fueran independientes. El estadístico Q mide cuánto se parecen las frecuencias observadas a las esperadas. Cuando Q tome valores grandes será poco creíble que ambas frecuencias se parecen.

El test Chi-cuadrado utilizado puntos de corte a lo largo de toda la variable continua, dará una idea de cuáles son los puntos que mejor diferencian a la población en función de la tasa de mora.

Weight of Evidence (WOE)

El peso de la evidencia o WOE indica el poder predictivo de cada categoría, Nieto (2010). Mide la diferencia entre la proporción de buenos y malos en cada grupo.

Se define de la siguiente manera:

$$WOE_i = \ln \left(\frac{\% \text{ Buenos}_i}{\% \text{ Malos}_i} \right)$$

Donde: $\% \text{ Buenos}_i = \frac{\text{Buenos}_i}{\text{Total Buenos}}$ y $\% \text{ Malos}_i = \frac{\text{Malos}_i}{\text{Total Malos}}$

Por definición, no puede haber una categoría que esté formada únicamente por buenos o malos.

El uso de WOES es particularmente adecuado para modelar usando regresión logística. Concretamente, la regresión logística realizará una regresión lineal de los WOES (introducidos como variables independientes) para predecir la transformación logit de la variable binaria. Por tanto, al usar los WOES como predictores, todas las variables tienen la misma magnitud y escala, y sus estimadores pueden ser comparados directamente.

Information Value (IV)

El Information Value es una medida de entropía muy popular en la construcción de scorecards. Con este estadístico se puede medir el poder de predicción de agrupar los atributos de una variable. Además, es un buen indicador a la hora de seleccionar variables para un modelo de regresión logística binario, como es el caso de un modelo de scoring.

Se define de la siguiente manera, para un número k de categorías:

$$IV = \sum_{i=1}^k (\% \text{ Buenos}_i - \% \text{ Malos}_i) * WOE_i$$

Generalmente, se establecen los siguientes criterios en cuanto al IV (Brotherton, 2013):

- Menor a 0,02: La variable no es predictiva.
- Entre 0,02 y 0,1: La variable es débilmente predictiva.

- Entre 0,1 y 0,3: La variable es predictiva media.
- Más de 0,3: La variable es fuertemente predictiva.

No se tendrán en cuenta para el modelo aquellas variables cuyo IV sea menor a 0,02. Sin embargo, un IV mayor que 0,5 significa que la variable es sobre predictiva y no se debe tener en cuenta igualmente.

En el presente trabajo, las variables han sido agrupadas de tal manera que se maximice el Information Value, asegurando que cada categoría contiene un número suficiente de defaults.

4.1.2 Variables numéricas

Puntuación TRIAD

TRIAD es un modelo de scoring externo⁵. Es un servicio proporcionado por FICO que permite automatizar decisiones basadas en la exposición al riesgo de la compañía a nivel cliente. Si bien es un buen indicador de la probabilidad de impago de un cliente, este modelo fue creado de forma genérica y no está calibrado para esta muestra en concreto. Por ello se utilizará como variable independiente en el modelo.

A continuación se presentan los estadísticos descriptivos principales de la variable TRIAD, así como su distribución sin tener en cuenta los outliers:

Observaciones	16.973
Media	525
Máximo	766
Q3	679
Mediana	652
Q1	601
Mínimo	0

Tabla 2: Estadísticos de la variable TRIAD

⁵ Para más información sobre el TRIAD: <http://www.fico.com/en/products/fico-triad-customer-manager>

Distribución TRIAD

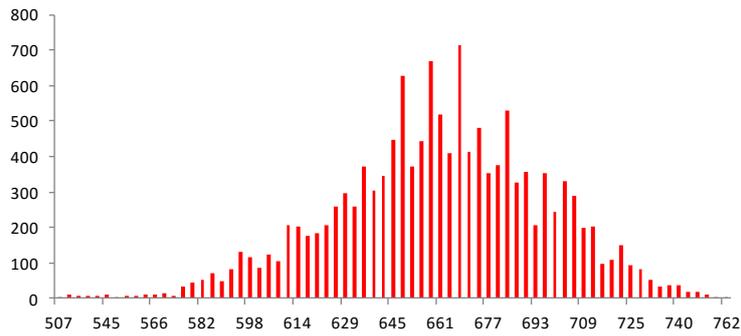


Gráfico 1: Distribución de TRIAD

Una vez que se conoce la distribución de los clientes a lo largo de la variable, se ha realizado la prueba chi-cuadrado para identificar posibles puntos de corte:

Chi-cuadrado y Information Value - TRIAD

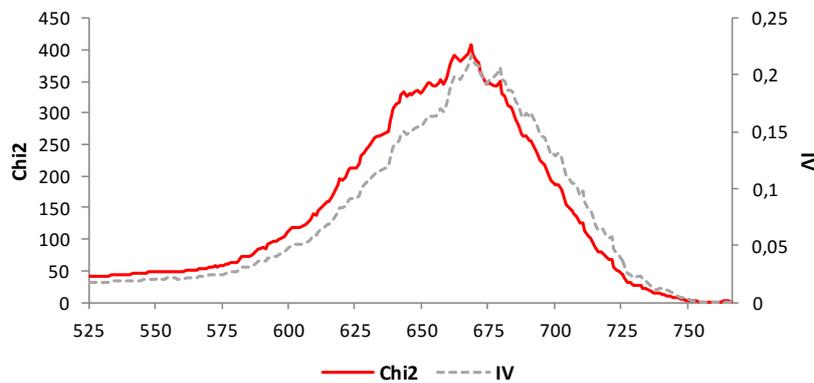


Gráfico 2: Valores del estadístico Chi2 a lo largo de TRIAD

En base a esta prueba, se han establecido puntos de corte en aquellos puntos donde el estadístico Chi2 alcance máximos relativos y se han ido ajustando hasta asegurar que los grupos establecidos presentan tasas de mora diferenciadas, malos suficientes y un buen poder predictivo. Además, se ha intentado que la mora presente un comportamiento monótono a lo largo de los grupos propuestos:

Variable	Tramo	Clientes	Buenos	Malos	% Población	Tasa Mora	% Buenos	% Malos	WOE	IV
Triad	-650	8.304	6.610	1.694	48,92%	20,40%	45,93%	65,63%	-0,357	0,2507
	650-675	3.967	3.406	561	23,37%	14,14%	23,67%	21,74%	0,085	
	675-700	2.848	2.605	243	16,78%	8,53%	18,10%	9,41%	0,654	
	700+	1.854	1.771	83	10,92%	4,48%	12,31%	3,22%	1,342	

Tabla 3: Tramos propuestos para variable TRIAD

A continuación, es test ANOVA demuestra que los cuatro grupos definidos son diferentes en media:

Fuente	DF	Suma de cuadrados	Cuadrado de la media	F-Valor	Pr > F
Among	3	56,877	18,959	150,923	<.0001
Within	16969	2131,643	0,126		

Tabla 4: Tabla ANOVA para variable TRIAD

Gráficamente, se observan las tasas de mora por score, así como los grupos definidos anteriormente. Se observa una tasa de mora decreciente con la puntuación del score y la variable tiene un poder predictivo moderadamente fuerte:

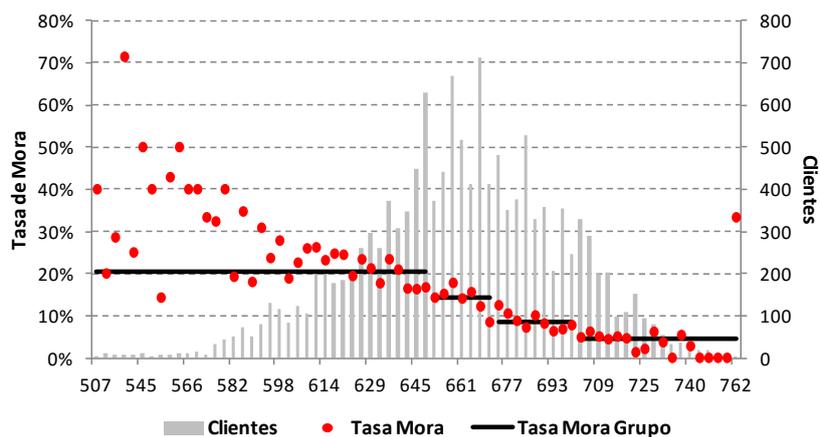


Tabla 5: Tasa de mora y distribución de clientes de TRIAD

Pasivo del cliente

El pasivo del cliente refleja la cantidad de dinero que tiene depositada el cliente en el banco. Es decir, es pasivo desde el punto de vista de la entidad.

A continuación se presentan los estadísticos principales y la distribución del Pasivo:

Observaciones	16.973
Media	4.563
Máximo	9.155.856
Q3	2.067
Mediana	718
Q1	230
Mínimo	0

Tabla 6: Estadísticos de la variable PASIVO

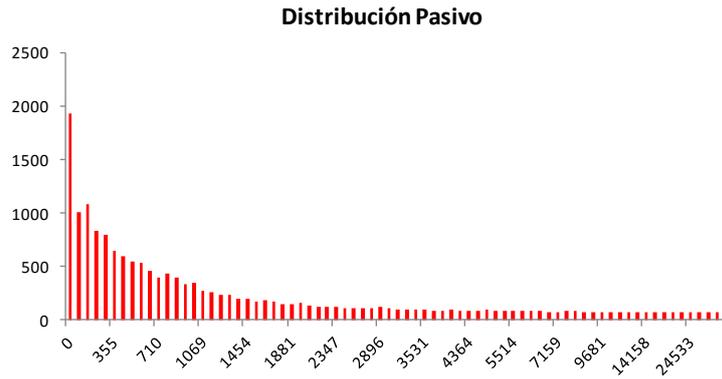


Gráfico 3: Distribución de PASIVO

Se comprueba que la distribución tiene una cola muy larga, así que se ha observado la distribución sin tener en cuenta el 25% de los clientes con mayor pasivo:

Observaciones	12.730
Media	587
Máximo	2.067
Q3	902
Mediana	433
Q1	155
Mínimo	0

Tabla 7: Estadísticos de la variable PASIVO, teniendo en cuenta el 75% de la muestra

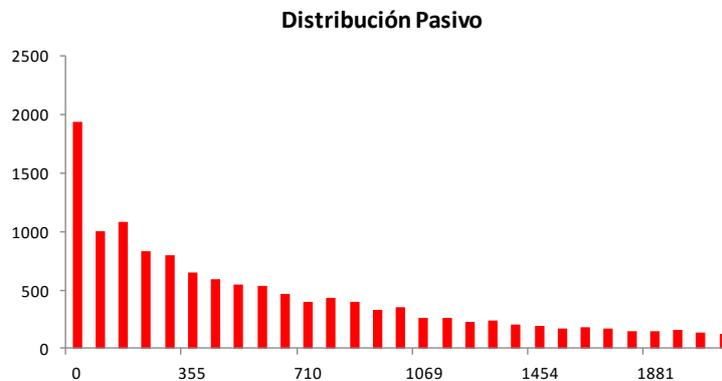


Gráfico 4: Distribución de PASIVO del 75% de la muestra

Con la prueba chi-cuadrado se analizan posibles puntos de corte:

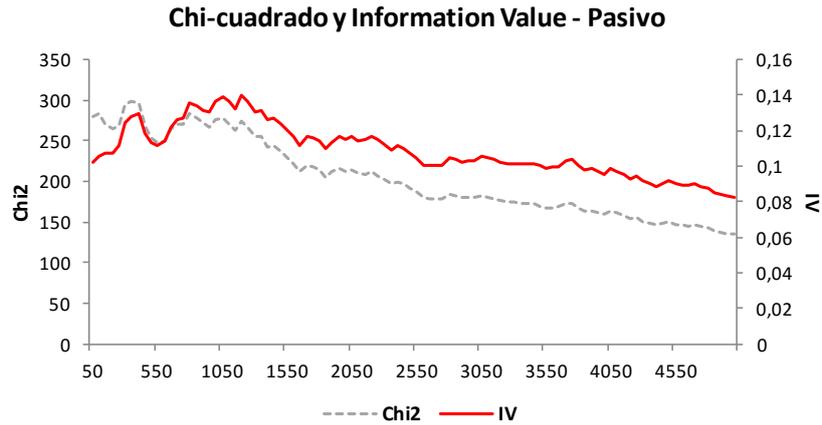


Gráfico 5: Valores del estadístico Chi2 a lo largo de PASIVO

En base a los resultados de la prueba chi-cuadrado se han establecido unos puntos de corte que se han ajustado manualmente para asegurar tasa de mora diferenciadas y monótonas, grupos suficientemente poblados y un buen poder predictivo:

Variable	Tramo	Clientes	Buenos	Malos	% Población	Tasa Mora	% Buenos	% Malos	WOE	IV
Pasivo	-100	2.160	1.568	592	12,73%	27,41%	10,89%	22,94%	-0,744	0,2134
	100-800	6.782	5.619	1.163	39,96%	17,15%	39,04%	45,06%	-0,143	
	800-2000	3.689	3.225	464	21,73%	12,58%	22,41%	17,98%	0,220	
	2000-3700	1.669	1.489	180	9,83%	10,78%	10,35%	6,97%	0,394	
	3700+	2.673	2.491	182	15,75%	6,81%	17,31%	7,05%	0,898	

Tabla 8: Tramos propuestos para variable PASIVO

Mediante el test ANOVA, se confirma que los grupos son estadísticamente diferentes en relación a su tasa de mora:

Fuente	DF	Suma de cuadrados	Cuadrado de la media	F-Valor	Pr > F
Among	4	59,374	14,843	118,293	<,0001
Within	16968	2129,147	0,125		

Tabla 9: Tabla ANOVA para variable PASIVO

Gráficamente, se representa el número de clientes por nivel de pasivo, la tasa de mora y la tasa de mora de los grupos definidos anteriormente. Se observa una tasa decreciente con la cantidad de pasivo y dado el Information Value, la variable tiene un buen poder predictivo:

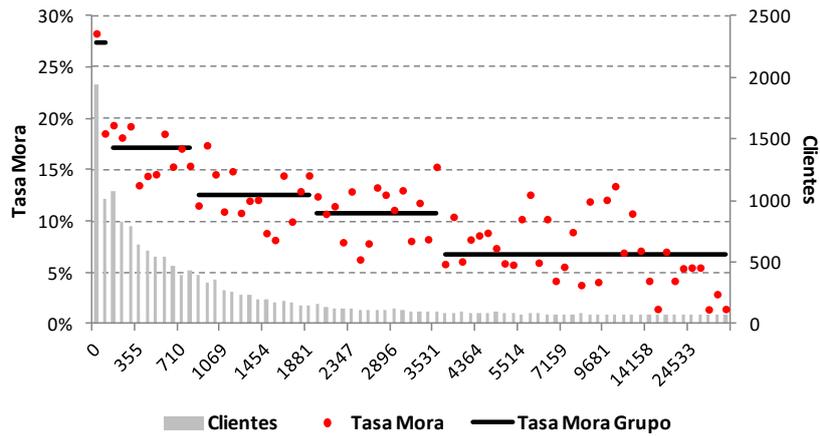


Gráfico 6: Tasa de mora y distribución de clientes de PASIVO

Antigüedad del cliente

Representa el tiempo, en meses, que el solicitante del préstamo lleva siendo cliente del banco.

Si se observan los estadísticos principales y la distribución se observa que cerca del 90% de los individuos se concentran en antigüedades menores a los 6 años, con una media de 28 meses:

Observaciones	16.973
Media	28
Máximo	485
Q3	32
Mediana	13
Q1	6
Mínimo	3

Tabla 10: Estadísticos de la variable ANTIGÜEDAD

Distribución Antigüedad

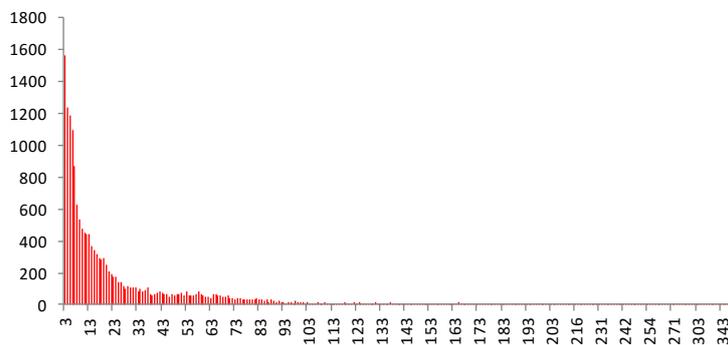


Gráfico 7: Distribución de ANTIGÜEDAD

Sin embargo, observando la distribución del 90% de la muestra:

Observaciones	16.101
Media	21
Máximo	100
Q3	26
Mediana	12
Q1	6
Mínimo	3

Tabla 11: Estadísticos de la variable ANTIGÜEDAD, teniendo en cuenta el 90% de la muestra

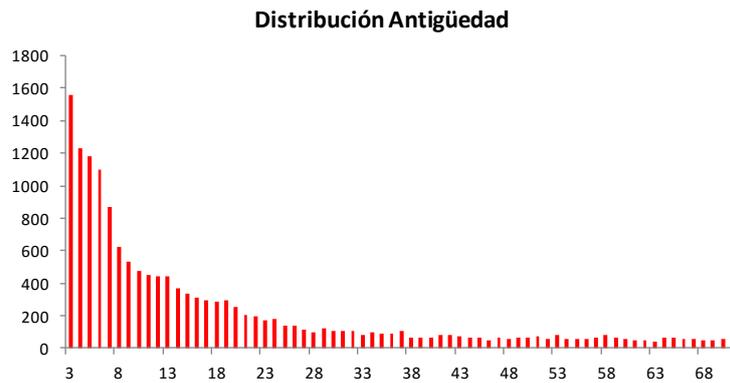


Gráfico 8: Distribución de ANTIGÜEDAD del 90% de la muestra

Mediante la prueba chi-cuadrado se identifican posibles puntos de corte que permitan categorizar la variable de forma adecuada:

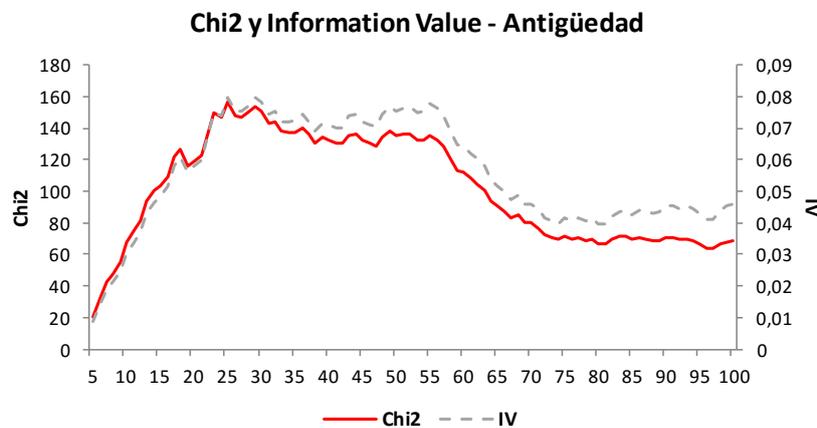


Gráfico 9: Valores del estadístico Chi2 a lo largo de ANTIGÜEDAD

De esta forma, se definen los siguientes grupos, que presentan tasas de mora diferenciadas y suficiente población:

Variable	Tramo	Cientes	Buenos	Malos	% Población	Tasa Mora	% Buenos	% Malos	WOE	IV
Antigüedad	-5	3.977	3.258	719	23,43%	18,08%	22,64%	27,86%	-0,207	0,0940
	5-13	4.929	4.060	869	29,04%	17,63%	28,21%	33,67%	-0,177	
	13-29	3.508	2.954	554	20,67%	15,79%	20,53%	21,46%	-0,045	
	29-53	1.865	1.639	226	10,99%	12,12%	11,39%	8,76%	0,263	
	53+	2.694	2.481	213	15,87%	7,91%	17,24%	8,25%	0,737	

Tabla 12: Tramos propuestos para variable ANTIGÜEDAD

Un análisis ANOVA permite contrastar que los grupos definidos presentan tasas de mora estadísticamente diferentes:

Fuente	DF	Suma de cuadrados	Cuadrado de la media	F-Valor	Pr > F
Among	4	22,433	5,608	43,933	<.0001
Within	16968	2166,087	0,128		

Tabla 13: Tabla ANOVA para variable ANTIGÜEDAD

Los grupos propuestos se presentan a continuación gráficamente, junto con la tasa de mora y la cantidad de clientes por mes de antigüedad. Se observa una tendencia negativa de la tasa de mora a lo largo de la variable y que la variable tiene un poder predictivo débil, pero suficiente:

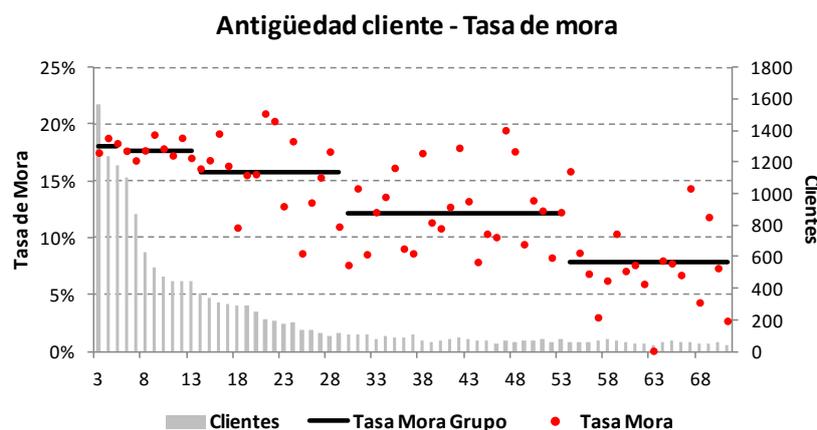


Gráfico 10: Tasa de mora y distribución de clientes de ANTIGÜEDAD

Antigüedad en el empleo

Mide los años de antigüedad del cliente en el empleo actual, medido en meses.

A continuación se muestran los estadísticos descriptivos principales de la antigüedad en el empleo y su distribución. Se observa que es una distribución con cola larga, que acumula gran cantidad de la muestra en los primeros años de antigüedad. Concretamente, el 95% de los clientes tiene menos de 6 años de antigüedad en su empleo actual, con una media ligeramente superior a los 2 años.

Observaciones	16.973
Media	27
Máximo	720
Q3	24
Mediana	12
Q1	12
Mínimo	0

Tabla 14: Estadísticos de la variable ANTIGÜEDAD EN EL EMPLEO

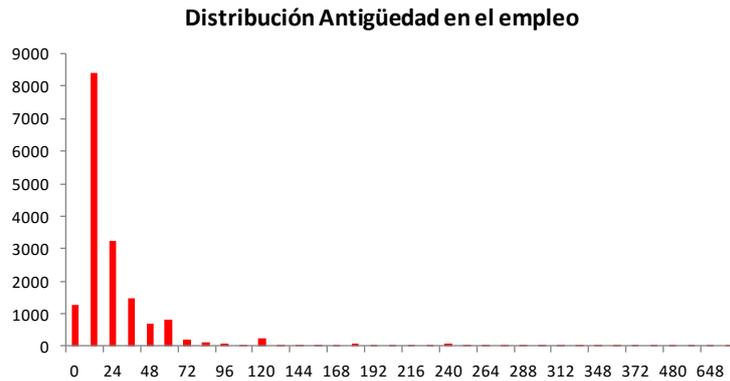


Gráfico 11: Distribución de ANTIGÜEDAD EN EL EMPLEO

Sin embargo, observando de cerca el 95% de la muestra, concentrada en antigüedades menores a 72 meses, la media se sitúa en menos de 2 años, y se pueden identificar clientes sin antigüedad, que corresponde a situación de paro.

Observaciones	16.174
Media	21
Máximo	72
Q3	24
Mediana	12
Q1	12
Mínimo	0

Tabla 15: Estadísticos de ANTIGÜEDAD EN EL EMPLEO, del 95% de la muestra

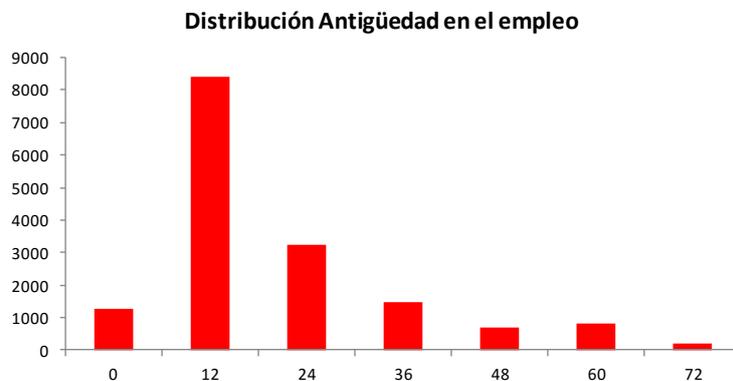


Gráfico 12: Distribución de ANTIGÜEDAD EN EL EMPLEO, del 95% de la muestra

De nuevo, la prueba de chi-cuadrado permite identificar potenciales puntos de corte. En este caso, dada la alta concentración de clientes en unos mismos valores, se ha dado prioridad a formar grupos con suficiente población:

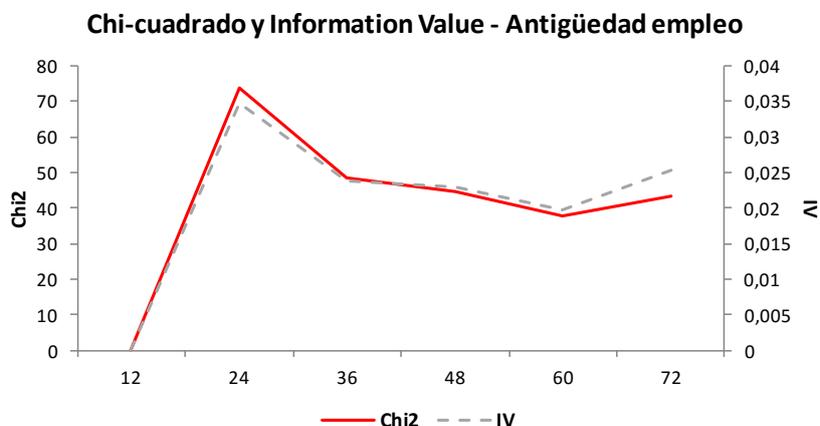


Gráfico 13: Valores del estadístico Chi2 a lo largo de ANTIGÜEDAD EN EL EMPLEO

Entonces, los grupos formados son los siguientes:

Variable	Tramo	Clientes	Buenos	Malos	% Población	Tasa Mora	% Buenos	% Malos	WOE	IV
Antig_Emp	-12	9.693	8.020	1.673	17,26%	17,26%	55,73%	64,82%	-0,151	
	12-36	4.722	4.091	631	13,36%	13,36%	28,43%	24,45%	0,151	0,0397
	36+	2.558	2.281	277	10,83%	10,83%	15,85%	10,73%	0,390	

Tabla 16: Tramos propuestos para ANTIGÜEDAD EN EL EMPLEO

El test ANOVA permite contrastar la hipótesis de igualdad de medias respecto a la tasa de mora de los diferentes grupos propuestos. EL valor del estadístico es suficientemente grande, por lo tanto, los grupos son diferentes en tasa de mora, con comportamiento monótono de ésta, con población suficiente.

Fuente	DF	Suma de cuadrados	Cuadrado de la media	F-Valor	Pr > F
Among	2	10,594	5,297	41,273	<,0001
Within	16970	2177,926	0,128		

Tabla 17: Tabla ANOVA para ANTIGÜEDAD EN EL EMPLEO

Gráficamente, se representa a continuación la cantidad de clientes por año de antigüedad, así como la tasa de mora y los grupos propuestos. Existe una relación inversa entre la tasa de mora y los años que el cliente lleva en el empleo. La variable tiene un poder predictivo débil pero suficiente para ser considerada en el modelo.

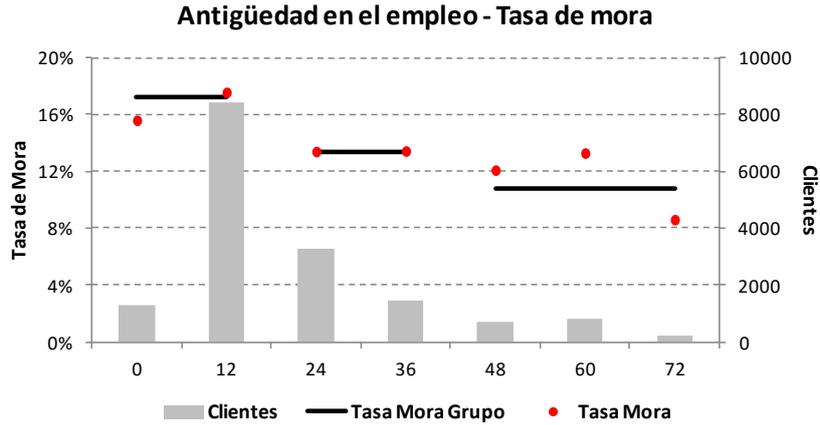


Gráfico 14: Tasa de mora y distribución de clientes de ANTIGÜEDAD EN EL EMPLEO

4.1.3 Variables categóricas

Nacionalidad

Dado que la muestra contiene solicitudes de tarjeta de crédito únicamente de clientes extranjeros, se dispone de la nacionalidad de solicitante.

Hay clientes de 135 nacionalidades distintas. Para agrupar esta variable se han ordenado las nacionalidades en base a su tasa de mora media y se han formado bloques de nacionalidades que posean tasas de mora medias similares y que contuvieran un número razonable de clientes totales y malos. El grupo con menor tasa de mora contiene un 5,37% de malos sobre el total de malos

Los grupos formados son los siguientes:

Variable	Tramo	Clientes	Buenos	Malos	% Población	Tasa Mora	% Buenos	% Malos	WOE	IV
Nacionalidad	1	1.014	766	248	5,97%	24,46%	5,32%	9,61%	-0,591	0,1235
	2	1.370	1.092	278	8,07%	20,29%	7,59%	10,77%	-0,350	
	3	6.264	5.211	1.053	36,91%	16,81%	36,21%	40,80%	-0,119	
	4	4.216	3.598	618	24,84%	14,66%	25,00%	23,94%	0,043	
	5	1.955	1.718	237	11,52%	12,12%	11,94%	9,18%	0,262	
	6	2.154	2.007	147	12,69%	6,82%	13,95%	5,70%	0,895	

Tabla 18: Grupos propuestos para NACIONALIDAD

Las nacionalidades que conforman cada grupo se han incluido en una tabla en el Anexo 1.

El test ANOVA demuestra que los grupos formados presentan tasas de mora medias diferenciadas, puesto que el p-valor del estadístico es menor a 0,05.

Fuente	DF	Suma de cuadrados	Cuadrado de la media	F-Valor	Pr > F
Among	5	30,952	6,190	48,681	<,0001
Within	16967	2157,568	0,127		

Tabla 19: Tabla ANOVA para NACIONALIDAD

A continuación, se puede observar el número de clientes y la tasa de mora por nacionalidad, así como los distintos grupos propuestos:

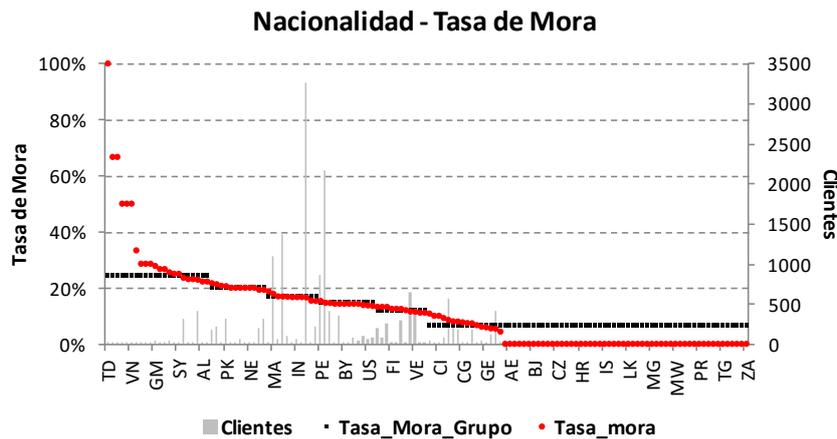


Gráfico 15: Tasa de mora y distribución de clientes de NACIONALIDAD

Se observa que hay ciertos países con una gran concentración de clientes. Este hecho se ha tenido en cuenta a la hora de formar los grupos.

Ratio Activo-Pasivo

El ratio Activo-Pasivo separa a los clientes en función de la relación entre el activo y el pasivo de los clientes. La variable ha sido construida de la siguiente manera:

- El ratio activo pasivo es igual al cociente Activo/Pasivo
- La variable tomará valor 1 si el pasivo del cliente es igual a cero, y el activo es positivo.
- Tomará valor 2 si el ratio Activo/Pasivo es mayor a 1, es decir si tanto el pasivo del cliente como su activo son positivos, pero el activo es mayor.
- Tomará valor 3 si el ratio Activo/Pasivo toma valores entre 0,5 y 1, es decir, el pasivo es igual al activo o hasta dos veces mayor.
- Tomará valor 4 si el ratio Activo/Pasivo toma valores entre 0 y 0,5, es decir, el pasivo de cliente es más de dos veces mayor que el activo.

A continuación, se muestran los grupos propuestos, su tasa de mora y el poder predictivo de la variable:

Variable	Tramo	Clientes	Buenos	Malos	% Población	Tasa Mora	% Buenos	% Malos	WOE	IV
Ratio_AC_PAS	1	779	566	213	4,59%	27,34%	3,93%	8,25%	-0,741	0,1150
	2	3.917	3.084	833	23,08%	21,27%	21,43%	32,27%	-0,410	
	3	572	477	95	3,37%	16,61%	3,31%	3,68%	-0,105	
	4	11.705	10.265	1.440	68,96%	12,30%	71,32%	55,79%	0,246	

Tabla 20: Grupos propuestos para RATIO ACTIVO PASIVO

Se comprueba que a medida que el activo es más pequeño en relación al pasivo del cliente, la tasa de mora media disminuye. Además esta variable artificial tiene un poder predictivo medio.

El test ANOVA permite observar que los grupos definidos tienen tasas de mora medias diferentes:

Fuente	DF	Suma de cuadrados	Cuadrado de la media	F-Valor	Pr > F
Among	3	35,841	11,947	94,176	<.0001
Within	16969	2152,679	0,127		

Tabla 21: Tabla ANOVA para RATIO ACTIVO PASIVO

Profesión

Se dispone de información sobre el tipo de profesión que lleva a cabo el cliente. Se han ordenado los tipos según su tasa de mora y se han agrupado de forma que cada grupo contenga suficientes malos, obteniendo los siguientes resultados:

Variable	Tramo	Clientes	Buenos	Malos	% Población	Tasa Mora	% Buenos	% Malos	WOE	IV
Profesion	1	513	403	110	3,02%	21,44%	2,80%	4,26%	-0,420	0,0433
	2	8.324	6.910	1.414	49,04%	16,99%	48,01%	54,78%	-0,132	
	3	2.677	2.287	390	15,77%	14,57%	15,89%	15,11%	0,050	
	4	4.348	3.778	570	25,62%	13,11%	26,25%	22,08%	0,173	
	5	1.111	1.014	97	6,55%	8,73%	7,05%	3,76%	0,628	

Tabla 22: Grupos propuestos para PROFESIÓN

Como se observa en la tabla, esta variable tiene un poder predictivo aceptable. Las profesiones que conforman cada grupo se han incluido en el Anexo 2.

El test ANOVA confirma que los grupos definidos son diferentes en tasa de mora media, puesto que el p-valor es menor a 0,05:

Fuente	DF	Suma de cuadrados	Cuadrado de la media	F-Valor	Pr > F
Among	4	11,314	2,828	22,043	<.0001
Within	16968	2177,206	0,128		

Tabla 23: Tabla ANOVA para PROFESIÓN

A continuación, el gráfico muestra los clientes y la tasa de mora por profesión, además de la tasa de mora de cada grupo formado anteriormente:

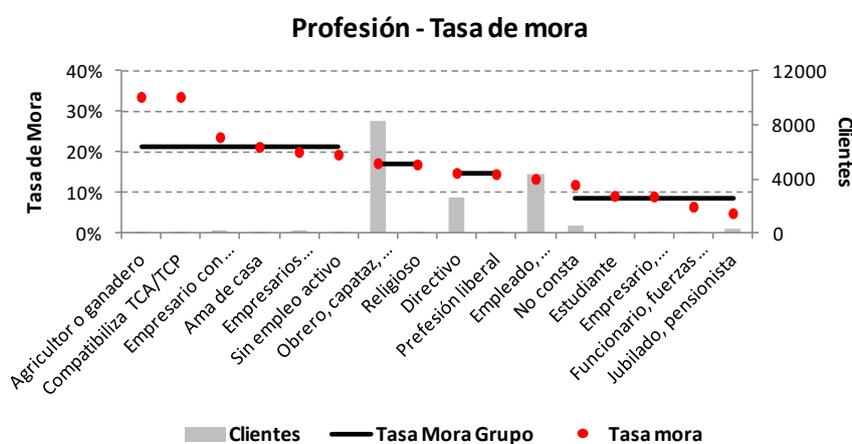


Gráfico 16: Tasa de mora y distribución de clientes de PROFESIÓN

Incidencias

La variable incidencias clasifica a los clientes en cinco grupos principales según su desempeño en créditos anteriores. Son los siguientes:

- A: Sin incidencias
- B: Incidencias leves
- C: Incidencias medias
- D: Incidencias graves
- E: Incidencias muy graves

Estos grupos presentan las siguientes tasas de mora:

Variable	Tramo	Cientes	Buenos	Malos	% Población	Tasa Mora	% Buenos	% Malos	WOE	IV
Incidencias	A	15.606	13.354	2.252	91,95%	14,43%	92,79%	87,25%	0,062	0,0362
	B	625	483	142	3,68%	22,72%	3,36%	5,50%	-0,494	
	C	397	292	105	2,34%	26,45%	2,03%	4,07%	-0,696	
	D	72	51	21	0,42%	29,17%	0,35%	0,81%	-0,831	
	E	273	212	61	1,61%	22,34%	1,47%	2,36%	-0,473	

Tabla 24: Grupos iniciales para INCIDENCIAS

La variable tiene un poder predictivo aceptable, y tasas de mora diferenciadas. Sin embargo, determinados grupos no son suficientemente poblados, por ello, se ha realizado la siguiente transformación:

- C, D, E → Grupo 1: Incidencias medias y graves

- B → Grupo 2: Incidencias leves
- A → Grupo 3: Sin incidencias

De esta forma, se agrupa a los clientes en función a su comportamiento, dando lugar a los siguientes resultados:

Variable	Tramo	Cientes	Buenos	Malos	% Población	Tasa Mora	% Buenos	% Malos	WOE	IV
Incidencias	INCIDENCIAS MEDIAS Y GRAVES	742	555	187	4,37%	25,20%	3,86%	7,25%	-0,631	0,0354
	INCIDENCIAS LEVES	625	483	142	3,68%	22,72%	3,36%	5,50%	-0,494	
	SIN INCIDENCIAS	15.606	13.354	2.252	91,95%	14,43%	92,79%	87,25%	0,062	

Tabla 25: Grupos propuestos para INCIDENCIAS

Estos grupos sí dan como resultado categorías con población suficiente, además de comportamientos diferenciados. La variable, como se puede observar, sigue manteniendo un poder predictivo aceptable y puede ser considerada en el modelo.

El test ANOVA muestra que los grupos tienen tasas de mora medias diferenciadas:

Fuente	DF	Suma de cuadrados	Cuadrado de la media	F-Valor	Pr > F
Among	4	11,314	2,828	22,043	<.0001
Within	16968	2177,206	0,128		

Tabla 26: Tabla ANOVA para INCIDENCIAS

Provincia en la que se solicita el préstamo

Se dispone de información referente a la provincia en la que el cliente ha solicitado la tarjeta. Se ha analizado si la zona geográfica en la que se solicita la tarjeta puede determinar la tasa de mora futura del cliente. Para ello, de nuevo, se han ordenado las provincias según su tasa de mora media, y se han agrupado en función del número de clientes, dando lugar a los siguientes resultados:

Variable	Tramo	Cientes	Buenos	Malos	% Población	Tasa Mora	% Buenos	% Malos	WOE	IV
Provincia	1	1.451	1.172	279	8,55%	19,23%	8,14%	10,81%	-0,283	0,0322
	2	9.146	7.672	1.474	53,89%	16,12%	53,31%	57,11%	-0,069	
	3	3.408	2.905	503	20,08%	14,76%	20,18%	19,49%	0,035	
	4	2.968	2.643	325	17,49%	10,95%	18,36%	12,59%	0,377	

Tabla 27: Grupos propuestos para PROVINCIA

Se observa que los grupos tienen suficiente población y tasas de mora distintas. La variable tiene un poder predictivo aceptable. Las provincias que contiene cada grupo se han incluido en el Anexo 3.

Para comprobar que los grupos tienen tasas de mora diferentes se ha llevado a cabo un test ANOVA:

Fuente	DF	Suma de cuadrados	Cuadrado de la media	F-Valor	Pr > F
Among	3	8,549	2,850	22,182	<,0001
Within	16969	2179,971	0,128		

Tabla 28: Tabla ANOVA para PROVINCIA

A continuación, se muestran los clientes y la tasa de mora de cada grupo definido, junto a los distintos grupos definidos:

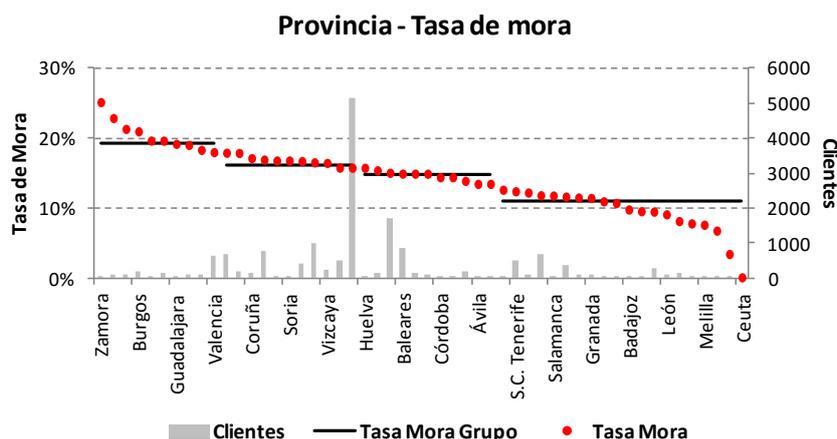


Gráfico 17: Tasa de mora y distribución de clientes de PROVINCIA

Estado Civil

Se dispone de información relativa al estado civil del solicitante de la tarjeta de crédito. Se ha analizado si esta característica puede ser determinante del evento de impago. La tasa de mora y la distribución de esta variable a lo largo de sus categorías es la siguiente:

Variable	Tramo	Clientes	Buenos	Malos	% Población	Tasa Mora	% Buenos	% Malos	WOE	IV
Est_Civil	B	921	790	131	5,43%	14,22%	5,49%	5,08%	0,0783	0,0329
	C	6.099	5.307	792	35,93%	12,99%	36,88%	30,69%	0,1837	
	D	321	281	40	1,89%	12,46%	1,95%	1,55%	0,2310	
	E	391	357	34	2,30%	8,70%	2,48%	1,32%	0,6329	
	F	60	48	12	0,35%	20,00%	0,33%	0,47%	-0,3322	
	H	191	160	31	1,13%	16,23%	1,11%	1,20%	-0,0773	
	N	516	426	90	3,04%	17,44%	2,96%	3,49%	-0,1639	
	S	8.338	6.904	1.434	49,13%	17,20%	47,97%	55,56%	-0,1469	
	V	136	119	17	0,80%	12,50%	0,83%	0,66%	0,2274	

Tabla 29: Grupos iniciales para ESTADO CIVIL

El estado civil por si mismo tiene un poder predictivo adecuado para ser tenido en cuenta en el modelo. Sin embargo, algunas categorías no contienen población suficiente, por ello, es necesario agruparlas. Utilizando la misma metodología que con el resto de grupos, el resultado es como se muestra a continuación:

Variable	Tramo	Cientes	Buenos	Malos	% Población	Tasa Mora	% Buenos	% Malos	WOE	IV
Est_Civil	1	8.338	6.904	1.434	49,13%	17,20%	47,97%	55,56%	-0,147	0,0296
	2	1.688	1.424	264	9,95%	15,64%	9,89%	10,23%	-0,033	
	3	6.099	5.307	792	35,93%	12,99%	36,87%	30,69%	0,184	
	4	848	757	91	5,00%	10,73%	5,26%	3,53%	0,400	

Tabla 30: Grupos propuestos para ESTADO CIVIL

Como resultado de la agrupación, la variable no ha perdido gran parte de su poder explicativo. Los tipos de estado civil que componen cada grupo pueden ser consultados en el Anexo 4.

El test ANOVA confirma que los grupos establecidos son diferentes en tasa de mora media, ya que el estadístico es suficientemente alto:

Fuente	DF	Suma de cuadrados	Cuadrado de la media	F-Valor	Pr > F
Among	3	8,046	2,682	20,873	<,0001
Within	16969	2180,474	0,128		

Tabla 31: Tabla ANOVA para ESTADO CIVIL

Debajo, se puede observar el número de clientes y la tasa de mora de cada estado civil, así como los distintos grupos propuestos:

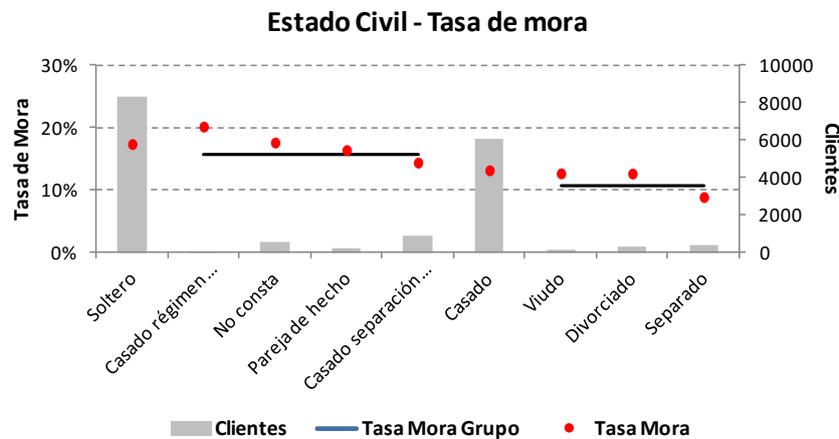


Gráfico 18: Tasa de mora y distribución de clientes de ESTADO CIVIL

4.2 Análisis multivariante

Una vez que la adecuación de incluir cada variable en el modelo ha sido analizada por separado, es necesario analizar todas las variables en conjunto. Esto se hará a través de la matriz de correlaciones, tanto de las variables numéricas tal cual, como de los WOE's calculados en la sección anterior.

Si bien en el apartado anterior sólo se han incluido los análisis de las variables finalmente introducidas en el modelo, en la matriz de correlaciones aparecerán todas las variables estudiadas:

Matriz de correlaciones de variables numéricas:

	Edad	ANTIGUEDAD	IMPORTE	TRIAD	PASIVO	ACTIVO	ING	ANTIG_EMP
Edad	100,00%	35,59%	24,16%	7,51%	5,58%	6,29%	3,67%	31,93%
ANTIGUEDAD	35,59%	100,00%	25,09%	18,77%	6,84%	13,21%	4,61%	37,33%
IMPORTE	24,16%	25,09%	100,00%	4,59%	7,52%	19,34%	25,14%	28,08%
TRIAD	7,51%	18,77%	4,59%	100,00%	1,08%	6,69%	1,60%	6,93%
PASIVO	5,58%	6,84%	7,52%	1,08%	100,00%	6,03%	2,75%	7,47%
ACTIVO	6,29%	13,21%	19,34%	6,69%	6,03%	100,00%	9,60%	5,79%
ING	3,67%	4,61%	25,14%	1,60%	2,75%	9,60%	100,00%	4,37%
ANTIG_EMP	31,93%	37,33%	28,08%	6,93%	7,47%	5,79%	4,37%	100,00%

Tabla 32: Matriz de correlaciones variables numéricas

La matriz de correlaciones de los WOE's calculados en el análisis univariante:

	WOE_EDAD	WOE_ANTIGUEDAD	WOE_TRIAD	WOE_PASIVO	WOE_ACTIVO	WOE_ratio_AC_PAS
WOE_EDAD	100,00%	23,38%	11,75%	10,41%	-2,51%	-1,29%
WOE_ANTIGUEDAD	23,38%	100,00%	16,71%	11,74%	-8,51%	-8,02%
WOE_TRIAD	11,75%	16,71%	100,00%	34,80%	14,49%	19,60%
WOE_PASIVO	10,41%	11,74%	34,80%	100,00%	33,25%	54,01%
WOE_ACTIVO	-2,51%	-8,51%	14,49%	33,25%	100,00%	79,33%
WOE_ratio_AC_PAS	-1,29%	-8,02%	19,60%	54,01%	79,33%	100,00%
WOE_ANTIG_EMP	18,36%	26,15%	8,08%	6,23%	-4,20%	-5,00%
WOE_ING	13,99%	16,48%	7,34%	17,75%	-4,73%	-2,33%
WOE_IMPORTE	20,66%	21,25%	7,11%	17,70%	-5,53%	-4,54%
WOE_INCIDENCIAS	-6,83%	-12,48%	9,30%	18,49%	23,36%	28,63%
WOE_NACIONALIDAD	17,91%	18,57%	11,82%	13,71%	-2,54%	-0,85%
WOE_PROVINCIA	5,04%	8,76%	7,05%	7,17%	0,76%	2,77%
WOE_NIV_ESTUD	4,88%	10,55%	8,67%	9,88%	-1,44%	-0,10%
WOE_EST_CIVIL	31,40%	10,45%	2,54%	3,74%	-3,98%	-4,29%
WOE_VIV_HAB	14,93%	16,18%	8,90%	9,50%	0,33%	1,66%
WOE_PROFESION	11,87%	19,72%	9,61%	8,76%	-2,09%	-1,92%
WOE_TIPO_CONT	9,09%	13,42%	4,20%	2,72%	-7,27%	-8,10%

	WOE_ANTIG_EMP	WOE_ING	WOE_IMPORTE	WOE_INCIDENCIAS	WOE_NACIONALIDAD	WOE_PROVINCIA
WOE_EDAD	18,36%	13,99%	20,66%	-6,83%	17,91%	5,04%
WOE_ANTIGUEDAD	26,15%	16,48%	21,25%	-12,48%	18,57%	8,76%
WOE_TRIAD	8,08%	7,34%	7,11%	9,30%	11,82%	7,05%
WOE_PASIVO	6,23%	17,75%	17,70%	18,49%	13,71%	7,17%
WOE_ACTIVO	-4,20%	-4,73%	-5,53%	23,36%	-2,54%	0,76%
WOE_ratio_AC_PAS	-5,00%	-2,33%	-4,54%	28,63%	-0,85%	2,77%
WOE_ANTIG_EMP	100,00%	11,72%	26,53%	-9,62%	9,53%	5,06%
WOE_ING	11,72%	100,00%	70,16%	-6,68%	20,22%	3,82%
WOE_IMPORTE	26,53%	70,16%	100,00%	1,82%	19,61%	5,29%
WOE_INCIDENCIAS	-9,62%	-6,68%	1,82%	100,00%	-2,32%	0,08%
WOE_NACIONALIDAD	9,53%	20,22%	19,61%	-2,32%	100,00%	12,08%
WOE_PROVINCIA	5,06%	3,82%	5,29%	0,08%	12,08%	100,00%
WOE_NIV_ESTUD	2,72%	19,79%	16,79%	-0,12%	22,33%	3,06%
WOE_EST_CIVIL	6,67%	7,02%	15,42%	-5,96%	6,33%	3,56%
WOE_VIV_HAB	13,76%	9,69%	12,74%	-1,84%	10,77%	6,22%
WOE_PROFESION	8,80%	11,36%	13,84%	-3,87%	23,76%	8,18%
WOE_TIPO_CONT	32,76%	9,21%	19,97%	-8,66%	11,80%	1,95%

	WOE_NIV_ESTUD	WOE_EST_CIVIL	WOE_VIV_HAB	WOE_PROFESION	WOE_TIPO_CONT
WOE_EDAD	4,88%	31,40%	14,93%	11,87%	9,09%
WOE_ANTIGUEDAD	10,55%	10,45%	16,18%	19,72%	13,42%
WOE_TRIAD	8,67%	2,54%	8,90%	9,61%	4,20%
WOE_PASIVO	9,88%	3,74%	9,50%	8,76%	2,72%
WOE_ACTIVO	-1,44%	-3,98%	0,33%	-2,09%	-7,27%
WOE_ratio_AC_PAS	-0,10%	-4,29%	1,66%	-1,92%	-8,10%
WOE_ANTIG_EMP	2,72%	6,67%	13,76%	8,80%	32,76%
WOE_ING	19,79%	7,02%	9,69%	11,36%	9,21%
WOE_IMPORTE	16,79%	15,42%	12,74%	13,84%	19,97%
WOE_INCIDENCIAS	-0,12%	-5,96%	-1,84%	-3,87%	-8,66%
WOE_NACIONALIDAD	22,33%	6,33%	10,77%	23,76%	11,80%
WOE_PROVINCIA	3,06%	3,56%	6,22%	8,18%	1,95%
WOE_NIV_ESTUD	100,00%	-2,60%	4,00%	19,07%	7,78%
WOE_EST_CIVIL	-2,60%	100,00%	6,74%	1,31%	3,98%
WOE_VIV_HAB	4,00%	6,74%	100,00%	12,72%	7,94%
WOE_PROFESION	19,07%	1,31%	12,72%	100,00%	18,01%
WOE_TIPO_CONT	7,78%	3,98%	7,94%	18,01%	100,00%

Tabla 33: Matriz de correlaciones WOE

Se debe tener especial cuidado con la variable pasivo y el ratio activo/Pasivo, sin embargo, la correlación de estas variables no alcanza el 70%, que podría considerarse problemático. El resto de variables no presentan correlaciones altas, por lo tanto, se introducirán en la regresión logística stepwise.

4.3 Regresión logística

Una vez que se han analizado las variables tanto numéricas como categóricas de forma separada y conjunta, serán introducidas en un modelo de regresión logística con un método de selección por pasos (stepwise).

La regresión logística es una de las metodologías más extendidas en la construcción de scorecards (Mester, 1997). La razón por la que se utiliza este tipo de regresión es que la variable dependiente es una probabilidad, por lo que debe estar acotada al intervalo [0,1], mientras que en una regresión lineal, la combinación lineal de variables independientes puede tomar valores en toda la recta $(-\infty, +\infty)$. Mediante la función de enlace logit, la variable dependiente de la regresión será una función no lineal de la probabilidad (Thomas, 2000).

$$\text{logit}(p) = \ln(\text{Odds}) = \ln\left(\frac{p}{1-p}\right) = \pi$$

La variable dependiente representa la ocurrencia o no de un suceso, por tanto, está codificada con 1 o 0. La regresión logística estima la variable dependiente en términos de probabilidad utilizando la función logit (García et al., 2005).

Una de las ventajas que presenta esta regresión es que permite que las variables independientes sean cuantitativas, discretas o continuas; o categóricas, sin que hay ninguna restricción. No requiere el supuesto de normalidad de los errores (Nieto, 2010).

La capacidad predictiva se mide mediante la comparación entre el grupo de pertenencia observado y estimado por el modelo, que clasifica a los individuos en función de un punto de corte establecido, que por defecto será 0,5. Es decir, si la probabilidad estimada es mayor a 0,5, el individuo será clasificado como 1, y viceversa. La capacidad predictiva se mide por el porcentaje de individuos que siendo 0 o 1, el modelo los clasifica como tal.

En este caso, la variable dependiente representa el evento de impago:

- Default=1 si el individuo no ha atendido correctamente a los pagos. El cliente se clasifica como malo.
- Default=0 si el individuo ha pagado correctamente el crédito. El cliente se clasifica como bueno.

Es decir, el modelo de regresión logística pretenderá predecir la probabilidad de default.

La ecuación del modelo será la siguiente:

$$\pi = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_i x_i$$

donde $p=P(\text{Default}=1|X)$

Mediante la ecuación anterior, β_i representa el efecto que se produce en la variable $\text{logit}(p)$ cuando la variable independiente i aumenta en una unidad. Entonces, la probabilidad estimada de default:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_i x_i)}}$$

Como variables independientes se utilizarán los WOE's calculados anteriormente y como método de selección de variables se utilizará el método stepwise, que consiste en una secuencia de estadísticos F que controlan la inclusión o exclusión de variables (Leung et al., 2008).

Los resultados de la regresión logística son los siguientes:

Variable	Beta	P-Valor
Intercept	-1,73	<,0001
WOE_ANTIGUEDAD	-0,6497	<,0001
WOE_TRIAD	-0,7064	<,0001
WOE_PASIVO	-0,4126	<,0001
WOE_ratio_AC_PAS	-0,5214	<,0001
WOE_ANTIG_EMP	-0,6422	<,0001
WOE_INCIDENCIAS	-0,6907	<,0001
WOE_NACIONALIDAD	-0,7528	<,0001
WOE_PROVINCIA	-0,5361	<,0001
WOE_EST_CIVIL	-0,8826	<,0001
WOE_PROFESION	-0,5139	<,000

Tabla 34: Resultados del modelo de regresión logística inicial

Asociación de probabilidades predichas y respuestas observadas			
Concordancia de porcentaje	70,2	D de Somers	0,403
Discordancia de porcentaje	29,8	Gamma	0,403
Porcentaje ligado	0	Tau-a	0,104
Pares	37.145.752	c	0,702

Tabla 35: Estadísticos del modelo de regresión logística inicial

Se observa una relación negativa entre los WOE's y la función logit. Es el resultado esperado, puesto que al aumentar la proporción de buenos sobre la de malos, y por tanto, aumentar el WOE, disminuye el cociente entre la probabilidad de hacer default y la probabilidad de no hacer default.

Por ejemplo, en el caso de la puntuación TRIAD, cuyo β es igual a -0,7064:

Variable	Tramo	Clientes	Buenos	Malos	% Población	Tasa Mora	% Buenos	% Malos	WOE	IV
Triad	-650	8.304	6.610	1.694	48,92%	20,40%	45,93%	65,63%	-0,357	0,2507
	650-675	3.967	3.406	561	23,37%	14,14%	23,67%	21,74%	0,085	
	675-700	2.848	2.605	243	16,78%	8,53%	18,10%	9,41%	0,654	
	700+	1.854	1.771	83	10,92%	4,48%	12,31%	3,22%	1,342	

Tabla 36: Grupos propuestos para la variable TRIAD, Ejemplo

Un individuo que esté en el grupo de mayor puntuación TRIAD, entrará en la ecuación con un mayor WOE, y tendrá como resultado una menor probabilidad de default.

El modelo tiene un buen poder predictivo, su estadístico c (Curva ROC) es igual a 0,702. Es decir, el modelo es capaz de clasificar correctamente al 70,2% de los individuos. Un modelo

perfecto tendría un valor de c de 100%, mientras que un modelo que clasificara aleatoriamente a los individuos tendría un valor de c de 50%.

La D de Somers, que puede tomar valores entre -1 (todos los pares discordantes) y 1 (todos los pares concordantes), muestra un buen valor de asociación. Se calcula como el porcentaje de pares concordantes menos el porcentaje de pares discordantes dividido entre el total de pares con distintas respuestas. En este caso $D=(70,2-29,8)/100=0,403$.

El test de Goodman-Kruskal Gamma puede tomar valores entre -1 (no asociación) y 1 (asociación perfecta). Al igual que la D de Somers, muestra un buen valor de asociación.

A continuación, se muestra la matriz de confusión:

		Predicho	
		Buenos	Malos
Observado	Buenos	10.477 (61,73%)	3.915 (23,07%)
	Malos	1.143 (6,73%)	1.438 (8,47%)

Tabla 37: Matriz de confusión del modelo inicial

Se observa un 6,73% de malos clientes que el modelo ha clasificado como buenos.

Las variables que no han sido incluidas en el modelo por el algoritmo stepwise pueden consultarse en el Anexo 5.5.

4.4 Obtención de los scores y Scorecard

Una vez que las variables han sido categorizadas, los WOE's han sido calculados y los betas del modelo han sido obtenidos, es posible reescalar las probabilidades estimadas en puntuaciones, de acuerdo a la siguiente fórmula (Nieto, 2010):

$$Score = \frac{Offset}{n} + \frac{Factor}{\ln(2)} \cdot \left(\frac{\beta_0}{n} - \beta_{ij} \cdot WOE_{ij} \right)$$

Donde i =Variable, j =Categoría, n es el número de variables, $Offset=500$, $Factor=20$

Offset es un término de traslación y factor un término de reescalamiento. Son valores impuestos por la entidad que permite estandarizar las puntuaciones y comparar unos modelos de scoring con otros. El offset determina la puntuación equivalente a un Odds 1:1 (igual proporción de buenos y de malos), y el factor determina el incremento en la puntuación

necesario para doblar los Odds. En este caso, una puntuación de 500 equivalen a unos Odds 1:1, y un incremento de 20 puntos sobre los 500 equivalen a unos Odds 2:1.

Dados los WOES calculados, las betas calculados y los parámetros definidos, se ha obtenido la siguiente tarjeta de puntuación o scorecard:

Variable	Tramo	Tasa Mora	WOE	Score	Beta	IV	SCR
Triad	-650	20,40%	-0,357	38	-0,7064	0,2507	35,32%
	650-675	14,14%	0,085	47			
	675-700	8,53%	0,654	58			
	700+	4,48%	1,342	72			
Pasivo	-100	27,41%	-0,744	36	-0,4126	0,2134	9,51%
	100-800	17,15%	-0,143	43			
	800-2000	12,58%	0,220	48			
	2000-3700	10,78%	0,394	50			
	3700+	6,81%	0,898	56			
Nacionalidad	1	24,46%	-0,591	32	-0,7528	0,1235	18,77%
	2	20,29%	-0,350	37			
	3	16,81%	-0,119	42			
	4	14,66%	0,043	46			
	5	12,12%	0,262	51			
	6	6,82%	0,895	64			
Ratio_AC_PAS	1	27,34%	-0,741	34	-0,5214	0,1150	6,72%
	2	21,27%	-0,410	39			
	3	16,61%	-0,105	43			
	4	12,30%	0,246	49			
Antigüedad	-5	18,08%	-0,207	41	-0,6497	0,0940	10,98%
	5-13	17,63%	-0,177	42			
	13-29	15,79%	-0,045	44			
	29-53	12,12%	0,263	50			
	53+	7,91%	0,737	59			
Profesion	1	21,44%	-0,420	39	-0,5139	0,0433	2,95%
	2	16,99%	-0,132	43			
	3	14,57%	0,050	46			
	4	13,11%	0,173	48			
	5	8,73%	0,628	54			
Antig_Emp	-12	17,26%	-0,151	42	-0,6422	0,0397	4,13%
	12-36	13,36%	0,151	48			
	36+	10,83%	0,390	52			
Incidencias	INCIDENCIAS MEDIAS Y GRAVES	25,20%	-0,631	32	-0,6907	0,0354	3,30%
	INCIDENCIAS LEVES	22,72%	-0,494	35			
	SIN INCIDENCIAS	14,43%	0,062	46			
Provincia	1	19,23%	-0,283	41	-0,5361	0,0322	2,28%
	2	16,12%	-0,069	44			
	3	14,76%	0,035	46			
	4	10,95%	0,377	51			
Est_Civil	1	17,20%	-0,147	41	-0,8826	0,0296	6,03%
	2	15,64%	-0,033	44			
	3	12,99%	0,184	50			
	4	10,73%	0,400	55			

Tabla 38: Scorecard modelo inicial

Se puede observar cómo a mayor WOE, (que no es más que la transformación logarítmica del ratio entre la proporción de buenos y la proporción de malos) mayor puntuación. Esta relación

se cumple de igual manera con la tasa de mora y la puntuación. La tasa de mora está calculada como número malos entre total de individuos en cada categoría.

La máxima puntuación posible es de 558, y la mínima 376.

El SCR (Score Contribution Ratio) pretende medir la contribución de cada variable al modelo de scoring. Se calcula de la siguiente manera:

$$SCR_i = \frac{\sum_{j=1}^{m_i} \left((Score_{ij}^2 \cdot \% Población_{ij}) - (Score_{ij} \cdot \% Población_{ij})^2 \right)}{\sum_{i=1}^n \sum_{j=1}^{m_i} \left((Score_{ij}^2 \cdot \% Población_{ij}) - (Score_{ij} \cdot \% Población_{ij})^2 \right)}$$

Donde i =Variable, j =Categoría, n =número de variables, m_j =número de categorías de la variable

$$i, \text{ y } \% Población_{ij} = \frac{Individuos_j}{Total Población}$$

La variable que más contribuye a la puntuación del individuo es la puntuación en el Scoring TRIAD, a continuación su nacionalidad y su antigüedad como cliente.

4.5 Poder predictivo del modelo inicial

Una vez que cada cliente ha sido asignado una puntuación, es posible observar cómo el modelo ha sido capaz de separar a los clientes buenos de los malos.

Para evaluar el poder predictivo se utilizan habitualmente el estadístico KS y el Gini (Rezác et al., 2011).

4.5.1 Test de Kolmogorov-Smirnov

Es un test no paramétrico de bondad de ajuste que permite comprobar si dos muestras independientes tienen la misma distribución. El estadístico se calcula como la máxima diferencia absoluta entre las funciones de distribución acumuladas empíricas de ambas muestras.

Las funciones de distribución empíricas vienen dadas por las fórmulas:

$$F_{BUENOS}(a) = \frac{1}{n} \sum_{i=1}^n I(s_i \leq a \wedge Default = 0) \quad \forall a \in [L, H]$$

$$F_{MALOS}(a) = \frac{1}{m} \sum_{i=1}^m I(s_i \leq a \wedge Default = 1) \quad \forall a \in [L, H]$$

Donde n es el número total de clientes buenos, m es el número de clientes malos, s_i es el score del cliente i , $I(\text{verdadero})=1$, $I(\text{Falso})=0$ es la función condicional. L es el mínimo score y H el máximo score que el modelo puede proporcionar. a es el score cuya función de distribución se quiere calcular.

El estadístico de Kolmogorov-Smirnov viene dado por la expresión:

$$KS = \max_{a \in [L, H]} |F_{MALOS}(a) - F_{BUENOS}(a)|$$

4.5.2 Índice de GINI

Es una prueba estadística para medir la desigualdad entre dos poblaciones. En este caso, para medir la desigualdad entre las poblaciones de buenos y malos. La curva de Lorenz de las distribuciones F_{BUENOS} y F_{MALOS} es la siguiente:

$$\mathcal{L}(F_{MALOS}, F_{BUENOS}) = \{(u, v) | u = F_{MALOS}(a) \text{ y } v = F_{BUENOS}(a); a \in (L, H)\}$$

Si las distribuciones de buenos y malos son diferentes, la curva de Lorenz será cóncava y hacia arriba. Si ambas distribuciones son iguales, la curva de Lorenz es la recta $u=v$, donde $u \in (0,1)$.

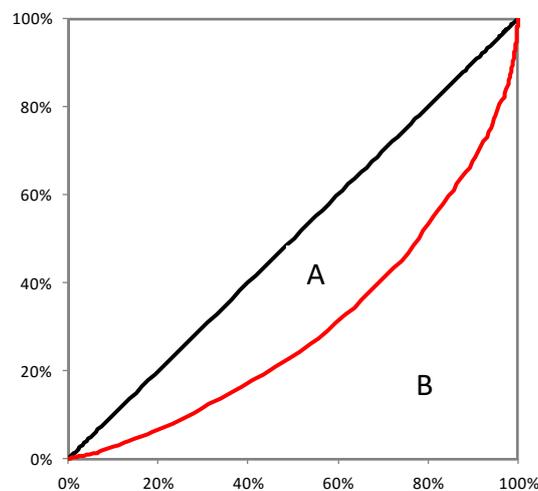


Gráfico 19: Curva de Lorenz y Gini

Por tanto, cuanto más se separe la curva de Lorenz de la recta $u=v$ mayor será la diferencia entre ambas distribuciones. El índice de GINI se describe con la siguiente expresión:

$$Gini = \frac{A}{A + B} = 2A$$

Cuando no se dispone de las funciones de distribución, pero se dispone de una muestra de ambas poblaciones, el índice de Gini puede aproximarse de la siguiente forma:

$$Gini = 1 - \sum_{k=2}^{n+m} [(F_{MALOS}(a) - F_{MALOS}(a - 1)) \cdot (F_{BUENOS}(a) + F_{BUENOS}(a - 1))]$$

4.5.3 Poder predictivo en la muestra de Desarrollo

La muestra con la que se ha desarrollado el modelo, compuesta de 16.973 clientes, presenta las siguientes distribuciones de buenos y malos:

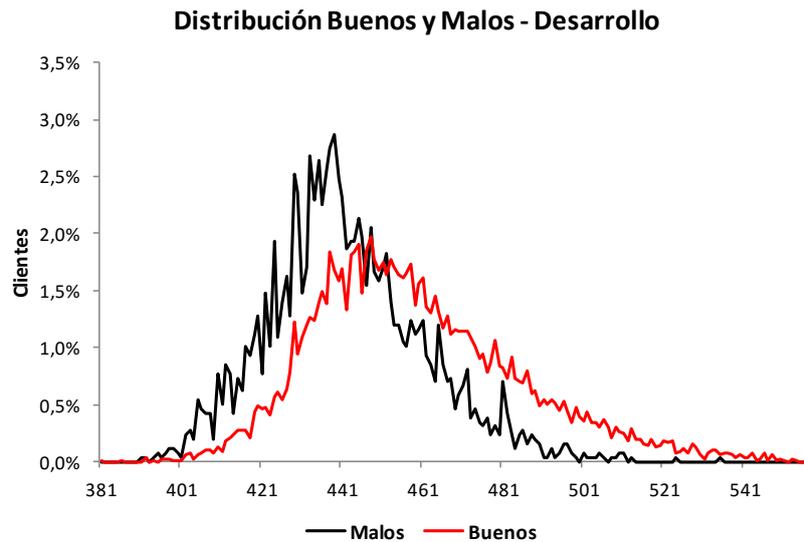


Gráfico 20: Distribución clientes en muestra de DESARROLLO

El modelo aplicado a la muestra de desarrollo presenta un KS de 29,35%. El score que maximiza la diferencia entre la distribución de buenos y malos es 446.

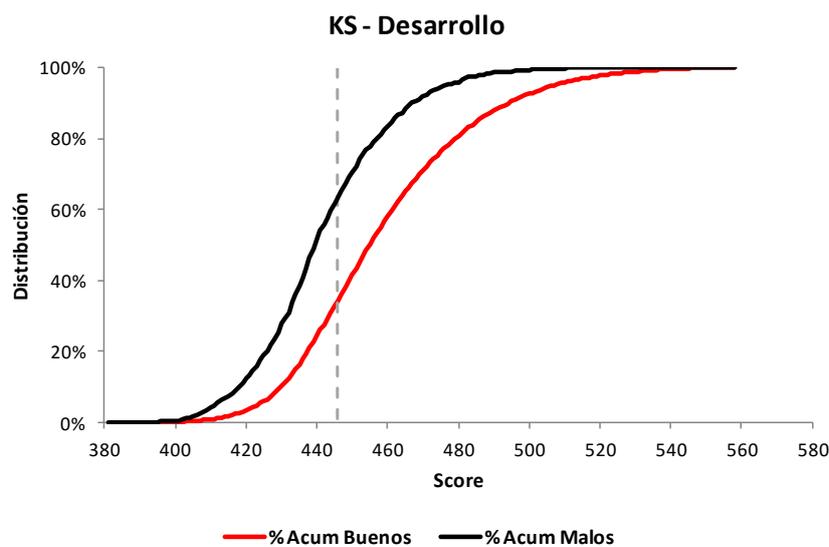


Gráfico 21: Distribuciones acumuladas clientes en muestra de DESARROLLO

A continuación, podemos ver la curva de Lorenz, cuyo índice de Gini es 40,37%:

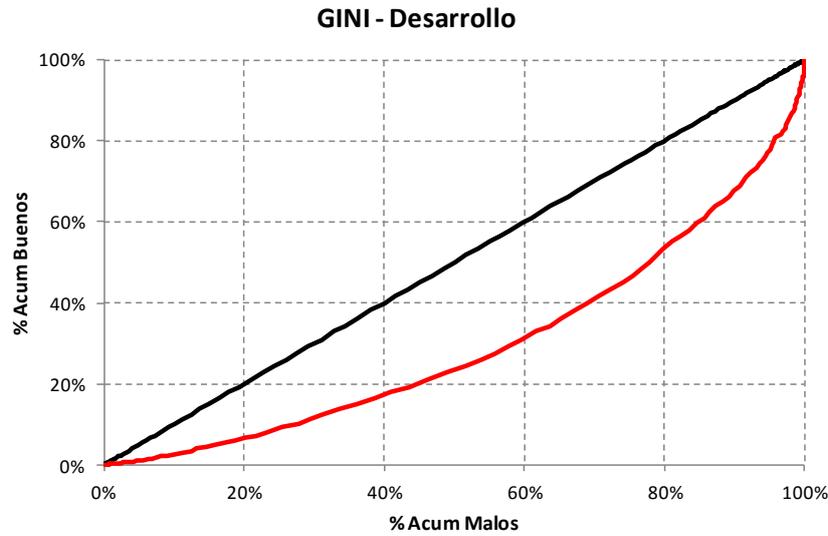


Gráfico 22: Curva de Lorenz en muestra de DESARROLLO

A continuación se muestra la distribución total de individuos por score, así como la tasa de mora. Se observa una tendencia decreciente muy pronunciada, por lo que se puede deducir que el score proporcionado es un buen indicador de la tasa de mora:

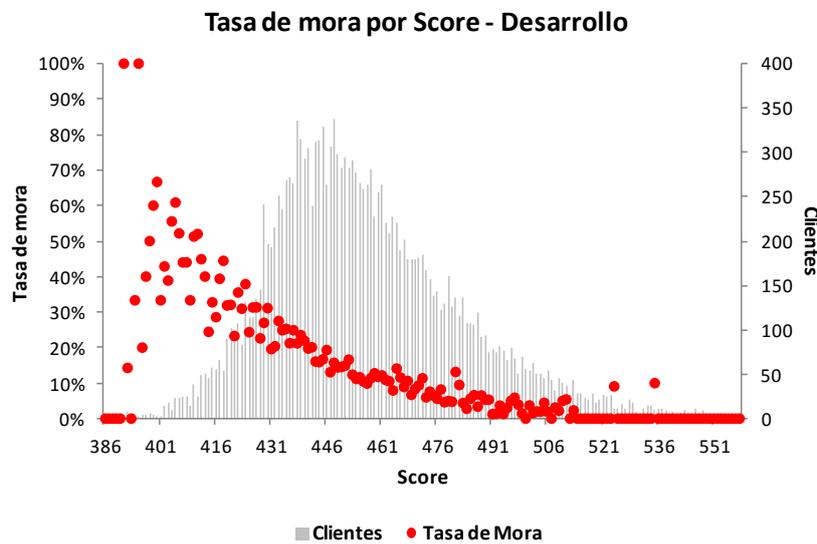


Gráfico 23: Tasa de mora por score del modelo inicial en muestra de DESARROLLO

4.5.4 Poder predictivo en la muestra Hold-Out

La muestra Hold-Out se compone del 30% del total de individuos a los que les fue concedida la tarjeta de crédito, un total de 7.274. El objetivo del modelo es predecir la tasa de mora y poder tomar una decisión en cuanto a aceptar o denegar un préstamo solicitado por clientes de

características similares a los incluidos en la muestra de desarrollo. Por ello, el modelo se va a validar con esta muestra, que ha sido mantenida a parte en todo momento.

Un buen modelo de scoring, que sea robusto, debe ser capaz de separar el tipo de cliente en cualquier muestra.

El siguiente gráfico muestra la distribución de la población de clientes buenos y clientes malos:

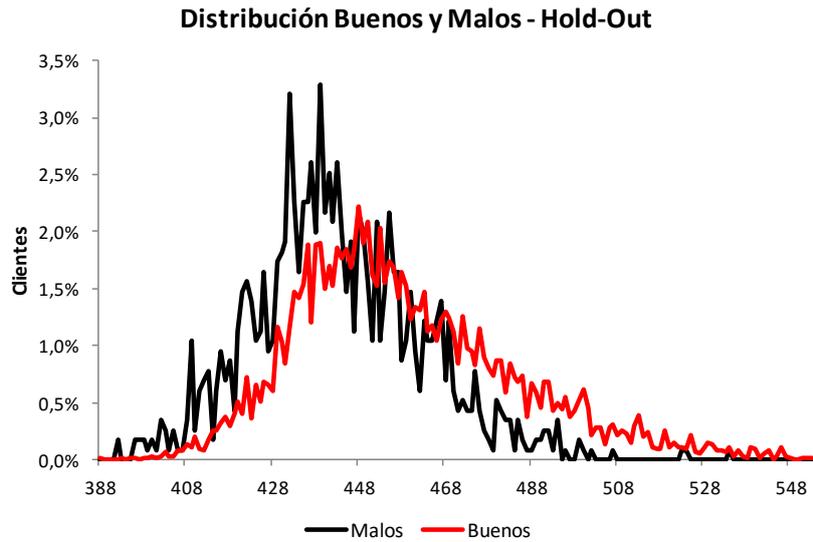


Gráfico 24: Dsitribución de clientes en muestra HOLDOUT

El KS del modelo para esta muestra es de 25,14%. Como es de esperar, el modelo a perdido parte de su poder predictivo respecto a la muestra de desarrollo:

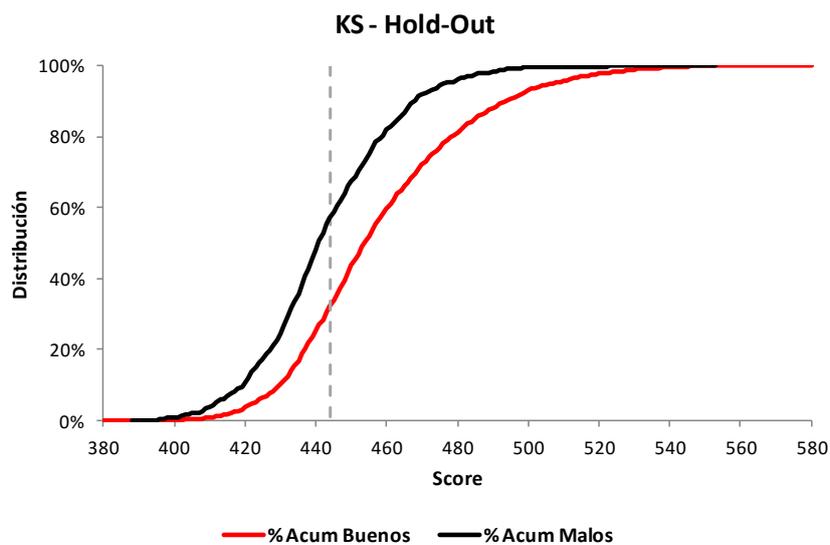


Gráfico 25: Distribuciones acumuladas de clientes en muestra HOLDOUT

El modelo también ha perdido poder predictivo en lo que al índice Gini se refiere, puesto que en este caso es de un 35,32%. En el gráfico siguiente se observa la curva de Lorenz:

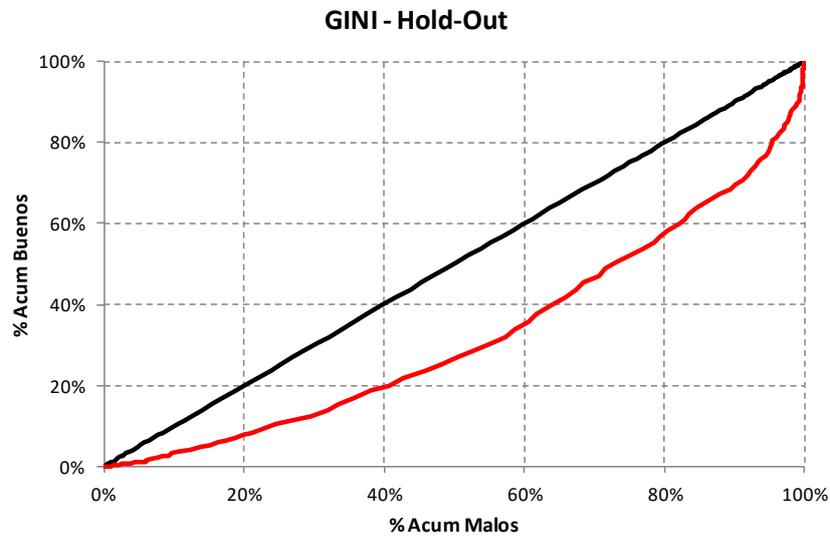


Gráfico 26: Curva de Lorenz en muestra HOLDOUT

A pesar de esa leve pérdida de poder predictivo mostrada en ambos estadísticos, se observa una tendencia decreciente muy pronunciada de la tasa de mora respecto al score, por lo que este modelo está prediciendo adecuadamente:

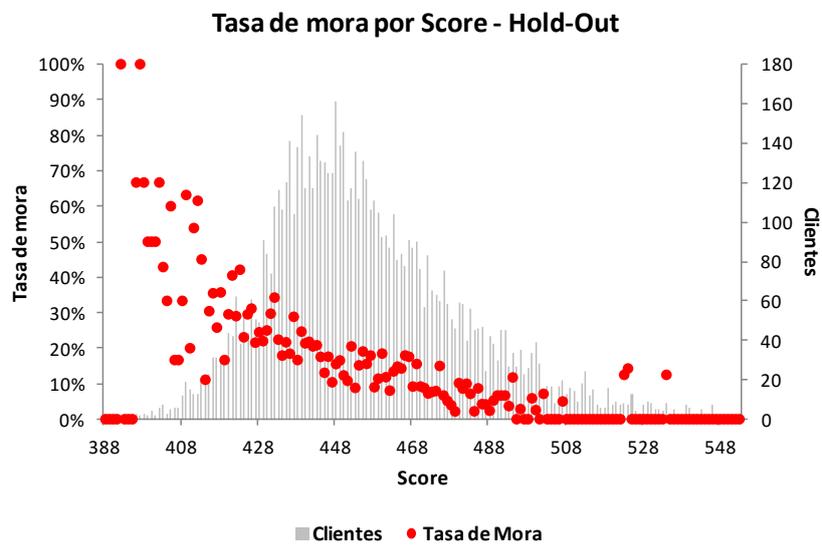


Gráfico 27: Tasa de mora por score del modelo inicial en muestra HOLDOUT

4.5.5 Poder predictivo en la muestra Out of Time

La muestra Out of Time consiste en solicitudes posteriores a la muestra de desarrollo. En este caso, se compone de 1.258 clientes que solicitaron una tarjeta un año después de los clientes con los que se ha desarrollado el modelo.

La distribución de clientes buenos y malos es la siguiente:

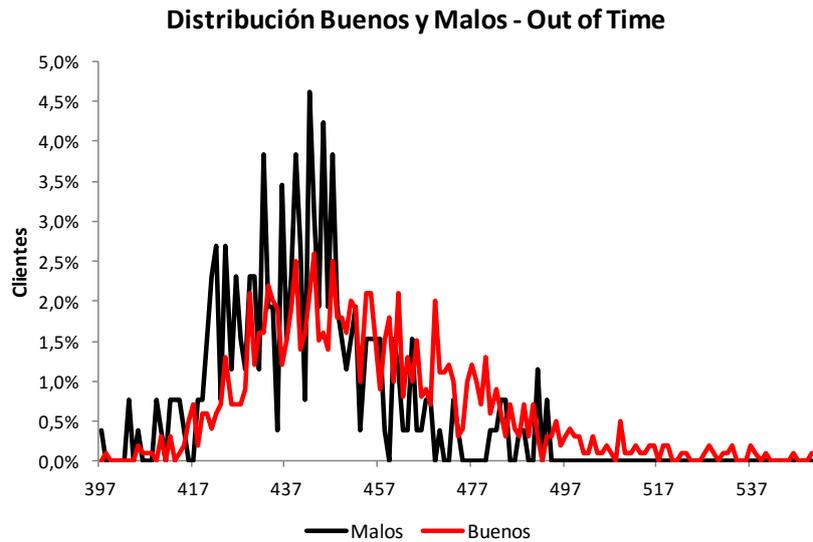


Gráfico 28: Dsitribución de clientes en muestra OUT OF TIME

De nuevo, el modelo ha perdido parte de su poder predictivo respecto a la muestra desarrollo, puesto que presenta un KS de 26,68%. En el siguiente gráfico, se pueden las distribuciones acumuladas de ambas poblaciones:

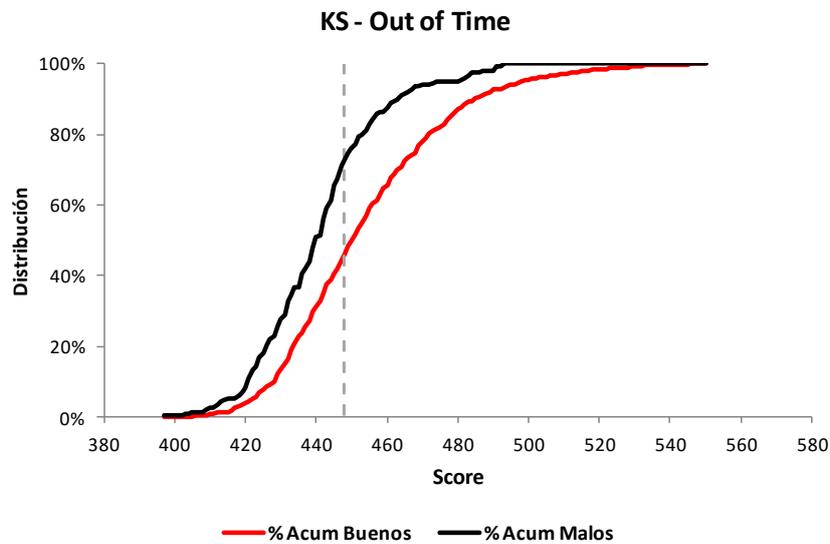


Gráfico 29: Distribuciones acumuladas de clientes en muestra OUT OF TIME

El índice de Gini sufre un descenso respecto a la muestra de desarrollo, siendo 32,73%. La curva de Lorenz es la siguiente:

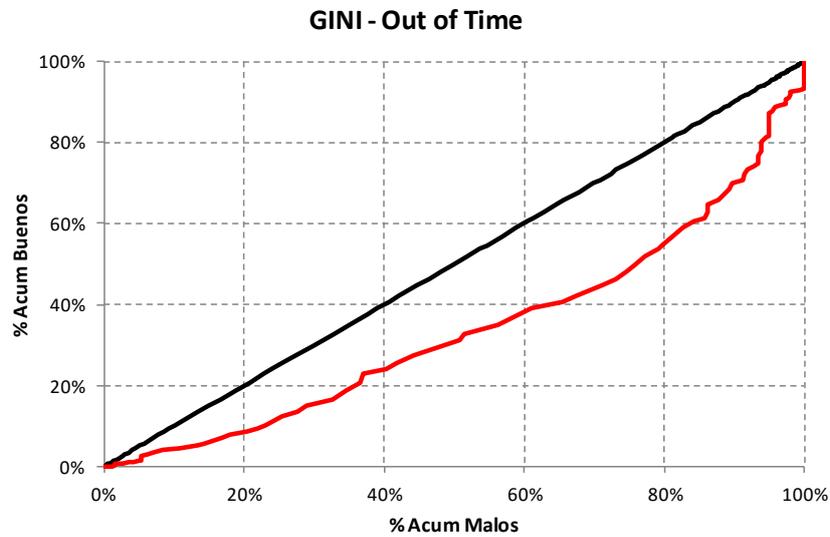


Gráfico 30: Curva de Lorenz en muestra OUT OF TIME

A pesar de la pérdida de poder predictivo, el modelo sigue discriminando bien el tipo de cliente, pero la muestra Out of Time presenta una tendencia menos definida que el resto de muestras. En la siguiente sección se comprobará si esto se debe a un cambio en la distribución de buenos y malos a lo largo del score.

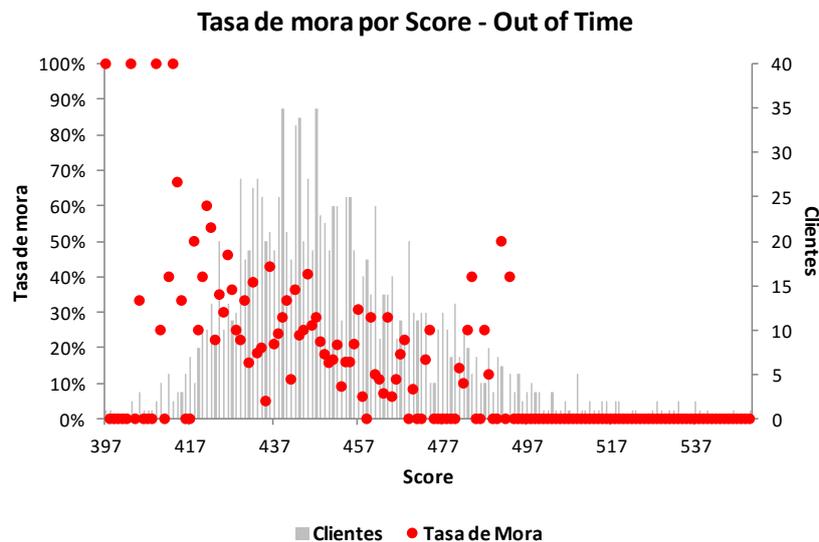


Gráfico 31: Tasa de mora por score del modelo inicial en muestra OUT OF TIME

La siguiente tabla muestra un resumen de los resultados de poder predictivo obtenidos:

Muestra	Observaciones	Tasa Mora	KS	GINI
Desarrollo	20.118	15,21%	30,99%	42,53%
Hold-Out	7.274	15,86%	24,00%	34,88%
Out of Time	1.258	20,67%	25,73%	31,42%

Gráfico 32: Resumen de resultados poder predictivo del modelo inicial

La muestra Out of Time tiene un volumen considerablemente más bajo que el resto de muestras, y su tasa de mora es significativamente más alta.

4.6 Estabilidad del modelo inicial

Dado que se dispone de una muestra de clientes posterior al tiempo en el que el modelo ha sido desarrollado, es posible verificar que la estabilidad del modelo. Consiste en comprobar si la distribución de buenos y malos es similar a lo largo de los scores establecidos. Para ello se utilizará de nuevo el test KS y el índice de Gini, con la salvedad de que en esta ocasión, el objetivo es obtener valores lo más bajos posibles.

La distribución de clientes en ambas muestras es la siguiente:

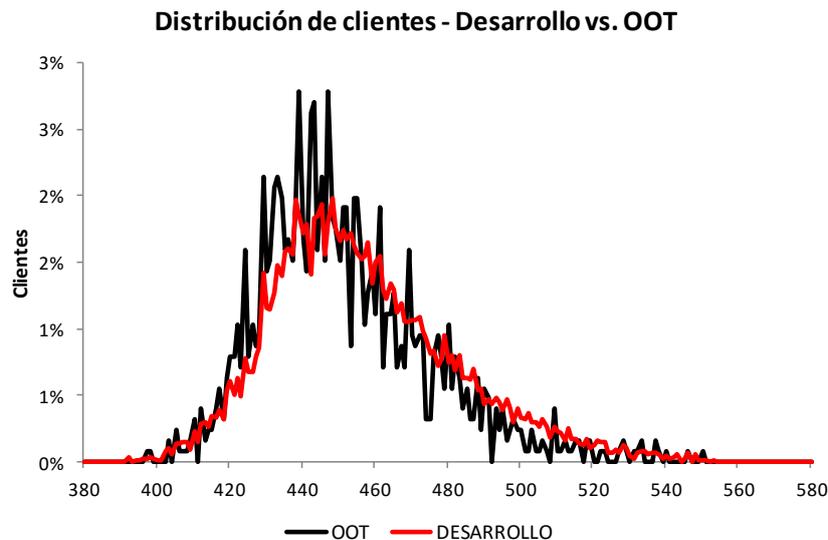


Gráfico 33: Distribución de clientes en muestras de DESARROLLO y OUT OF TIME

Las distribuciones acumuladas de la población total de clientes para cada muestra es la mostrada en el siguiente gráfico. El KS es de 9,52%, es decir, las distribuciones son muy parecidas, pero difieren en algunos scores:

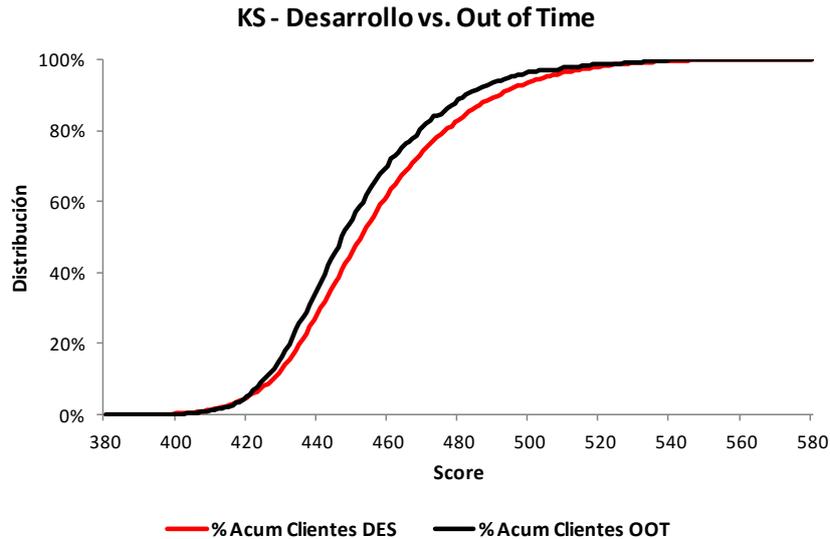


Gráfico 34: Distributions acumuladas de clientes en muestras de DESARROLLO y OUT OF TIME

El índice de Gini, con un valor de 11,94%, arroja resultados similares:

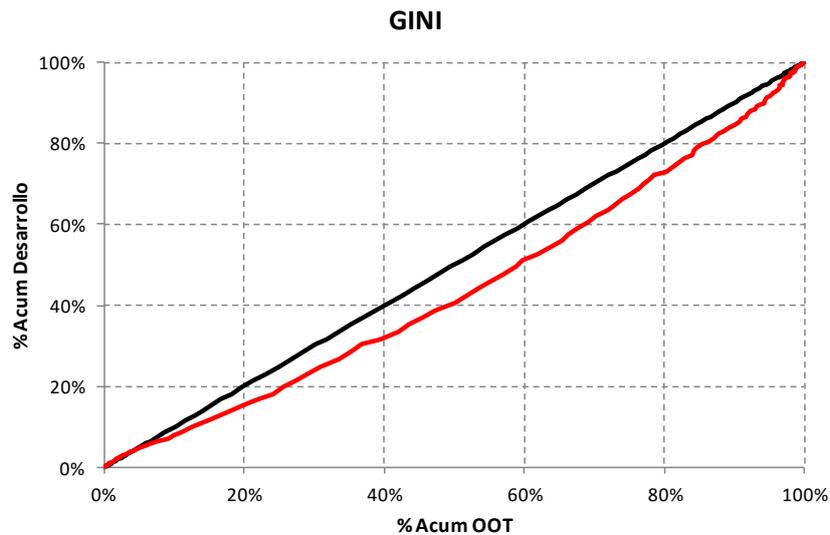


Gráfico 35: Curva de Lorenz en muestras de DESARROLLO y OUT OF TIME

Un estadístico utilizado habitualmente para medir la estabilidad de una muestra a lo largo del tiempo es el Population Stability Index. Es muy útil para comprobar si la distribución actual de la población es similar a la distribución de la población con la que un modelo fue construido (Gadidov et al. 2015). Se calcula de la siguiente manera:

$$PSI = \sum_{i=L}^H \ln \left(\frac{f_{Observada_i}}{f_{Esperada_i}} \right) \cdot (f_{Observada_i} - f_{Esperada_i})$$

Donde $f_{Observada}$ o $f_{Esperada}$ son las funciones de probabilidad.

Generalmente, se establecen los siguientes criterios a la hora de valorar el PSI:

- PSI menor de 0,1: cambios mínimos en la población
- PSI entre 0,1 y 0,25: pequeños cambios en la población.
- PSI mayor de 0,25: cambio significativo en la población.

Las muestras de desarrollo y Out of Time presentan un PSI de 13,61%, lo que indica un ligero cambio en la distribución de la población en el score, por lo tanto, el modelo necesita algún ajuste. Dado que se dispone de una muestra de denegados, de los cuales se puede inferir el comportamiento, éstos procederán a incorporarse a la muestra, y el modelo será reestimado teniendo en cuenta este cambio. Esto se explicará en la siguiente sección.

4.7 Inferencia de denegados y reestimación del modelo

La muestra de denegados consiste en 3.145 solicitudes de clientes a los que se les denegó el crédito. Por lo tanto, no se ha podido observar si son clientes buenos o malos. Sin embargo, sí se tiene información del resto de características del cliente. El objetivo de este proceso es inferir el comportamiento de esos clientes, y añadirlos a la muestra de desarrollo.

El método utilizado se conoce como *Fuzzy Augmentation* (Montrichard, 2008), (Zeng et al., 2014). Esta metodología de inferencia de denegados consiste en utilizar el modelo calibrado con la muestra de desarrollo para obtener una probabilidad de default estimada. Los pasos que se han seguido se explican detalladamente a continuación:

Primero, a cada cliente denegado se le asignan los WOE's calculados anteriormente, dependiendo de a qué categoría pertenezcan en cada variable.

Con los WOE's asignados, a cada cliente se le calcula una probabilidad de default en base al modelo calculado con la muestra de desarrollo, utilizando la siguiente ecuación:

$$PD = \frac{1}{1 + e^{-(\beta_0 + \beta_i x_i)}}$$

Donde los β_i son los siguientes, para cada una de las variables:

Variable	Beta	P-Valor
Intercept	-1,73	<,0001
WOE_ANTIGUEDAD	-0,6497	<,0001
WOE_TRIAD	-0,7064	<,0001
WOE_PASIVO	-0,4126	<,0001
WOE_ratio_AC_PAS	-0,5214	<,0001
WOE_ANTIG_EMP	-0,6422	<,0001
WOE_INCIDENCIAS	-0,6907	<,0001
WOE_NACIONALIDAD	-0,7528	<,0001
WOE_PROVINCIA	-0,5361	<,0001
WOE_EST_CIVIL	-0,8826	<,0001
WOE_PROFESION	-0,5139	<,000

Gráfico 36: Resultados del modelo inicial

Una vez que todos los clientes denegados tienen asignada una Probabilidad de Default (PD), cada observación se desdobra en dos registros: uno malo y otro bueno, ponderados según su probabilidad.

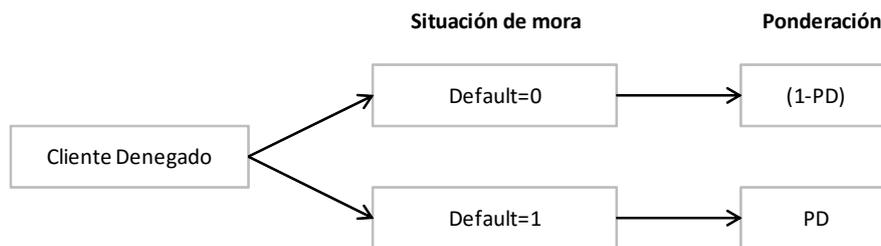


Ilustración 2: Descripción metodología de inferencia de denegados

Es decir, un cliente denegado cuya PD estimada es de 20%, será desdoblado en dos cliente de iguales características, salvo por su desempeño. Éste contará como 0,8 clientes buenos y 0,2 clientes malos. Esta PD es tan sólo un factor por el cual se ponderará el número de buenos y de malos de las muestras de desarrollo y denegados, por lo tanto, para la muestra de desarrollo, este factor toma valor 1 para todos los individuos.

Como resultado de aplicar este tratamiento se obtienen 2.273,21 clientes buenos y 871,79 clientes malos. La muestra de denegados tiene una tasa de mora de 27,73%, significativamente superior a la del resto de muestras, lo cual tiene sentido, puesto que son clientes cuyas características determinaron no ser solicitudes adecuadas según un modelo de scoring anterior.

4.7.1 Análisis univariante

Una vez que la muestra de denegados ponderada por su probabilidad de default se incorpora a la muestra de desarrollo, es necesario llevar a cabo de nuevo el análisis univariante. Para simplificar, se han intentado mantener la mismas categorías que en el modelo inicial. Si alguna variable presenta deficiencias en cuanto al número de individuos en cada grupo, tasas de mora no monótonas o tasas de mora no diferenciadas, se han llevado a cabo pequeños ajustes para solucionarlas. Los grupos que finalmente se han introducido en el modelo son los siguientes:

Variable	Tramo	Cientes	Buenos	Malos	% Poblacion	Tasa mora	% Buenos	% Malos	WOE	IV
Pasivo	-100	3.486	2.389,44	1.096,56	17,33%	31,46%	14,34%	31,81%	-0,797	0,2775
	100-800	7.837	6.434,19	1.402,81	38,97%	17,90%	38,61%	40,70%	-0,053	
	800-2000	4.087	3.546,33	540,67	20,32%	13,23%	21,28%	15,69%	0,305	
	2000-3700	1.804	1.603,69	200,31	8,97%	11,10%	9,62%	5,81%	0,504	
	3700+	2.898	2.691,56	206,44	14,41%	7,12%	16,15%	5,99%	0,992	

Tabla 39: Tramos propuestos para PASIVO con denegados

Variable	Tramo	Cientes	Buenos	Malos	% Poblacion	Tasa mora	% Buenos	% Malos	WOE	IV
Triad	-650	10.425	8.054,11	2.370,89	51,83%	22,74%	48,33%	68,79%	-0,353	0,2667
	650-675	4.575	3.875,40	699,60	22,75%	15,29%	23,25%	20,30%	0,136	
	675-700	3.128	2.842,97	285,03	15,55%	9,11%	17,06%	8,27%	0,724	
	700+	1.984	1.892,73	91,27	9,86%	4,60%	11,36%	2,65%	1,456	

Tabla 40: Tramos propuestos para TRIAD con denegados

Variable	Tramo	Cientes	Buenos	Malos	% Poblacion	Tasa mora	% Buenos	% Malos	WOE	IV
Ratio_AC_PAS	1	1.765	1.161,68	603,32	8,78%	34,18%	6,97%	17,50%	-0,921	0,1871
	2	4.662	3.620,35	1.041,65	23,18%	22,34%	21,72%	30,22%	-0,330	
	3	659	547,68	111,32	3,28%	16,89%	3,29%	3,23%	0,017	
	4	13.026	11.335,51	1.690,49	64,77%	12,98%	68,02%	49,05%	0,327	

Tabla 41: Grupos propuestos para RATIO ACTIVO PASIVO con denegados

Variable	Tramo	Cientes	Buenos	Malos	% Poblacion	Tasa mora	% Buenos	% Malos	WOE	IV
Nacionalidad	1	978	709,26	268,74	4,86%	27,48%	4,26%	7,80%	-0,605	0,1246
	2	1.939	1.505,29	433,71	9,64%	22,37%	9,03%	12,58%	-0,332	
	3	8.519	6.921,31	1.597,69	42,36%	18,75%	41,53%	46,35%	-0,110	
	4	4.260	3.557,28	702,72	21,18%	16,50%	21,35%	20,39%	0,046	
	5	1.960	1.700,25	259,75	9,75%	13,25%	10,20%	7,54%	0,303	
	6	2.456	2.271,82	184,18	12,21%	7,50%	13,63%	5,34%	0,937	

Tabla 42: Grupos propuestos para NACIONALIDAD con denegados

Destaca la variable incidencias, que ha sido agrupada en dos categorías: Con incidencias y sin incidencias.

Variable	Tramo	Cientes	Buenos	Malos	% Poblacion	Tasa mora	% Buenos	% Malos	WOE	IV
Incidencias	CON INCIDENCIAS	4.076	2.959,33	1.116,67	20,27%	27,40%	17,76%	32,40%	-0,601	0,1167
	SIN INCIDENCIAS	16.036	13.705,88	2.330,12	79,73%	14,53%	82,24%	67,60%	0,196	

Tabla 43: Grupos propuestos para INCIDENCIAS con denegados

Variable	Tramo	Clientes	Buenos	Malos	% Poblacion	Tasa mora	% Buenos	% Malos	WOE	IV
Antigüedad	-11	9.144	7.350,92	1.793,08	45,47%	19,61%	44,11%	52,02%	-0,165	0,0802
	11-29	5.346	4.338,65	1.007,35	26,58%	18,84%	26,03%	29,23%	-0,116	
	29-53	2.313	1.984,41	328,59	11,50%	14,21%	11,91%	9,53%	0,222	
	53+	3.309	2.991,23	317,77	16,45%	9,60%	17,95%	9,22%	0,666	

Tabla 44: Tramos propuestos para ANTIGÜEDAD con denegados

Variable	Tramo	Clientes	Buenos	Malos	% Poblacion	Tasa mora	% Buenos	% Malos	WOE	IV
Profesion	1	666	514,26	151,74	3,31%	22,78%	3,09%	4,40%	-0,355	0,0432
	2	9.973	8.058,23	1.914,77	49,59%	19,20%	48,35%	55,55%	-0,139	
	3	3.024	2.534,36	489,64	15,04%	16,19%	15,21%	14,21%	0,068	
	4	5.158	4.395,83	762,17	25,65%	14,78%	26,38%	22,11%	0,176	
	5	1.291	1.162,52	128,48	6,42%	9,95%	6,98%	3,73%	0,627	

Tabla 45: Grupos propuestos para PROFESIÓN con denegados

Variable	Tramo	Clientes	Buenos	Malos	% Poblacion	Tasa mora	% Buenos	% Malos	WOE	IV
Antig_Emp	-12	11.619	9.369,64	2.249,36	57,77%	19,36%	56,22%	65,26%	-0,149	0,0419
	12-36	5.616	4.758,04	857,96	27,92%	15,28%	28,55%	24,89%	0,137	
	36+	2.877	2.537,53	339,47	14,30%	11,80%	15,23%	9,85%	0,436	

Tabla 46: Tramos propuestos para ANTIGÜEDAD EN EL EMPLEO con denegados

Variable	Tramo	Clientes	Buenos	Malos	% Poblacion	Tasa mora	% Buenos	% Malos	WOE	IV
Est_Civil	1	10.165	8.180,67	1.984,33	50,54%	19,52%	49,09%	57,57%	-0,159	0,0351
	2	1.924	1.589,12	334,88	9,57%	17,41%	9,54%	9,72%	-0,019	
	3	6.947	5.952,09	994,91	34,54%	14,32%	35,72%	28,86%	0,213	
	4	1.076	943,34	132,66	5,35%	12,33%	5,66%	3,85%	0,386	

Tabla 47: Grupos propuestos para ESTAdO CIVIL con denegados

Variable	Tramo	Clientes	Buenos	Malos	% Poblacion	Tasa mora	% Buenos	% Malos	WOE	IV
Provincia	1	3.306	2.636,28	669,72	16,44%	20,26%	15,82%	19,43%	-0,206	0,0319
	2	9.443	7.738,91	1.704,09	46,95%	18,05%	46,44%	49,44%	-0,063	
	3	3.617	3.019,15	597,85	17,98%	16,53%	18,12%	17,35%	0,044	
	4	2.110	1.822,26	287,74	10,49%	13,64%	10,93%	8,35%	0,270	
	5	1.636	1.448,62	187,38	8,13%	11,45%	8,69%	5,44%	0,469	

Tabla 48: Grupos propuestos para PROVINIA con denegados

4.7.2 Análisis multivariante

De nuevo, se analiza la matriz de correlaciones de los nuevos WOE's calculados

	WOE_EDAD	WOE_ANTIGUEDAD	WOE_TRIAD	WOE_PASIVO	WOE_ACTIVO	WOE_ratio_AC_PAS
WOE_EDAD	100,00%	23,19%	11,79%	11,33%	1,52%	3,17%
WOE_ANTIGUEDAD	23,19%	100,00%	13,42%	9,84%	-3,74%	-2,03%
WOE_TRIAD	11,79%	13,42%	100,00%	36,37%	19,33%	24,27%
WOE_PASIVO	11,33%	9,84%	36,37%	100,00%	46,90%	58,36%
WOE_ACTIVO	1,52%	-3,74%	19,33%	46,90%	100,00%	78,29%
WOE_ratio_AC_PAS	3,17%	-2,03%	24,27%	58,36%	78,29%	100,00%
WOE_ANTIG_EMP	17,35%	23,91%	7,47%	6,90%	-1,26%	-0,90%
WOE_ING	13,12%	16,89%	7,59%	15,69%	-2,11%	-0,27%
WOE_IMPORTE	23,20%	12,92%	12,09%	23,57%	8,68%	13,48%
WOE_INCIDENCIAS	1,68%	-13,08%	19,22%	35,32%	31,98%	43,10%
WOE_NACIONALIDAD	19,06%	17,75%	11,91%	13,93%	-0,55%	1,92%
WOE_PROVINCIA	2,69%	6,80%	6,23%	5,82%	2,02%	3,47%
WOE_NIV_ESTUD	5,09%	10,09%	9,96%	9,74%	0,75%	2,44%
WOE_EST_CIVIL	30,74%	9,03%	2,95%	4,87%	-1,06%	-0,84%
WOE_VIV_HAB	21,25%	22,96%	10,99%	10,68%	-0,20%	1,84%
WOE_PROFESION	10,36%	17,58%	9,82%	8,12%	-0,83%	-0,31%
WOE_TIPO_CONT	19,83%	20,21%	8,73%	7,81%	-2,95%	-2,32%

	WOE_ANTIG_EMP	WOE_ING	WOE_IMPORTE	WOE_INCIDENCIAS	WOE_NACIONALIDAD	WOE_PROVINCIA
WOE_EDAD	17,35%	13,12%	23,20%	1,68%	19,06%	2,69%
WOE_ANTIGUEDAD	23,91%	16,89%	12,92%	-13,08%	17,75%	6,80%
WOE_TRIAD	7,47%	7,59%	12,09%	19,22%	11,91%	6,23%
WOE_PASIVO	6,90%	15,69%	23,57%	35,32%	13,93%	5,82%
WOE_ACTIVO	-1,26%	-2,11%	8,68%	31,98%	-0,55%	2,02%
WOE_ratio_AC_PAS	-0,90%	-0,27%	13,48%	43,10%	1,92%	3,47%
WOE_ANTIG_EMP	100,00%	12,73%	27,97%	-1,47%	10,80%	4,22%
WOE_ING	12,73%	100,00%	49,93%	-3,44%	20,63%	2,45%
WOE_IMPORTE	27,97%	49,93%	100,00%	36,85%	18,50%	4,09%
WOE_INCIDENCIAS	-1,47%	-3,44%	36,85%	100,00%	4,10%	3,69%
WOE_NACIONALIDAD	10,80%	20,63%	18,50%	4,10%	100,00%	8,55%
WOE_PROVINCIA	4,22%	2,45%	4,09%	3,69%	8,55%	100,00%
WOE_NIV_ESTUD	3,29%	18,33%	12,00%	3,78%	20,50%	3,56%
WOE_EST_CIVIL	5,71%	6,97%	18,50%	2,53%	8,25%	1,99%
WOE_VIV_HAB	19,35%	15,92%	18,55%	0,34%	17,04%	5,67%
WOE_PROFESION	7,81%	11,98%	11,05%	1,52%	23,53%	5,92%
WOE_TIPO_CONT	24,36%	13,14%	17,86%	0,38%	20,35%	3,39%

	WOE_NIV_ESTUD	WOE_EST_CIVIL	WOE_VIV_HAB	WOE_PROFESION	WOE_TIPO_CONT
WOE_EDAD	5,09%	30,74%	21,25%	10,36%	19,83%
WOE_ANTIGUEDAD	10,09%	9,03%	22,96%	17,58%	20,21%
WOE_TRIAD	9,96%	2,95%	10,99%	9,82%	8,73%
WOE_PASIVO	9,74%	4,87%	10,68%	8,12%	7,81%
WOE_ACTIVO	0,75%	-1,06%	-0,20%	-0,83%	-2,95%
WOE_ratio_AC_PAS	2,44%	-0,84%	1,84%	-0,31%	-2,32%
WOE_ANTIG_EMP	3,29%	5,71%	19,35%	7,81%	24,36%
WOE_ING	18,33%	6,97%	15,92%	11,98%	13,14%
WOE_IMPORTE	12,00%	18,50%	18,55%	11,05%	17,86%
WOE_INCIDENCIAS	3,78%	2,53%	0,34%	1,52%	0,38%
WOE_NACIONALIDAD	20,50%	8,25%	17,04%	23,53%	20,35%
WOE_PROVINCIA	3,56%	1,99%	5,67%	5,92%	3,39%
WOE_NIV_ESTUD	100,00%	-1,56%	5,74%	18,36%	6,82%
WOE_EST_CIVIL	-1,56%	100,00%	11,16%	2,34%	7,26%
WOE_VIV_HAB	5,74%	11,16%	100,00%	17,89%	23,24%
WOE_PROFESION	18,36%	2,34%	17,89%	100,00%	43,34%
WOE_TIPO_CONT	6,82%	7,26%	23,24%	43,34%	100,00%

Tabla 49: Matriz de correlaciones WOE

Se observa una correlación cercana al 60% entre los WOES de la variable Pasivo y la variable ratio Activo Pasivo. Dado que no alcanza el 70%, es una correlación alta pero no llega a ser problemática.

4.7.3 Regresión logística

De nuevo se lanza el algoritmo de selección de variables stepwise de la regresión logística. Las variables seleccionadas son las mismas que en el modelo inicial, dados los ligeros cambios que se han introducido. En este caso, cada observación introducida en el modelo está ponderada por la variable PD, mencionada anteriormente. Los clientes de la muestra de desarrollo entran en el modelo con peso igual a 1, y cada cliente denegado entra en el modelo dos veces: como cliente bueno y malo, ponderados por sus valores de (1-PD) y PD respectivamente. Los resultados son los siguientes, en comparación con el modelo inicial:

Variable	Denegados		Desarrollo	
	Beta	P-Valor	Beta	P-Valor
Intercept	-1,5844	<,0001	-1,73	<,0001
WOE_ANTIGUEDAD	-0,6715	<,0001	-0,6497	<,0001
WOE_TRIAD	-0,6724	<,0001	-0,7064	<,0001
WOE_PASIVO	-0,4032	<,0001	-0,4126	<,0001
WOE_ratio_AC_PAS	-0,3718	<,0001	-0,5214	<,0001
WOE_ANTIG_EMP	-0,6211	<,0001	-0,6422	<,0001
WOE_INCIDENCIAS	-0,5177	<,0001	-0,6907	<,0001
WOE_NACIONALIDAD	-0,7051	<,0001	-0,7528	<,0001
WOE_PROVINCIA	-0,6165	<,0001	-0,5361	<,0001
WOE_EST_CIVIL	-0,7968	<,0001	-0,8826	<,0001
WOE_PROFESION	-0,4904	<,000	-0,5139	<,000
Estadístico C	0,690		0,702	

Tabla 50: Resultados regresión logística del modelo con denegados

Se observan pequeños cambios en los estimadores de las variables como resultado de haber incluido la muestra de denegados.

Los estadísticos del modelo son los siguientes:

Asociación de probabilidades predichas y respuestas observadas			
Concordancia de porcentaje	69	D de Somers	0,380
Discordancia de porcentaje	31	Gamma	0,380
Porcentaje ligado	0	Tau-a	0,141
Pares	100.277.320	c	0,690

Tabla 51: Estadísticos del modelo de regresión logística con denegados

El modelo posee un buen valor del estadístico c, clasifica correctamente al 69% de los individuos. Además posee buenos valores de la D de Somers y Gamma de Goodman-Kruskal, de 0,38. Sin embargo todos estos valores han disminuido respecto al modelo inicial.

A continuación, se muestra la matriz de confusión:

		Predicho	
		Buenos	Malos
Observado	Buenos	11.838,77 (58,86%)	4.826,44 (24,00%)
	Malos	1.409,23 (7,01%)	2.037,56 (10,13%)

Tabla 52: Matriz de confusión del modelo con denegados

Se observa un 7,01% de clientes malos que han sido clasificados como buenos, frente al 6,73% del modelo inicial.

4.7.4 Scorecard

Una vez obtenidos los resultados de la regresión, es posible calcular las puntuaciones. De nuevo utilizando la formula:

$$Score = \frac{Offset}{n} + \frac{Factor}{\ln(2)} \cdot \left(\frac{\beta_0}{n} - \beta_{ij} \cdot WOE_{ij} \right)$$

Donde Offset=500 y Factor=20.

Variable	Tramo	Tasa mora	WOE	Score	Beta	IV	SCR
Pasivo	-100	31,46%	-0,797	36	-0,4032	0,2775	10,99%
	100-800	17,90%	-0,053	45			
	800-2000	13,23%	0,305	49			
	2000-3700	11,10%	0,504	51			
	3700+	7,12%	0,992	57			
Triad	-650	22,74%	-0,353	39	-0,6724	0,2667	35,91%
	650-675	15,29%	0,136	48			
	675-700	9,11%	0,724	59			
	700+	4,60%	1,456	74			
Ratio_AC_PAS	1	34,18%	-0,921	36	-0,3718	0,1871	5,16%
	2	22,34%	-0,330	42			
	3	16,89%	0,017	46			
	4	12,98%	0,327	49			
Nacionalidad	1	27,48%	-0,605	33	-0,7051	0,1246	16,86%
	2	22,37%	-0,332	39			
	3	18,75%	-0,110	43			
	4	16,50%	0,046	46			
	5	13,25%	0,303	52			
	6	7,50%	0,937	64			
Incidencias	CON INCIDENCIAS	27,40%	-0,601	36	-0,5177	0,1167	6,56%
	SIN INCIDENCIAS	14,53%	0,196	48			
Antigüedad	-11	19,61%	-0,165	42	-0,6715	0,0802	9,92%
	11-29	18,84%	-0,116	43			
	29-53	14,21%	0,222	50			
	53+	9,60%	0,666	58			
Profesion	1	22,78%	-0,355	40	-0,4904	0,0432	2,92%
	2	19,20%	-0,139	43			
	3	16,19%	0,068	46			
	4	14,78%	0,176	48			
	5	9,95%	0,627	54			
Antig_Emp	-12	19,36%	-0,149	43	-0,6211	0,0419	3,75%
	12-36	15,28%	0,137	48			
	36+	11,80%	0,436	53			
Est_Civil	1	19,52%	-0,159	42	-0,7968	0,0351	4,80%
	2	17,41%	-0,019	45			
	3	14,32%	0,213	50			
	4	12,33%	0,386	54			
Provincia	1	20,26%	-0,206	42	-0,6165	0,0319	3,13%
	2	18,05%	-0,063	44			
	3	16,53%	0,044	46			
	4	13,64%	0,270	50			
	5	11,45%	0,469	54			

Tabla 53: Scorecard modelo con denegados

AL observar los SCR se concluye, de nuevo, que las variables que más aportan al score son la puntuación TRIAD, la nacionalidad y el pasivo. Esta scorecard tiene una relación adecuada entre la tasa de mora, los WOE's y las puntuaciones. Permite obtener puntuaciones entre 389 y 565. El siguiente paso es comprobar la efectividad del modelo a la hora de separar clientes buenos y malos.

4.7.5 Poder predictivo del modelo

Muestra de Desarrollo con Denegados

La distribución de buenos y malos en la muestra que contiene ambas poblaciones es la siguiente:

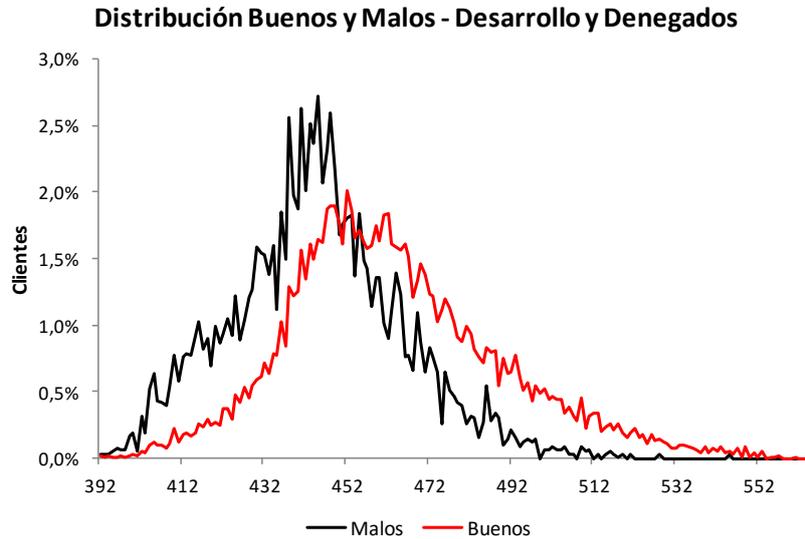


Gráfico 37: Distribución de clientes en muestra de DESARROLLO CON DENEGADOS

Observando la distribución acumulada de los clientes buenos y malos, se obtiene un estadístico KS de 30,99%. Presenta una ligera mejora respecto al modelo inicial:

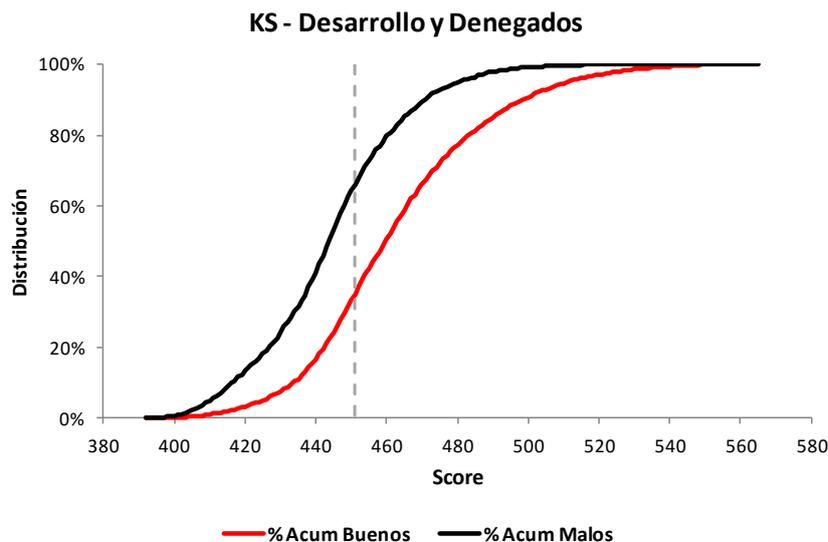


Gráfico 38: Distribuciones acumuladas de cliente en muestra de DESARROLLO CON DENEGADOS

EN el siguiente gráfico se muestra la curva de Lorenz, cuyo índice de GINI es 42,53%:

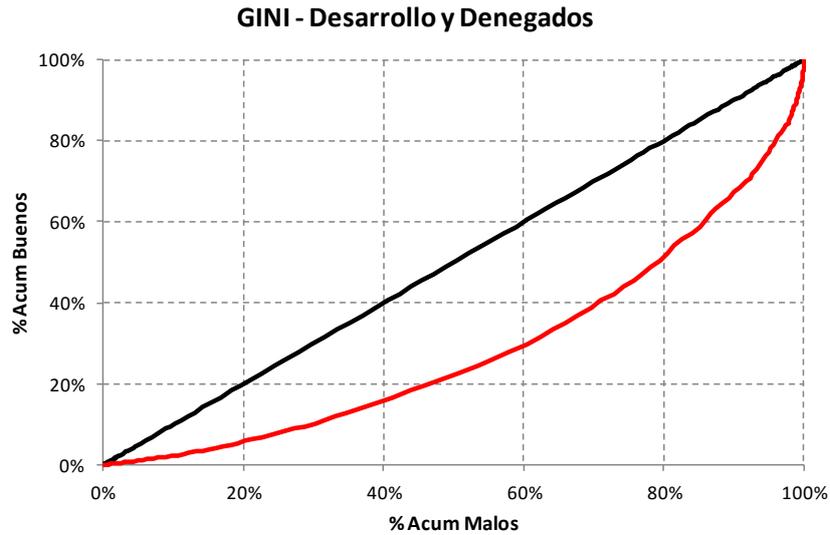


Gráfico 39: Curva de Lorenz en muestra de DESARROLLO CON DENEGADOS

Al observar la tasa de mora y la distribución de clientes por score, se muestra claramente una tendencia muy pronunciada. Los resultados mejoran los del modelo inicial.

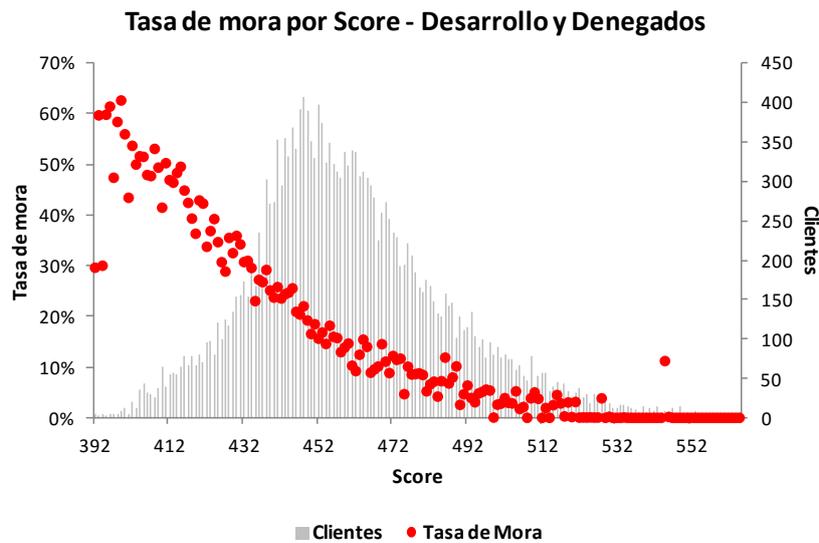


Gráfico 40: Tasa de mora por score del modelo final en muestra de DESARROLLO CON DENEGADOS

Muestra Hold-Out

La distribución de las poblaciones de clientes buenos y malos en la muestra de validación es la siguiente:

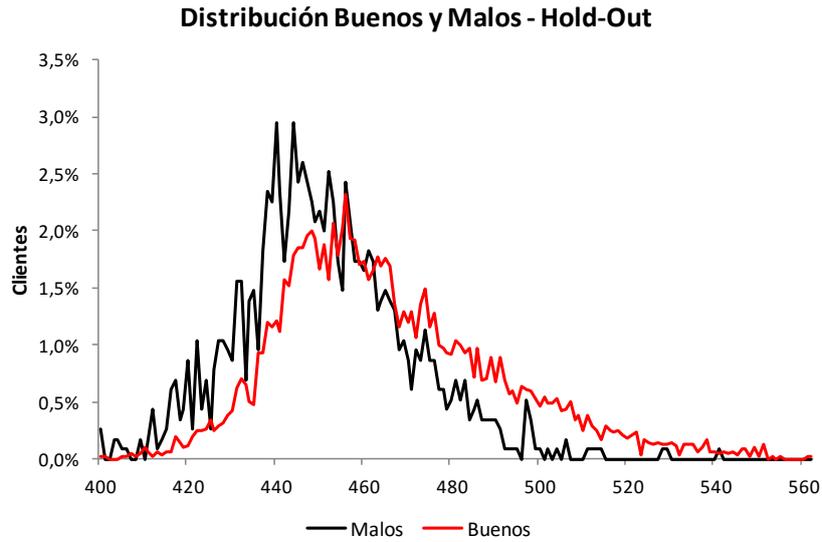


Gráfico 41: Distribución de clientes en muestra HOLDOUT

La distribución acumulada de ambas poblaciones muestra un KS de 24%, ligeramente inferior que el mismo estadístico arrojado por el modelo inicial.

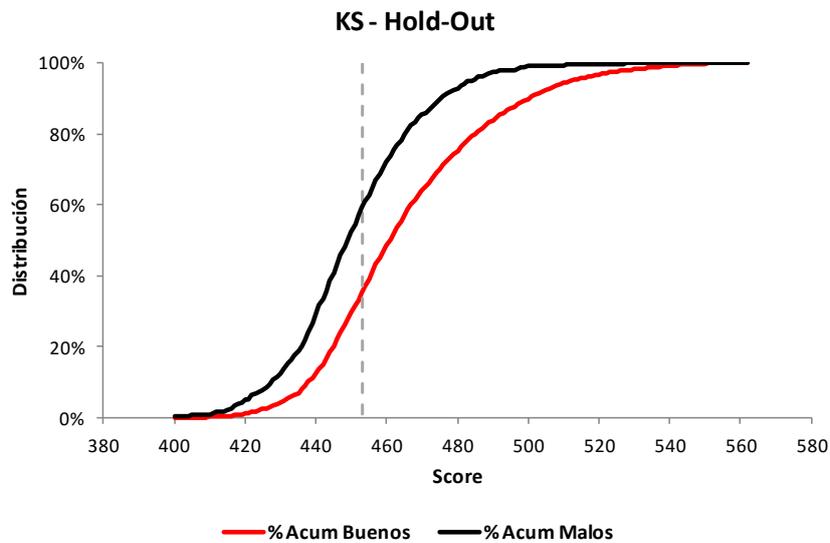


Gráfico 42: Distribuciones acumuladas de clientes en muestra HOLDOUT

La curva de Lorenz de la distribución de clientes presenta un índice de Gini de 34,88%, algo inferior al índice obtenido con el modelo inicial:

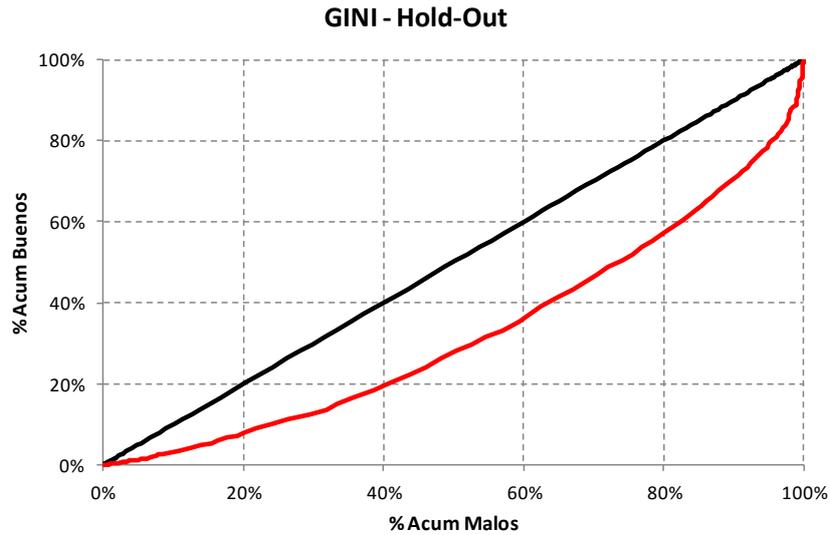


Gráfico 43: Curva de Lorenz en muestra HOLDOUT

La distribución de clientes y tasa de mora por score muestra una tendencia decreciente de la tasa de mora según aumenta el score:

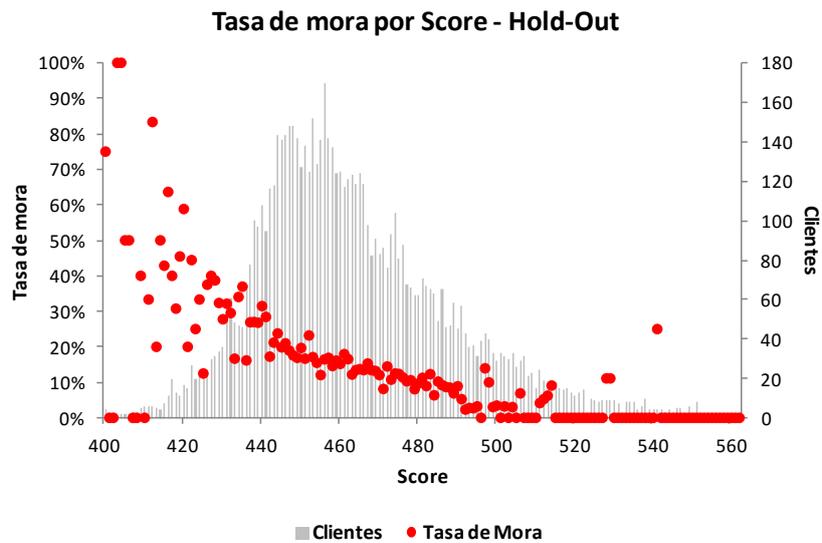


Gráfico 44: Tasa de mora por score del modelo final en muestra HOLDOUT

Muestra Out of Time

A continuación, se muestra la distribución de clientes buenos y malos a lo largo del score:

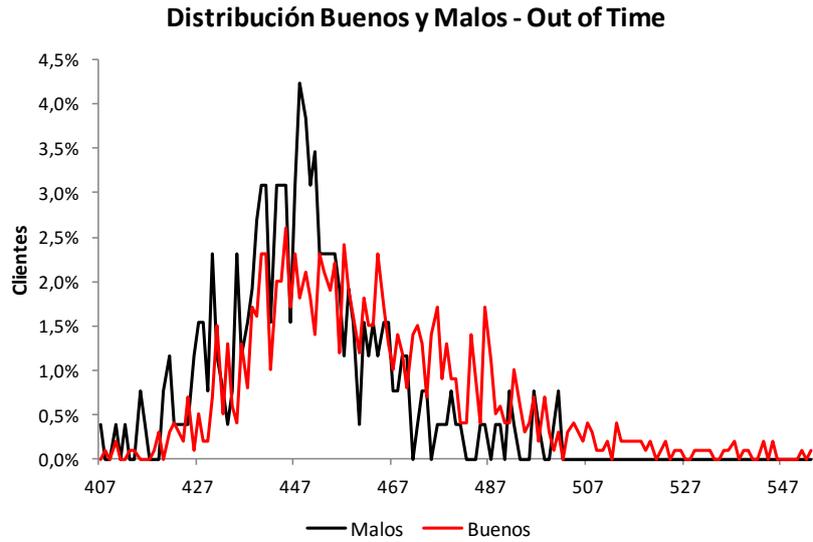


Gráfico 45: Distribución de clientes en muestra OUT OF TIME

La distribución acumulada de las poblaciones de buenos y malos presenta un estadístico KS de 25,73%:

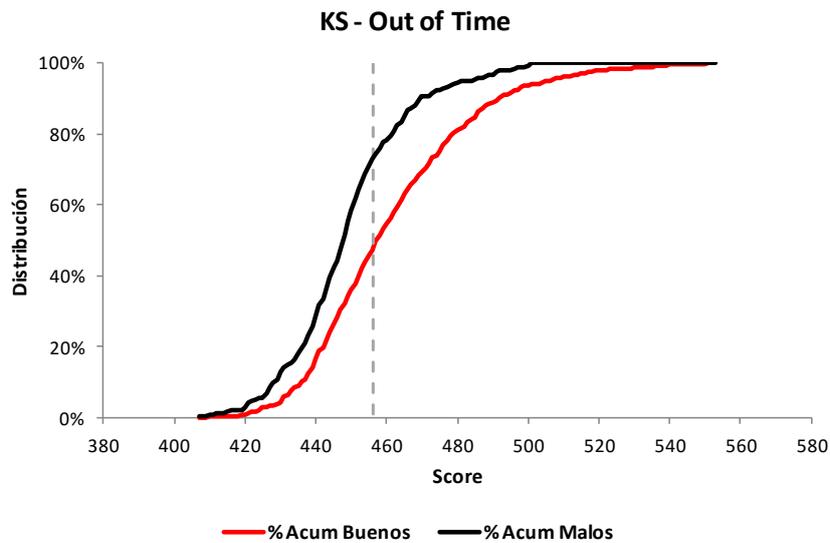


Gráfico 46: Distribuciones acumuladas de clientes en muestra OUT OF TIME

La distribución de clientes y tasa de mora por score se muestran en el gráfico a continuación. Se observa una tasa de mora decreciente no muy definida a lo largo del score, posiblemente debido al reducido tamaño de la muestra:

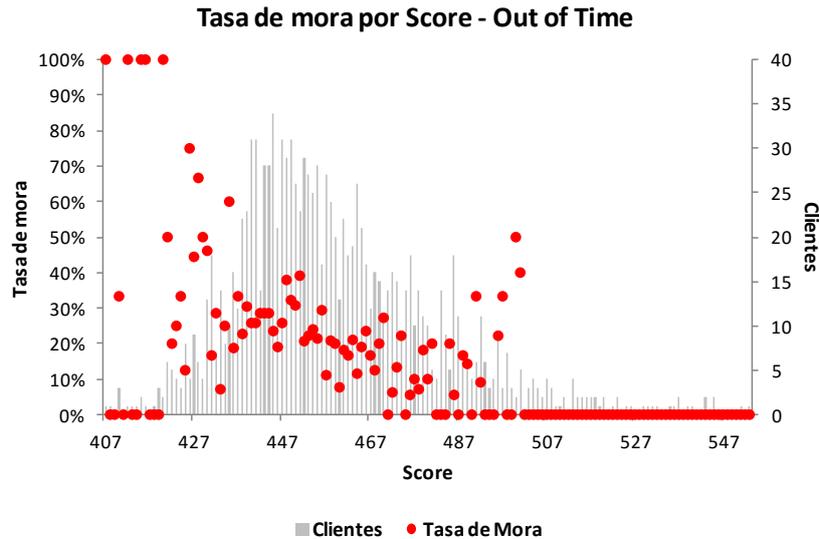


Gráfico 47: Tasa de mora por score del modelo final en muestra OUT OF TIME

La siguiente tabla muestra un resumen de los resultados de poder predictivo obtenidos con el modelo calibrado con clientes denegados:

Muestra	Observaciones	Tasa Mora	KS	GINI
Desarrollo	16.973	15,21%	30,99%	42,53%
Hold-Out	7.274	15,86%	24,00%	34,88%
Out of Time	1.258	20,67%	25,73%	31,42%

Tabla 54: Resumen de resultados poder predictivo del modelo final

4.7.6 Estabilidad del modelo

En esta sección se evalúa la estabilidad de la distribuciones de clientes a lo largo del score de la muestra de desarrollo con denegados respecto a la muestra Out of Time.

Las distribuciones de clientes son similares en ambas muestras:

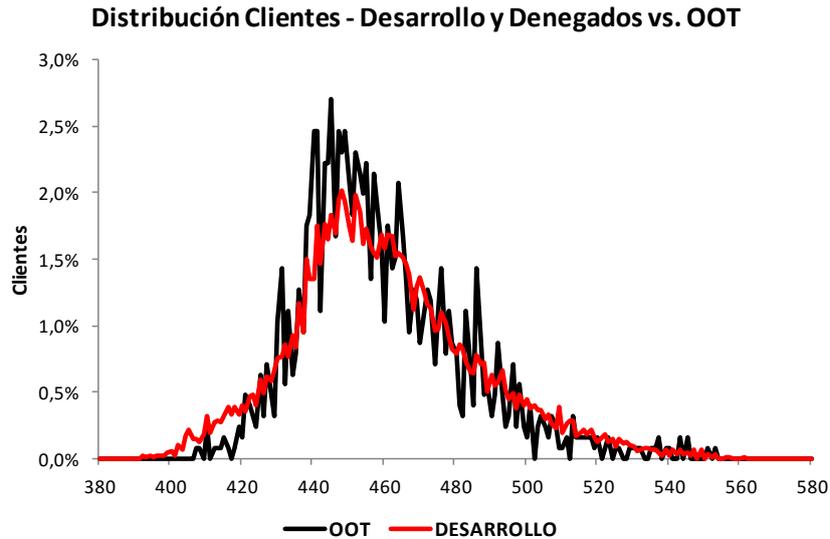


Gráfico 48: Distribución de clientes en muestras de DESARROLLO CON DENEGADOS y OUT OF TIME

Al comparar las distribuciones acumuladas de los clientes en ambas muestras, se obtiene un KS de estabilidad de 4,57%. El gráfico siguiente muestra que ambas distribuciones son muy parecidas:

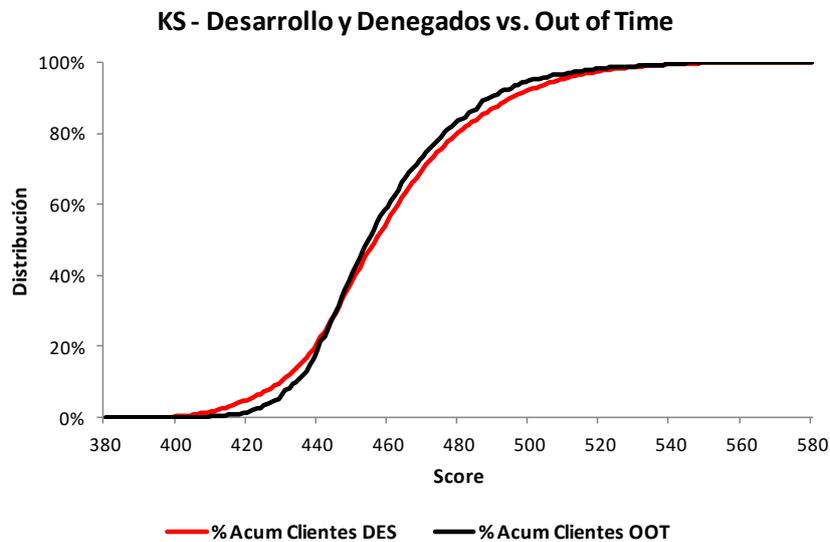


Gráfico 49: Distribuciones acumuladas de clientes en muestras de DESARROLLO CON DENEGADOS y OUT OF TIME

El mismo resultado se obtiene al evaluar la estabilidad por medio del Gini. La curva de Lorenz se aproxima mucho a la igualdad absoluta, obteniéndose un índice de Gini de 2,74%:

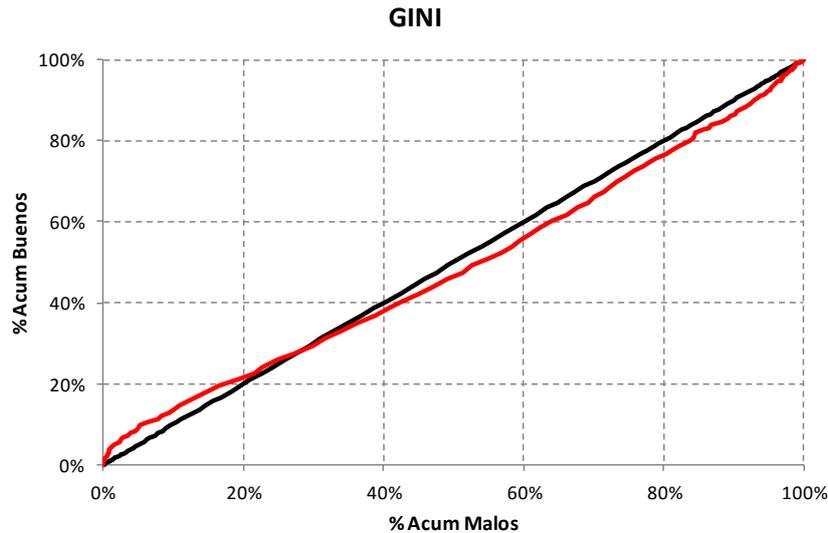


Gráfico 50: Curva de Lorenz en muestras de DESARROLLO CON DENEGADOS y OUT OF TIME

Sin embargo, el Population Stability Index, con un valor de 12,13%, muestra que la muestra Out of Time ha sufrido pequeños cambios respecto a la de Desarrollo. Esto puede deberse a la diferencia en el tamaño de ambas muestras.

4.8 Validación Bootstrapping

Bootstrapping es un método de validación recursivo en el que se simula un elevado número de muestras aleatorias de tamaño establecido provenientes de la muestra de desarrollo (Wu, 2008). En cada una de esas N muestras se evaluará el KS y el índice Gini de poder predictivo.

Una vez que se genera una muestra aleatoria, la siguiente se generará a partir de la muestra de desarrollo completa. Es decir, un mismo elemento puede estar incluido en varias muestras.

El modelo de scoring debería conseguir resultados similares de poder predictivo en muestras de características similares a la muestra de desarrollo, independientemente del tamaño de la muestra aleatoria.

Los resultados de generar 50 muestras con 15.000 observaciones a partir de la muestra de desarrollo son los siguientes:

Repeticiones	50
Tamaño muestras	15.000

#	Tasa de Mora	Score KS	KS	GINI
Desarrollo	17,14%	451	30,99%	42,53%
1	17,12%	455	30,20%	41,84%
2	16,74%	450	31,96%	43,05%
3	17,16%	451	29,68%	40,74%
4	16,64%	449	31,66%	43,39%
5	16,86%	449	30,99%	41,80%
6	17,15%	449	32,32%	43,45%
7	16,97%	449	31,40%	43,28%
8	17,09%	453	31,17%	42,58%
9	17,11%	449	31,26%	42,72%
10	17,19%	457	31,11%	42,53%
11	16,99%	452	31,75%	43,44%
12	17,13%	449	30,52%	41,52%
13	17,30%	451	30,87%	42,03%
14	17,01%	451	31,68%	42,75%
15	17,11%	451	30,75%	42,53%
16	16,98%	451	31,26%	42,83%
17	17,32%	450	31,65%	42,63%
18	17,14%	451	30,79%	42,40%
19	17,29%	449	31,86%	43,38%
20	17,17%	449	31,02%	42,30%
21	16,95%	451	31,13%	42,47%
22	17,38%	451	30,19%	41,41%
23	16,98%	451	31,14%	42,51%
24	17,13%	451	30,47%	42,00%
25	16,96%	450	30,96%	42,73%
26	17,13%	451	31,45%	42,83%
27	17,04%	448	31,42%	42,76%
28	17,35%	451	30,38%	42,01%
29	17,14%	451	30,71%	42,05%
30	16,92%	449	32,15%	43,69%

#	Tasa de Mora	Score KS	KS	GINI
Desarrollo	17,14%	451	30,99%	42,53%
31	17,47%	449	31,32%	42,40%
32	17,00%	453	30,64%	42,51%
33	17,25%	451	30,56%	41,74%
34	17,06%	451	31,85%	43,17%
35	17,16%	449	30,25%	41,64%
36	17,38%	448	31,58%	42,76%
37	17,13%	451	31,08%	43,00%
38	16,91%	449	30,37%	41,89%
39	17,07%	451	32,29%	43,70%
40	17,52%	449	31,73%	43,05%
41	17,13%	451	31,66%	42,57%
42	17,27%	455	31,45%	42,93%
43	17,04%	450	31,90%	43,45%
44	17,50%	449	31,27%	43,22%
45	16,93%	453	30,98%	42,39%
46	17,24%	451	31,13%	43,03%
47	17,48%	449	30,55%	41,85%
48	17,32%	451	30,46%	41,76%
49	17,57%	453	30,39%	41,87%
50	17,36%	453	30,95%	41,98%

Tabla 55: Muestras aleatorias de 15.000 clientes

La tasa de mora de las 50 muestras generadas aleatoriamente es muy similar, con una máxima diferencia de 0,50% (en la muestra aleatoria nº 4). Se observan valores similares del estadístico KS, con un máximo de 32,32% y un mínimo de 29,68%, que son muy próximos a 30,99% obtenido a partir de la muestra de desarrollo. En cuanto al índice GINI, los valores máximo y mínimo observados en las 50 muestras aleatorias son, respectivamente, 43,70% y 40,74%. Es decir, al estudiar el poder predictivo del modelo en muestras aleatorias generadas a partir de la muestra de desarrollo, los estadísticos no difieren de los obtenidos anteriormente.

El mismo análisis ha sido realizado para 50 muestras aleatorias de 10.000 clientes, y para 50 muestras aleatorias de 5.000 clientes. A medida que la muestra contiene un número menor de clientes, los estadísticos de KS y Gini difieren más de los obtenidos con la muestra de desarrollo. Los resultados de ambos análisis se muestran a continuación:

Repeticiones	50
Tamaño muestras	10.000

#	Tasa de Mora	Score KS	KS	GINI
Desarrollo	17,14%	451	30,99%	42,53%
1	17,20%	451	30,61%	41,40%
2	16,82%	449	30,32%	41,92%
3	17,06%	457	30,38%	42,06%
4	17,25%	451	31,21%	41,93%
5	17,62%	450	32,25%	43,86%
6	17,33%	451	31,02%	42,28%
7	17,43%	451	30,77%	42,19%
8	16,83%	448	30,26%	41,64%
9	17,09%	452	30,00%	41,20%
10	17,28%	449	31,79%	43,48%
11	17,27%	452	29,75%	40,65%
12	16,75%	451	30,47%	41,22%
13	17,51%	451	31,23%	42,36%
14	17,05%	451	30,76%	42,37%
15	16,74%	448	30,68%	41,75%
16	17,10%	453	30,16%	41,66%
17	17,15%	451	31,04%	42,53%
18	16,74%	455	32,24%	43,66%
19	17,03%	455	31,22%	42,52%
20	16,99%	452	31,84%	43,19%
21	16,50%	449	30,67%	41,68%
22	17,41%	451	31,59%	43,11%
23	17,71%	451	30,83%	42,49%
24	16,82%	449	30,47%	42,10%
25	16,87%	450	32,37%	43,69%
26	16,79%	449	32,16%	43,61%
27	16,96%	455	31,01%	42,22%
28	17,25%	451	30,34%	41,09%
29	17,64%	451	31,45%	43,00%
30	17,56%	451	31,32%	42,78%

#	Tasa de Mora	Score KS	KS	GINI
Desarrollo	17,14%	451	30,99%	42,53%
31	17,34%	452	30,94%	42,87%
32	16,69%	455	31,19%	42,57%
33	17,45%	451	32,52%	44,21%
34	17,48%	456	31,99%	42,92%
35	17,53%	451	32,08%	42,95%
36	17,18%	456	32,08%	43,46%
37	17,16%	453	31,22%	43,19%
38	17,23%	451	30,72%	41,90%
39	17,48%	450	30,65%	42,54%
40	17,51%	451	31,65%	43,51%
41	17,27%	449	29,70%	40,96%
42	17,27%	451	31,10%	42,76%
43	17,27%	448	32,06%	42,75%
44	17,67%	454	30,83%	41,98%
45	17,08%	450	30,79%	42,33%
46	17,28%	451	32,22%	43,51%
47	16,77%	456	30,10%	42,07%
48	17,22%	457	28,82%	40,76%
49	17,52%	456	29,18%	40,63%
50	16,84%	451	30,73%	42,15%

Tabla 56: Muestras aleatorias de 10.000 clientes

Repeticiones	50
Tamaño muestras	5.000

#	Tasa de Mora	Score KS	KS	GINI
Desarrollo	17,14%	451	30,99%	42,53%
1	16,46%	453	31,72%	41,89%
2	17,17%	449	29,75%	39,86%
3	17,65%	451	32,16%	43,50%
4	16,98%	449	30,21%	41,99%
5	17,62%	453	32,23%	43,32%
6	17,72%	457	28,45%	40,23%
7	17,53%	449	32,11%	42,76%
8	17,98%	451	31,83%	41,83%
9	17,37%	448	30,82%	41,60%
10	17,64%	453	32,09%	43,18%
11	17,60%	448	34,67%	44,75%
12	17,54%	451	29,59%	40,82%
13	16,88%	452	32,36%	44,22%
14	17,93%	453	29,88%	41,18%
15	17,79%	449	33,28%	43,54%
16	17,27%	457	28,42%	41,21%
17	17,48%	452	30,74%	42,14%
18	17,06%	456	31,84%	42,96%
19	17,30%	457	30,57%	42,35%
20	16,82%	448	33,03%	45,01%
21	17,19%	450	31,56%	42,38%
22	16,85%	452	30,54%	40,58%
23	17,37%	450	28,57%	39,32%
24	16,45%	454	31,17%	42,28%
25	18,12%	455	32,92%	44,37%
26	17,56%	452	30,54%	41,42%
27	17,80%	452	31,00%	41,31%
28	17,55%	448	31,56%	42,55%
29	17,66%	457	34,03%	44,50%
30	16,75%	448	31,11%	43,16%

#	Tasa de Mora	Score KS	KS	GINI
Desarrollo	17,14%	451	30,99%	42,53%
31	16,57%	455	31,26%	42,44%
32	17,16%	449	27,93%	38,89%
33	16,32%	451	32,21%	43,76%
34	17,53%	451	32,77%	43,76%
35	17,15%	457	30,55%	41,62%
36	16,14%	449	32,16%	43,05%
37	17,32%	449	31,74%	42,26%
38	17,13%	452	32,42%	43,15%
39	17,30%	456	32,26%	43,66%
40	17,41%	454	31,89%	42,49%
41	16,85%	451	31,65%	42,90%
42	16,77%	449	31,71%	42,87%
43	17,43%	451	33,45%	44,37%
44	18,20%	449	31,11%	42,25%
45	17,26%	451	33,90%	43,59%
46	17,07%	448	31,97%	43,53%
47	17,49%	455	29,72%	41,23%
48	15,74%	458	30,97%	41,70%
49	16,34%	455	30,85%	41,18%
50	16,13%	452	34,59%	46,06%

Tabla 57: Muestras aleatorias de 5.000 clientes

Las mayores diferencias respecto a los estadísticos de desarrollo se observan en las muestras de 5.000 clientes. La mayor diferencia en el estadístico KS es de 3,60%, en la muestra 11, cuyo KS es 34,67%. La mayor diferencia en el Gini es de 3,65%, en la muestra 32, cuyo índice de Gini es 38,89%. El poder predictivo del modelo no es muy sensible al tamaño de la muestra siempre que esta tenga características similares a aquella con la que ha sido desarrollado.

5 Resultados y Conclusiones

En esta sección se compararán los resultados obtenidos a partir del modelo inicial, con los obtenidos a partir del modelo que incorpora la población de denegados.

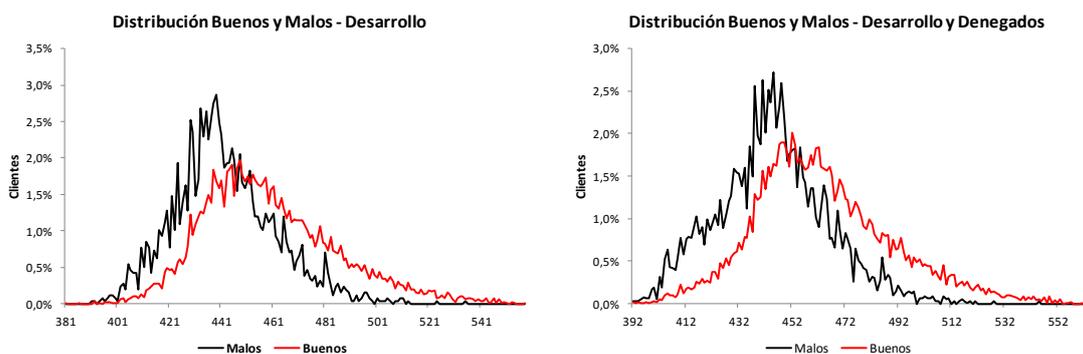
En cuanto al modelo de regresión, se observan cambios menores en los estimadores de cada variable. Al añadir la población de denegados, se observa una pequeña disminución en el

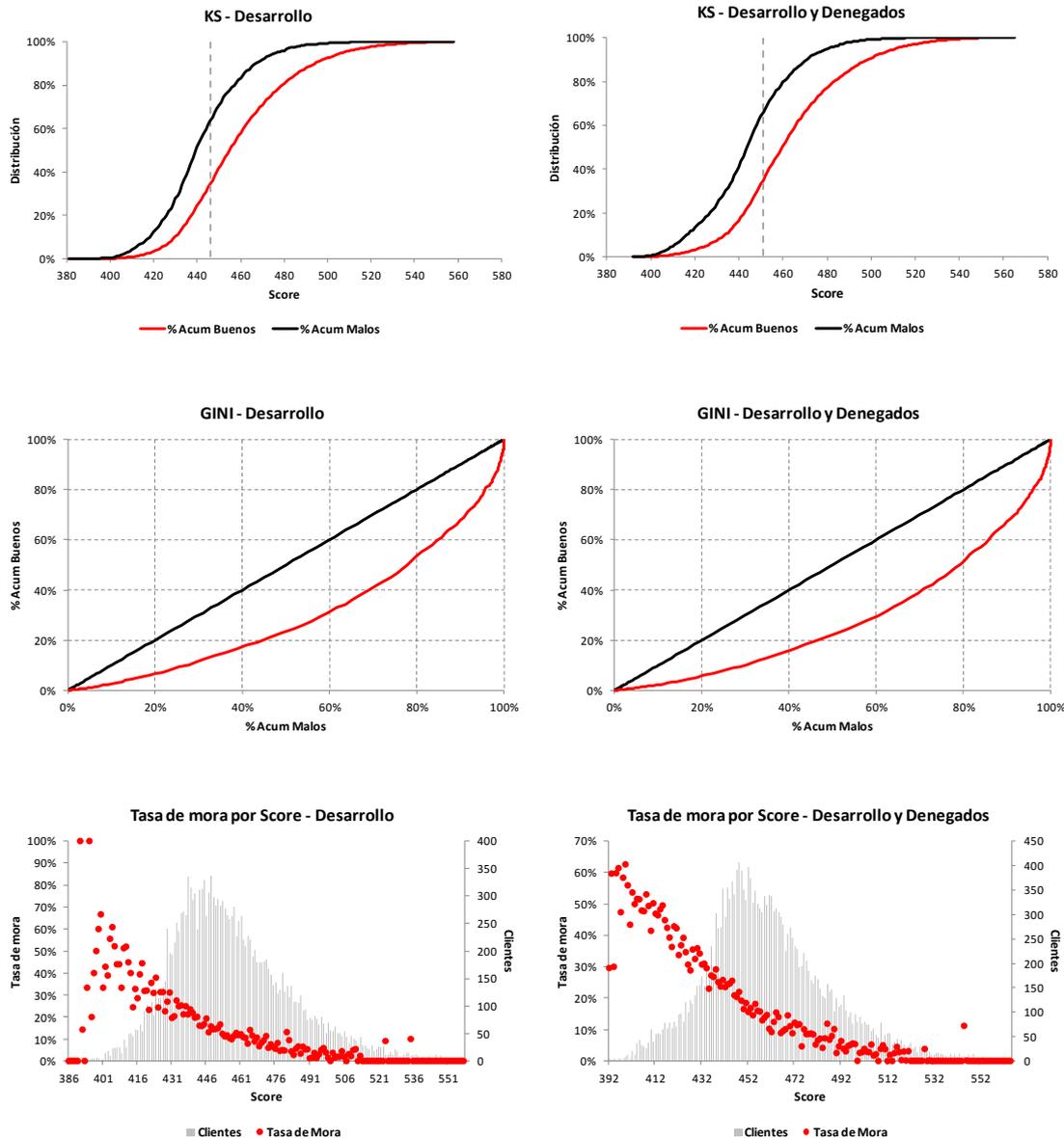
desempeño del modelo, puesto que el estadístico c, que mide el porcentaje de individuos que el modelo clasifica correctamente, varía de un 70,2% a un 69%, como se muestra en la tabla:

Variable	Desarrollo		Denegados	
	Beta	P-Valor	Beta	P-Valor
Intercept	-1,73	<,0001	-1,5844	<,0001
WOE_ANTIGUEDAD	-0,6497	<,0001	-0,6715	<,0001
WOE_TRIAD	-0,7064	<,0001	-0,6724	<,0001
WOE_PASIVO	-0,4126	<,0001	-0,4032	<,0001
WOE_ratio_AC_PAS	-0,5214	<,0001	-0,3718	<,0001
WOE_ANTIG_EMP	-0,6422	<,0001	-0,6211	<,0001
WOE_INCIDENCIAS	-0,6907	<,0001	-0,5177	<,0001
WOE_NACIONALIDAD	-0,7528	<,0001	-0,7051	<,0001
WOE_PROVINCIA	-0,5361	<,0001	-0,6165	<,0001
WOE_EST_CIVIL	-0,8826	<,0001	-0,7968	<,0001
WOE_PROFESION	-0,5139	<,000	-0,4904	<,000
Estadístico C	0,702		0,690	

Tabla 58: Resultados regresión logística de los modelos inicial y final

En cuanto al poder predictivo del modelo en la propia muestra en la que ha sido desarrollado, se observan ligeras diferencias. Las distribución de clientes buenos y malos, la distribución acumulada de los mismos, y la curva de Lorenz del modelo se muestra a continuación para ambas muestras:

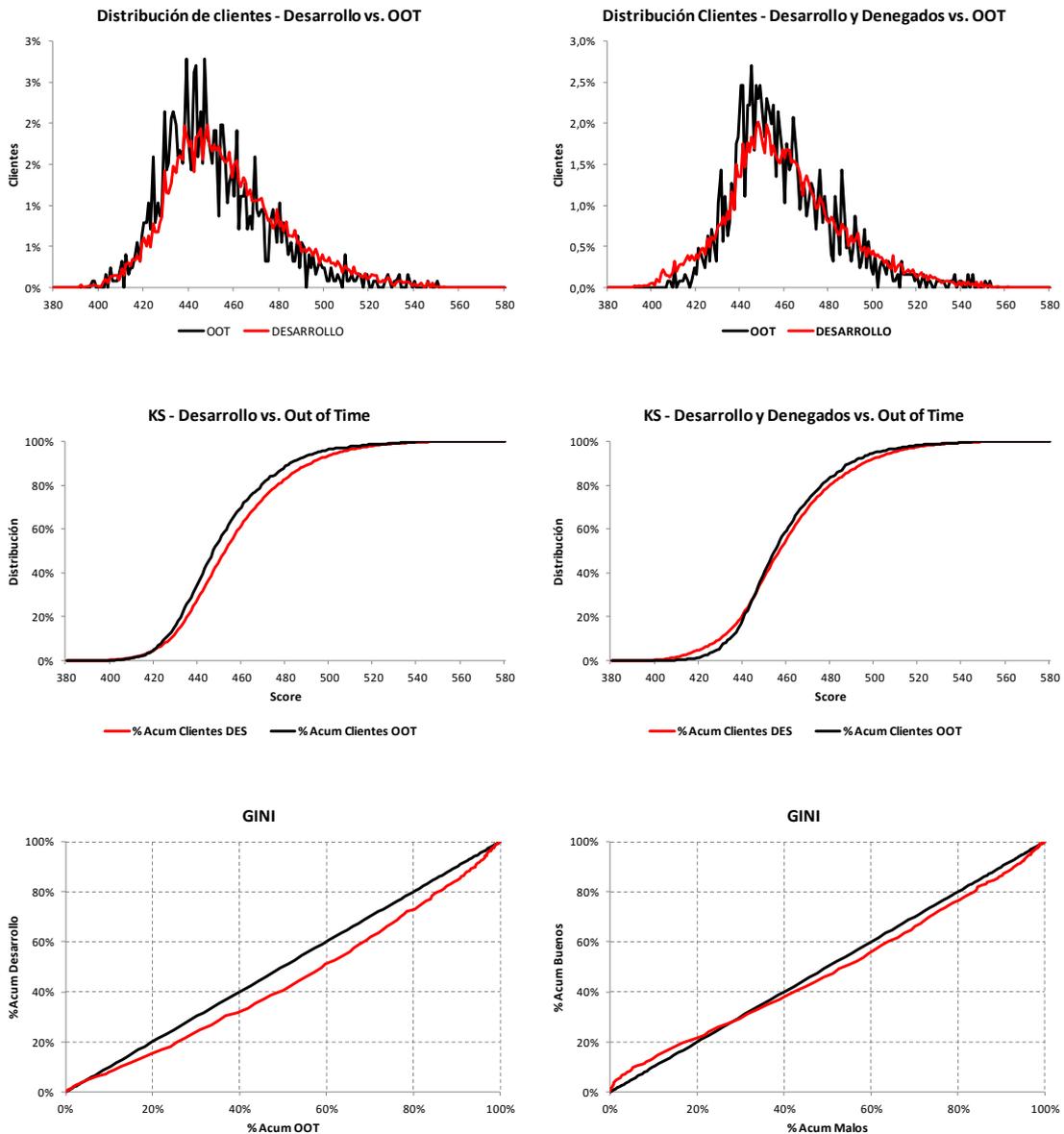




Se observan resultados similares en ambos modelos. El modelo desarrollado a partir de la muestra con denegados tiene mejor poder discriminatorio, presenta mejores estadísticos KS y Gini, como demuestra la siguiente tabla. Sin embargo, el poder predictivo disminuye cuando se aplica a muestras diferentes. Esto puede deberse a que, tanto la muestra Hold-Out como la Out of Time contienen, únicamente, clientes cuya solicitud fue admitida, por lo que la distribución de clientes puede ser diferente. Además, los clientes denegados, lo fueron en base a un modelo de scoring anterior, que podría haber rechazado clientes que el nuevo modelo acepta, y viceversa (Huang et al., 2007).

Muestra	Modelo inicial				Modelo con denegados			
	Observaciones	Tasa Mora	KS	GINI	Observaciones	Tasa Mora	KS	GINI
Desarrollo	16.973	15,21%	29,35%	40,37%	20.118	17,16%	30,99%	42,53%
Hold-Out	7.274	15,86%	25,14%	35,32%	7.274	15,86%	24,00%	34,88%
Out of Time	1.258	20,67%	26,68%	32,73%	1.258	20,67%	25,73%	31,42%

En cuanto a la estabilidad de los dos modelos desarrollados comparando las muestras de desarrollo y la muestra Out of Time, los resultados son los siguientes:



	Estabilidad		
	KS	GINI	PSI
Modelo inicial	9,52%	11,94%	13,61%
Modelo con denegados	4,57%	2,74%	12,13%

Tanto los gráficos como la tabla anteriores, muestran mejores resultados de estabilidad del modelo desarrollado a partir de la muestra que incluye denegados respecto a la muestra Out of Time. Puesto que la muestra Out of Time es relativamente pequeña (1.258 observaciones), y su tamaño difiere mucho de las muestras de desarrollo, estos resultados no son concluyentes.

Teniendo en cuenta el desempeño global de ambos candidatos, el modelo seleccionado sería el realizado a partir de la muestra con denegados. Su mayor tamaño permite obtener resultados más robustos, y sus mejores resultados de estabilidad determinan la elección final. Si bien muestra es muy antigua, y los resultados se han obtenido con datos anteriores a la actual crisis económica, por lo tanto, no podrían ser aplicados en ese momento; los estudios realizados si son aplicables a muestras más actualizadas. Por tanto, aplicando estos mismos estudios a una muestra actual de solicitudes de tarjeta de crédito de clientes extranjeros, o cualquier otro tipo de clientes, se podría obtener una scorecard de admisión útil en la gestión del riesgo de crédito. Los modelos de scoring son esenciales para la gestión del riesgo de crédito de una entidad financiera, puesto que los parámetros exigidos por la regulación bancaria (PD, LGD, EAD) vendrán segmentados, en parte, por el score del cliente.

6 Agradecimientos

Quiero agradecer a mis tutoras de prácticas Laura Romero, Ruth Manso y Covadonga Fernández por sus esfuerzos para proporcionarme los datos necesarios para la realización de este trabajo.

A Rubén Sánchez, María Jesús Espinosa, Diego Alejo y Noelia Alonso por asesorarme en el desarrollo del modelo y por su ayuda en la resolución de las dudas que han ido surgiendo por el camino.

Por último, quiero agradecer a José Miguel Rodríguez-Pardo y Jesús Ramón Simón su labor como tutores de este trabajo.

7 Referencias Bibliográficas

- [1]. Altman, E. I. y Saunders A. (1998). Credit risk measurement: Development over the last 20 years. *Journal of Banking & Finance*, 21, 1721-1742.
- [2]. Avery, R. B., Bostic, R. W., Calem, P. S. y Canner, G. B. (1996). Credit Risk, Credit Scoring, and the Performance of Home Mortgages. *Federal Reserve Bulletin*, 82 (7), 621-648.
- [3]. Boj, E., Claramunt, M., Esteve, A. y Fortiana, J. (2009). Criterios de selección de modelo en credit scoring. Aplicación del análisis discriminante basado en distancias. *Anales del Instituto de Actuarios Españoles*, 15, 209-231.
- [4]. Boj, E., Costa, T. y Fortiana, J. (2012). Bondad de ajuste y elección del punto de corte en regresión logística basada en distancias. Aplicación al problema de credit scoring. *Anales del Instituto de Actuarios Españoles*, 18, 19-40.
- [5]. Brotherton, D. y Lund, B. (2013). Information Value Statistic. En MidWest SAS Users Group 2013 Annual Conference. Paper AA-14-2013. Recuperado de:
<http://www.mwsug.org/proceedings/2013/AA/MWSUG-2013-AA14.pdf>
- [6]. Dilsha, M. y Kiruthika (2014). Model Validation of a Credit Scorecard Using Bootstrap Method. *IOSR Journal of Economics and Finance*, 3 (3), 64-68.
- [7]. Duarte, J., Escobar, C. y Reyes J. (2007). Una aplicación del modelo de regresión logística en la predicción del rendimiento estudiantil. *Estudios pedagógicos XXXIII*, 2, 101-120.
- [8]. Fuentes, L. E. (2013). Metodología para la elección de punto de corte óptimo para dicotomizar covariables continuas. *Revista cubana de genética comunitaria*, 7 (3), 36-42.
- [9]. Gadidov, B. y McBurnett, B. (2015). Population Stability and Model Performance Metrics Replication for Business Model at SunTrust Bank. SAS Conference Proceedings: SouthEast SAS Users Group. Paper 132. Recuperado de:
http://www.lexjansen.com/sesug/2015/132_Final_PDF.pdf
- [10]. García, A., Mures, M. J. y Vallejo, M. E. (2005). Aplicación del Análisis Discriminante y Regresión Logística en el estudio de la morosidad en las entidades financieras. Comparación de resultados. *Pecvnia. Revista de la Facultad de Ciencias Económicas y Empresariales. Universidad de León*, 1, 175-199.
- [11]. Huang, E. y Scott, C. (2007). Credit Risk Scorecard Design, Validation and User Acceptance. En *Credit Scoring and Credit Control conference*. Editor: HBoS. Recuperado de:

- <http://www.business-school.ed.ac.uk/crc/conferences/conference-archive?a=45569>
- [12]. Laruzzia, E. y Lazzari, V. (2000). Credit risk measurement for a bank loans' portfolio. Recuperado de:
http://www.sascommunity.org/seugi/SEUGI2000/laruzzia_creditrisk.pdf
- [13]. Leung, K., Cheong, C. y Cheong, F. (2008). Building a Scorecard in Practice. En *The 7th International Conference on Computational Intelligence in Economics and Finance*. Taiwan. Recuperado de: <https://researchbank.rmit.edu.au/view/rmit:13594>
- [14]. Majer, I. (2006). Application scoring: logit model approach and the divergence method compared. (Working Paper, Warsaw School of Economics, Institute of Econometrics, Department of Applied Econometrics, Warsaw, Poland). Recuperado de:
http://kolegia.sgh.waw.pl/pl/KAE/struktura/IE/struktura/ZES/Documents/Working_Papers/aewp10-06.pdf
- [15]. Merton, R. C. (1974). On the pricing of corporate debt: the risk structure of interest rates. *Journal of Finance*, 29 (2), 449-470.
- [16]. Mester, L. (1997). What's the Point of Credit Scoring? *Federal reserve bank of Philadelphia: Business review*, September/October 1997. Recuperado de:
<https://www.philadelphiafed.org/research-and-data/publications/business-review/1997>
- [17]. Montrichard, D. (2014). Reject Inference Methodologies in Credit Risk Modeling. En SAS Conference Proceedings: SouthEast SAS Users Group. ST-160.
- [18]. Nieto, S. (2010). Crédito al consume: La estadística aplicada a un problema de riesgo crediticio. (Tesis, Universidad Autónoma Metropolitana, Departamento de Matemáticas aplicadas e industriales, Ciudad de México, México). Recuperado de:
http://www.academia.edu/8454174/Proyecto_de_Tesis_Cr%C3%A9dito_al_Consumo_La_Estad%C3%ADstica_aplicada_a_un_problema_de_Riesgo_Crediticio
- [19]. Obrová, V. (2012). Construction and application of scoring models. En 30th International Conference Mathematical Methods in Economics. Paper 113. Recuperado de:
http://mme2012.opf.slu.cz/proceedings/pdf/113_Obrova.pdf
- [20]. Rezac, F. y Rezac, M. (2011). How to Measure the Quality of Credit Scoring Models. *Czech Journal of Economics and Finance*, 61 (5), 486-507.
- [21]. Schreiner, M. (2002). Ventajas y Desventajas del Scoring Estadístico para las Microfinanzas. Recuperado de:
http://www.microfinance.com/Castellano/Documentos/Scoring_Ventajas_Desventajas.pdf

- [22]. Thomas, L. C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, 16, 149-172.
- [23]. Wu, X. (2008). Credit Scoring Model Validation (Tesis Máster, Universiteit Van Amsterdam, Faculty of Science, Korteweg-de Vries Institute for Mathematics, Amsterdam, Netherlands). Recuperado de:
<https://esc.fnwi.uva.nl/thesis/centraal/files/f1668180724.pdf>
- [24]. Zeng, G. y Zhao, Q. (2014). A rule of thumb for reject inference in credit scoring. *Mathematical Finance Letters*, 2014 (2), 13 páginas.
- [25]. Federal Deposit Insurance Corporation (2007). Risk Management Examination Manual for Credit Card Activities
- [26]. Equal Credit Opportunity Act (1974). United States Code, Title 15, Chapter 1691.

8 Anexos

8.1 Anexo 1: Nacionalidades que componen cada grupo en el modelo inicial

Grupo	NACIONALIDAD	Clientes	Tasa Mora	Grupo	NACIONALIDAD	Clientes	Tasa Mora				
1	Albania	1014	24,46%	6	Emiratos Árabes Unidos	2154	6,82%				
	Armenia				Afganistán						
	Azerbaiyán				Antillas Holandesas						
	Bangladesh				Australia						
	Dominica				Aruba						
	República Dominicana				Bélgica						
	Gambia				Benín						
	Guinea Ecuatorial				Belice						
	Grecia				Canadá						
	Hungría				República Democrática del Congo						
	Lituania				República del Congo						
	Letonia				Suiza						
	Macedonia				Costa de Marfil						
	Nicaragua				China						
	Noruega				Checoslovaquia						
	Portugal				Cabo Verde						
	Eslovenia				Chipre						
	Sierra Leona				República Checa						
	Siria				Alemania						
Chad	Estonia										
Tailandia	Egipto										
Vietnam	Francia										
2	Austria	1370	20,29%	Reino Unido	2154	6,82%					
	Burkina Faso			Georgia							
	Brasil			Gibraltar							
	Camerún			Guatemala							
	Ghana			Guinea-Bissau							
	Guinea			Croacia							
	Líbano			Haití							
	Níger			Indonesia							
	Nigeria			Israel							
	Pakistán			Irak							
	Senegal			Islandia							
	Yugoslavia			Italia							
	3			Bosnia-Herzegovina			6264	16,81%	Jordania	2154	6,82%
Dinamarca		Japón									
Ecuador		Kenia									
India		Corea del Sur									
Marruecos		Sri Lanka									
Mali		Liberia									
Mauritania		Luxemburgo									
Polonia		Mónaco									
Rumanía		Montenegro									
El Salvador		Madagascar									
Túnez	Mongolia										
4	Angola	4216	14,66%	Macao	2154	6,82%					
	Bulgaria			Martinica							
	Bolivia			Mauricio							
	Bielorrusia			Malawi							
	Colombia			México							
	Argelia			Malasia							
	Irán			Mozambique							
	Perú			Holanda							
	Filipinas			Nepal							
	Paraguay			Panamá							
	Suecia			Puerto Rico							
	Estados Unidos de América			Arabia Saudí							
	5			Argentina			1955	12,12%	Islas Seychelles	2154	6,82%
				Chile					Eslovaquia		
Costa Rica		Togo									
Cuba		Turquía									
Finlandia		Ucrania									
Honduras		Uganda									
Irlanda		Uzbequistán									
Moldavia		Vaticano									
Rusia		Sudáfrica									
Uruguay											
Venezuela											

8.2 Anexo 2: Profesiones que conforman cada grupo en el modelo inicial

Grupo	PROFESION	Cientes	Tasa Mora
1	Empresario con asalariado Agricultor o ganadero Compatibiliza TCA/TCP Ama de casa Sin empleo activo Empresarios individuales	513	21,44%
2	Obrero, capataz, peón, aprendiz Religioso	8324	16,99%
3	Directivo Profesión liberal	2677	14,57%
4	Empleado, administrativo	4348	13,11%
5	Jubilado, pensionista Funcionario, fuerzas armadas Empresario, comerciante Estudiante No consta	1111	8,73%

8.3 Anexo 3: Provincias que conforman cada grupo en el modelo inicial

Grupo	PROVINCIA	Cientes	Tasa Mora
1	Álava Burgos Guadalajara La Rioja Lugo Segovia Teruel Valencia Valladolid Zamora	1451	19,23%
2	Alicante Almería Castellón Coruña Jaén Las Palmas Madrid Murcia Soria Vizcaya Zaragoza	9146	16,12%
3	Asturias Ávila Baleares Barcelona Córdoba Cuenca Huelva Lérida Pontevedra Tarragona Toledo	3408	14,76%

Grupo	PROVINCIA	Cientes	Tasa Mora
4	Albacete Badajoz Cáceres Cádiz Cantabria Ceuta Ciudad Real Gerona Granada Guipúzcoa Huesca León Málaga Melilla Navarra Orense Palencia S.C. Tenerife Salamanca Sevilla	2968	10,95%

8.4 Anexo 4: Tipos de estado civil que componen cada grupo en el modelo inicial

Grupo	PROFESION	Cientes	Tasa Mora
1	Soltero	8338	17,20%
2	Casado separación bienes	1688	15,64%
	Casado régimen particular		
	Pareja de hecho		
	No consta		
3	Casado	6099	12,99%
4	Divorciado	848	10,73%
	Separado		
	Viudo		

8.5 Variables no incluidas en el modelo inicial

Estas son las variables analizadas y no incluidas en el modelo inicial por el algoritmo de regresión stepwise.

Variable	Tramo	Cientes	Buenos	Malos	% Poblacion	Tasa Mora	% Buenos	% Malos	WOE	IV
Edad	-25	1.932	1.507	425	11,38%	22,00%	10,47%	16,47%	-0,453	0,0534
	25-33	5.943	5.018	925	35,01%	15,57%	34,87%	35,84%	-0,028	
	33-40	4.208	3.574	634	24,79%	15,07%	24,83%	24,56%	0,011	
	40-50	3.383	2.927	456	19,93%	13,48%	20,34%	17,67%	0,141	
	50+	1.507	1.366	141	8,88%	9,36%	9,49%	5,46%	0,552	

Variable	Tramo	Cientes	Buenos	Malos	% Poblacion	Tasa Mora	% Buenos	% Malos	WOE	IV
Activo	-50	10.744	9.394	1.350	63,30%	12,57%	65,27%	52,31%	0,221	0,0789
	50-700	1.537	1.176	361	9,06%	23,49%	8,17%	13,99%	-0,538	
	700-1500	1.742	1.392	350	10,26%	20,09%	9,67%	13,56%	-0,338	
	1500+	2.950	2.430	520	17,38%	17,63%	16,88%	20,15%	-0,177	

Variable	Tramo	Cientes	Buenos	Malos	% Poblacion	Tasa Mora	% Buenos	% Malos	WOE	IV
Ingresos	-1500	13.838	11.648	2.190	81,53%	15,83%	80,93%	84,85%	-0,047	0,0193
	1500-2100	1.848	1.581	267	10,89%	14,45%	10,99%	10,35%	0,060	
	2100+	1.287	1.163	124	7,58%	9,64%	8,08%	4,80%	0,520	

Variable	Tramo	Cientes	Buenos	Malos	% Poblacion	Tasa Mora	% Buenos	% Malos	WOE	IV
Importe Solicitado	-1000	623	517	106	3,67%	17,01%	3,59%	4,11%	-0,134	0,0249
	1000-4000	11.503	9.628	1.875	67,77%	16,30%	66,90%	72,65%	-0,082	
	4000-5500	2.632	2.272	360	15,51%	13,68%	15,79%	13,95%	0,124	
	5500+	2.215	1.975	240	13,05%	10,84%	13,72%	9,30%	0,389	

Variable	Tramo	Cientes	Buenos	Malos	% Poblacion	Tasa Mora	% Buenos	% Malos	WOE	IV
Nivel Estudios	1	5.095	4.346	749	30,02%	14,70%	30,20%	29,02%	0,040	0,0286
	2	2.055	1.846	209	12,11%	10,17%	12,83%	8,10%	0,460	
	3	4.480	3.760	720	26,40%	16,07%	26,13%	27,90%	-0,066	
	4	5.343	4.440	903	31,48%	16,90%	30,85%	34,99%	-0,126	

Grupo	Nivel Estudios	Cientes	Tasa Mora
1	No consta	5.095	14,70%
2	Estudios superiores Estudios universitarios	2.055	10,17%
3	Estudios secundarios	4.480	16,07%
4	Sin estudios Estudios primarios	5.343	16,90%

Variable	Tramo	Cientes	Buenos	Malos	% Poblacion	Tasa Mora	% Buenos	% Malos	WOE	IV
Vivienda Habitual	1	2.568	2.284	284	15,13%	11,06%	15,87%	11,00%	0,366	0,0225
	2	2.034	1.741	293	11,98%	14,41%	12,10%	11,35%	0,064	
	3	9.006	7.551	1.455	53,06%	16,16%	52,47%	56,37%	-0,072	
	4	3.365	2.816	549	19,83%	16,32%	19,57%	21,27%	-0,084	

Grupo	Vivenda Habitual	Cientes	Tasa Mora
1	Prop, libre de cargas Otros	2.568	11,06%
2	Prop, Hipotecada	2.034	14,41%
3	Alquiler	9.006	16,16%

Variable	Tramo	Cientes	Buenos	Malos	% Poblacion	Tasa Mora	% Buenos	% Malos	WOE	IV
Tipo Contrato	1	8.609	7.443	1.166	50,72%	13,54%	51,72%	45,18%	0,135	0,0183
	2	487	417	70	2,87%	14,37%	2,90%	2,71%	0,066	
	3	7.877	6.532	1.345	46,41%	17,08%	45,39%	52,11%	-0,138	

Grupo	Tipo Contrato	Cientes	Tasa Mora
1	No Informado Fijo Jubilado	8.609	13,54%
2	Temporero Otros	487	14,37%
3	Temporal Autonomo	7.877	17,08%

8.6 Índice de tablas y gráficos

8.6.1 Índice de tablas

Tabla 1: Muestras utilizadas.....	19
Tabla 2: Estadísticos de la variable TRIAD.....	23
Tabla 3: Tramos propuestos para variable TRIAD.....	24
Tabla 4: Tabla ANOVA para variable TRIAD.....	25
Tabla 5: Tasa de mora y distribución de clientes de TRIAD.....	25
Tabla 6: Estadísticos de la variable PASIVO.....	25
Tabla 7: Estadísticos de la variable PASIVO, teniendo en cuenta el 75% de la muestra.....	26
Tabla 8: Tramos propuestos para variable PASIVO.....	27
Tabla 9: Tabla ANOVA para variable PASIVO.....	27
Tabla 10: Estadísticos de la variable ANTIGÜEDAD.....	28
Tabla 11: Estadísticos de la variable ANTIGÜEDAD, teniendo en cuenta el 90% de la muestra.....	29

Tabla 12: Tramos propuestos para variable ANTIGÜEDAD	30
Tabla 13: Tabla ANOVA para variable ANTIGÜEDAD	30
Tabla 14: Estadísticos de la variable ANTIGÜEDAD EN EL EMPLEO	31
Tabla 15: Estadísticos de ANTIGÜEDAD EN EL EMPLEO, del 95% de la muestra	31
Tabla 16: Tramos propuestos para ANTIGÜEDAD EN EL EMPLEO	32
Tabla 17: Tabla ANOVA para ANTIGÜEDAD EN EL EMPLEO.....	32
Tabla 18: Grupos propuestos para NACIONALIDAD.....	33
Tabla 19: Tabla ANOVA para NACIONALIDAD.....	34
Tabla 20: Grupos propuestos para RATIO ACTIVO PASIVO.....	35
Tabla 21: Tabla ANOVA para RATIO ACTIVO PASIVO	35
Tabla 22: Grupos propuestos para PROFESIÓN	35
Tabla 23: Tabla ANOVA para PROFESIÓN	35
Tabla 24: Grupos iniciales para INCIDENCIAS	36
Tabla 25: Grupos propuestos para INCIDENCIAS.....	37
Tabla 26: Tabla ANOVA para INCIDENCIAS	37
Tabla 27: Grupos propuestos para PROVINCIA.....	37
Tabla 28: Tabla ANOVA para PROVINCIA	38
Tabla 29: Grupos iniciales para ESTADO CIVIL	38
Tabla 30: Grupos propuestos para ESTADO CIVIL.....	39
Tabla 31: Tabla ANOVA para ESTADO CIVIL	39
Tabla 32: Matriz de correlaciones variables numéricas.....	40
Tabla 33: Matriz de correlaciones WOE.....	41
Tabla 34: Resultados del modelo de regresión logística inicial.....	43
Tabla 35: Estadísticos del modelo de regresión logística inicial	43
Tabla 36: Grupos propuestos para la variable TRIAD, Ejemplo	43
Tabla 37: Matriz de confusión del modelo inicial	44
Tabla 38: Scorecard modelo inicial	45
Tabla 39: Tramos propuestos para PASIVO con denegados	58
Tabla 40: Tramos propuestos para TRIAD con denegados	58
Tabla 41: Grupos propuestos para RATIO ACTIVO PASIVO con denegados	58
Tabla 42: Grupos propuestos para NACIONALIDAD con denegados	58
Tabla 43: Grupos propuestos para INCIDENCIAS con denegados.....	58
Tabla 44: Tramos propuestos para ANTIGÜEDAD con denegados	59
Tabla 45: Grupos propuestos para PROFESIÓN con denegados.....	59
Tabla 46: Tramos propuestos para ANTIGÜEDAD EN EL EMPLEO con denegados.....	59

Tabla 47: Grupos propuestos para ESTAdO CIVIL con denegados.....	59
Tabla 48: Grupos propuestos para PROVINIA con denegados.....	59
Tabla 49: Matriz de correlaciones WOE.....	60
Tabla 50: Resultados regresión logística del modelo con denegados	61
Tabla 51: Estadísticos del modelo de regresión logística con denegados	61
Tabla 52: Matriz de confusión del modelo con denegados	62
Tabla 53: Scorecard modelo con denegados	63
Tabla 54: Resumen de resultados poder predictivo del modelo final	69
Tabla 55: Muestras aleatorias de 15.000 clientes.....	72
Tabla 56: Muestras aleatorias de 10.000 clientes.....	73
Tabla 57: Muestras aleatorias de 5.000 clientes.....	74
Tabla 58: Resultados regresión logística de los modelos inicial y final.....	75

8.6.2 Índice de gráficos

Gráfico 1: Distribución de TRIAD.....	24
Gráfico 2: Valores del estadístico Chi2 a lo largo de TRIAD	24
Gráfico 3: Distribución de PASIVO.....	26
Gráfico 4: Distribución de PASIVO del 75% de la muestra	26
Gráfico 5: Valores del estadístico Chi2 a lo largo de PASIVO	27
Gráfico 6: Tasa de mora y distribución de clientes de PASIVO	28
Gráfico 7: Distribución de ANTIGÜEDAD.....	28
Gráfico 8: Distribución de ANTIGÜEDAD del 90% de la muestra	29
Gráfico 9: Valores del estadístico Chi2 a lo largo de ANTIGÜEDAD	29
Gráfico 10: Tasa de mora y distribución de clientes de ANTIGÜEDAD	30
Gráfico 11: Distribución de ANTIGÜEDAD EN EL EMPLEO	31
Gráfico 12: Distribución de ANTIGÜEDAD EN EL EMPLEO, del 95% de la muestra	31
Gráfico 13: Valores del estadístico Chi2 a lo largo de ANTIGÜEDAD EN EL EMPLEO	32
Gráfico 14: Tasa de mora y distribución de clientes de ANTIGÜEDAD EN EL EMPLEO.....	33
Gráfico 15: Tasa de mora y distribución de clientes de NACIONALIDAD.....	34
Gráfico 16: Tasa de mora y distribución de clientes de PROFESIÓN.....	36
Gráfico 17: Tasa de mora y distribución de clientes de PROVINCIA	38
Gráfico 18: Tasa de mora y distribución de clientes de ESTADO CIVIL	39
Gráfico 19: Curva de Lorenz y Gini	47
Gráfico 20: Distribución clientes en muestra de DESARROLLO	48
Gráfico 21: Distribuciones acumuladas clientes en muestra de DESARROLLO.....	48

Gráfico 22: Curva de Lorenz en muestra de DESARROLLO	49
Gráfico 23: Tasa de mora por score del modelo inicial en muestra de DESARROLLO	49
Gráfico 24: Dsitribución de clientes en muestra HOLDOUT.....	50
Gráfico 25: Distribuciones acumuladas de clientes en muestra HOLDOUT	50
Gráfico 26: Curva de Lorenz en muestra HOLDOUT	51
Gráfico 27: Tasa de mora por score del modelo inicial en muestra HOLDOUT	51
Gráfico 28: Dsitribución de clientes en muestra OUT OF TIME	52
Gráfico 29: Distribuciones acumuladas de clientes en muestra OUT OF TIME	52
Gráfico 30: Curva de Lorenz en muestra OUT OF TIME	53
Gráfico 31: Tasa de mora por score del modelo inicial en muestra OUT OF TIME.....	53
Gráfico 32: Resumen de resultados poder predictivo del modelo inicial	54
Gráfico 33: Distribución de clientes en muestras de DESARROLLO y OUT OF TIME.....	54
Gráfico 34: Dsitribuciones acumuladas de clientes en muestras de DESARROLLO y OUT OF TIME	55
Gráfico 35: Curva de Loorenz en muestras de DESARROLLO y OUT OF TIME	55
Gráfico 36: Resultados del modelo inicial	57
Gráfico 37: Distribución de clientes en muestra de DESARROLLO CON DENEGADOS.....	64
Gráfico 38: Distribuciones acumuladas de cliente en muestra de DESARROLLO CON DENEGADOS	64
Gráfico 39: Curva de Lorenz en muestra de DESARROLLO CON DENEGADOS.....	65
Gráfico 40: Tasa de mora por score del modelo final en muestra de DESARROLLO CON DENEGADOS	65
Gráfico 41: Distribución de clientes en muestra HOLDOUT.....	66
Gráfico 42: Distribuciones acumuladas de clientes en muestra HOLDOUT	66
Gráfico 43: Curva de Lorenz en muestra HOLDOUT	67
Gráfico 44: Tasa de mora por score del modelo final en muestra HOLDOUT	67
Gráfico 45: Distribución de clientes en muestra OUT OF TIME	68
Gráfico 46: Distribuciones acumuladas de clientes en muestra OUT OF TIME	68
Gráfico 47: Tasa de mora por score del modelo final en muestra OUT OF TIME	69
Gráfico 48: Distribución de clientes en muestras de DESARROLLO CON DENEGADOS y OUT OF TIME	70
Gráfico 49: Distribuciones acumuladas de clientes en muestras de DESARROLLO CON DENEGADOS y OUT OF TIME.....	70
Gráfico 50: Curva de Lorenz en muestras de DESARROLLO CON DENEGADOS y OUT OF TIME.	71

8.6.3 Índice de ilustraciones

Ilustración 1: Explicación de la situación de default	20
Ilustración 2: Descripción metodología de inferencia de denegados	57