Innovare
Revista de ciencia y tecnología

Análisis de riesgo crediticio en un banco hondureño aplicando regresión logística y árbol de decisión.

Walter Jeremías López Flores[1]

Universidad Tecnológica Centroamericana UNITEC, San Pedro Sula, Honduras.

(Enviado: diciembre, 2019)

Resumen:

El propósito de este estudio es proponer dos modelos para la medición de riesgo crediticio en un banco de Honduras estimando la probabilidad de que un préstamo caiga en mora: uno Logit empleando una regresión logística, y un árbol de decisión implementando un algoritmo CART. Los resultados sugieren que ambos modelos son bastante adecuados para esta tarea, pero el más confiable, aunque ligeramente, fue el modelo Logit con una precisión y sensibilidad alrededor del 98%, una exactitud de 96.7% y un ROC con una curva muy pronunciada casi sin solapamiento con mucha área superior. La muestra de datos es la población completa de préstamos reportados a la Central de Información Crediticia al mes de julio de 2019 compuesta por 50,829 registros, a los cuales se les aplicó minería de datos con algoritmos de aprendizaje de máquina supervisado para su procesamiento en lenguaje R.

***Palabras Claves**:* Riesgo de crédito, banca, regresión logística, árbol de decisión, aprendizaje de máquina.

Clasificación JEL: C01, C02, C35, C51, C55, C58, C63, C81, D81, G21, G32.

Abstract:

In this study two models are proposed for the measurement of credit risk at a bank in Honduras estimating the probability that a loan will fall into default: a Logit using a logistic regression, and a decision tree implementing a CART algorithm. The results suggest that both models are quite suitable for this task, but the most reliable, although slightly, was the Logit model with an accuracy and sensitivity around 98%, a precision of 96.7% and an ROC with a very pronounced curve almost without overlap with much upper area. The data sample is the entire population of loans reported to the Credit Information Center as of July 2019, consisting of 50,829 records, to which data mining was applied with supervised machine learning algorithms for processing it in R language.

***Keywords**:* Credit Risk, banking, logistic regression, decision tree, machine learning.

JEL codes: C01, C02, C35, C51, C55, C58, C63, C81, D81, G21, G32.

1. Introducción.

La gestión integral del riesgo se ha convertido en los últimos años en un área importante y prioritaria dentro de la banca comercial y sus entes reguladores a nivel mundial, ya que el negocio principal de los bancos es, en su forma más simplista, captar recursos financieros para luego prestarlos a cambio de una tasa de interés dentro de un plazo establecido, actividad que vuelve los conceptos de riesgo y crédito como inseparables. Por ello, es vital para las entidades bancarias contar con un modelo pertinente de medición y control de riesgo crediticio que les

[1] Autor para correspondencia. Email: wjlopez@unitec.edu

permita evitar asumir niveles de riesgo difíciles de enfrentar, o en su defecto, evitar estrategias conservadoras que puedan frenar su colocación de recursos que puedan generarles buenos rendimientos hasta el punto de elevar su capital ocioso en detrimento de la rentabilidad.

El objetivo general de este estudio es proponer una metodología de evaluación crediticia de los clientes de un banco en Honduras basada en modelos de distribuciones probabilísticas, empleando regresión logística y árboles de decisión que permitan mejorar la toma de decisiones en la gestión de su cartera de préstamos. Dicha metodología permite predecir la capacidad de pago y riesgo de caer en mora de los clientes, tomando como base su información proveniente de la Central de Información Crediticia (CIC). A partir de lo anterior, se pretende responder las siguientes preguntas de investigación: ¿cuál es la probabilidad de que un préstamo con las variables observadas dentro de los modelos caiga en impago o *default*? y ¿cuál modelo de clasificación permite evaluar el riesgo de la cartera crediticia de manera más adecuada?. Las respuestas a estas preguntas conllevan una implicación práctica de mucho interés para el sector financiero, así como un aporte y contribución teórica al campo de la inteligencia de negocios dentro de este tipo de organizaciones.

2. Marco Teórico.

En las entidades financieras, el riesgo se entiende como la posibilidad de que pueda producirse una pérdida (Velandia, 2013). La globalización de los mercados y la desregulación de las economías, son tendencias mundiales que han afectado la operación de las instituciones financieras, cuya esencia de la actividad es la toma de riesgos, por lo que este es un componente inevitable en su operación. La importancia del conocimiento y análisis del riesgo radica en el hecho de que existe una relación directa entre el grado de riesgo asumido por una institución y el potencial de utilidades a ser generadas (De la Fuente, 2004).

2.1 La gestión integral del riesgo (Risk Management).

El comité de supervisión bancaria de Basilea ha sido precursor de la reglamentación de la medición integral de riesgos y el adecuado provisionamiento de capitales para sobrellevar los posibles riesgos incurridos y evitar la quiebra de las instituciones financieras (Cardona Hernández, 2004).

Un marco para la gestión de riesgos brinda las políticas, procedimientos y disposiciones organizacionales que incluirán la gestión de riesgos en toda la organización y en todos los niveles (ISO, 2009). Dentro del análisis de riesgo se deben identificar las principales fuentes de exposición, por lo que los tipos de riesgo se clasifican así (Cardona Hernández, 2004):

- *Riesgo de liguidez*: es la posibilidad de que una institución financiera no pueda cumplir un compromiso financiero con un cliente o mercado en algún lugar, moneda o momento determinado.
- *Riesgo legal*: es la contingencia de pérdida derivada de situaciones de orden legal o normativo impuestas por el gobierno o ente regulador.
- *Riesgo operativo*: es la posibilidad de pérdida como resultado de deficiencias a causa de fallas en los sistemas de información, fallas en procesos, errores humanos, mala fe de los funcionarios y fallas en el control gerencial.

- *Riesgo de mercado*: es la contingencia de pérdida o ganancia de una posición de la entidad financiera, como resultado de un cambio en el nivel o la volatilidad de las tasas de interés (*riesgo de interés*), tasas de cambio (*riesgo cambiario*) o precios.
- *Riesgo de contraparte*: es la posibilidad de incumplimiento de las obligaciones contractuales entre la entidad financiera y el sector real o financiero.
- *Riesgo de crédito*: es la posibilidad de que una entidad incurra en pérdidas y se disminuya el valor de sus activos como consecuencia de que sus deudores fallen en el cumplimiento oportuno o cumplan imperfectamente los términos acordados. Este es el tipo de riesgo en el que se centrará este estudio.

2.2 El riesgo de crédito.

Saavedra García (2010) define el riesgo de crédito como la probabilidad de que, a su vencimiento, una entidad no haga frente, en parte o en su totalidad, a su obligación de devolver una deuda o rendimiento, acordado sobre un instrumento financiero, debido a quiebra, iliquidez o alguna otra razón (Chorafas, 2000), y que puede analizarse en tres dimensiones (Galicia, 2003): riesgo de incumplimiento, exposición y recuperación.

Para Basso (2013) el riesgo de crédito se define como la posibilidad de incurrir en pérdidas producto del incumplimiento, por falta de solvencia, de las obligaciones contractuales asumidas por una contraparte, el cuál debe gestionarse mediante políticas conservadoras y diseño de procedimientos adecuados de admisión, seguimiento y recuperación, soporte de herramientas de proceso de información, así como sistemas propios de calificación de *rating* y herramientas automáticas de decisión (*credit scoring*, sistemas expertos), teniendo dos niveles de análisis mediante modelos:

1. *Individual*: posee tres parámetos coincidentes con las dimensiones de (Galicia, 2003):
    a. *Probabilidad de inclumplimiento o default*: es la frecuencia relativa con la que pueda ocurrir el no pago del deudor sobre la obligación contraída.
    b. *Tasa de Recuperación*: es la proporción de deuda que se podrá recuperar una vez ocurrido el incumplimiento.
    c. *Exposición Crediticia*: principal remanente más los intereses acumulados.
2. *Portafolio*: considera dos puntos:
    a. *Participación*: de cada crédito u operación en el portafolio total.
    b. *Correlación*: entre los diferentes activos que lo componen.

También define que para el manejo e identificación del riesgo crediticio existen dos enfoques: el tradicional, que se basa en la experiencia de los oficiales de crédito, donde la decisión de crédito surge de la reflexión de los funcionarios sobre la capacidad de pago del cliente utilizando medias de riesgo arbitrarias; y el actual, que plantea la necesidad de contar con técnicas de manejo de riesgos más sofisticadas, acordes con mercados financieros competitivos y productos financieros complejos y diversos, eliminando la selección adversa e incrementando la sensibilidad al riesgo individual y de portafolio (Basso, 2013).

2.3 Métodos de medición del riesgo crediticio.

El propósito de la valoración del riesgo es suministrar información y análisis con base en evidencias para tomar decisiones informadas sobre la manera de tratar los riesgos particulares y seleccionar entre diferentes opciones (ISO, 2009). El riesgo crediticio se puede evaluar a través de métodos estadísticos mediante técnicas de *Credit Scoring* tanto

paramétricas, que suponen conocida una función y distribución y entre las que se encuentran el análisis discriminante y los modelos de probabilidad lineal, Logit y Probit; así mismo están las no paramétricas, que son métodos de distribución libre que buscan parámetros de una función y distribución conocida, como la programación lineal, redes neuronales y árboles de decisión (Pantoja Vilchez, 2016).

Una clasificación de modelos para el riesgo crediticio tomando como base la probabilidad de inclumplimiento del deudor es la propuesta por Basso (2013):

1. *Modelos expertos*: basados en criterios subjetivos y el juicio o experiencia del analista de cartera. Entre estos se encuentran las 5 C del crédito (Carácter, Capital, Capacidad, Colateral y Ciclo económico).
2. *Modelos paramétricos*: calculan las probabilidades de incumplimiento utilizando la información de un conjunto de variables que caracterizan a los individuos sujetos de crédito, sin pretender conocer las causas que las generan y toman como base calificaciones de riesgo de clasificadoras, combinaciones de apalancamiento, distancias al vencimiento del crédito, etc. Entre ellos enumera:
    a. *Modelos de Scoring*: técnica estadística que clasifica las observaciones en grupos definidos a priori, según un conjunto de variables que caracterizan al individuo que se desea clasificar, que pueden ser de Análisis Discriminante (Z-score, Z-model, EMS- Emerging Markets Corporate Bond System) o de elección cualitativa (Probabilidad lineal, Probit, Logit).
    b. *Matrices de transición*: el método Credimetrics.
    c. *Modelos de frecuencias esperadas de incumplimiento EDF*: "Portafolio Manager" y "Credit Monitor" de KMV Corporation.
    d. *Análisis actuarial*: "Credit Risk+" de CSFP.
    e. *Modelos RAROC*: Permite determinar el "rendimiento óptimo" de una facilidad crediticia sujeto a que éste cubra las pérdidas esperadas y algún margen deseado de las pérdidas no esperadas, de manera tal que el retorno sobre el capital ajustado por riesgo (RAROC) sea como mínimo superior al costo de oportunidad del capital.
3. *Modelos condicionales*: Son metodologías que pretenden conocer las causas del incumplimiento sobre un análisis basado en un modelo con relaciones de causalidad entre las diferentes variables financieras, sectoriales y macroeconómicas, entre ellos: Credit Portafolio View de McKinsey, "Algo Credit" de Algoritmics, y "CredScoRisk" de AIS.

Pantoja Vilchez (2016) explica en base a Rayo et al. (2010) que los modelos de probabilidad lineal emplean el método de regresión por mínimos cuadrados ordinarios (MCO) donde la variable dependiente toma el valor de uno si el cliente cae en mora y de cero si cumple con el pago en función lineal de las variables explicativas, pero según Kim (2005) tienen la desventaja de que las probabilidades estimadas podrían quedar fuera del intervalo (0,1). Esta situación se corrige en los modelos Logit que usan una regresión logística para calcular la probabilidad de caer o no en impago sin necesidad de plantear hipótesis de partida, mejorando el tratamiento de las variables cualitativas categóricas y siempre dentro del rango de variación entre 0 y 1.

También explica en cuanto a los modelos Probit basado en Lara (2010) que, a menos que las muestras sean grandes, desde una perspectiva teórica ofrecerán resultados similares en términos de probabilidad, dado que la distribución normal y la logística acumulada están muy próximas entre sí, excepto en los extremos. En cuanto a la programación lineal permite programar plantillas o sistemas de asignación de rating sin perder de vista el criterio de optimización de clientes correctamente clasificados, pero son inexactos en la predicción, no

estiman probabilidades de impago y son de difícil comprensión, estas dos últimas desventajas al igual que la no estimación directa de parámetros aplican a las redes neuronales, que tratan de imitar el sistema nervioso mediante un sistema de nodos con cierto grado de inteligencia.

Sobre los árboles de decisión, su principal ventaja es que no están sujetos a supuestos estadísticos de distribuciones o funciones y presentan relaciones visuales entre las variables, grupos de la variable respuesta y el riesgo (Pantoja Vilchez, 2016). Estos son categorizados por Saavedra García (2010) como metodología usada por los sistemas expertos, quien clasifica los modelos de valuación del riesgo crediticio en tradicionales y modernos, quedando dentro de los primeros junto a los sistemas de calificación. Entre los modernos menciona los siguientes modelos: KMV, valuación de Merton, Credimetrics de J.P. Morgan, Credit Risk+, Retorno sobre capital ajustado al riesgo y CyRCE. Para efectos de este estudio se profundizará en la regresión logística (Logit) y el árbol de decisión.

2.4 El modelo Logit: evaluando el riesgo crediticio mediante la regresión logística.

En el modelo Logit, la probabilidad de inclumplimiento de un deudor se distribuye como una función logística de acuerdo con la siguiente fórmula (Basso, 2013):

$$f(Z_i) = \frac{1}{1 + e^{-Z_i}} \qquad \textbf{(1)}$$

Donde: $Z_i = \sum \beta_j X_{ij} + U_j$ siendo que U_j es un error que se distribuye normalmente.

La variable dependiente binaria (dummy) como es dicotómica se ajusta a una distribución binomial discreta, cuyos casos aplicados al *default* de un préstamos se establecerán de la siguiente forma tomando como base el estudio de Támara Ayús, Aristizábal, & Velásquez (2010):

$$y_i = \begin{cases} 0, & El\ préstamo\ no\ está\ en\ mora \\ 1, & El\ préstamo\ \ sí\ \ está\ en\ mora \end{cases} \qquad \textbf{(2)}$$

Las variables explicativas, denominadas covariables pueden ser de cualquier naturaleza: dicotómicas o politómicas, ya sean nominales u ordinales, y contínuas. El proceso que se sigue para estimar los parámetros del modelo es complejo, ya que en esta situación particular se realiza por el método de máxima verosimilitud (Velandia, 2013).

En el lenguaje R se utiliza la función glm (Generalized Linear Models) para ajustar modelos lineales generalizados, especificados mediante una descripción simbólica del predictor lineal y una descripción de la distribución del error, y la optimización por máxima verosimilitud del modelo la hace mediante el algoritmo de Fisher Scoring.

2.5 El árbol de decisión aplicado al análisis de riesgo de crédito.

Los árboles de decisión (Decision Trees, DT) son una popular herramienta utilizada en análisis estadístico y minería de datos, ideales para realizar clasificación y predicción mediante una estructura de árbol donde cada nodo representa una "prueba" o condición sobre el valor de un atributo, las ramas representan el resultado de la evaluación del atributo y las hojas (finales en el árbol) son las clases o variables dependientes. Estos permiten dividir un extenso conjunto de datos relacionados entre sí, en conjuntos más pequeños de datos mediante la aplicación secuencial de sencillas reglas de decisión (Tello, Eslava, & Tobías, 2013).

Cardona (2004) explica que son un método no paramétrico que no requiere supuestos distribucionales, permite detectar interacciones, modela relaciones no lineales y no es sensible a la presencia de datos faltantes y outliers (Breiman et al., 1984; Kass, 1980) y cuyo principio básico es generar particiones recursivas por reglas de clasificación hasta llegar a una final, tal que es posible identificar perfiles (nodos terminales) en los que la proporción de clientes malos es muy alta o baja, y de esta forma asignar su probabilidad.

Por su parte, Feldman & Gross (2005) acentúan la importancia de los árboles de clasificación y regresión (Classification and Regression Trees - CART) por ser apropiados para trabajar con cantidades grandes de datos (Big Data), que contienen alta dimensionalidad, tipos de datos mezclados, datos perdidos, diferentes relaciones entre las variables en diferentes partes del espacio de medición y valores atípicos (outliers).

El método del árbol de decisión es una técnica predictiva usada en aprendizaje de máquinas (machine learning) para ambos: clasificación y regresión, y su implementación del algoritmo CART en el lenguaje R se llama *Recursive Partitioning And Regression Trees* (RPART, de sus siglas en inglés), el cuál funciona con una técnica de partición recursiva dividiendo repetidamente los datos en múltiples subespacios, de modo que los resultados en cada subespacio final sean lo más homogéneos posible. El resultado producido consiste en un conjunto de reglas generadas por el modelo y visualizadas como un árbol binario para predecir la variable resultado, que puede ser una variable contínua para árboles de regresión o una variable categórica para árboles de clasificación (Kassambara, 2018).

A continuación se presentan las fórmulas y relaciones con las que el algoritmo RPART calcula la probabilidad de cada nodo A para futuras observaciones (3), su riesgo (4), el riesgo de todo el modelo o árbol completo (5) y el criterio de división (6) (Therneau & Atkinson, 2019):

$$P(A) = \sum_{i=1}^{C} \pi_i P\{x \in A | \tau(x) = i\} \approx \sum_{i=1}^{C} \pi_i n_{iA}/n_i \quad \textbf{(3)}$$

Donde π_i son las probabilidades previas de cada clase; $\tau(x)$ es la verdadera clase de una observación x, donde x está en el vector de variables predictoras; n_i es el número de observaciones en la muestra que son clase i; y n_A es el número de observaciones en el nodo A.

$$R(A) = \sum_{i=1}^{C} p(i|A) L(i, \tau(A)) \quad \textbf{(4)}$$

Donde $\tau(A)$ es la clase asignada a A, si A fuese a ser tomado como el nodo final y es elegido para minimizar este riesgo.

$$R(T) = \sum_{j=1}^{k} P(A_j) R(A_j) \quad \textbf{(5)}$$

Donde A_j son los nodos terminales del árbol.

Para construir el árbol, el criterio de división es el siguiente: si se parte un nodo A en dos hijos A_L y A_R (hijo izquierdo y derecho) se tendrá la siguiente relación:

$$P(A_L) r(A_L) + P(A_R) r(A_R) \leq P(A) r(A) \quad \textbf{(6)}$$

2.6 Descripción de la banca comercial dentro del sistema financiero hondureño.

El sistema financiero hondureño está compuesto por bancos comerciales, estatales, de segundo piso, aseguradoras, fondos de pensiones y sociedades financieras, regulados todos por la Comisión Nacional de Bancos y Seguros (CNBS); otro ente regulador es el Banco Central de Honduras (BCH) quien dicta la política monetaria, crediticia y cambiaria del país. El sector bancario está compuesto por quince bancos comerciales autorizados para realizar intermediación financiera de manera habitual y sistemática mediante operaciones de financiamiento a terceros con recursos captados del público en forma de depósitos, préstamos u otras obligaciones (CNBS, 2019).

Los activos del sistema bancario comercial, son conformados principalmente por la cartera de créditos y las inversiones. La cartera crediticia está segmentada en cuatro tipos de créditos que durante el periodo 2014-2018 han mantenido en promedio el 66.0% los créditos comerciales, 18.5% consumo y tarjetas de crédito, vivienda el 14.6% y microcrédito representa el 0.9% del total de la cartera (CNBS, 2019).

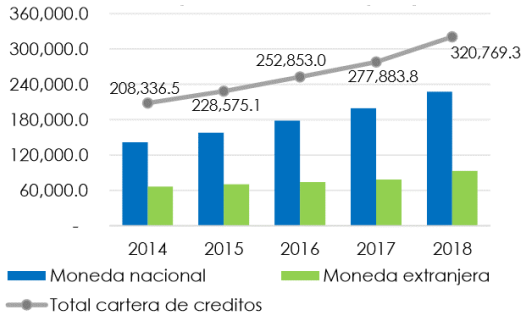


Figura 1. Cartera crediticia bancos comerciales de Honduras (en millones de Lps).
Fuente: (CNBS, 2019)

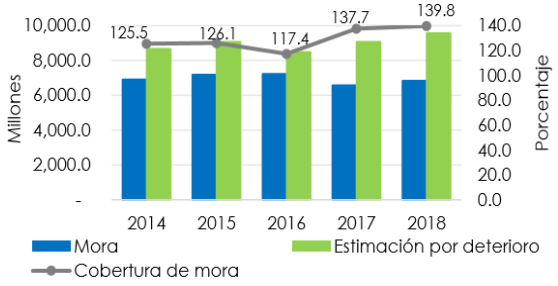


Figura 2. Cobertura de mora de la cartera crediticia de bancos comerciales de Honduras.
Fuente: (CNBS, 2019)

El crédito otorgado por la banca comercial denotó una expansión interanual del 15.4% (L42,885.5 millones) a diciembre de 2018, variación que supera el 9.9% (L25,030.8 millones) registrado en el año 2017. Este comportamiento es impulsado por el crédito otorgado en moneda nacional, cuyo crecimiento pasa de 11.7% (L20,854.1 millones) registrado en el año 2017 a 14.1% (L28,128.5 millones) en el año 2018. La cobertura de préstamos en mora durante el periodo 2014-2018 se mantuvo en un rango mayor al 110% requerido según normativa vigente, manteniendo un promedio de 129.3%. Para el año 2018 aumentó 2.1 puntos porcentuales en comparación al año anterior, debido al aumento de las estimaciones por deterioro de un 5.5% y los créditos en mora (CNBS, 2019).

2.7 Estudios previos sobre riesgo de crédito aplicando modelos Logit y CART.

Támara Ayús, Aristizábal, & Velásquez (2010) hicieron un estudio en Colombia donde estimaron las provisiones esperadas en una institución financiera usando modelos Logit, Probit y árbol de decisión en una muestra de 1,500 clientes clasificados dentro del portafolio de cartera comercial midiendo las siguientes variables: actividad, edad, ingresos, activos, margen operativo neto, endeudamiento y razón pasivo/ingreso, cuyo árbol permitió agruparlos en niveles bajo, medio y alto con respecto a su grado de endeudamiento. En cuanto a los modelos Logit y Probit, comprobaron que al ser su naturaleza muy similar, generan resultados parecidos notándose la diferencia entre la provisión de los modelos de la entidad y el ente regulador.

Muy similar, aunque no aplicado dentro de un banco, sino en una entidad comercial, es el estudio de Ruiz (2017), donde se diseñó un modelo de calificación de clientes morosos usando árboles de decisión, análisis discriminante y regresión logística. Lo valioso de su aporte es que compara los resultados entre las tres metodologías, demostrando que cualquiera de ellas es factible, pero que la regresión registró un mayor poder de discriminación con un K-S[2] del 24.3%, mientras que los árboles obtienen el menor K-S entre los tres modelos con el 19.7%, así mismo, las metodologías paramétricas (Regresión Logística y Análisis Discriminante) alcanzan el 95% de coincidencia en calificar a los clientes por niveles de riesgo, pero a pesar de que los árboles obtienen menor poder de discriminación que los dos métodos paramétricos estudiados, estos tienen la ventaja de que sus segmentos tienen una descripción explicita de sus perfiles, mientras que las otras técnicas asignan un puntaje o una probabilidad.

Feldman (2005) realizó un estudio de impago de hipotecas aplicando análisis con árboles de clasificación incorporando un conjunto de datos muy grande con variables contínuas y categóricas describiendo las características de tamaño y tipo de préstamo, así como a los deudores. Al igual que un estudio de Galindo & Tamayo (2000) donde hacen una evaluación de riesgo crediticio apoyados en *machine learning* combinando modelos Probit, CART y Redes Neuronales.

3. Metodología.

3.1 Diseño del estudio y datos de entrada.

El enfoque del estudio es cuantitativo de alcance explicativo y diseño no experimental de corte transversal. La fuente de datos es secundaria y consistirá en los datos de la cartera de préstamos completa de un banco de Honduras que reporta a la CNBS, consistente en 50,829 observaciones. Todo el conjunto de datos se obtuvo de la CIC en formato de hoja de cálculo electrónica con saldos a julio de 2019, la estructura de las variables que componen el *dataset* se explican en la *Tabla 1*, en la siguiente página.

3.2 Identificación de variables y procesamiento de la información.

La variable respuesta para los modelos será una variable binaria llamada *EnMora*, que denotará si el préstamo está en *default* o no, en función de las demás variables independientes que se seleccionen para predecirla. Se dispone de 50 variables tanto cualitativas como cuantitativas que pueden ser usadas como predictoras, y se seleccionarán según su pertinencia y con diferentes combinaciones hasta determinar cuáles conforman el modelo óptimo.

La preparación preliminar y limpieza de los datos se hará en Excel, y su análisis explotario y descriptivo en conjunto con Minitab. Una vez preparados se construirán los modelos Logit y CART en el lenguaje R utilizando Jupyter Notebook 6.0 bajo el framework Anaconda para Python 3.7 procesándolos con técnicas de minería de datos empleando algoritmos y rutinas de aprendizaje de máquina supervisado. Los datos se dividirán en dos subconjuntos: el 70% (35,580 registros) para entrenamiento de la máquina y el restante 30% (15,249 préstamos) para prueba y comprobación de la capacidad predictora de estos.

[2] Prueba Kolmogorov-Smirnov (K-S): Es una prueba que se basa en medir la separación de las distribuciones acumuladas de los buenos y malos para cada rango percentil del puntaje.

Tabla 1. Estructura de los datos de la población de préstamos

No	Variable	Descripción	Tipo de dato	Categorías
1	CODSUC	Código de la sucursal del banco a la que pertenece el préstamo.	Alfanumérico, nominal.	01: Santa Rosa, 02: San Pedro, 03: La Esperanza, 04: Tegucigalpa, 05: Siguatepeque, 06: Comayagua, 07: Choluteca, 08: Juticalpa, 09: La Ceiba.
2	CODOFI	Código de la oficina donde fue otorgado el préstamo.	Alfanumérico, nominal.	Hay 106 códigos para cada oficina.
3	Zona	Zona geográfica del país. donde se encuentra la oficina que otorgó el préstamo.	Texto, nominal.	"Centro", "Norte", "Occidente".
4	FechaOtorgado	Fecha en que se otorgó el préstamo, originalmente el archivo tiene un campo FECOTO alfanumérico que se separó en tres columnas que contienen el año, mes y día: ANIOTO, MESOTO y DIAOTO para calcular este nuevo campo.	Fecha, discreta, serie de tiempo.	N/A.
5	FechaVencimiento	Fecha en que vence el préstamo, originalmente el archivo tiene un campo FECVEN alfanumérico que se separó en tres columnas que contienen el año, mes y día: ANIVEN, MESVEN y DIAVEN para calcular este nuevo campo.	Fecha, discreta, serie de tiempo.	N/A.
6	DuracionMeses	Campo calculado como la diferencia en meses entre las fechas de otorgado y de vencimiento.	Entero, discreta.	N/A es de conteo.
7	SALVIG	Saldo Vigente.	Decimal.	NA es contínua.
8	ESTAOP	Estado actual del capital en seguimiento a normas contables vigentes.	Texto, nominal.	EJE: Ejecución, MOR: Mora, VAN: Vencido anticipado, VEN: Vencido, VIG: Vigente.
9	SaldoVigente y SaldoMora	Estas columnas son mutuamente excluyentes a partir de los datos de SALVIG y ESTAOP, los clasifica y en la que no corresponde el saldo será cero.	Decimal.	N/A son contínuas.
10	**EnMora** (*Variable Dependiente*)	Calculada a partir de SaldoMora, si este es > 0 se considera que el cliente cayó en *default*, impago o mora.	Dicotómica (dummy)	0: No está en mora, 1: En mora
11	INTXCO	Saldo de intereses por cobrar.	Decimal.	N/A es contínua.
12	TotalCartera	Suma de SALVIG + INTXCO.	Decimal.	N/A es contínua.
13	MONOTO	Monto otorgado del préstamo.	Decimal.	N/A es contínua.
14	VALGAR	Valor de la garantía.	Decimal.	N/A es contínua.
15	NODIAA	Número de días de atraso.	Entero, conteo.	NA es discreta.
16	RangoDiasAtraso	Categorías de rangos de días de atraso según el valor de NODIAA	Texto, ordinal.	Cero días; Hasta 30 días; De 31 a 90 días; De 91 a 180 días; De 181 a 360 días; De 361 a 1,000 días; Más de 1,000 días.
17	TIPGAR	Tipo de Garantía que presenta el préstamo según su naturaleza.	Entero, conteo, nominal.	16 tipos en la tabla 9, manual (CNBS, 2015, pág. 112).

18	Garantía	Nombre de la garantía en base a TIPGAR (Fiduciaria, Accesoria, Prendaria, etc).	Texto, nominal.	16 tipos en la tabla 9, manual (CNBS, 2015, pág. 112).
19	TIPOPE	Código del tipo de operación (Préstamos a la vista, documentos descontados, garantías bancarias, etc).	Alfanumérico, nominal.	38 tipos en la tabla 10, manual (CNBS, 2015, pág. 113).
20	TIPMON	Tipo de moneda (hay hasta para 9 monedas pero en los datos solo hay 2).	Entero, nominal.	1: Nacional, 2: Dólares EUA.
21	Moneda	Nombre de la moneda según TIPMON.	Texto, nominal.	"Lps": Lempiras, "Dls": Dólares.
22	TIPCRE	Código del tipo de crédito según los criterios de las normas para clasificación de cartera vigentes.	Entero, nominal.	1: Comercial, 2: Consumo, 3: Vivienda, 4: Microcrédito.
23	TipoCrédito	Nombre del tipo de crédito (aparecen 4 en el conjunto de datos, a pesar de que solo salen códigos del 1 al 3).	Texto, nominal.	Agropecuario, consumo, comercial y vivienda.
24	ORIFON	Código del origen de los fondos que el banco usa para colocar los créditos.	Entero, nominal.	9 tipos en la tabla 16, manual (CNBS, 2015, pág. 118).
25	Origen	Nombre del origen según ORIFON (en los datos aparecen 6 de los 9).	Texto, nominal.	Propios, Banhprovi, Disp Inmediata, BCIE, RAP, Otros
26	DESCRE	Códigos del destino del crédito.	Alfanumérico, nominal.	223 tipos en la tabla 22, manual (CNBS, 2015, págs. 121-126).
27	Destino	Descripción del destino según DESCRE.	Texto, nominal.	
28	CREDES	Crédito Especial, identifica características particulares del crédito.	Carácter, nominal.	12 tipos en la tabla 15, manual (CNBS, 2015, pág. 118).
29	REFINA	Código del tipo de renegociación.	Carácter, nominal.	A: Readecuación, F: Refinanciamiento, N: Renovación, Z: No aplica.
30	Renegociación	Descripción del tipo según REFINA.	Texto, nominal.	
31	FORPAG	Código de la forma de pago del capital del préstamo.	Alfanumérico, nominal.	28 tipos en la tabla 14, manual (CNBS, 2015, pág. 117).
32	FormaPagoCapital	Forma de pago de capital según FORPAG.	Texto, nominal.	
33	INTSUS	Saldo de intereses en suspenso.	Decimal.	N/A es contínua.
34	TASINT	Tasa de interés anual nominal.	Decimal.	N/A es contínua.
35	GDIVIS	Identifica al deudor si es generador de divisas independientemente de la moneda en que se otorgó el crédito.	Texto, nominal.	8 tipos en la tabla 5, manual (CNBS, 2015, pág. 86).
36	Divisas	Descripción según GDIVIS (aparecen 3 de los 8 tipos definidios).	Texto, nominal.	GA: Generador, NG: No Generador, ZZ: No Aplica.
37	CIIU	Clasificación Industrial Internacional Uniforme: identifica la actividad económica del deudor.	Alfanumérico, nominal.	Grupos de la tabla 8, manual (CNBS, 2015, págs. 93-112).
38	UBICAC	Ubicación geográfica: identifica al departamento de Honduras donde será utilizado el crédito.	Entero, discreto, nominal.	19 tipos en la tabla 18, manual (CNBS, 2015, pág. 119).
39	REGION	Ubicación o área en donde será utilizado el crédito.	Carácter, dicotómica.	U: Urbana, R: Rural.

40	CATEGF	Corresponde a la categoría definitiva asignada a las operaciones del deudor en cumplimiento de las regulaciones vigentes en la materia.	Texto, nominal.	Grupos de la tabla 12, manual (CNBS, 2015, págs. 115-116).
41	Categoria	Categoría de Riesgo asignada según catálogo en base a los diferentes criterios de reseerva y mora según la clasificación de CATEGF del crédito.	Texto, ordinal.	I: Bueno, II: Especialmente mencionado, III: Bajo Norma, IV: Dudosa recuperación, V: Pérdida.
42	TipoCliente	Categorías de clientes según su tamaño y tipo de préstamo.	Texto, nominal.	
43	RESERV	Porcentaje de reserva o provisión, asignado en cumplimiento a las Normas de Evaluación y Clasificación de Cartera Crediticia vigentes.	Decimal.	N/A es contínua.
44	FUPACA	Fecha de último pago a capital	Alfanumérico.	No es necesaria.
45	COATPO	Costo Anual Total Porcentual: tasa anual que incluye la totalidad de los costos y gastos inherentes a los créditos, es la tasa de interés efectiva.	Decimal.	N/A es contínua.
46	TAINMO	Tasa de Interés Moratoria: tasa anual nominal de sobrecargo que se cobra en caso de que el deudor incumpla con las condiciones contratadas.	Decimal.	N/A es contínua.
47	GRACIA	Periodo de gracia: se reporta en meses y corresponde al plazo de la operación dentro del cual se pactó originalmente con el deudor no amortizar capital de la obligación.	Entero, discreta.	N/A es de conteo.
48	SALCMO	Saldo de Capital en Mora.	Decimal.	N/A es contínua.
49	MTOCTA	Monto de la cuota a pagar en el período: pago periódico para la amortización de la deuda.	Decimal.	N/A es contínua.
50	MANCOMUNADO	Si el préstamo es individual o a nombre de dos o más clientes.	Dicotómica (dummy)	0: No, 1: Sí.
51	SALARIO	Registra el ingreso declarado por el cliente en el préstamo.	Decimal.	N/A es contínua.

Fuente: Elaboración propia.

4. Resultados.

4.1 Análisis preliminar de los datos.

La cartera de préstamos analizada es bastante saludable, ya que solo un 8.4% de esta se encuentra en mora. Al analizar la mora por las categorías de riesgo de los préstamos que componen la cartera se obtienen los resultados siguientes resumidos en la *Tabla 2*:

Tabla 2. Préstamos en mora según su categoría de riesgo asignada.

Categoría de Riesgo	En Mora				Total general	
	No		Sí			
I	L 38,596,193,707.66	73.65%	L 431,873,700.34	0.82%	L 39,028,067,408.00	74.47%
II	L 6,165,128,428.65	11.76%	L 272,500,621.52	0.52%	L 6,437,629,050.17	12.28%
III	L 1,821,772,194.31	3.48%	L 433,816,406.96	0.83%	L 2,255,588,601.27	4.30%
IV	L 516,486,578.84	0.99%	L 833,452,490.74	1.59%	L 1,349,939,069.58	2.58%
V	L 902,429,066.64	1.72%	L 2,430,987,820.30	4.64%	L 3,333,416,886.94	6.36%
Totales:	L 48,002,009,976.10	91.60%	L 4,402,631,039.86	8.40%	L 52,404,641,015.96	100 %

Fuente: Elaboración propia.

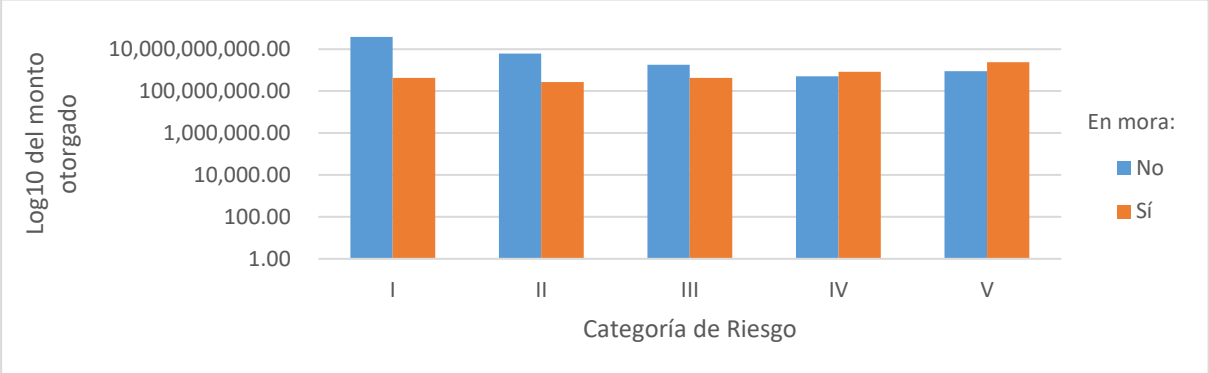


Figura 3. Comparación de la mora con el riesgo asignado en escala logarítmica.

Fuente: Elaboración propia.

En la *Figura 3*, como en la categoría I las diferencias son muy grandes con respecto a las demás, se usa una escala logarítmica base 10 para apreciar mejor las barras de las categorías restantes que se verían muy pequeñas con escala normal, pues lo que se quiere visualizar es la tendencia que sigue la mora a medida que va avanzando la categoría ordinal de riesgo asignada, la cuál es de la siguiente manera: los préstamos que van cayendo en mora aumentan a medida que se avanza a la siguiente categoría de riesgo, y los créditos que no caen en *default* van disminuyendo a medida que se avanza de categoría a excepción de la última, que es la más riesgosa, donde el monto de los préstamos que no están en mora es mayor que el de la categoría anterior. De no ser por este último caso, el comportamiento sería el esperado.

Similar situación se presenta en la categoría II, donde los préstamos que sí están en mora son menores a los de la categoría ordinal anterior, y se esperaría que fuesen superiores. Muy probablemente estas situaciones se den por una mala clasificación que se hace en base a la Norma para la Evaluación y Clasificación de Cartera Crediticia de la CNBS del año 2014 con respecto a estos créditos, ya sea por parte del banco o por deficiencias en la normativa.

Para el análisis crediticio son de importancia las variables relacionadas con tiempo, capital, interés, monto y tasa; que son las involucradas en la fórmula financiera general de valor futuro, y también se consideran los valores de las garantías que respaldan dichos préstamos. De la *Tabla 1* se seleccionan las siguientes variables que cumplen dichas condiciones y dado que son de naturaleza numérica, de escala de medición continua en el caso de las tasas y valores monetarios, y de medición discreta en el caso del tiempo[3], se analizan en la *Tabla 3* las principales medidas de tendencia central y dispersión para entender de qué manera se distribuyen los datos.

Tabla 3. Estadística descriptiva de las variables numéricas de interés para el estudio.

Variable	EnMora	N	N*	Media	Desv.Est.	Mínimo	Q1	Mediana	Q3	Máximo
DuracionMeses	0	48,084	0	69.448	49.528	0	36	60	84	372
	1	2,745	0	72.04	53.36	0	36	60	96	245
NODIAA	0	48,084	0	3.426	12.477	0	0	0	0	505
	1	2,745	0	285.20	373.07	0	100	162	315	3,684
GRACIA	0	48,084	0	0.3324	2.6369	0	0	0	0	60
	1	2,745	0	0.4827	3.0770	0	0	0	0	45
MONOTO	0	48,084	0	998,295	9,188,847	275	75,000	200,000	503,365	972,000,000
	1	2,745	0	1,603,873	22,213,808	3000	116,566	300,000	729,500	1,081,041,021
INTXCO	0	48,084	0	10,420	239,952	0	51	538	2,506	30,982,454
	1	2,745	0	20,677	169,821	0	0	0	2,968	5,308,411
TotalCartera	0	48,084	0	749,795	8,103,318	0	31,993	116,840	378,294	930,039,081
	1	2,745	0	857,474	4,772,328	15	52,654	188,549	503,590	136,254,802
VALGAR	0	48,084	0	4,122,075	22,934,413	0	0	427,768	1,603,772	829,715,520
	1	2,745	0	4,654,321	28,786,011	0	0	676,497	2,160,044	539,500,399
TASINT	0	48,084	0	16.940	9.917	0.000	12.210	14.000	16.210	46.000
	1	2,745	0	15.591	5.609	0.000	13.320	14.900	16.000	46.000
COATPO	0	48,084	0	11.081	7.265	0.000	0.0000	13.800	15.640	87.590
	1	2,745	0	12.106	7.946	0.000	0.000	15.040	17.300	86.640

Fuente: Elaboración propia.

Se aprecian diferencias sustanciales entre la mediana y la media de cada variable, así como bastante margen entre su rango de valores mínimos y máximos, acompañado de una alta desviación típica, lo cuál da indicios de una elevada dispersión en las distribuciones de cada una de ellas, por lo cuál se identificará si la muestra contiene *outliers* y si los datos se ajustan a una distribución normal.

4.2 Pruebas de datos atípicos y de normalidad.

Los gráficos de caja de las variables que se presentan en la *Figura 4* sugieren que en todas ellas, tanto para los préstamos en mora o al día, existen uno o más datos atípicos, por lo que se usará una prueba de Dixon ya que están diseñadas para superar el efecto enmascaramiento que pueden causar múltiples valores atípicos posibles, y como las muestras

[3] Ya que se cuentan los días o meses transcurridos desde cierta fecha de la operación.

más grandes de una población tienen mayor probabilidad de incluir valores extremos, según el criterio de Dixon para un $n \geq 14$ se recomienda la prueba $r22$ (Minitab, s.f.) con la cuál se obtuvieron los siguientes resultados[4] según el estado de la mora para las categorías de riesgo de cada préstamo con un nivel de significancia de 0.05:

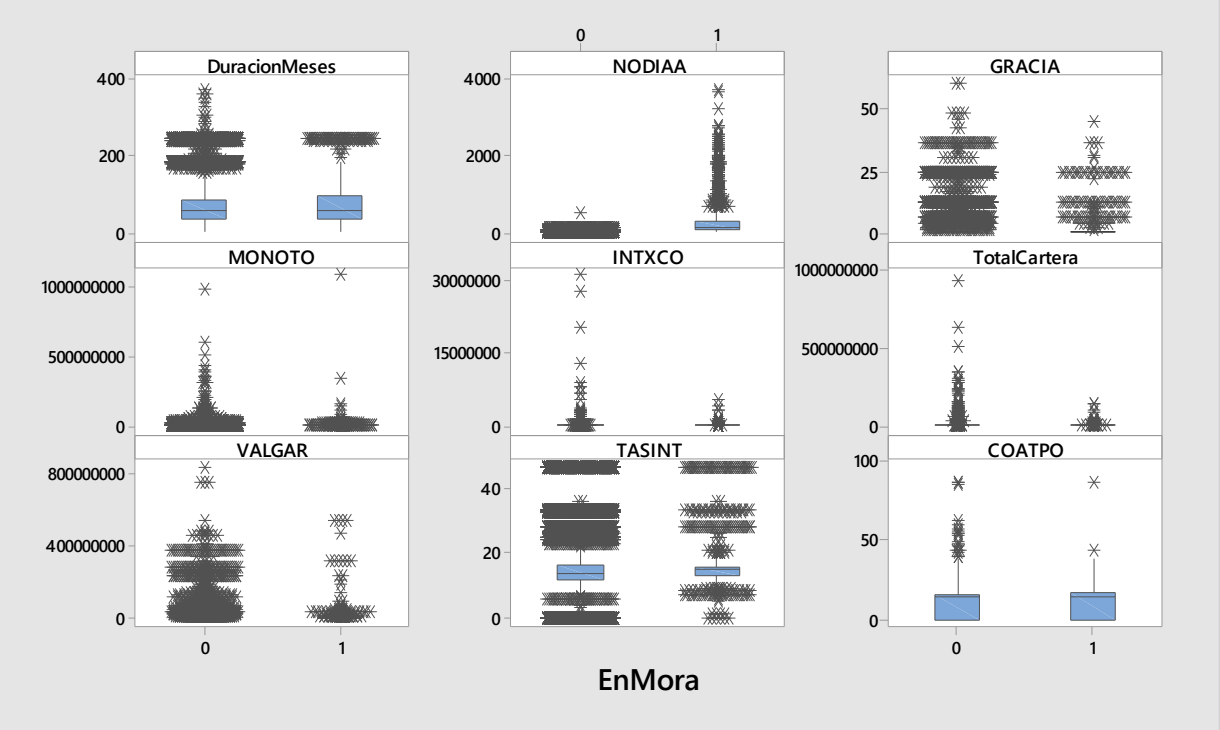


Figura 4. Gráficas de caja y bigote de las variables contínuas y discretas del estudio.

Fuente: Elaboración propia.

En cuanto a las variables que miden temporalidad, para la duración en meses de se encontraron dos valores atípicos en los préstamos al día en las categorías III (348 meses) y V (240 meses), no se encontraron *outliers* en la duración de los que están en mora. En cuanto al número de días atrasados solo se encontraron dos y ambos en la categoría V: 505 días para los que están al día y 3,684 días para los que están en mora. Para ver las gráficas y resultados consultar el *Anexo 1* en la página 23. El período de gracia no se consideró por ser una condición muy variable y especial en las condiciones de los préstamos.

Para las variables monetarias, en el monto otorgado se encontraron atípicos en mora en las categorías I, II, IV y V con sumas bastante elevadas de capital, y en los préstamos al día, resultaron cuatro atípicos en las categorías I, III, IV y V. Al analizar los intereses por cobrar de los préstamos al día se obtuvieron resultados similares en las mismas categorías de los montos otorgados a diferencia de los que están en mora que presentan *outliers* en todas las cinco categorías de riesgo, al igual que el total de cartera en mora; la cartera al día tiene atípicos para riesgo I, III, IV y V. Finalmente los valores de las garantías para los préstamos buenos tienen dos *outliers* en la categoría III y IV respectivamente y tres atípicos en los que están en impago cada uno en las tres primeras categorías de riesgo. Para mayor información ver *Anexo 2* en la página 24.

[4] Solo se detallan los resultados con *p-values* ≤ 0.05 con los que se rechaza la H_0 de que no hay datos atípicos.

En cuanto a las variables que miden las tasas de interés, la nominal solo presentó un valor atípico en la categoría III de riesgo de los créditos que se encuentran en mora, dicha tasa es del 36% anual; sin embargo, el costo anual total porcentual que representa la tasa de interés efectiva, presentó cuatro *outliers*: dos para los préstamos al día en las categorías IV y V (42.62 y 57.46% respectivamente) y dos para los créditos en mora en las categorías de riesgo I y III con tasas del 86.64 y 43.66% en cada uno. Para mayor detalle ver *Anexo 3* en la página 26.

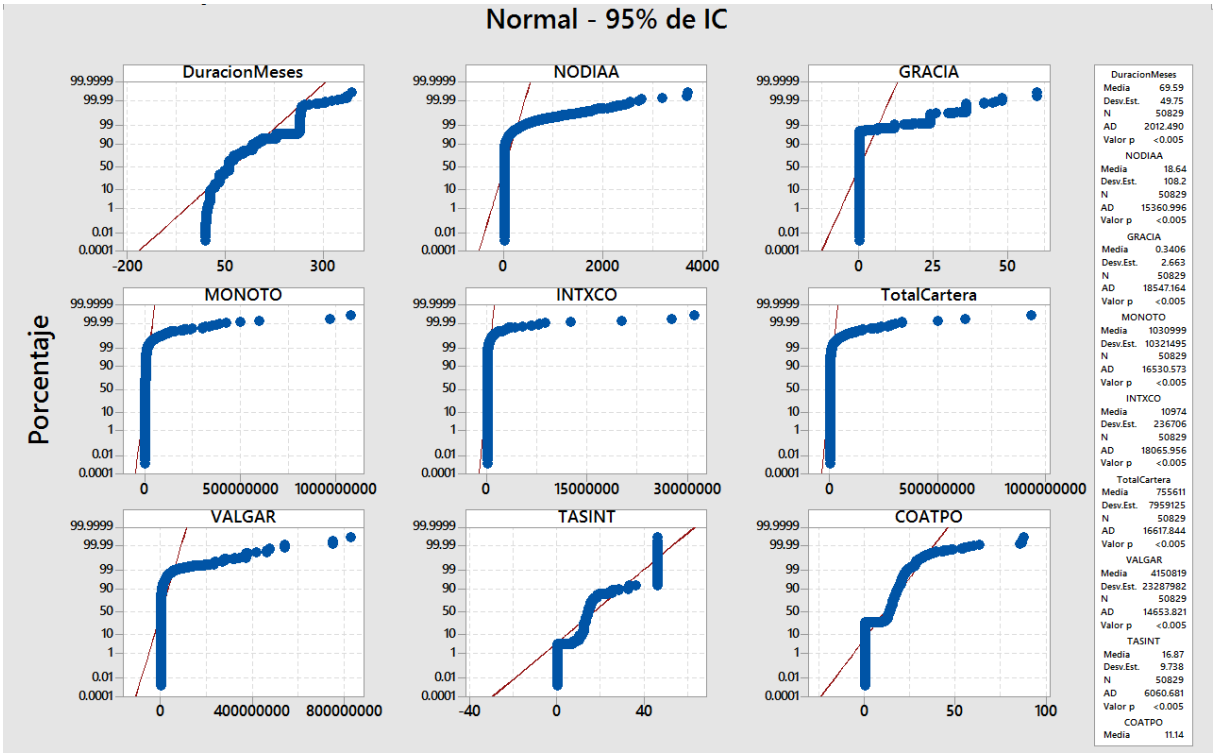


Figura 5. Gráficas de normalidad de las variables contínuas y discretas del estudio.

Fuente: Elaboración propia.

Como se puede apreciar en la *Figura 5*, tanto visualmente como en los resultados del costado derecho, ninguna de las variables se ajusta a una distribución bajo la curva normal, todos los coeficientes de la prueba AD (Anderson-Darling) obtuvieron *p-values* menores de 0.05, por lo que se rechaza la H_0 de normalidad en la distribución al 95% de confianza.

Al hacer el mismo análisis sin los datos atípicos identificados se obtuvieron resultados similares, por lo que bajo el supuesto estadístico de que se pueden obtener resultados adecuados con datos no normales si la muestra es lo suficientemente grande[5] y dado el tamaño y en consideración que el *dataset* no es una muestra, sino la población completa, entonces se continuará con la construcción del modelo paramétrico Logit. En el caso del modelo CART, como es no paramétrico, no afecta que los datos no sean normales.

[5] Basado en el teorema del límite central que demuestra que la distribución media de los datos de cualquier distribución se acerca a la curva normal a medida que aumenta el tamaño de la muestra, por lo que, para hacer inferencias sobre una media de población, el supuesto de normalidad no es fundamental si la muestra es lo suficientemente grande.

4.3 Modelo Logit para predicción del impago de préstamos.

Se hicieron pruebas con las variables categóricas más relevantes de la *Tabla 1* y diferentes combinaciones de las cuantitativas analizadas previamente para predecir la variable dependiente, que sería si el préstamo cae en mora o no. Como la variable respuesta es dicotómica se usa una distribución binomial para el modelo de regresión logística, el cuál quedó especificado de la siguiente manera con la combinación que mejores indicadores presentó:

```
Call:
glm(formula = EnMora ~ MONOTO + TASINT + DuracionMeses + SALARIO + CODSUC + Zona +
    REGION + Oficina + Garantia + TIPMON + TipoCredito + Origen + Destino +
    Categoria + TipoCliente + MANCOMUNADO, family = binomial, data =
    prestamos.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4679  -0.1085  -0.0773  -0.0453   3.6275

Coefficients: (12 not defined because of singularities)

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 14950.9  on 35579  degrees of freedom
Residual deviance:  5920.6  on 35505  degrees of freedom
AIC: 6070.6

Number of Fisher Scoring iterations: 18
```

La salida completa de los resultados de la consola de R se muestran en el *Anexo 4* en la página 27, de estos se tomaron solo las variables que arrojaron coeficientes significativamente distintos de cero y se volvió a correr el modelo obteniendo la siguiente salida resumida:

```
Call:
glm(formula = EnMora ~ MONOTO + TASINT + DuracionMeses + SALARIO + CODSUC + Oficina
    + Garantia + Origen + Destino + Categoria + TipoCliente + MANCOMUNADO, family =
    binomial, data = prestamos.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4614  -0.1086  -0.0774  -0.0455   3.6409

Coefficients: (8 not defined because of singularities)

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 14950.9  on 35579  degrees of freedom
Residual deviance:  5921.5  on 35508  degrees of freedom
AIC: 6065.5

Number of Fisher Scoring iterations: 18
```

Este modelo depurado muestra resultados más consistentes, tanto el intercepto como todos los coeficientes de las variables seleccionadas presentan *p-values* ≤ 0.05, tanto en las variables cualitativas al menos en dos o más de sus categorías y algunas en todas. Además se obtuvo un índice AIC de 6,065.5, inferior al del modelo inicial de 6,070.6; por lo que según criterio de información de Akaike se selecciona este segundo modelo como preferido. Los resultados obtenidos se resumen en la siguiente *Tabla 4*.

Tabla 4. Matríz de confusión del modelo Logit seleccionado.

	Predicciones				
Observaciones	**Buenos**	**En Mora**	**Total**	**% Aciertos**	**% Error**
Buenos	14,214	211	14,425	98.54%	1.46%
En mora	292	532	824	64.56%	35.44%
Total	14,506	743	15,249	96.70%	3.30%

Fuente: Elaboración propia.

La exactitud del modelo es de 96.7%, su sensibilidad del 98.54%, la precisión es del 97.99%, la especificidad del 71.6%. Entre más se aleje de cero hacia la derecha el resultado del pronóstico, mayor será la probabilidad de el préstamo caiga en mora. La curva ROC (Receiver Operating Characteristic curve) quedó de la siguiente forma, donde el eje *y* representa la sensibilidad y el eje *x* la especificidad del modelo:

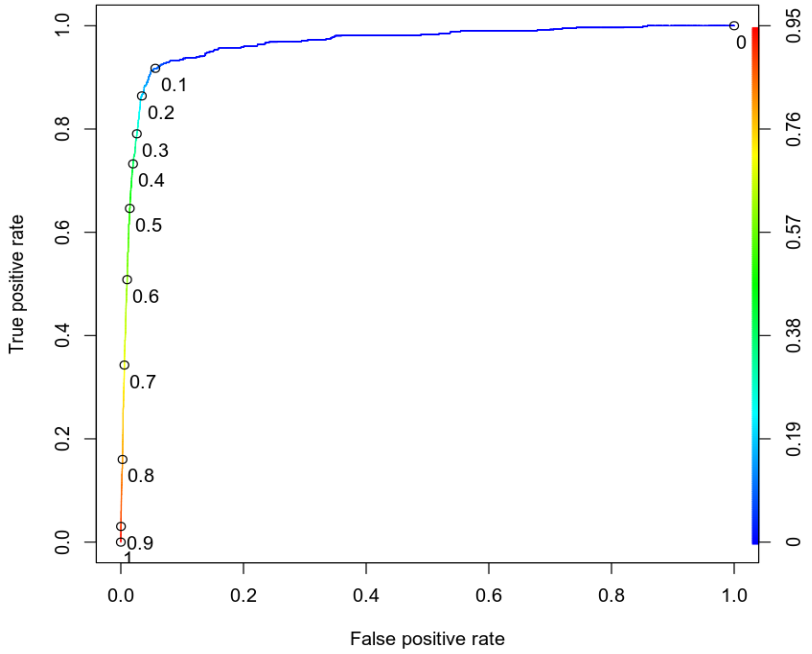


Figura 6. Curva ROC del modelo Logit seleccionado.

Fuente: Elaboración propia.

La *Figura 6* muestra que la curva superior intermedia es muy pronunciada casi sin solapamiento, lo cuál es característica de un excelente modelo, ya que si ambas variables coincidieran en cuanto a la proporción de los verdaderos positivos fuese igual a la de falsos positivos la curva sería una línea diagonal, por lo que la prueba sería inutil para predecir dado que tendría igual cantidad de aciertos y errores.

4.4 Modelo CART para predicción del impago de préstamos.

Para este modelo se utilizaron las mismas variables predictoras del modelo Logit seleccionado. En el *Anexo 5* de la página 31 se muestran los resultados completos, a continuación se detalla un resumen y el árbol generado a partir de los datos de entrenamiento:

```
Call:
rpart(formula = EnMora ~ MONOTO + TASINT + DuracionMeses + SALARIO + CODSUC +
      Oficina + Garantia + Origen + Destino + Categoria + TipoCliente +
      MANCOMUNADO, data = prestamos.train, method = "class")

n= 35580
          CP nsplit rel error    xerror       xstd
1 0.17907340      0 1.0000000 1.0000000 0.02219136
2 0.14731910      1 0.8209266 0.8209266 0.02020897
3 0.02628839      2 0.6736075 0.6736075 0.01838210
4 0.01336110      4 0.6210307 0.6371681 0.01789623
5 0.01000000      7 0.5809474 0.6095783 0.01751798

Variable importance
  Categoria TipoCliente    Garantia     Destino      TASINT     Oficina
         79           6           4           4           3           3
     CODSUC      MONOTO
          1           1
```

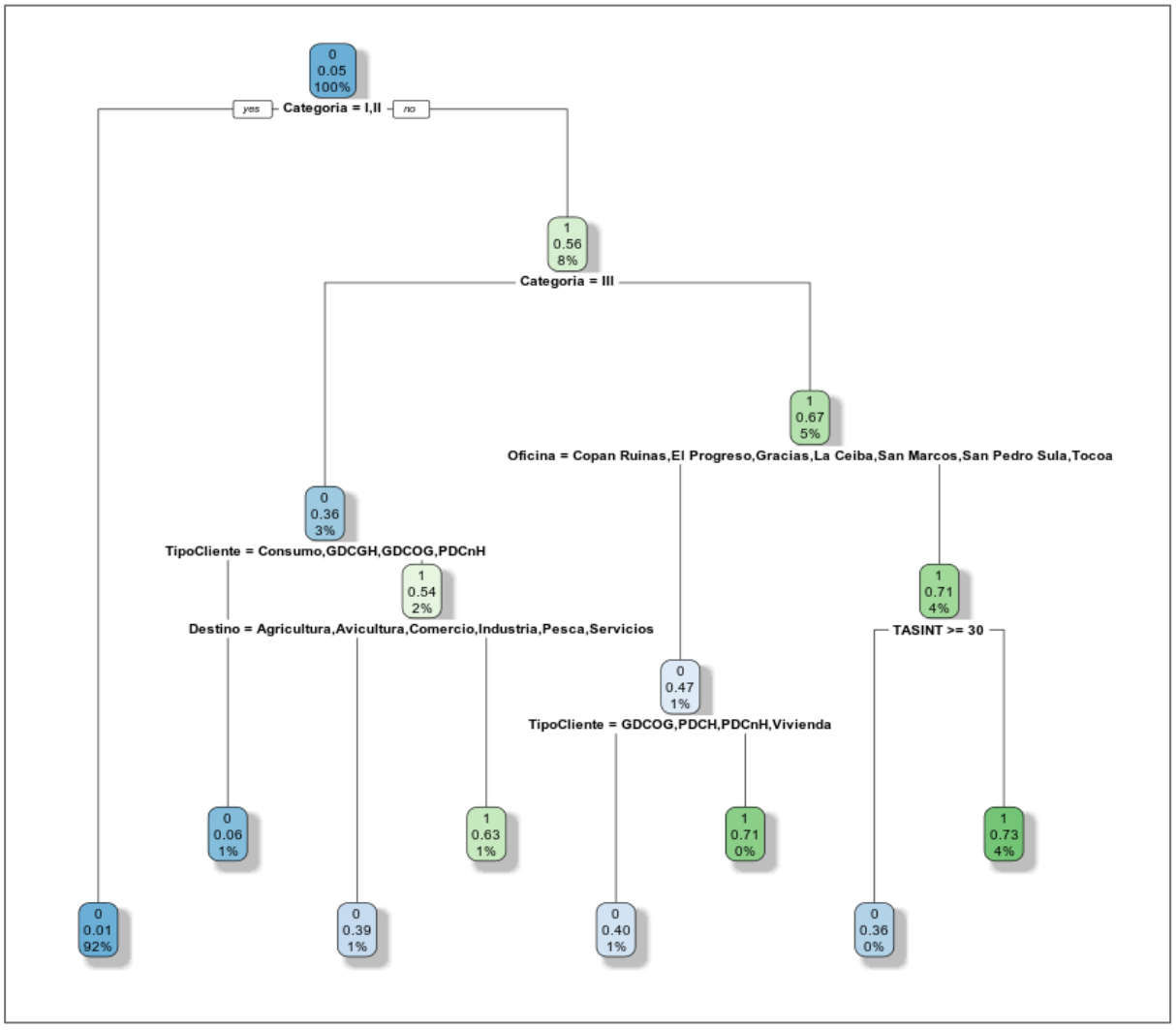


Figura 7. Árbol de clasificación del modelo CART de préstamos.

Fuente: Elaboración propia.

Como se puede apreciar en la *Figura 7*, la categoría de riesgo resultó ser la variable más importante de clasificación, un 92% de los préstamos se categorizan como I o II (riesgo bajo), de no estar en estas categorías la probabilidad de caer en *default* es de 0.56, donde el 8% de los préstamos cumplen esta condición de malos, siendo aún la categoría el siguiente criterio de peso

en la clasificación, en este caso III; de caer en ella tienen una probabilidad de pago del 0.36, un 3% de los préstamos cae en esta, los que no, son un 5% y tienen un 67% de probabilidad de caer en impago, de los cuáles el siguiente criterio de clasificación es la oficina en la que se asigna el préstamo. De no ser de las que aparecen en el listado la probabilidad de *default* aumenta al 71%, donde por última instancia el riesgo de impago dependerá de la tasa de interés, si esta es mayor o igual al 30% anual nominal es probable en un 0.73 que sean préstamos que caigan en mora, 4% de las observaciones cumplen esta condición.

Tabla 5. Matríz de confusión del modelo CART generado.

	Predicciones				
Observaciones	**Buenos**	**En Mora**	**Total**	**% Aciertos**	**% Error**
Buenos	14,192	233	14,425	98.38%	1.62%
En mora	295	529	824	64.20%	35.80%
Total	14,487	762	15,249	96.54%	3.46%

Fuente: Elaboración propia.

Como se puede apreciar en la *Tabla 5*, la exactitud del modelo es de 96.54%, su sensibilidad del 98.38%, la precisión es del 97.96%, la especificidad del 69.42%. Ambas matrices de confusión presentan valores muy parecidos, lo que indica que la capacidad predictiva de ambos modelos es bastante similar. La curva ROC de este modelo se muestra en la *Figura 6*, su forma indica de igual manera que el modelo es bastante bueno para predecir, aunque no de la misma manera que lo hace la regresión logística, ya que es menor el área sobre la diagonal.

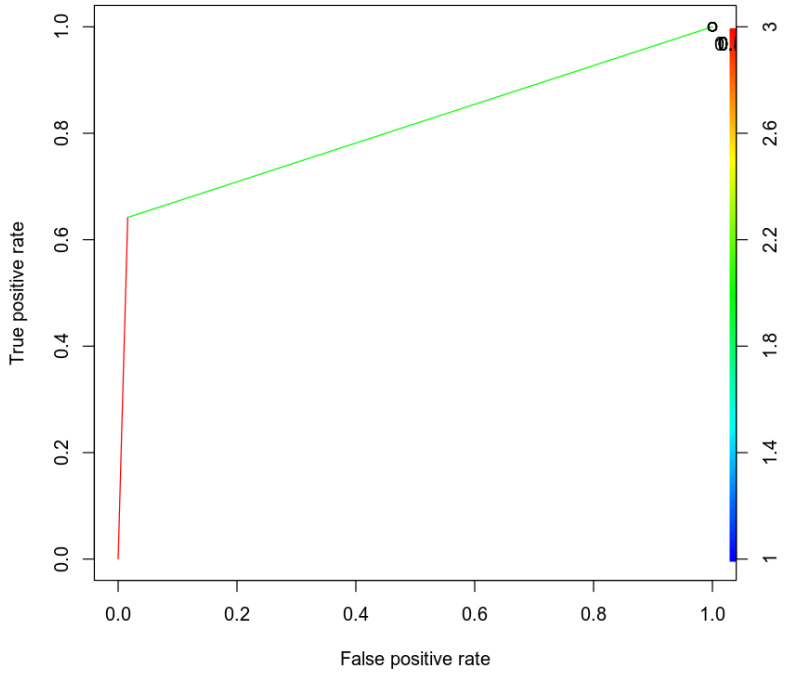


Figura 6. Curva ROC del modelo del árbol de decisión de préstamos.

Fuente: Elaboración propia.

5. Conclusiones.

Ambas técnicas demuestran ser herramientas muy útiles, eficaces y eficientes para construir modelos que puedan calcular la probabilidad de que un cliente del banco caiga en impago de algún préstamo. Sin embargo, los resultados obtenidos demuestran que en este caso específico la técnica de regresión logística presentó un modelo más adecuado para la tarea de medición de riesgo de crédito, con mejores indicadores estadísticos y menor margen de error, con una precisión y sensibilidad alrededor del 98%, una exactitud de 96.7% y un ROC con una curva muy pronunciada casi sin solapamiento con mucha área superior. Sin embargo, el árbol provee una visualización gráfica de la clasificación en base a las variables más importantes y su probabilidad que no tiene el otro modelo y es más fácil de entender e interpretar sus resultados.

Basándose en lo anterior, se puede concluir que en cuanto a la teoría, los modelos estadísticos que incluyen la probabilidad estimada de inclumplimiento de pago tienden a ofrecer mayor precisión que los modelos de riesgo tradicionales basados en la experiencia o en un enfoque cualitativo de clasificación y estimación del riesgo de crédito. El estudio contribuye a reforzar los resultados de otras investigaciones previas entre las metodologías para predecir variables binarias, donde la regresión logística presenta mayor robustez a la hora de generar modelos predictivos que los árboles binarios, pero que las reglas de decisión de estos últimos reflejan de mejor manera el comportamiento general del riesgo de mora, aún si no se tiene acceso a datos de la información confidencial de los clientes, y que es válido constuir estos modelos basándose únicamente en datos transaccionales.

Medir el riesgo de crédito de manera cuantitativa respaldado en cálculos de probabilidad para apoyar el juicio experto, le da mayor confiabilidad y objetividad a esta tarea, por lo que la incorporación de estos modelos como herramientas de trabajo por parte de los analistas de créditos, para que puedan evaluar sus solicitudes y carteras de manera más precisa y fundamentada, es crucial para mejorar el desempeño organizacional, por lo cuál, se recomienda la implementación de ambos modelos en la práctica.

6. Limitaciones e investigaciones futuras.

La investigación se limitó a los datos de un banco, por lo que podría replicarse este mismo modelo tomando la base de datos completa de la CIC que consolida los datos de los préstamos de todo el sistema bancario a nivel nacional; si bien es cierto que se necesitan permisos de seguridad especiales para acceder a esta información, puede ser conducido internamente por el ente regulador para todo el sistema bancario.

Otro punto notable en cuanto a la composición de datos de la muestra, es que no se tuvo acceso a variables demográficas de los clientes de los préstamos como tipo de persona (natural o jurídica), de la cuál se derivarían otras como edad, sexo, ingresos promedio, activos y utilidad, (dependiendo del caso); con las cuáles podría mejorarse el nivel de predicción del modelo, pero a pesar de que la restricción de acceso a una base de datos más amplia de información fue una limitante importante de este estudio, la metodología propuesta puede repetirse con una actualización y nuevas variables propuestas a incluir; además, en estudios futuros se pueden incluir variables macroeconómicas en los modelos a fin de expandir el entorno y no solo limitarlo a lo interno de la institución financiera.

Bibliografía.

Basso Winffel, O. (2013). Modelos de Gestión del Riesgo de Crédito. En S. d. Guatemala (Ed.), *XVI Conferencia sobre supervisión financiera.* Guatemala.

Bastos, J. A. (2010). Forecasting bank loans loss-given-default. *Journal of Banking & Finance*(34), 2510-2517.

Cardona Hernández, P. A. (2004). Aplicación de árboles de decisión en modelos de riesgo crediticio. *Revista Colombiana de Estadística, XXVII*(2), 139-151.

CNBS. (2014). *Normas para la Evaluación y Clasificación de la Cartera Crediticia.* Circular 036/2014, Comisión Nacional de Bancos y Seguros., Tegucigalpa M.D.C.

CNBS. (2015). *Manual Reporte de Datos Crédito.* Manual, Comisión Nacional de Bancos y Seguros, Central de Información Crediticia CIC, Tegucigalpa, Honduras.

CNBS. (2019). *Evolución del Sistema Supervisado.* Informe, Comisión Nacional de Bancos y Seguros, Departamento de Estadísticas y Publicaciones, Gerencia de Estudios, Tegucigalpa.

Feldman, D., & Gross, S. (2005). Mortgage Default: Classification Trees Analysis. *The Journal of Real Estate Finance and Economics, 30*(4), 369-396.

Fernández Castaño, H., & Pérez Ramírez, F. O. (2005). El modelo logístico: una herramienta estadística para evaluar el riesgo de crédito. *Revista Ingenierías Universidad de Medellín, IV*(6), 55-75.

Galindo, J., & Tamayo, P. (2000). Credit Risk Assessment using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications. *Computational Economics*(15), 107-143.

García Sanchez, M., & Sánchez Barradas, C. (2005). *Riesgo de crédito en México: aplicación del modelo CreditMetrics.* Tesis, Escuela de Negocios Universidad de las Américas Puebla, Departamento de Contaduría y Finanzas., México.

Gujarati, D. N., & Porter, D. C. (2010). *Econometría* (5a ed.). México: McGraw Hill Educación.

ISO. (2009). *Norma IEC/ISO 31010:2009 Gestión de Riesgos. Técnicas de Valoración del Riesgo.* International Standards Organization. Bogotá: Instituto Colombiano de Normas Técnicas y Certificación (ICONTEC).

Joos, P., Vanhoof, K., Ooghe, H., & Sierens, N. (1998). Credit Classification: a comparison of Logit models and Decision Trees. *Applications of machine learning and data mining in finance*, 59-73.

Kassambara, A. (2018). *CART model: Decision Tree Essentials*. Recuperado el 11 de septiembre de 2019, de Statistical tools for high-throughput data analysis: http://www.sthda.com/english/articles/35-statistical-machine-learning-essentials/141-cart-model-decision-tree-essentials/

Lee, T. S., Chiu, C. C., Chou, Y. C., & Lu, C. J. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Computational Statistics & Data Analysis*(50), 1113-1130.

Mileris, R. (2010). Estimation of loan applicants default probability applying discriminant analysis and simple Bayesian Classifier. *Ekonomika ir Vadyba*(15), 1078-1084.

Minitab. (s.f.). *Seleccionar las opciones de análisis para Prueba de valores atípicos*. Recuperado el 16 de septiembre de 2019, de Soporte de Minitab 18:

https://support.minitab.com/es-mx/minitab/18/help-and-how-to/statistics/basic-statistics/how-to/outlier-test/perform-the-analysis/select-the-analysis-options/

Mora Araya, C. E. (2014). *Mejora del proceso de evaluación de riesgo crediticio para Bancoestado Microempresas.* Tesis, Universidad de Chile, Facultad de Ciencias Físicas y Matemáticas, Santiago.

Nie, G., Rowe, W., Zhang, L., Tian, Y., & Shi, Y. (2011). Credit card churn forecasting by logistic regression and decision tree. *Expert Systems with Applications*(38), 15273-15285.

Pantoja Vilchez, P. M. (2016). *Propuesta de un Modelo Logit para evaluar el Riesgo Crediticio en las Cajas Municipales de Ahorro y Crédito: Caso de la Caja Municipal de Huancayo, período 2011-2015.* Tesis, Universidad San Ignacio de Loyola, Facultad de Ciencias Empresariales, Lima, Perú.

Puertas, R., & Martí, M. L. (2013). Análisis del Credit Scoring. *RAE-Revista de Administración de Empresas, 53*(3), 303-315.

Ruiz, H. R. (2017). Diseño de un modelo matemático para la calificación de clientes morosos en una entidad comercial mediante las metodologías de árboles de decisión, análisis discriminante y regresión logística. *INNOVA Research Journal, 2*(7), 176-188.

Saavedra García, M. L., & Saavedra García, M. J. (2010). Modelos para medir el riesgo de crédito de la banca. *Cuadernos de Administración, XXIII*(40), 295-319.

Salinas Flores, J. W. (2005). *Reconocimiento de patrones de morosidad para un producto crediticio usando la técnica de árbol de clasificación CART.* Tesis, Universidad Nacional Mayor de San Marcos, Facultad de Ingeniería Industrial, Lima, Perú.

Támara Ayús, A. L., Aristizábal, R. E., & Velásquez, H. (2010). Estimación de las provisiones esperadas en una institución financiera utilizando modelos LOGIT y PROBIT. *Revista Ciencias Estratégicas, 18*(24), 259-270.

Tello, M. L., Eslava, H. J., & Tobías, L. B. (2013). Análisis y evaluación del nivel de riesgo en el otorgamiento de créditos financieros utilizando técnicas de minería de datos. *Revista Visión Electrónica*(1), 13-26.

Therneau, T. M., & Atkinson, E. J. (2019). *An Introduction to Recursive Partitioning using the RPART Routines.* Recuperado el 12 de septiembre de 2019, de The Comprehensive R Archive Network: https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf

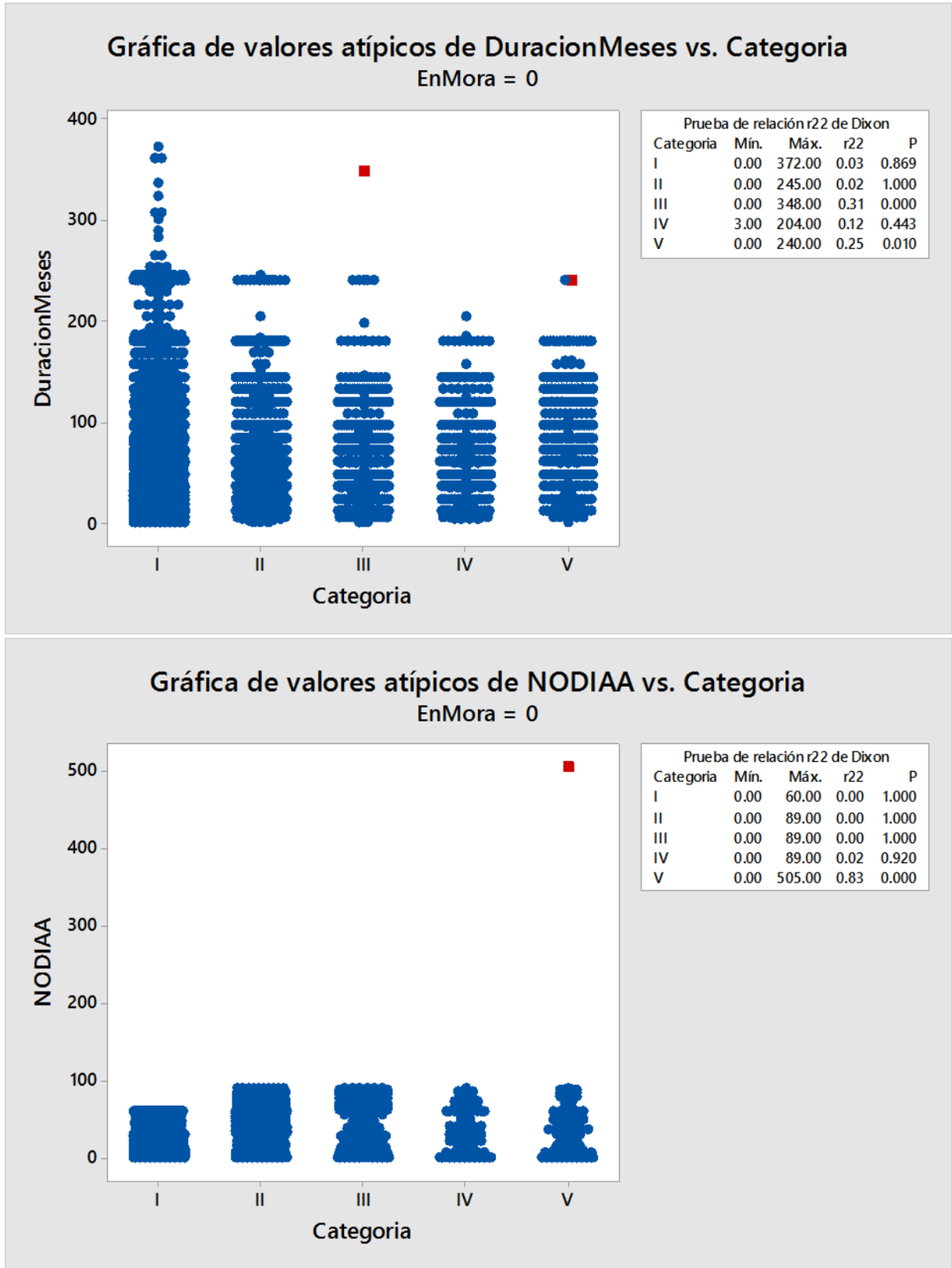
Thomas, L. C. (2000). A survey of credit and behavioural scoring: forecasting financial fisk of lending to consumers. *International Journal of Forecasting*(16), 149-172.

Ubipaipovic, B., & Durkovic, E. (2011). Application of Business Intelligence in the Banking Industry. *Management Information Systems, XI*(4), 23-30.

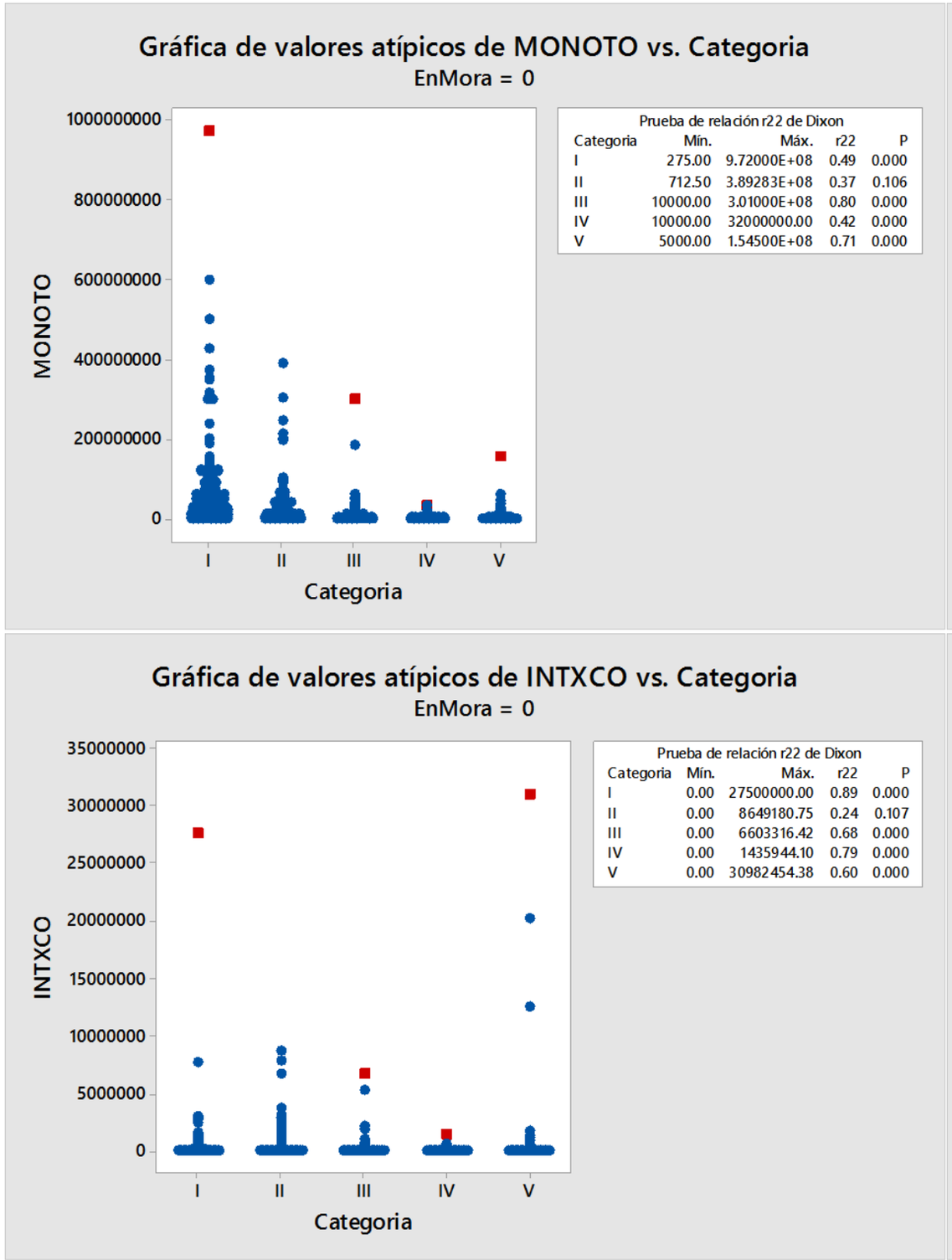
Velandia, N. (2013). *Establecimiento de un Modelo Logit para la Medición del Riesgo de Incumplimiento en Créditos para una Entidad Financiera del Municipio de Arauca, Departamento de Arauca.* Tesis, Universidad Nacional de Colombia, Manizales.

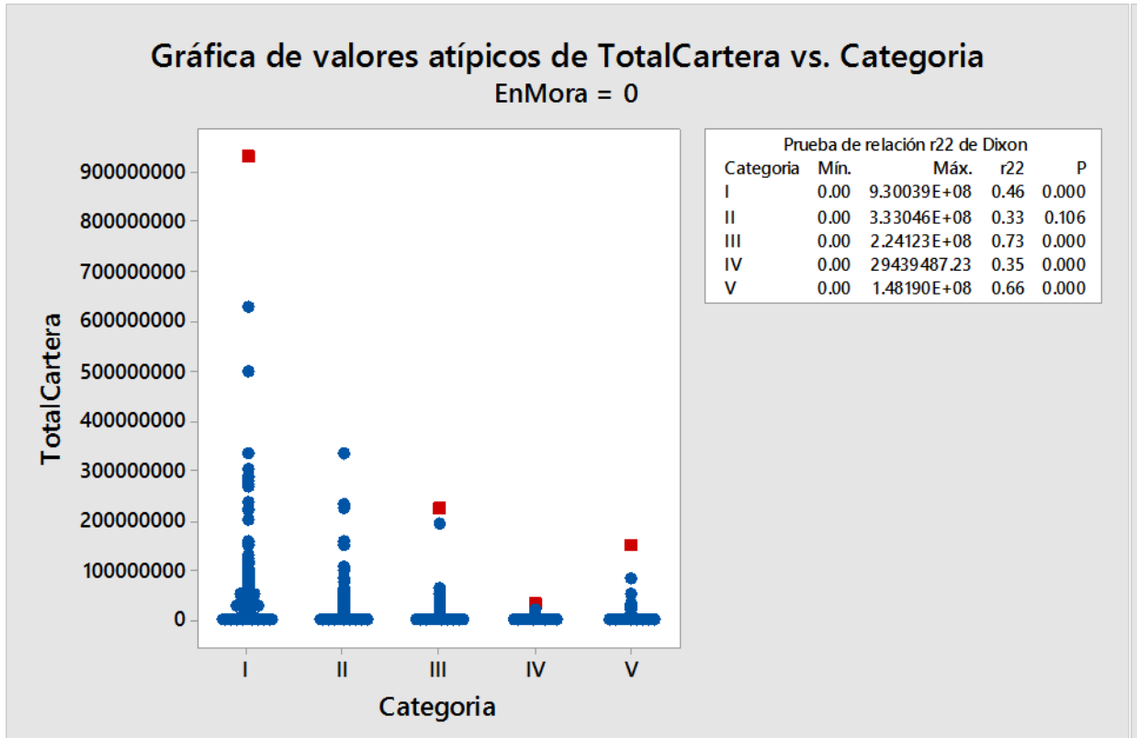
Anexos:

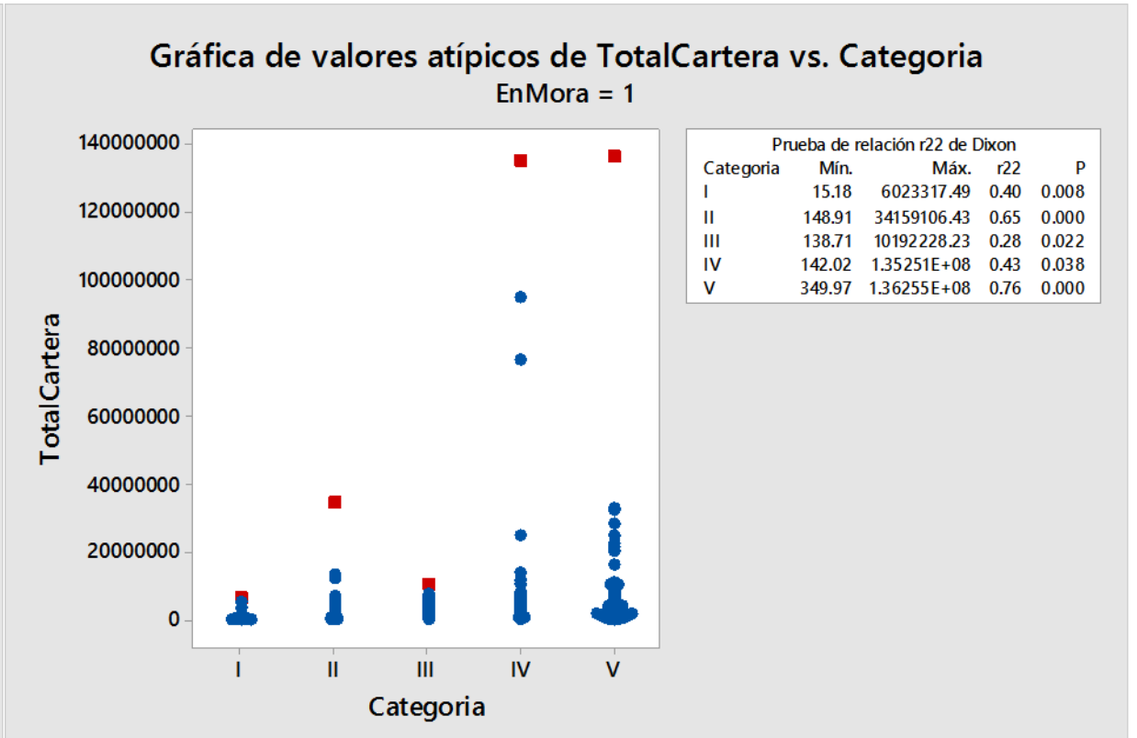
Anexo 1. Pruebas de valores atípicos de las variables predictoras relacionadas con tiempo.

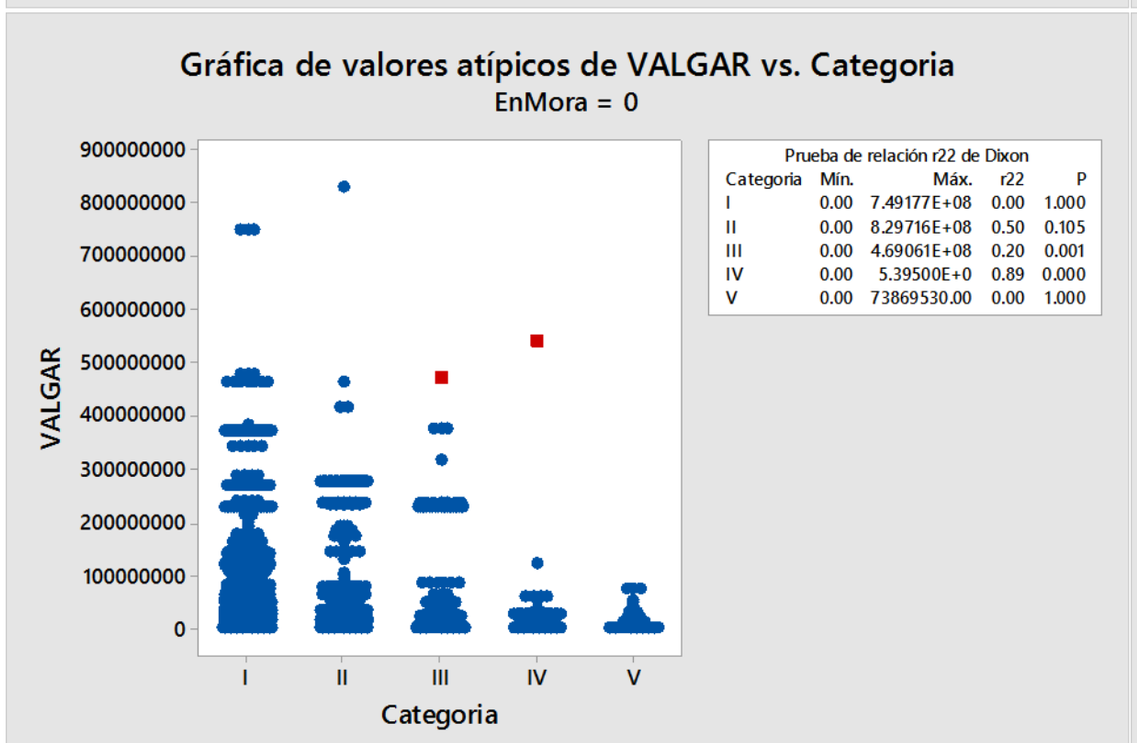


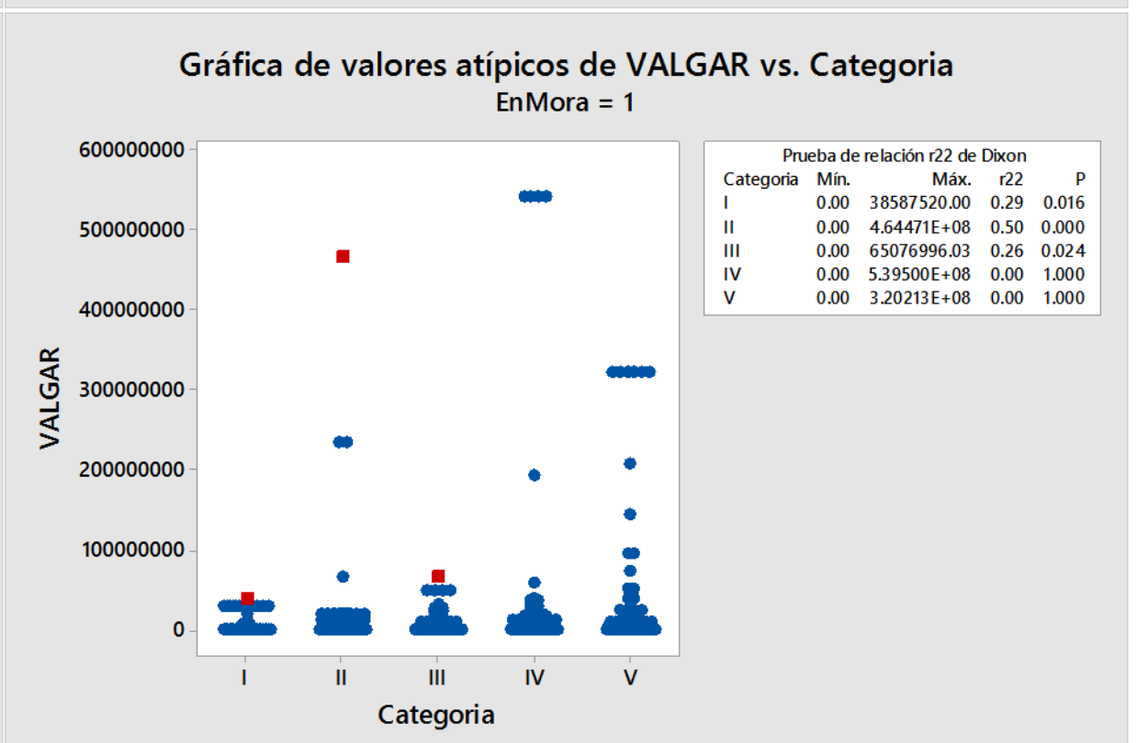
Anexo 2. Pruebas de valores atípicos de las variables predictoras relacionadas con valores monetarios: capital, interés, monto y garantías.



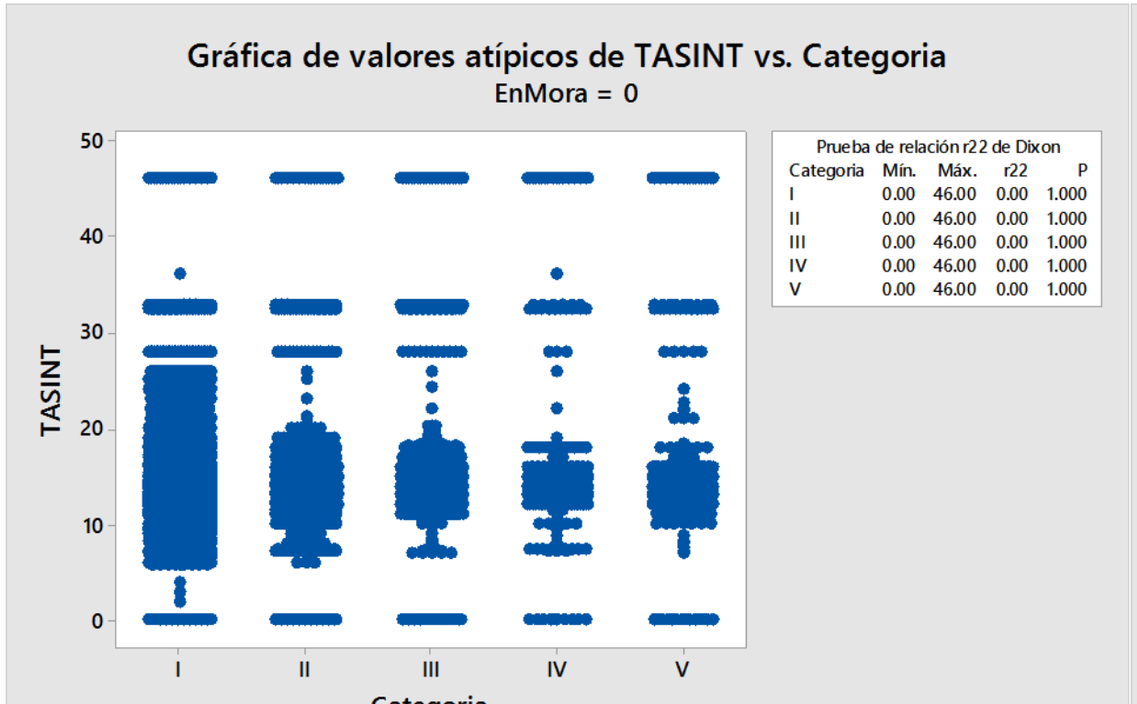

Gráfica de valores atípicos de TotalCartera vs. Categoria
EnMora = 0
TotalCartera
Categoria
I
II
III
IV
V
0
100000000
200000000
300000000
400000000
500000000
600000000
700000000
800000000
900000000
Prueba de relación r22 de Dixon
Categoria Mín. Máx. r22 P
I 0.00 9.30039E+08 0.46 0.000
II 0.00 3.33046E+08 0.33 0.106
III 0.00 2.24123E+08 0.73 0.000
IV 0.00 29439487.23 0.35 0.000
V 0.00 1.48190E+08 0.66 0.000

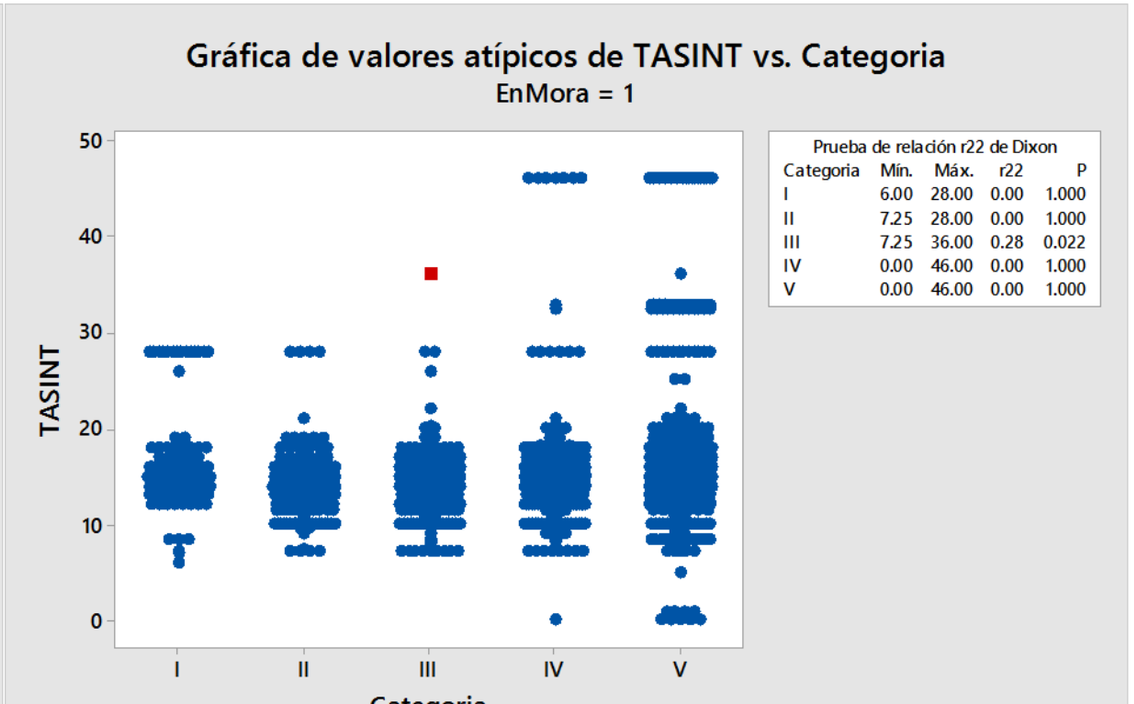

Gráfica de valores atípicos de TotalCartera vs. Categoria
EnMora = 1
TotalCartera
Categoria
I
II
III
IV
V
0
20000000
40000000
60000000
80000000
100000000
120000000
140000000
Prueba de relación r22 de Dixon
Categoria Mín. Máx. r22 P
I 15.18 6023317.49 0.40 0.008
II 148.91 34159106.43 0.65 0.000
III 138.71 10192228.23 0.28 0.022
IV 142.02 1.35251E+08 0.43 0.038
V 349.97 1.36255E+08 0.76 0.000

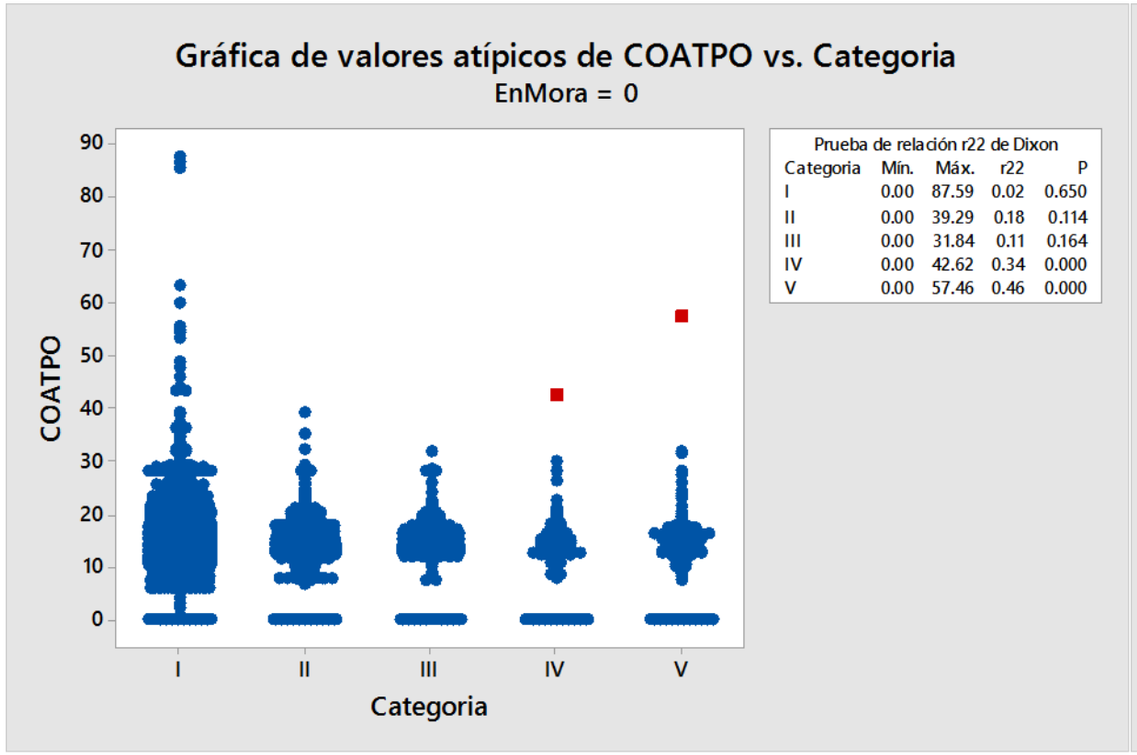

Gráfica de valores atípicos de VALGAR vs. Categoria
EnMora = 0
VALGAR
Categoria
I
II
III
IV
V
0
100000000
200000000
300000000
400000000
500000000
600000000
700000000
800000000
900000000
Prueba de relación r22 de Dixon
Categoria Mín. Máx. r22 P
I 0.00 7.49177E+08 0.00 1.000
II 0.00 8.29716E+08 0.50 0.105
III 0.00 4.69061E+08 0.20 0.001
IV 0.00 5.39500E+0 0.89 0.000
V 0.00 73869530.00 0.00 1.000

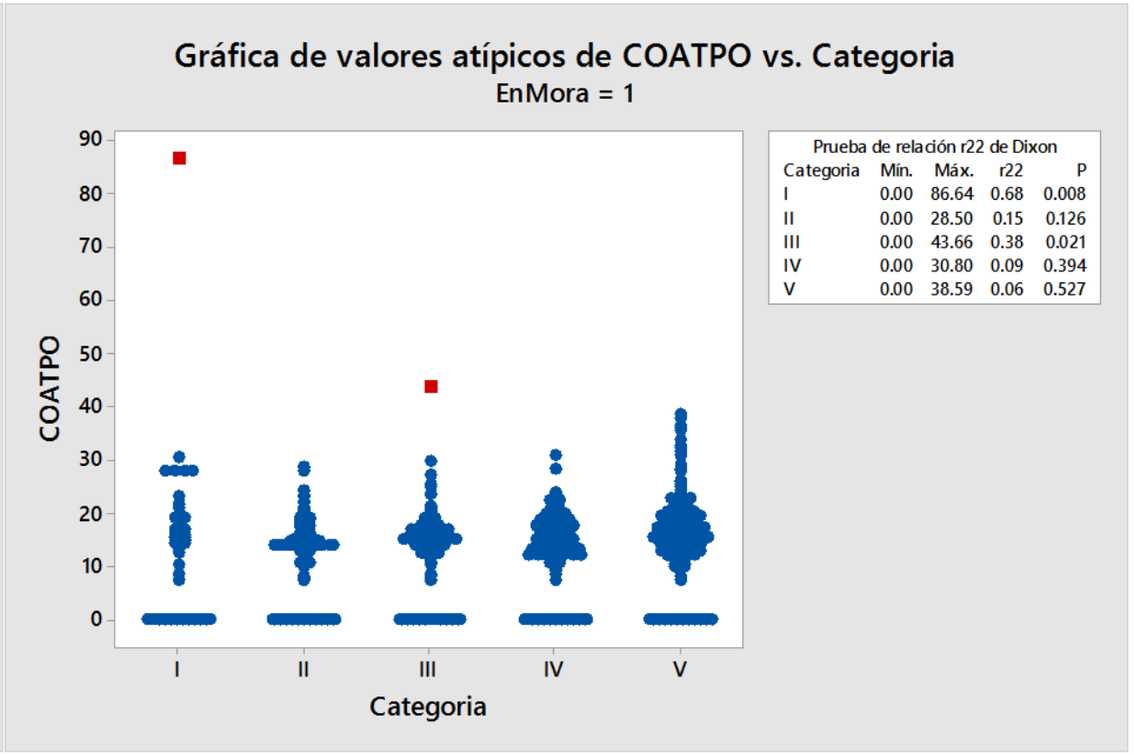

Gráfica de valores atípicos de VALGAR vs. Categoria
EnMora = 1
VALGAR
Categoria
I
II
III
IV
V
0
100000000
200000000
300000000
400000000
500000000
600000000
Prueba de relación r22 de Dixon
Categoria Mín. Máx. r22 P
I 0.00 38587520.00 0.29 0.016
II 0.00 4.64471E+08 0.50 0.000
III 0.00 65076996.03 0.26 0.024
IV 0.00 5.39500E+08 0.00 1.000
V 0.00 3.20213E+08 0.00 1.000

Anexo 3. Pruebas valores atípicos de las variables predictoras relacionadas con tasas (nominal y efectiva).









Anexo 4. Resultados de los modelos Logit inicial y depurado.

```
Call:
glm(formula = EnMora ~ MONOTO + TASINT + DuracionMeses + SALARIO +
    CODSUC + Zona + REGION + Oficina + Garantia + TIPMON + TipoCredito +
    Origen + Destino + Categoria + TipoCliente + MANCOMUNADO,
    family = binomial, data = prestamos.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4679  -0.1085  -0.0773  -0.0453   3.6275

Coefficients: (12 not defined because of singularities)
```

	Estimate	Std. Error	value	Pr(>|z|)	
(Intercept)	-1.86E+01	2.49E+03	-0.007	0.994035	
MONOTO	1.58E-08	3.58E-09	4.422	0.000010	***
TASINT	-2.56E-02	6.01E-03	-4.255	0.000021	***
DuracionMeses	-7.24E-03	1.11E-03	-6.551	0.000000	***
SALARIO	-3.51E-07	1.60E-07	-2.200	0.027802	*
CODSUC02	2.27E-01	4.72E-01	0.480	0.630980	
CODSUC03	3.29E-01	2.08E-01	1.581	0.113989	
CODSUC04	4.60E-01	1.24E-01	3.700	0.000215	***
CODSUC05	4.36E-01	1.51E-01	2.880	0.003971	**
CODSUC06	8.47E-01	2.10E-01	4.025	0.000057	***
CODSUC07	9.61E-02	3.04E-01	0.316	0.751640	
CODSUC08	9.29E-02	2.57E-01	0.362	0.717527	
CODSUC09	-2.11E-01	4.52E-01	-0.466	0.641567	
ZonaNorte	NA	NA	NA	NA	
ZonaOccidente	NA	NA	NA	NA	
REGIONU	-6.24E-02	1.59E-01	-0.392	0.694782	
OficinaCholuteca	2.78E-01	3.16E-01	0.880	0.379093	
OficinaComayagua	NA	NA	NA	NA	
OficinaCopan Ruinas	-8.80E-01	2.75E-01	-3.195	0.001396	**
OficinaDanli	1.49E-01	2.72E-01	0.547	0.584683	
OficinaEl Paraiso	8.33E-02	3.50E-01	0.238	0.811942	
OficinaEl Progreso	6.32E-01	5.22E-01	1.211	0.225835	
OficinaGracias	-2.72E-01	3.13E-01	-0.869	0.384742	
OficinaGuaimaca	4.60E-01	4.54E-01	1.013	0.311154	
OficinaJuticalpa	-9.63E-03	3.00E-01	-0.032	0.974381	
OficinaLa Ceiba	-9.93E-02	4.91E-01	-0.202	0.839750	
OficinaLa Entrada	2.67E-01	1.87E-01	1.431	0.152526	
OficinaLa Esperanza	4.00E-01	2.45E-01	1.634	0.102300	
OficinaMarcala	NA	NA	NA	NA	
OficinaNacaome	3.25E-01	3.78E-01	0.859	0.390328	
OficinaOcotepeque	7.57E-01	3.02E-01	2.508	0.012138	*
OficinaPuerto Cortes	-5.60E-01	6.44E-01	-0.869	0.384762	
OficinaSan Lorenzo	NA	NA	NA	NA	
OficinaSan Marcos	-2.42E-01	4.12E-01	-0.588	0.556423	
OficinaSan Pedro Sula	-2.43E-01	4.82E-01	-0.504	0.614569	
OficinaSanta Barbara	-5.79E-01	5.42E-01	-1.069	0.285136	

OficinaSanta Rosa de Copan	NA	NA	NA	NA	
OficinaSiguatepeque	NA	NA	NA	NA	
OficinaTegucigalpa	NA	NA	NA	NA	
OficinaTela	NA	NA	NA	NA	
OficinaTocoa	NA	NA	NA	NA	
GarantiaBonos en Prenda	-5.32E-01	1.38E+00	-0.386	0.699583	
GarantiaFiduciaria	2.29E-02	1.87E-01	0.122	0.902805	
GarantiaHip. S/Bienes Inmuebles	-1.01E-02	1.90E-01	-0.053	0.957826	
GarantiaPrendaria	-9.28E-01	2.54E-01	-3.651	0.000261	***
GarantiaPrendas S/Depositos	-1.53E+00	5.13E-01	-2.981	0.002875	**
GarantiaReciprocas S/Cobertura	1.67E-02	1.94E+00	0.009	0.993106	
GarantiaTit. Inst. Financ.	-1.45E+01	7.59E+03	-0.002	0.998478	
TIPMON2	-1.55E+01	2.79E+03	-0.006	0.995570	
TipoCreditoComercial	1.52E+01	2.49E+03	0.006	0.995116	
TipoCreditoConsumo	1.49E+01	2.49E+03	0.006	0.995211	
TipoCreditoVivienda	1.60E+01	2.49E+03	0.006	0.994877	
OrigenBcie	-1.67E+01	4.42E+03	-0.004	0.996988	
OrigenDisp. Inmediata	1.43E+01	2.79E+03	0.005	0.995910	
OrigenOtras Inst.	-8.72E-02	2.04E+00	-0.043	0.965909	
OrigenPropios	-9.60E-01	2.30E-01	-4.167	0.000031	***
OrigenRap	1.61E-02	3.75E-01	0.043	0.965804	
DestinoApicultura	1.61E+01	2.49E+03	0.006	0.994836	
DestinoAvicultura	1.49E+01	2.49E+03	0.006	0.995231	
DestinoComercio	-4.43E-01	2.00E-01	-2.219	0.026492	*
DestinoConsumo	-6.42E-01	2.24E-01	-2.869	0.004123	**
DestinoElectricidadAguaGasServicios	-1.78E+01	1.77E+03	-0.010	0.991987	
DestinoExportacion	-1.45E+00	6.44E-01	-2.253	0.024243	*
DestinoGanaderia	1.47E+01	2.49E+03	0.006	0.995293	
DestinoGob. Central	-2.77E+01	1.08E+04	-0.003	0.997949	
DestinoGob. Local	-1.62E+01	1.88E+03	-0.009	0.993110	
DestinoIndustria	-1.04E+00	3.28E-01	-3.167	0.001538	**
DestinoInst. Descentralizadas	-2.09E+01	6.39E+03	-0.003	0.997397	
DestinoMinas	7.28E-03	1.01E+00	0.007	0.994248	
DestinoPesca	1.52E+01	2.49E+03	0.006	0.995122	
DestinoPolizas	1.42E+01	7.59E+03	0.002	0.998513	
DestinoPropiedad Raiz	-4.49E-01	2.00E-01	-2.241	0.025041	*
DestinoSector Financiero	-1.30E+01	1.46E+03	-0.009	0.992869	
DestinoServicios	-1.05E+00	2.53E-01	-4.157	0.000032	***
DestinoSilvicultura	1.48E+01	2.49E+03	0.006	0.995254	
DestinoTransporte	-9.82E-01	2.47E-01	-3.976	0.000070	***
CategoriaII	3.33E+00	1.24E-01	26.795	< 2e-16	***
CategoriaIII	5.18E+00	1.24E-01	41.787	< 2e-16	***
CategoriaIV	6.11E+00	1.32E-01	46.420	< 2e-16	***
CategoriaV	6.79E+00	1.27E-01	53.404	< 2e-16	***
TipoClienteConsumo TC	-1.29E+01	1.56E+02	-0.082	0.934249	
TipoClienteGDCGH	-2.48E+00	5.05E-01	-4.915	0.000001	***
TipoClienteGDCOG	-2.66E+00	3.79E-01	-7.003	0.000000	***

```
 TipoClientePDCH                -1.35E-01    2.11E-01  -0.638  0.523571
 TipoClientePDCnH                     NA          NA      NA        NA
 TipoClienteVivienda                  NA          NA      NA        NA
 MANCOMUNADO1                   4.56E-01    1.88E-01   2.421  0.015489  *

---
Signif. codes:  0 “***” 0.001 “**” 0.01 “*” 0.05 “.” 0.1 “ ” 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 14950.9  on 35579  degrees of freedom
Residual deviance:  5920.6  on 35505  degrees of freedom
AIC: 6070.6

Number of Fisher Scoring iterations: 18
```

MODELO DEPURADO SOLO CON LAS VARIABLES CON SIGNIFICANCIA ESTADÍSTICA:

```
Call:
glm(formula = EnMora ~ MONOTO + TASINT + DuracionMeses + SALARIO +
    CODSUC + Oficina + Garantia + Origen + Destino + Categoria +
    TipoCliente + MANCOMUNADO, family = binomial, data = prestamos.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4614  -0.1086  -0.0774  -0.0455   3.6409

Coefficients: (8 not defined because of singularities)

                                Estimate   Std. Error  z value   Pr(>|z|)
 (Intercept)                -3.682E+00    3.597E-01  -10.236    < 2e-16  ***
 MONOTO                      1.579E-08    3.583E-09    4.406   1.05E-05  ***
 TASINT                     -2.550E-02    6.005E-03   -4.247   2.16E-05  ***
 DuracionMeses              -7.276E-03    1.102E-03   -6.600   4.11E-11  ***
 SALARIO                    -3.523E-07    1.598E-07   -2.205   2.75E-02  *
 CODSUC02                    2.339E-01    4.707E-01    0.497   6.19E-01
 CODSUC03                    3.306E-01    2.080E-01    1.589   1.12E-01
 CODSUC04                    4.572E-01    1.243E-01    3.679   2.34E-04  ***
 CODSUC05                    4.331E-01    1.512E-01    2.865   4.17E-03  **
 CODSUC06                    8.442E-01    2.102E-01    4.015   5.94E-05  ***
 CODSUC07                    9.675E-02    3.037E-01    0.319   7.50E-01
 CODSUC08                    9.025E-02    2.568E-01    0.351   7.25E-01
 CODSUC09                   -2.238E-01    4.509E-01   -0.496   6.20E-01
 OficinaCholuteca            2.767E-01    3.164E-01    0.874   3.82E-01
 OficinaComayagua                   NA           NA       NA        NA
 OficinaCopan Ruinas        -8.759E-01    2.754E-01   -3.180   1.47E-03  **
 OficinaDanli                1.544E-01    2.716E-01    0.568   5.70E-01
 OficinaEl Paraiso           8.944E-02    3.499E-01    0.256   7.98E-01
 OficinaEl Progreso          6.137E-01    5.190E-01    1.182   2.37E-01
 OficinaGracias             -2.699E-01    3.122E-01   -0.865   3.87E-01
 OficinaGuaimaca             4.642E-01    4.543E-01    1.022   3.07E-01
 OficinaJuticalpa           -6.996E-03    3.000E-01   -0.023   9.81E-01
```

OficinaLa Ceiba	-9.518E-02	4.905E-01	-0.194	8.46E-01	
OficinaLa Entrada	2.673E-01	1.868E-01	1.431	1.52E-01	
OficinaLa Esperanza	3.996E-01	2.450E-01	1.631	1.03E-01	
OficinaMarcala	NA	NA	NA	NA	
OficinaNacaome	3.146E-01	3.778E-01	0.833	4.05E-01	
OficinaOcotepeque	7.681E-01	3.007E-01	2.554	1.06E-02	*
OficinaPuerto Cortes	-5.747E-01	6.427E-01	-0.894	3.71E-01	
OficinaSan Lorenzo	NA	NA	NA	NA	
OficinaSan Marcos	-2.388E-01	4.124E-01	-0.579	5.63E-01	
OficinaSan Pedro Sula	-2.578E-01	4.802E-01	-0.537	5.91E-01	
OficinaSanta Barbara	-5.995E-01	5.388E-01	-1.113	2.66E-01	
OficinaSanta Rosa de Copan	NA	NA	NA	NA	
OficinaSiguatepeque	NA	NA	NA	NA	
OficinaTegucigalpa	NA	NA	NA	NA	
OficinaTela	NA	NA	NA	NA	
OficinaTocoa	NA	NA	NA	NA	
GarantiaBonos en Prenda	-5.347E-01	1.379E+00	-0.388	6.98E-01	
GarantiaFiduciaria	2.203E-02	1.872E-01	0.118	9.06E-01	
GarantiaHip.S/BienesInmuebles	-9.860E-03	1.901E-01	-0.052	9.59E-01	
GarantiaPrendaria	-9.294E-01	2.543E-01	-3.655	2.57E-04	***
GarantiaPrendas S/Depositos	-1.533E+00	5.132E-01	-2.987	2.82E-03	**
GarantiaReciprocas S/Cobertura	4.894E-01	1.558E+00	0.314	7.53E-01	
GarantiaTit. Inst. Financ.	-1.448E+01	7.592E+03	-0.002	9.98E-01	
OrigenBcie	-1.665E+01	4.468E+03	-0.004	9.97E-01	
OrigenDisp. Inmediata	-1.183E+00	3.013E-01	-3.928	8.58E-05	***
OrigenOtras Inst.	-6.247E-01	1.602E+00	-0.390	6.97E-01	
OrigenPropios	-9.559E-01	2.301E-01	-4.155	3.25E-05	***
OrigenRap	2.221E-02	3.747E-01	0.059	9.53E-01	
DestinoApicultura	9.060E-01	2.490E+00	0.364	7.16E-01	
DestinoAvicultura	-3.650E-01	4.821E-01	-0.757	4.49E-01	
DestinoComercio	-4.924E-01	1.488E-01	-3.309	9.36E-04	***
DestinoConsumo	-6.908E-01	1.813E-01	-3.809	1.39E-04	***
DestinoElectricidadAguaGasServ	-1.784E+01	1.775E+03	-0.010	9.92E-01	
DestinoExportacion	-1.495E+00	6.319E-01	-2.366	1.80E-02	*
DestinoGanaderia	-5.529E-01	2.012E-01	-2.748	5.99E-03	**
DestinoGob. Central	-2.767E+01	1.075E+04	-0.003	9.98E-01	
DestinoGob. Local	-1.625E+01	1.879E+03	-0.009	9.93E-01	
DestinoIndustria	-1.091E+00	3.023E-01	-3.610	3.06E-04	***
DestinoInst. Descentralizadas	-2.089E+01	6.395E+03	-0.003	9.97E-01	
DestinoMinas	-2.694E-02	1.002E+00	-0.027	9.79E-01	
DestinoPesca	-2.840E-02	7.059E-01	-0.040	9.68E-01	
DestinoPolizas	1.410E+01	7.592E+03	0.002	9.99E-01	
DestinoPropiedad Raiz	-4.858E-01	1.716E-01	-2.832	4.63E-03	**
DestinoSector Financiero	-1.308E+01	1.458E+03	-0.009	9.93E-01	
DestinoServicios	-1.099E+00	2.155E-01	-5.102	3.37E-07	***
DestinoSilvicultura	-4.606E-01	1.247E+00	-0.369	7.12E-01	
DestinoTransporte	-1.031E+00	2.095E-01	-4.919	8.70E-07	***

```
 CategoriaII                      3.327E+00    1.241E-01   26.799    < 2e-16  ***
 CategoriaIII                     5.179E+00    1.239E-01   41.804    < 2e-16  ***
 CategoriaIV                      6.108E+00    1.316E-01   46.431    < 2e-16  ***
 CategoriaV                       6.790E+00    1.271E-01   53.406    < 2e-16  ***
 TipoClienteConsumo TC           -1.288E+01    1.562E+02   -0.082   9.34E-01
 TipoClienteGDCGH                -2.190E+00    5.047E-01   -4.338   1.44E-05  ***
 TipoClienteGDCOG                -2.360E+00    4.048E-01   -5.830   5.55E-09  ***
 TipoClientePDCH                  1.625E-01    2.132E-01    0.762   4.46E-01
 TipoClientePDCnH                 2.984E-01    1.805E-01    1.653   9.83E-02  .
 TipoClienteVivienda              1.033E+00    2.975E-01    3.473   5.15E-04  ***

---
Signif. codes:  0 “***” 0.001 “**” 0.01 “*” 0.05 “.” 0.1 “ ” 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 14950.9  on 35579  degrees of freedom
Residual deviance:  5921.5  on 35508  degrees of freedom
AIC: 6065.5

Number of Fisher Scoring iterations: 18
```

Anexo 5. Resultados del modelo de árbol de clasificación.

```
Call:
rpart(formula = EnMora ~ MONOTO + TASINT + DuracionMeses + SALARIO + CODSUC +
      Oficina + Garantia + Origen + Destino + Categoria + TipoCliente +
      MANCOMUNADO, data = prestamos.train, method = "class")

n= 35580

          CP nsplit rel error    xerror       xstd
1 0.17907340      0 1.0000000 1.0000000 0.02219136
2 0.14731910      1 0.8209266 0.8209266 0.02020897
3 0.02628839      2 0.6736075 0.6736075 0.01838210
4 0.01336110      4 0.6210307 0.6371681 0.01789623
5 0.01000000      7 0.5809474 0.6095783 0.01751798

Variable importance
  Categoria TipoCliente    Garantia     Destino      TASINT     Oficina
         79           6           4           4           3           3
     CODSUC      MONOTO
          1           1

Node number 1: 35580 observations,    complexity param=0.1790734
  predicted class=0  expected loss=0.05399101  P(node) =1 class counts: 33659  1921
  probabilities: 0.946 0.054 left son=2 (32728 obs) right son=3 (2852 obs)
  Primary splits:
      Categoria   splits as  LLRRR, improve=1589.68800, (0 missing)
      TipoCliente splits as  RLLLRRR, improve=  33.78931, (0 missing)
      Oficina     splits as  RRRLRRLLRRLRRRRLLRLLLLRRLL, improve=  33.77196, (0
missing)
      CODSUC      splits as  LLRRRRRRL, improve=  30.69472, (0 missing)
      Destino     splits as  RRRRLLLRLLRLRRLRLLRL, improve=  30.67148, (0 missing)

Node number 2: 32728 observations
  predicted class=0  expected loss=0.009869225  P(node) =0.9198426
    class counts: 32405   323 probabilities: 0.990 0.010
Node number 3: 2852 observations,    complexity param=0.1473191
  predicted class=1  expected loss=0.4396914  P(node) =0.08015739
    class counts:  1254  1598 probabilities: 0.440 0.560
```

```
  left son=6 (995 obs) right son=7 (1857 obs)
  Primary splits:
      Categoria   splits as  --LRR, improve=125.35020, (0 missing)
      TipoCliente splits as  R-LLRRR, improve= 46.09770, (0 missing)
      TASINT      < 30.2     to the right, improve= 45.56852, (0 missing)
      Oficina     splits as  RRRLRRLLRRLRRRRRLRLLLLRRRR, improve= 32.56271, (0
missing)
      CODSUC      splits as  LLRRRRRRL, improve= 29.47544, (0 missing)
  Surrogate splits:
      TipoCliente splits as  R-LLRRR, agree=0.665, adj=0.039, (0 split)
      Oficina     splits as  RRRLRRRLRRRRRRRRLRRRLRRRLR, agree=0.657, adj=0.016, (0
split)
      Garantia    splits as  RRRRRLRR, agree=0.653, adj=0.006, (0 split)
      Destino     splits as  RLRRRRRR--R-RLRR-RRR, agree=0.653, adj=0.004, (0
split)

Node number 6: 995 observations,    complexity param=0.02628839
  predicted class=0  expected loss=0.3577889  P(node) =0.02796515
    class counts:   639   356 probabilities: 0.642 0.358
  left son=12 (380 obs) right son=13 (615 obs)
  Primary splits:
      TipoCliente splits as  L-LLRLR, improve=106.73790, (0 missing)
      Garantia    splits as  RRLRLLR-, improve= 74.36270, (0 missing)
      Destino     splits as  RRRLLLLR--L-RR-R-L-R, improve= 53.88569, (0 missing)
      Oficina     splits as  LRRLRRRLLRRRRRRRLLLLLLRLRR, improve= 26.87952, (0
missing)
      TASINT      < 20.23    to the right, improve= 25.37003, (0 missing)
  Surrogate splits:
      Garantia splits as  RLLRLRL-, agree=0.915, adj=0.776, (0 split)
      Destino  splits as  RLRRLLLR--R-RR-R-R-R, agree=0.813, adj=0.511, (0 split)
      TASINT   < 17.035   to the right, agree=0.713, adj=0.247, (0 split)
      Oficina  splits as  RRRRRRRRLRRRRRRRRRRRRRLRR, agree=0.686, adj=0.179, (0
split)
      MONOTO   < 122766.8 to the left,  agree=0.677, adj=0.155, (0 split)

Node number 7: 1857 observations,    complexity param=0.0133611
  predicted class=1  expected loss=0.3311793  P(node) =0.05219224
    class counts:   615  1242 probabilities: 0.331 0.669
  left son=14 (329 obs) right son=15 (1528 obs)
  Primary splits:
      Oficina     splits as  RRRLRRLLRRLRRRRRRRLLRRRRRL, improve=32.22279, (0
missing)
      TASINT      < 14.015   to the left,  improve=25.63426, (0 missing)
      CODSUC      splits as  LLRRRRRLL, improve=23.79706, (0 missing)
      TipoCliente splits as  R-LLLRR, improve=16.68560, (0 missing)
      Destino     splits as  R-RRRLLR--R-RRLR-RRL, improve=16.48608, (0 missing)
  Surrogate splits:
      CODSUC splits as  RLRRRRRRL, agree=0.957, adj=0.76, (0 split)

Node number 12: 380 observations
  predicted class=0  expected loss=0.06315789  P(node) =0.01068016
    class counts:   356    24 probabilities: 0.937 0.063

Node number 13: 615 observations,    complexity param=0.02628839
  predicted class=1  expected loss=0.4601626  P(node) =0.01728499
    class counts:   283   332
   probabilities: 0.460 0.540
  left son=26 (240 obs) right son=27 (375 obs)
  Primary splits:
      Destino     splits as  L-LLR--R--L-RL-R-L-R, improve=17.282630, (0 missing)
      Oficina     splits as  LRLLRRLL-RLLRLRRLLLLLLRLRR, improve=14.334610, (0
missing)
      Origen      splits as  R-L-LR, improve= 9.177241, (0 missing)
      TipoCliente splits as  ----L-R, improve= 8.553989, (0 missing)
      CODSUC      splits as  LLRRRRRRR, improve= 7.270503, (0 missing)
  Surrogate splits:
      DuracionMeses < 60.5     to the left,  agree=0.728, adj=0.304, (0 split)
      TASINT        < 16.06    to the right, agree=0.646, adj=0.092, (0 split)
      Oficina       splits as  RRRRRRRR-RRRRRRLLRLRLRRLRR, agree=0.646, adj=0.092,
(0 split)
```

```
      MONOTO        < 108500   to the left,  agree=0.639, adj=0.075, (0 split)
      TipoCliente   splits as  ----L-R, agree=0.634, adj=0.063, (0 split)

Node number 14: 329 observations,    complexity param=0.0133611
  predicted class=0  expected loss=0.4680851  P(node) =0.009246768
    class counts:   175   154 probabilities: 0.532 0.468
  left son=28 (259 obs) right son=29 (70 obs)
  Primary splits:
      TipoCliente splits as  R-RLLLL, improve=10.779590, (0 missing)
      Destino     splits as  L-LLRLLR--L--R-L-LLL, improve= 9.708433, (0 missing)
      Categoria   splits as  ---LR, improve= 9.391661, (0 missing)
      MONOTO      < 175773.4 to the right, improve= 5.077252, (0 missing)
      TASINT      < 15.2     to the left,  improve= 4.631280, (0 missing)
  Surrogate splits:
      Destino splits as  L-LLRLLL--L--L-L-LLL, agree=0.954, adj=0.786, (0 split)
      TASINT  < 30.2     to the left,  agree=0.881, adj=0.443, (0 split)
      MONOTO  < 88500    to the right, agree=0.821, adj=0.157, (0 split)
      Origen  splits as  L-R-LL, agree=0.796, adj=0.043, (0 split)

Node number 15: 1528 observations,    complexity param=0.0133611
  predicted class=1  expected loss=0.2879581  P(node) =0.04294547
    class counts:   440  1088
   probabilities: 0.288 0.712
  left son=30 (92 obs) right son=31 (1436 obs)
  Primary splits:
      TASINT        < 30.2     to the right, improve=24.444870, (0 missing)
      DuracionMeses < 59.5     to the right, improve=13.349370, (0 missing)
      TipoCliente   splits as  R-LLRRR, improve=11.326490, (0 missing)
      Oficina       splits as  LLR-RL--RL-LRLRRRR--LLRRL-, improve=10.656770, (0
missing)
      Destino       splits as  R-RRRLLR--R-RRLR-RRR, improve= 9.155077, (0 missing)
  Surrogate splits:
      Origen splits as  RRLRRR,       agree=0.948, adj=0.130, (0 split)
      MONOTO < 5800     to the left,  agree=0.941, adj=0.022, (0 split)

Node number 26: 240 observations
  predicted class=0  expected loss=0.3916667  P(node) =0.006745363
    class counts:   146    94  probabilities: 0.608 0.392

Node number 27: 375 observations
  predicted class=1  expected loss=0.3653333  P(node) =0.01053963
    class counts:   137   238   probabilities: 0.365 0.635

Node number 28: 259 observations
  predicted class=0  expected loss=0.4015444  P(node) =0.00727937
    class counts:   155   104   probabilities: 0.598 0.402

Node number 29: 70 observations
  predicted class=1  expected loss=0.2857143  P(node) =0.001967397
    class counts:    20    50   probabilities: 0.286 0.714

Node number 30: 92 observations
  predicted class=0  expected loss=0.3586957  P(node) =0.002585722
    class counts:    59    33   probabilities: 0.641 0.359

Node number 31: 1436 observations
  predicted class=1  expected loss=0.2653203  P(node) =0.04035975
    class counts:   381  1055   probabilities: 0.265 0.735
```