UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN INGENIERÍA
INGENIERIA DE SISTEMAS – OPTIMACION FINANCIERA

MODELO DE CALIFICACIÓN CREDITICIA PARA RETENCIÓN DE CLIENTES EN CRÉDITOS HIPOTECARIOS.

TESIS QUE PARA OPTAR POR EL GRADO DE:
MAESTRA EN INGENIERÍA

PRESENTA:
DANIELA YAKELINNE ALONSO PEÑA

TUTOR PRINCIPAL
JORGE RODRIGUEZ RUBIO - FACULTAD DE INGENIERIA

CIUDAD UNIVERSITARIA, CD. MX., ENERO DE 2022

Dedicado a mis padres y hermanos, los amo.

Agradecimientos

A mis padres, Adela y David

De quienes no solo he recibido el amor incondicional, sino también todo el apoyo, la comprensión, la paciencia y la orientación suficiente para salir adelante. Si algo bueno y bello he hecho en mi vida siempre ha sido atravesado por su luz.

A mis hermanos, Saúl, Fabiola y Valeria

Quienes siempre serán mis cómplices y mi mejor equipo, los amo.

A mis amigos, Juan, Arturo, Maricela, Isabel, Iván, Alberto, Sabino, Dante y Martín.

Con quienes he compartido momentos de alegría y siempre han estado escuchándome y alentándome con mis sueños, gracias por creer en mí y demostrar su amistad a pesar de la distancia.

A mi tutor, M. en I. Jorge Rodríguez Rubio

Quien además de brindarme los conocimientos requeridos para la elaboración de este trabajo, me ha impulsado a seguir creciendo profesionalmente y se ha convertido también en un gran amigo.

A mis sinodales, Dra. Magnolia Miriam Sosa Castro, M. en I. Yonahandy Malfavón Ruíz, Dra. Isabel Patricia Aguilar Juárez y Dr. Elio Agustín Martínez Miranda

Quienes me apoyaron con la revisión de este trabajo aun cuando sus actividades son demandantes, aportándome comentarios importantes para enriquecer mis análisis.

A mi novio

Quien ha sido un gran apoyo en todos los aspectos, gracias por la paciencia y tu compañía inigualable, te quiero amor.

Definitivamente no caben aquí las personas con las que he crecido y me han enseñado tanto, así como en lo académico como en lo personal, les agradezco inmensamente.

Y, por último, pero no menos importante, a la UNAM quien me brindó las herramientas para poder desarrollarme.

Índice General

Índice de Figuras.

Introducción

Actualmente las instituciones financieras con actividades crediticias tienen un papel muy importante en México para impulsar el desarrollo económico y la creación de valor, de ahí la necesidad de establecer principios de administración y gestión de riesgos que permitan identificar características para la mitigación del mismo (Banco de México).

Dado que el riesgo de crédito es una de las fuentes principales de pérdidas, su estudio requiere mayor relevancia por lo que se ha puesto en marcha la búsqueda de mejores prácticas para su administración. Aun cuando este tipo de riesgo ha estado presente en las actividades de las instituciones financieras, solo recientemente se han incorporado el desarrollo de metodologías y modelos más robustas orientados a la evaluación y cuantificación de sus efectos.

Debemos tener en cuenta que los riesgos han evolucionado y por tanto las técnicas y herramientas para medirlos no pueden ser las mismas. Afortunadamente, debido al creciente desarrollo tecnológico se han podido desarrollar nuevas herramientas las cuales contemplan diversas fuentes de información y a su vez diferente tipo de datos. Existen varios modelos multivariados que utilizan indicadores financieros o características de los clientes como insumos para la estimación de la probabilidad de *default* o incumplimiento. Los modelos multivariados se pueden agrupar en 3 categorías como los son el análisis discriminante (lineal y cuadrático), modelos de regresión (lineal, logit y probit) y modelos inductivos (redes neuronales, algoritmos genéticos y árboles de decisión) (Alpaydin, 2020).

Un punto clave en la medición del riesgo en carteras de crédito es la distinción de segmentos en la población que nos ayuden a crear estrategias más personalizadas de acuerdo a un nivel de riesgo otorgado.

Este trabajo se conforma por 3 capítulos a través de los cuales se plantean los antecedentes, metodologías, aplicación y resultados del análisis. El objetivo de la presente investigación es estimar a aquellos clientes propensos a mudar su crédito hipotecario en los primeros 3 años de crédito, se mostrará un modelo de Regresión Logística considerando

árboles de decisión con el fin de darle al lector un panorama más amplio de evaluación y su impacto en el desempeño financiero.

Antecedentes

A través de los años las instituciones financieras han buscado mejorar el conocimiento del cliente para ofertar mejores productos de crédito que se adapten a las necesidades del mismo o mitiguen pérdidas que podrían ser generadas a través del incumplimiento de pagos o anulación de relación con la entidad.

Se han desarrollado diferentes herramientas y métodos con bases estadísticas que permitan a través de la información del cliente discriminar de acuerdo a un objetivo y generen una toma de decisiones más eficiente y oportuna en materia de crédito.

El primero en utilizar métodos estadísticos para discriminar una población fue Ronald Fisher en 1936 el cual a través de reconocimiento de patrones y aprendizaje de máquinas encontró una combinación lineal que separaba dos o más clases de objetos. Años más tarde en 1941 David Durand escribe reportes y sugiere estadísticos para tomar decisiones de crédito, es aquí cuando empieza a surgir el concepto de credit scoring (Crook, 2007).

Las técnicas de credit scoring se comenzaron aplicar a partir de 1960 en los Estados Unidos cuyo objetivo era determinar si los individuos que solicitaban crédito podrían ser buenos candidatos utilizando una forma automatizada. Este tipo de técnicas se empezaron a usar debido a la alta demanda de tarjetas de crédito por lo que las técnicas tradicionales no eran eficientes para agilizar las solicitudes. El primer uso de Application Scoring[1] fue por la compañía American Investments.

Un componente clave en la evolución del sistema financiero en México ha sido la creación de las sociedades de información crediticia como Buró de crédito, este surge 1994 por la necesidad de contar con una institución sólida capaz de reunir y administrar información confiable y completa sobre el historial crediticio de la población para el uso de

[1] Es la puntuación otorgada de acuerdo a los riesgos asociados de una solicitud.

los proveedores de crédito en el país. Es así como dicha información ha ayudado a tener un mayor conocimiento integral del cliente respecto a todos sus créditos y ha sido muy importante para la evaluación del riesgo de un cliente.

La información del cliente o acreditado es de los principales componentes para la toma de decisiones, más no el único, otro componente importante a considerar son las herramientas tecnológicas.

Los avances tecnológicos han permitido la concentración de grandes volúmenes de información, así como su procesamiento. Esto ha permitido a las instituciones financieras poder aceptar más solicitudes de crédito, lo que a su vez disminuye y logra cuantificar el riesgo.

El aumento del poder de procesamiento ha permitido mejorar la potencialización de los modelos existentes o desarrollar nuevos modelos ya que anteriormente no era posible realizar cálculos de manera rápida y robusta.

Considerando estos aspectos es posible actuar de manera preventiva para las decisiones de crédito y dar una respuesta eficiente ante las necesidades de clasificación de cliente.

Problemática

Actualmente las instituciones financieras utilizan el método de credit scoring para automatizar la toma de decisiones partiendo de la información del acreditado cuyo método indicará la aprobación o no de una operación de financiación. En algunas instituciones estos métodos sirven como apoyo para la correcta toma de decisiones.

Cabe destacar que lo que se pretende realizando un credit scoring es estandarizar respuestas, procesando un gran volumen de clientes de crédito de un modo más ágil asegurando protocolos y principios de riesgo de cada entidad (Bishop, 2006).

A partir de la pandemia la inestabilidad económica de muchos clientes ha ocasionado que las instituciones financieras disminuyan su apetito al riesgo y por lo tanto se cierren criterios para la aceptación o formalización de créditos.

Es importante destacar que el riesgo de crédito no solo abarca la originación, sino el mantenimiento de los clientes y la retención de los mismos, por lo que se deben implementar estrategias en periodos de crisis para cuidar la cartera de crédito en todos estos sentidos.

Hay que denotar que en estos periodos se tiene poca solvencia e incertidumbre económica, lo que hace reconsiderar a los clientes sus necesidades financieras para afrontar las emergencias validando sus niveles de endeudamiento como la capacidad de pago.

Es aquí donde las estrategias de negocio benefician a las instituciones financieras, ya que se crean nuevas propuestas para la atracción de clientes, tanto internos como externos. Es así como se busca que este trabajo se enfoque en la retención de clientes para el mercado hipotecario identificando variables claves para su oportuna detección.

Alcance y justificación

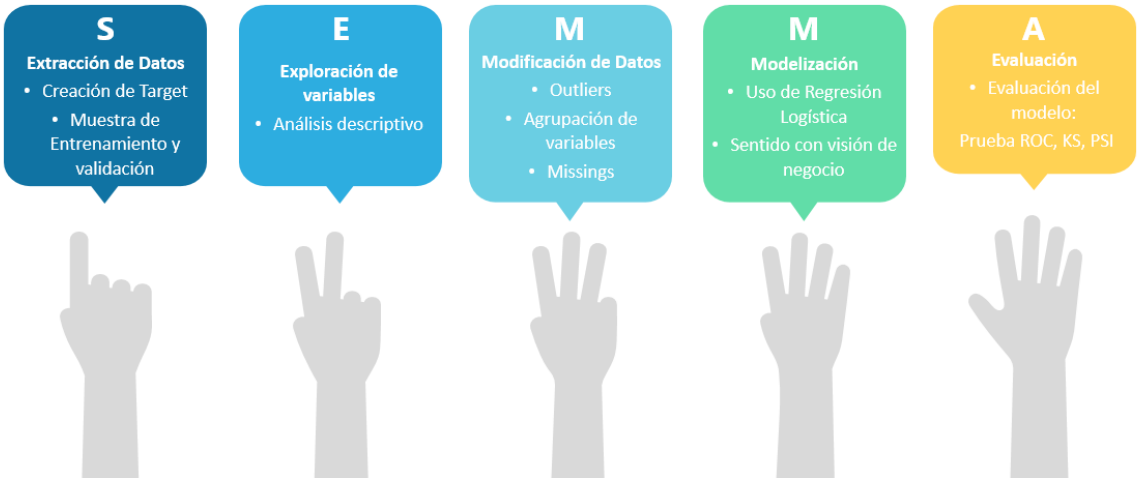
Derivado de lo anterior, el objetivo de esta tesis es realizar un modelo que nos permita analizar a los clientes con créditos hipotecarios con el fin de identificar a aquellos con mayor propensión a mudar su hipoteca, es decir, se creará un modelo de retención a partir de la información obtenida.

Sostenemos que al final de este estudio podremos distinguir las principales características de los clientes con mayor abandono en una institución financiera, lo que nos podrá otorgar mayor visión de negocio para crear estrategias de retención o acrecentar la vinculación con los clientes a consecuencia se traducirá en una mayor rentabilidad y mejor reputación.

Los pasos que seguiremos en esta tesis se desprenden de la metodología SEMMA la cual consiste en los siguiente:

- *SAMPLE*: Aquí obtendremos la población de la cual se busca deducir algún comportamiento relacionado con el objetivo, también se identificará el periodo de tiempo sobre el cual se quiere inferir. Una vez definido esto se hará el muestreo que consiste en crear dos poblaciones; una de entrenamiento y otra de validación.
- *EXPLORE*: En esta etapa se identificarán variables que permitan segmentar a los clientes empleando métodos estadísticos de acuerdo a nuestra variable objetivo.
- *MODIFY:* Se realizará una modificación a aquellas variables que mejor nos ayuden a describir el objetivo para obtener mejores estadísticos.
- *MODEL*: Se buscará crear un modelo el cual nos ayude a inferir mejor en la población de acuerdo a la variable objetivo.
- *ASSESS:* Consiste en evaluar los resultados obtenidos en el paso anterior y la estabilidad de variables.

Figura 1. Metodología SEMMA



Fuente: Elaboración propia.

Al seguir esta metodología buscamos explicar un comportamiento oportuno con el fin de que el área de negocio trabaje propuestas más personalizadas al cliente de un crédito hipotecario con mayor probabilidad de abandono.

Capítulo 1

1. Antecedentes del riesgo de crédito y su regulación

El presente capítulo se realiza para dar al lector un panorama general respecto a los conceptos que abarcan el riesgo de crédito, ya que es uno de los tipos de riesgos más importantes en el sector financiero. Se explican conceptos básicos para su entendimiento, así como el uso y las características de los principales modelos para su análisis.

Dado lo anterior, se espera que este marco conceptual permita mostrar la importancia de la medición del riesgo y el impacto que tiene en la calidad de crédito ya que, como objeto de estudio de esta tesis, deriva en la administración de riesgos empresarial procurando la evaluación temprana de comportamientos no deseados en los clientes.

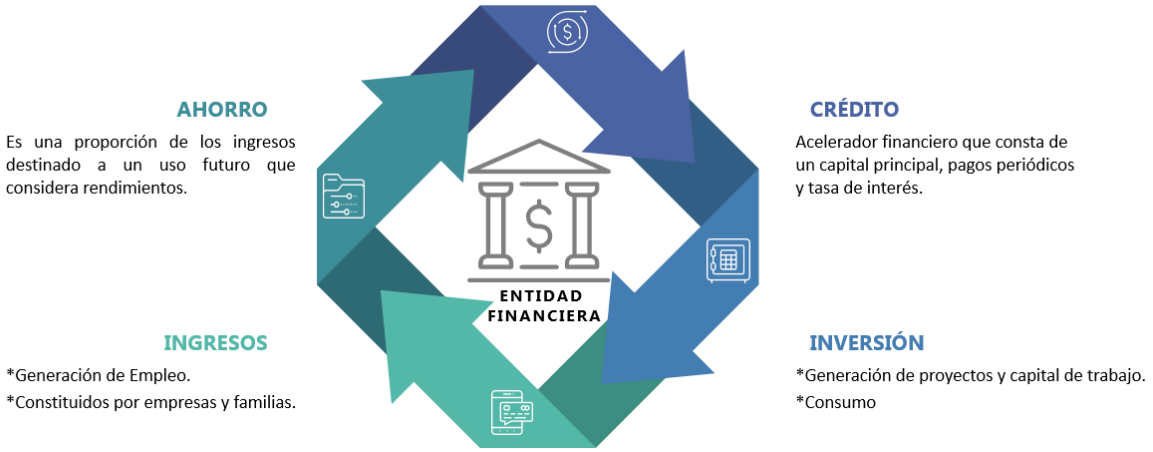
1.1. El crédito

La palabra crédito proviene del latín *credere* el cual tiene por significado confiar o creer. Siendo así, el crédito está definido como una transacción entre dos entes, estos son el acreedor cuya función consiste en otorgar bienes o servicios al acreditado o contraparte con el compromiso de devolución en próximos plazos adicionando una ganancia que se puede traducir en intereses o por el manejo de la operación (Banco de México, s.f.).

El crédito es una forma de financiamiento la cual se emplea para satisfacer necesidades de consumo, es decir, se utilizan recursos que no se disponen en el presente pero que un futuro se esperaría contar con estos. Es por eso que el crédito permite el desarrollo de nuevos proyectos y así mismo mejora los niveles de vida de las personas.

La función del crédito en la economía es impulsar el desarrollo económico y creación de valor, ya que propicia el incremento de intercambios comerciales funcionando como un acelerador financiero. (Figura 2.1).

Figura 2.1. Función del crédito en la economía



Fuente: Elaboración propia.

Dentro de las principales funciones de un banco está el otorgamiento de crédito, los cuales son conocidos como préstamos o financiamientos[1].

Es importante mencionar que el otorgamiento de un crédito será determinado[2] de acuerdo al criterio del acreedor o en este caso institución financiera. Ahora bien, de acuerdo a las necesidades de los clientes, las ofertas de crédito se deben orientar a diferentes formas de producción y consumo.

Las solicitudes de crédito que llegan a las instituciones financieras se evalúan de acuerdo a los criterios de cada una como se mencionó anteriormente, lo que se busca predecir es el comportamiento del cliente dentro de los primeros meses en la composición de la cartera, es por eso que la información juega un papel muy preponderante en la medición del riesgo ya que una mala estimación nos puede generar pérdidas y al solicitante un mal historial crediticio.

[2] Información de Banco de México. Divulgación Sistema Financiero.

Al clasificar adecuadamente al cliente se tiene como objetivo mitigar los efectos del riesgo económico ya que se reduce el montó de pérdida esperado y obtienes mayor conocimiento de tu cartera.

1.1.1. Tipos de Crédito

En México existen diferentes tipos de crédito los cuales están clasificados dependiendo de la necesidad del acreditado (Banco de México, s.f.). En las instituciones financieras los más comunes con los siguientes:

- **Consumo Revolvente:**

  Este tipo de crédito se denomina reutilizable o revolvente debido a que la parte utilizada de la línea de crédito puede renovarse automáticamente en cuanto el tarjetahabiente la liquide (Banco de México). El producto revolvente más utilizado es la tarjeta de crédito debido a su flexibilidad de uso, además de tener una cartera con mayor volumen que otros productos.

- **Hipotecario:**

  Se define como crédito hipotecario a los préstamos otorgados para la adquisición de bienes inmuebles no desplazables por si mismos ni por la acción de alguna persona, por ejemplo, terrenos, casas, departamentos, etc. Cabe mencionar que la garantía que el deudor otorga es en sí el bien inmueble al que se denominará hipoteca. Este tipo de crédito tiene una duración de entre 10 y 30 años por lo que en los pagos mensuales se cubrirá parte del capital y los intereses generados.

- **Automotriz:**

  Son créditos que se otorgan con la finalidad de adquirir un automóvil funcionando de manera prendaria, es decir, que el bien (en este caso el automóvil) quedará como garantía. Este tipo de créditos da la posibilidad a las instituciones financieras de otorgar a las distribuidoras de automóviles la facultad de generar el crédito a los clientes directamente.

- **Nómina**:

  Este tipo de crédito son otorgados a trabajadores cuyo salario es abonado por el mismo banco ya que se tiene una mayor certeza de pago al realizar el cargo respectivo a la cuenta del trabajador.

- **ABCD:**

  Los créditos ABCD como sus siglas lo indican, son créditos para adquisición de bienes de consumo duradero, es decir, bienes muebles que se consumen por varios años. Por ejemplo, se pueden considerar los aparatos electrodomésticos, computadoras, celulares, etc. ya que son bienes que se pueden desplazar por si mismos o por acción de alguna persona.

Como se ha mencionado previamente, la presente investigación se enfocará en el crédito hipotecario ya que el objetivo de este documento es determinar a los clientes con mayor propensión de mudar su hipoteca, adicional de que es un producto de suma importancia en instituciones financieras ya que la cartera cuenta con los montos de crédito más altos y una rentabilidad alta asociada con las tasas de interés.

1.1.2. Riesgo de Crédito

En el ámbito financiero existe la posibilidad de que ocurran acontecimientos los cuales generen pérdidas a partir de la concesión de un crédito, esto debido a que los acreditados incumplan con sus obligaciones de pago total o parcialmente. Es así, que el riesgo de crédito lo podemos definir como la posibilidad de impago por parte del acreditado al vencimiento. (De Lara Haro, 2008) Define el riesgo de crédito como la pérdida potencial que se registra con motivo del incumplimiento de una contraparte en una transacción financiera (o en alguno de los términos y condiciones de la transacción). También se concibe como un deterioro en la calidad crediticia de la contraparte o en la garantía o colateral pactado originalmente.

Algunos de los factores considerados para determinar el riesgo de crédito de un cliente prospecto son información crediticia, demográfica, de comportamiento de pago, financieros, entre otros, que tratarán de estimar la probabilidad de incumplimiento[2].

Otro de los factores considerados de gran relevancia es el credit scoring el cual consiste en asignarle una calificación o puntaje de crédito al cliente con base en su conducta crediticia previa, este representa una estimación del desempeño que además ayuda a automatizar la toma de decisiones en cuanto a conceder o no dicha operación de crédito. Este credit scoring es elaborado por sociedades de información crediticia como Buró de crédito y Círculo de Crédito las cuales recopilan, manejan y entregan información crediticia de personas físicas y morales. Este tipo de instituciones promueven minimizar el riesgo de crédito, que, a su vez, favorecen la cultura de crédito sana y responsable en la población.

Debido a lo anterior, es importante contemplar una evaluación profunda de aspectos cuantitativos y cualitativos del solicitante, con la finalidad de identificar y mitigar el riesgo crediticio, con el objetivo de garantizar el adecuado retorno de los recursos o mantener la calidad crediticia en dicha institución.

La medición del riesgo de crédito juega un papel sumamente importante en las instituciones crediticias ya que al tener un adecuado control de riesgo nos permite identificar en las carteras aquellos clientes con mayor probabilidad de pérdida, es así como las instituciones deben crear metodologías con el fin de calcular con mayor precisión las reservas capturando comportamientos actuales con enfoques prospectivos.

Actualmente en México, la estimación de las pérdidas esperadas se realizaría por las instituciones de crédito considerando 3 etapas dependiendo del nivel de deterioro crediticio de los activos:

- **Etapa 1**: incorpora los instrumentos financieros cuyo riesgo crediticio no se ha incrementado de manera significativa desde su reconocimiento inicial.
    - Consumo → Créditos con PV0 y PV1;
    - Comercial → Créditos con días de atraso menores o iguales a 30 días.

- **Etapa 2**: incorpora los instrumentos en los que se presenta un incremento significativo en el riesgo crediticio desde su reconocimiento inicial. Créditos PV>1 y PV<=3.
    - Consumo → Créditos con PV2 y PV3;
    - Comercial → Créditos con días atraso mayores a 30 días y menores a 90 días.
- **Etapa 3**: englobará los instrumentos en los que existe una evidencia objetiva de deterioro.
    - Consumo → Créditos con PV>3;
    - Comercial → Créditos con días atraso mayores o iguales a 90 días.

Derivado de la crisis de la pandemia en México, las instituciones deberán diseñar escenarios prospectivos (EP's) para realizar la estimación preventiva de riesgos crediticios considerando el nivel de riesgo de crédito previamente calculado. Dichos EP's deberán complementar los modelos históricos al incorporar escenarios que permitan a la Institución identificar situaciones potenciales futuras de forma prospectiva.

En el caso de créditos hipotecarios aquellos que cuenten con incremento significativo del riesgo, es decir, en Etapa 2 se les debe constituir una reserva de vida completa, considerando un pago teórico anual amortizable por el plazo remanente del crédito.

El cálculo de reserva se realiza de la siguiente forma:

$$Reservas\ Etapa\ 1\ o\ 3i = PIi\ x\ SPi\ x\ Eli \quad (1)$$

$$Reservas\ Etapa\ 2i = Max\ (Reservas\ Vida\ Completai, PIi\ x\ SPi\ x\ Eli) \quad (2)$$

Donde:

i = Índice de crédito.
PI = Probabilidad de Incumplimiento.[3]
SP = Severidad de la Pérdida.[4]
EI = Exposición al incumplimiento.[5]

Hay que enfatizar que el monto destinado a reservas no puede considerarse para otro tipo de transacciones, como lo son inversiones ya que como veremos más adelante Basilea se diseñó para tener un monto mínimo de respuesta ante obligaciones futuras.

1.1.2.1. Riesgo de Crédito Individual y Portafolio

Para la gestión del riesgo de crédito es importante determinar la probabilidad de incumplimiento del acreditado. Esta probabilidad puede analizarse a nivel individual y de portafolio.

El riesgo individual refiere a un análisis del acreditado en función de sus características propias como estructura financiera, datos demográficos, etc. En el estudio de este tipo de riesgo se consideran 3 factores importantes (Elizondo, y otros, 2012):

1. Probabilidad de incumplimiento: Es la medida de qué tan probable es que un acreditado deje de cumplir con sus obligaciones contractuales. Su mínimo valor es

[3] Medida de calificación crediticia que se otorga para estimar la posibilidad de que un acreditado incumpla en su obligación de deuda
[4] Mide la magnitud de la pérdida en caso de incumplimiento expresada como porcentaje de la exposición al incumplimiento.
[5] Se define como el monto al que la institución está expuesta al momento del incumplimiento de un crédito.

cero, lo cual indicaría que es imposible que incumpla con sus obligaciones, y su máximo valor es uno cuando es seguro que incumpla. [6]

2. Tasa de recuperación: Es la proporción de la deuda recuperada una vez que la contraparte ha caído en incumplimiento.

3. Migración del crédito: Es el grado con que la calidad o calificación del crédito puede mejorar o deteriorarse.

Por otro lado, el riesgo de portafolio está determinado por el conjunto de créditos que determinan el volumen de cartera. Se determina la calidad de su composición y distribución dentro de la cartera, incluyendo las correlaciones existentes entre los acreditados y su entorno. Es común encontrar grupos de acreditados que poseen características similares por lo que el estudio de estos ayuda a determinar comportamientos que simplifican el análisis.

Las correlaciones que existen entre los acreditados de un portafolio tienen como objetivo visualizar la diversificación de los créditos. Para este tipo de riesgo se consideran estos factores importantes (Elizondo, y otros, 2012):

1. La asociación entre la probabilidad de incumplimiento dado el producto de crédito otorgado.

2. Concentración del riesgo: Una vez determinados los segmentos de acreditados de acuerdo a su probabilidad de incumplimiento se busca encontrar la contribución de estos al riesgo total del portafolio.

[6] Información de Banco de México. Definiciones Básicas de Riesgo. México 2005

Es preciso señalar que nuestro objeto de estudio se centra en el riesgo de crédito de portafolio ya que se analizará la cartera de créditos hipotecarios del portafolio total.

1.1.3. Ciclo de vida del Crédito

Todo crédito debe gestionarse y analizarse para identificar, valorar y priorizar los riesgos que son inherentes a él por lo que es de suma importancia administrar los riesgos para coordinar acciones y enfrentarlos, controlando así los impactos negativos o identificando oportunidades.

Es por ello que la administración de riesgo es el conjunto de principios y herramientas que ayudarán a las instituciones financieras a gestionar el ciclo de vida del crédito de manera efectiva ya que estará fundamentado por las políticas, sistemas y modelos (Banco de México, s.f.).

Las principales etapas de vida de un crédito son:

- Originación: Es la etapa en la cual se prospectan los clientes que serán candidatos a un crédito. Es aquí donde se buscará perfilar a los clientes que conformaran nuestra cartera de crédito utilizando modelos estadísticos que incluyan información demográfica, historial crediticio, comportamiento financiero, ingresos, etc.

- Seguimiento: Es el periodo en el cual se observa el comportamiento del crédito, así como la migración a diferentes morosidades. Se analiza el deterioro o mejora del crédito a través de las distintas estrategias implementadas.

- Cobranza: Se considera como cobranza aquella gestión de la cuenta morosa que puede llevarse a cabo a partir del primer pago vencido y se busca obtener la máxima recuperación evitando así que las pérdidas sean mayores o totales.

Figura 2.2. Ciclo de vida del crédito

Fuente: Elaboración propia.

En la figura 2.2 observamos que el ciclo de vida de crédito presentado sirve principalmente para el crédito de consumo, en las cuales actualmente se trata de contestar las siguientes necesidades de negocio:

- Riesgo - Cálculos de medición de riesgo, principalmente infiere a pérdidas monetarias.
- Retención – Vinculación del cliente para evitar que migre con la competencia.
- Ingreso – Determinación del nivel de ingreso de un solicitante.
- Comportamiento – Generación de estrategias de mitigación o de crecimiento en el portafolio.
- Cobranza – Asignación de manera óptima de los recursos.
- Fraude – Identificación de perfiles fraudulentos sin afectar otra fase del ciclo de crédito.

De acuerdo al objetivo de este trabajo, nos centraremos en la fase de seguimiento de la cual se evaluará el comportamiento de los clientes, esta fase es de gran importancia ya que al ser clientes maduros[7] podemos observar comportamientos atípicos o estables de los clientes para las diferentes variables que nos ayuden a segmentarlos y así, a tomar acciones preventivas para mejorar la vinculación.

1.2. Marco Regulatorio Nacional e Internacional

1.2.1. Comité de Basilea

El Comité de Basilea se originó a partir de la crisis financiera suscitada por la insolvencia de los bancos *Bankhaus Herstatt* y *Franklin National Bank*, la caída del sistema de tasas fijas en materia de intereses y la creciente globalización de los mercados financieros lo que dio pauta al intercambio de información por parte de los bancos centrales para intervenir en los mercados (Banco de pagos Internacionales, s.f.).

El Comité de Basilea es la principal organización en materia de regulación y supervisión bancaria, cuyo principal objetivo es garantizar la estabilidad financiera mundial. Actualmente el comité está formado por miembros de más de 20 países, los cuales se encargan de la supervisión del sistema bancario de su respectivo país.

Dicho Comité anima a una convergencia de estándares comunes buscando la mejor adaptabilidad de ejecución en los diferentes sistemas financieros. Cabe destacar que las recomendaciones y directrices que se emiten son responsabilidad de las autoridades competentes de los países representados en el Comité, así como de aquellos que no forman parte y quieran introducir dichos acuerdos en su legislación (Comisión Nacional Bancaria y de Valores, s.f.).

[7] Se denomina cliente maduro aquellos que cumplan más de 6 meses en la cartera de crédito.

Entre los principales acuerdos que ha elaborado el Comité de Basilea destacan:

- **Basilea I**

  Originalmente conocido como Acuerdo de Capital de Basilea de 1988, en él se establece un sistema para medir el capital de los bancos en función del incumplimiento de pago de sus activos en el cual fijó un capital mínimo del 8% de los riesgos, en otras palabras, identificaba el capital para afrontar los límites máximos de riesgo, solventando así un correcto funcionamiento.
  A pesar de las mejoras en el sistema financiero a partir de este acuerdo se encontraron problemas significativos como se menciona a continuación:

  "En términos concisos, Basilea I define los requerimientos mínimos de capital de un banco en función del riesgo de sus activos y de los riesgos de mercado que afectan a la institución. Sin embargo, la principal limitación del acuerdo de Basilea I es que es insensible a las variaciones de riesgo y que ignora una dimensión esencial de la calidad crediticia y, por lo tanto, la diversa probabilidad de incumplimiento de los distintos prestatarios. Es decir, consideraba que todos los créditos tenían la misma probabilidad de incumplir." (Sierra Núñez, 2011).

- **Basilea II**

  Derivado de las inconsistencias de Basilea I se buscó redefinir el acuerdo y con esto surge Basilea II con el nombre de "Convergencia Internacional de Medición de Capital y Estándares de Capital: Un Marco Revisado", en el cual se constituye un nuevo marco de capital considerando mayor sensibilidad basado en 3 pilares:

  I. Requerimientos Mínimos de Capital
     Este pilar nos indica el capital necesario para que una institución bancaria soporte los riesgos asumidos y se considera por primera vez la calidad crediticia de los prestatarios. Está conformado por tres tipos de riesgo: Crédito, de Mercado y Operacional.

En el caso específico del riesgo de crédito se consideran los siguientes componentes para su control:

- *Probability of Default* (Probabilidad de incumplimiento).
- *Loss Given Default* (Pérdida ante el incumplimiento).
- *Exposure at Default* (Exposición ante el incumplimiento).

II. Supervisión de la Suficiencia de Capital

Ofrece un entorno mayor a riesgos que dadas sus características no se acoplan al pilar I por lo que otorga ciertas instrucciones para la regulación de capital. También permite al supervisor exigir capital adicional a las instituciones que incumplan la normativa.

III. Disciplina de Mercado

Este tercer pilar establece requerimientos para la divulgación de información certera y oportuna brindando a los usuarios la libertad de elección para optimizar sus recursos. Se enfoca en la transparencia informativa.

- **Basilea III**

Este conjunto de medidas surge tras la crisis financiera del 2008, la cual fue desencadenada por instituciones financieras que otorgaron hipotecas a clientes sin realizar un análisis detallado de la capacidad de crédito emitiendo valores para financiarse, los cuales al dejarse de pagar las hipotecas se declararon en quiebra. Esta crisis provocó una desaceleración económica a nivel global por lo que el objetivo de Basilea III es reforzar la regulación, supervisión y gestión de riesgo de los bancos para así lograr fortalecer el capital y cubrir pérdidas potenciales de su actividad.

Los tres pilares que lo conforman son los siguientes:

I. Medición de riesgos

Las principales características de la medición de riesgos son:

a. Capital de mayor calidad: El requerimiento mínimo de capital total sigue establecido en un 8% al igual que en Basilea I y II, el cual estará compuesto en su mayoría por capital de alta calidad, este es un elemento clave de los recursos propios de un banco.
b. Colchón de conservación de capital: Este punto refiere al aumento de capital fuera de periodos de tensión con el objetivo de hacer uso de esta reserva en caso de incurrir en pérdidas. Son sencillas normas diseñadas para evitar el incumplimiento de requerimientos mínimos de capital.
c. Colchón anticíclico: El objetivo es garantizar que las instituciones financieras tengan en cuenta el entorno en el que operan, es así, que se exigirá un colchón de capital que le proteja ante posibles pérdidas futuras cuando se estime un crecimiento excesivo del crédito aumentando así el riesgo sistémico.

II. Autoridad supervisora

Las autoridades financieras de cada país deben tener suficientes facultades para la intervención oportuna en instituciones que tengan evidencia de inestabilidad. También se pretende una mayor cooperación entre los entes de supervisión a nivel internacional.

III. Transparencia

Solicita el reforzamiento en la transparencia y divulgación de información resaltando la claridad en los riesgos de instrumentos complejos, así como en los instrumentos derivados y de esta manera conocer las operaciones efectuadas por las instituciones financieras.

1.2.2. Sistema Financiero Mexicano

El sistema financiero mexicano es el conjunto de instituciones y organismos dentro del cual se organiza la actividad financiera con el objetivo de movilizar el ahorro en sus usos más eficientes dentro del marco legal que corresponde en territorio nacional.

Este permite que la activad económica se desarrolle adecuadamente ya que facilita el flujo de recursos, mientras que los individuos que tienen suficientes recursos para aportar reciben un rendimiento de acuerdo a una tasa de interés otros necesitan en el corto plazo más recursos lo que origina préstamos con un costo de tasa de interés; por lo que se logra equilibrar las aspiraciones de unos con las necesidades de otros.

Este sistema tiene como principal órgano administrativo a la Secretaría de Hacienda y Crédito Público (SHCP). Esta dependencia gubernamental centralizada tiene por objetivo obtener recursos monetarios para financiar el desarrollo del país, con lo que se busca el equilibrio presupuestal del gobierno; también distribuye el presupuesto entre las diversas secretarías, dependencias del gobierno, organismos descentralizados, estados y municipios, conforme lo aprueban los diputados.

Junto con ella existen 6 instituciones más que logran la supervisión y regulación del sistema financiero mexicano[8]:

- **Banco de México (Banxico):** Esta institución su principal objetivo es mantener la inflación baja y estable mediante la creación y regulación de la moneda nacional esto con el fin de mantener un poder adquisitivo estable. Es constitucionalmente autónomo en sus funciones y administración[9].
- **Comisión Nacional Bancaria y de Valores (CNBV):** Este tiene como facultades la autorización, regulación y supervisión a las entidades financieras con el fin de procurar su estabilidad y correcto funcionamiento para que estas hagan buen uso de los ahorros y también cuidar la forma en la que se otorgan créditos.

[8] Información de portal SHCP https://www.gob.mx/shcp
[9] Información de portal https://www.banxico.org.mx/

- **Comisión Nacional de Seguros y Fianzas (CNSF):** Dentro de sus principales funciones están la inspección y vigilancia de las instituciones y de las sociedades mutualistas de seguros.
- **Comisión Nacional de Sistemas de Ahorro para el Retiro (CONSAR):** Se encarga de la regulación y supervisión de los sistemas de pensiones, buscando la protección de los intereses de los trabajadores asegurando una administración eficiente y transparente de su ahorro.
- **Comisión Nacional para la Protección y Defensa de los Usuarios de Servicios Financieros (CONDUSEF):** Su objetivo es proteger, asesorar y defender los derechos de las personas que utilizan un producto de una institución financiera; también busca promover la educación financiera.
- **Instituto para la Protección al Ahorro Bancario (IPAB):** Su objetivo es proteger los depósitos de los ahorradores para con ello, contribuir a salvaguardar la estabilidad del sistema financiero.

Figura 2.3 Estructura del Sistema Financiero Mexicano

Fuente: Secretaría de Hacienda y Crédito Público.

Las propuestas contenidas en este marco regulatorio nos proponen la incorporación de nuevos desarrollos dentro de la teoría financiera de cuantificación del riesgo considerando el capital regulatorio mínimo que es exigido a las entidades financieras.

El riesgo, al ser un ente que no puede desaparecer y solo puede ser mitigado, requiere de un profundo análisis; la administración del mismo ha evolucionado en las entidades financieras ya que permite conocer la exposición de sus portafolios de crédito y establecer las pérdidas a las cuales se exponen. Debemos tomar en cuenta que el pilar fundamental de la supervisión bancaria está en la gestión interna del riesgo y como tal, está orientado a fortalecer y mejorar esta gestión, generando criterios para mantener adecuadamente evaluados los riesgos de crédito implícitos en los activos de crédito.

Aunque estas regulaciones no obligan a las entidades financieras a usar un modelo específico en la evaluación de riesgo crediticio establecen algunos parámetros que sirven como guía para la construcción de un modelo, que serán consideradas de acuerdo al objetivo del mismo.

Cabe destacar que estas regulaciones buscan incluir medidas para controlar los riesgos sistémicos, en la actualidad México ha perseguido la evolución continua de la regulación y supervisión bancaria para adaptarse a nuevos contextos derivados de las innovaciones financieras. Dado esto, también busca disminuir las probabilidades de un colapso bancario y así el impacto que tendría a su vez en la economía, comenzando por las pérdidas monetarias y la posible parálisis del crédito en el sector económico.

Es importante considerar que en nuestro país el ente que supervisa a las instituciones bancarias en términos de crédito es la Comisión Nacional Bancaria y de Valores, quien exige que se dispongan de los adecuados procesos de gestión de riesgo de crédito y ofrezcan una visión integral de la exposición a ese riesgo. Este supervisor verifica que dichos procesos concuerden con el apetito de riesgo, el perfil de riesgo y la solidez del capital del banco, teniendo en cuenta la situación macroeconómica y de los mercados, dando como resultado unos criterios prudentes en materia de concesión, evaluación, administración y vigilancia de los créditos.

También exige que la estrategia se encuentra documentada y eficazmente aplicada, así como sólidas políticas y procesos para la asunción de riesgo de crédito sin la necesidad de evaluaciones de crédito externas.

Capítulo 2

2. Metodología y análisis de datos

De acuerdo a la metodología SEMMA, en este capítulo se describirá la metodología planteada para la resolución de un caso práctico y, por medio del análisis exploratorio se seleccionarán las mejores variables aplicando una serie de métricas descriptivas que nos permitan discernir aquellas variables con baja potencia estadística en relación al sentido de negocio.

Al final de este capítulo se habrá reducido la dimensionalidad del problema, en otras palabras, se reducirán el número de variables conteniendo las más significativas para la aplicación del modelo.

2.1. Medición de riesgo de crédito

Como se ha mencionado previamente el riesgo de crédito ha ido evolucionando y como consecuencia su estudio adquiere mayor relevancia en el ámbito empresarial ya que se espera el tratamiento oportuno para evitar pérdidas. Identificar los riesgos constituye una fuente de información de vital importancia para la gestión y competitividad empresarial, y eso requiere de herramientas que complementen la experiencia y ayuden a la eficiente toma de decisiones.

La evolución de la administración de riesgos ha permitido el desarrollo de estrategias que establezcan acciones para identificar, evaluar y monitorear riesgos que puedan afectar el cumplimiento de los objetivos de una empresa logrando responder con medidas más efectivas (De Lara Haro, 2008).

Actualmente adquiere una gran importancia la prevención de riesgos ya que si un riesgo no es identificado oportunamente implica afrontar las consecuencias de su materialización, es decir, no se pueden definir medidas para disminuir su probabilidad de

ocurrencia o minimizar el impacto financiero. Para ello es primordial contar con información actualizada y pertinente, con el fin de establecer mecanismos de alerta temprana que permitan identificar cambios en variables críticas[10] logrando determinar riesgos que puedan impactar a la organización y de esta manera crear procedimientos que garanticen el monitoreo de forma periódica y que contemplen las transformaciones del entorno económico.

Existen variedad de técnicas y metodologías que se pueden aplicar dependiendo de los recursos económicos, humanos y tecnológicos de la organización, así como de la implementación de la administración de riesgos. Es importante que en la aplicación de estos métodos se observe consistencia en su uso, de manera que garantice una administración en forma integral de acuerdo a los objetivos de la organización.

Estas técnicas se pueden clasificar dependiendo si existe o no una variable objetivo para modelar o predecir, se les denomina como de aprendizaje supervisado y no supervisado. Los primeros buscan realizar una clasificación de acuerdo a una variable objetivo mientras que los segundos intentan encontrar un patrón de comportamiento de acuerdo a las características de la población.

Gracias a los avances en la tecnología, el sector financiero ha mejorado en el tratamiento de altos volúmenes de información para la gestión de riesgos y la toma de decisiones oportuna.

2.2. Aprendizaje Automatizado (Machine Learning)

El concepto de *machine learning* surge en el siglo XX a finales de los años cincuenta coincidiendo con el inicio del desarrollo de la inteligencia artificial, al cual podemos definir como una disciplina científica y rama de la inteligencia artificial cuyo objetivo reside en que los sistemas aprendan automáticamente, es decir, sin intervención humana. En este sentido el término de aprendizaje automatizado se refiere a la identificación de patrones complejos de un gran volumen de información, sin embargo quien realiza el aprendizaje será el

[10] Se considera aquella variable que influye directamente en el proceso.

algoritmo[11] ya que este es capaz de predecir comportamientos adaptándose a la incorporación de información y recalibrando resultados.

Un punto importante de las técnicas de *machine learning* es que poseen un soporte estadístico lo cual permite obtener más información de los datos lo que representa un complemento en la resolución de problemas, es decir, generan una mayor relevancia en los datos ya que se pueden fusionar generando más aristas de un problema (Goldberg, 1988).

También se busca que la detección de patrones que realiza el aprendizaje automatizado sea interpretable ya que su objetivo es desarrollar algoritmos con valor práctico por lo que se busca que sean eficientes considerando el tiempo de ejecución y el espacio utilizado. Los dos puntos anteriores son de gran relevancia ya que al tratarse de un alto volumen de información generalmente son estos los que determinan si es de utilidad el modelo o no.

Actualmente la aplicación de estas técnicas ha permitido un mejor desempeño en la gestión de riesgo de crédito ya que surge derivado de las siguientes situaciones: mayor volumen de información, cambios socioeconómicos más frecuentes e identificación oportuna de créditos morosos.

Existen varios métodos dentro de *machine learning*, los más comúnmente adoptados son el aprendizaje supervisado y el no supervisado.

2.2.1. Aprendizaje supervisado

El aprendizaje supervisado consiste en partir de una muestra $D = \{(Xi|Yi)|i = 1,2, \dots n\}$ construida por n realizaciones o entrenamientos de un par de variables $(X|Y)$, con lo cual se construye un función $f: X \rightarrow Y$ que permite a partir de un vector de entrada X, se pueda predecir con cierto grado de confianza la variable $Y = f(x)$.

[11] Serie de pasos organizados que describen un proceso para obtener una solución de un problema específico.

Para cada observación $(Xi|Yi)$ de D, a la variable $Xi \in X$ se le llamará variable de entrada, explicativa o input y a $Yi \in Y$ variable dependiente u output. También se les conoce a las variables input como características o atributos.

En otras palabras, en el aprendizaje supervisado, el algoritmo crea una función de correspondencia entre las variables input y output. Este tipo de aprendizaje es ampliamente utilizado en la práctica ya que resuelve problemas de clasificación y regresión; en el caso de que la variable dependiente sea categórica o nominal de un conjunto finito se habla de un problema de clasificación y en el caso de que la variable dependiente sea continua estamos ante un problema de regresión.

2.2.2. Aprendizaje no supervisado

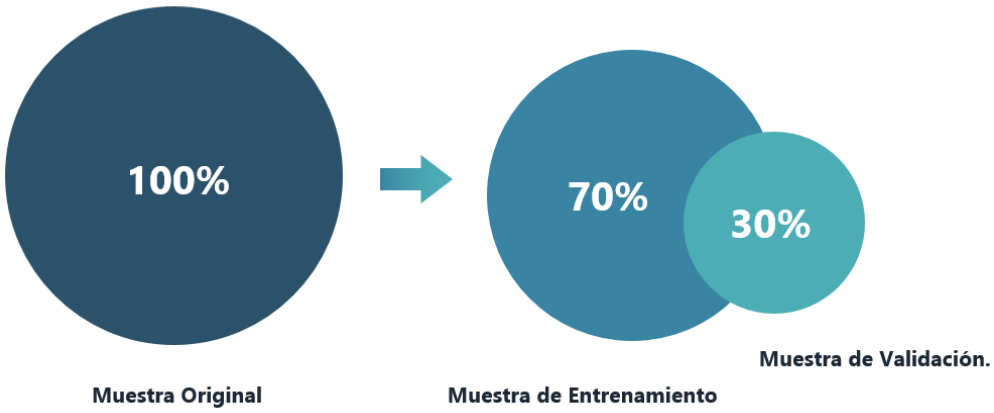
El segundo método principal es el aprendizaje no supervisado en el cual solo se tienen variables de entrada $D = \{Xi|i = 1,2, \dots n\}$, el objetivo principal es identificar patrones interesantes en la información que permitan clasificar o asociar de acuerdo a ciertas características.

Este es un problema mucho menos definido ya que no se dice que tipo de patrones buscar y no hay una métrica de error obvia para usar, a diferencia del aprendizaje supervisado donde se puede comparar la predicción de Y para un valor observado X.

2.3. Datos

Los datos utilizados en este estudio se obtuvieron del portafolio de créditos hipotecarios de una institución financiera y por cuestiones de confidencialidad de datos se consideró una muestra a la cual se le aplicó un factor para su presentación. La muestra cuenta con 179, 762 registros, cabe destacar que la forma de trabajo en ella será con 70% de entrenamiento y 30% de validación, lo cual nos permite validar la solidez del modelo.

Figura 3.1 División de conjunto de datos



Fuente: Elaboración propia.

Esta muestra tiene 166 variables las cuales se dividen de acuerdo a la siguiente tabla.

Tabla 3.1 Tipo de Dato

Tipo de Datos	# de Variables
Carácter	20
Numérica	146
Total	**166**

Fuente: Elaboración propia.

Considerando los resultados anteriores proseguimos a un primer análisis exploratorio de los posibles valores distintos que hay en cada una de las 166 variables con el fin de catalogar correctamente a cada una, es decir, se evaluó si correspondían verazmente a una variable de intervalo o a una variable categórica para análisis posteriores.

Se reconoció que existen variables numéricas con solo un valor para todos los registros; a estas las descartamos a partir de este punto ya que este tipo de variables son prescindibles. También hallamos variables numéricas con solo 2 valores, lo que nos indica que este tipo de variables deben tener un tratamiento diferente ya que nos permiten clasificar los datos por medio de valores fijos asociados a una categoría. Dado lo anterior obtuvimos los tipos de variable que existen en la información que estamos trabajando (Tabla 3.2).

Tabla 3.2 Tipo de Variables

Tipo de Variable	# de Variables
Categórica	127
Intervalo	35
Unaria	4
Variables Restantes	**162**

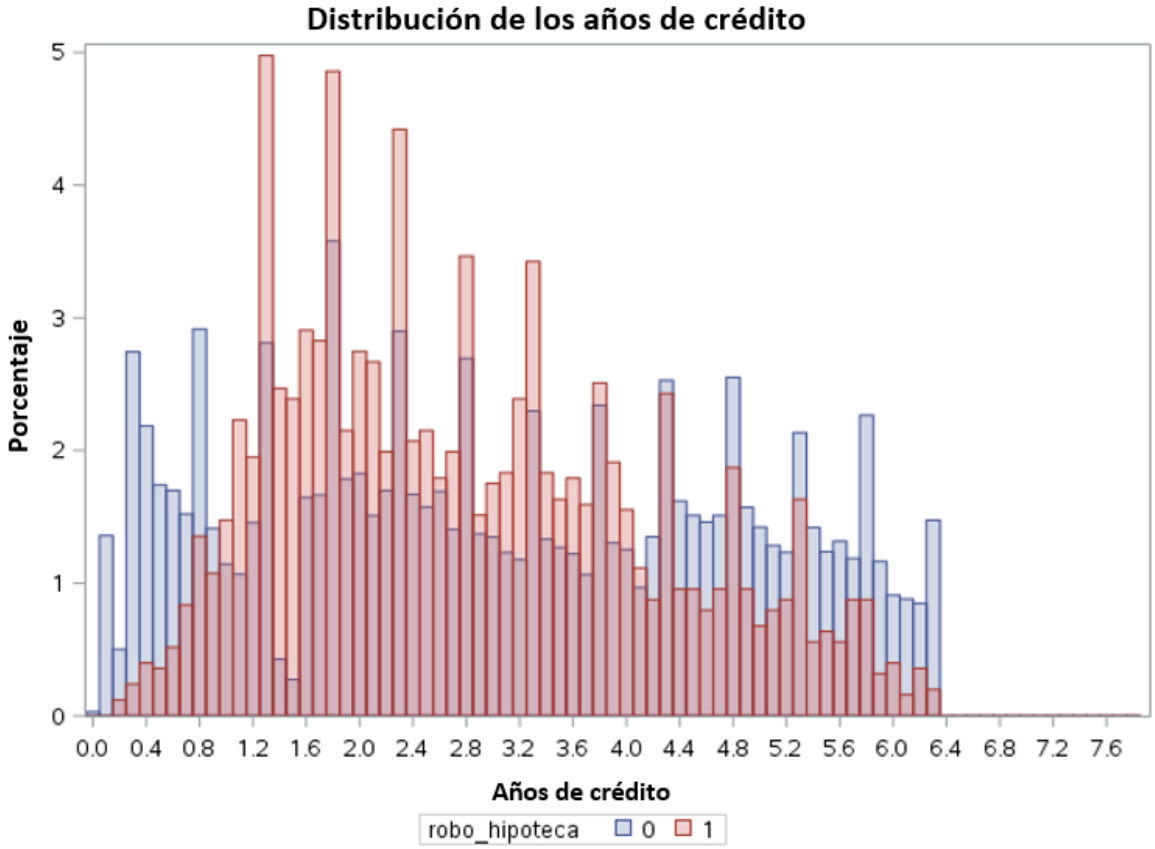
Fuente: Elaboración propia.

Al tener las variables clasificadas de esta forma nos ayuda a determinar el tipo de análisis y métodos al que pueden ser sometidas estas.

2.3.1. Variable Objetivo

Como lo hemos comentado a lo largo de este trabajo nuestro objetivo es identificar aquellos clientes con mayor propensión a mudar su crédito hipotecario, dentro de las variables mencionadas en el cuadro anterior se consideraron 2 variables numéricas para definir el *target* (variable objetivo) las cuales son *robo_hipoteca* y *anios_credito*. La primera contiene valores [1, 0] y nos describe si el cliente ha mudado su crédito hipotecario o no respectivamente, la segunda determina los años del crédito que han trascurrido.

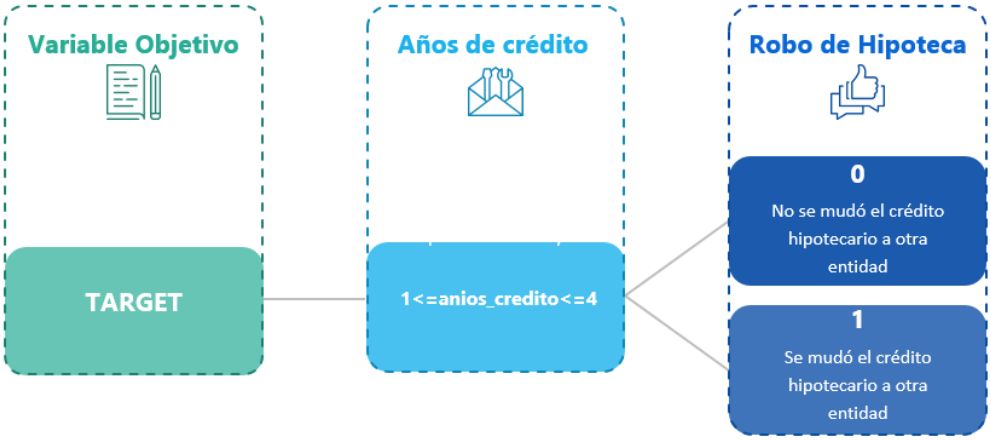
Figura 3.2 Gráfica de distribución de variable objetivo



Fuente: Elaboración propia

Como observamos en la gráfica anterior de acuerdo a la distribución de la variable *robo_hipoteca* en *anios_credito*, se observa una mayor recurrencia entre los años 1 y 4 del crédito por lo que decidimos definir la variable objetivo como se muestra en la figura 2.3.

Figura 3.3. Definición de variable objetivo



Fuente: Elaboración propia

2.4. Análisis Descriptivo

La exploración de datos debe empezar por el análisis descriptivo contemplando estadísticas simples como lo son la media, mediana, moda, percentiles, etc. (Siddiqi, 2006) lo que nos permitirá tener una mejor visión del objetivo.

Realizamos la evaluación de valores ausentes en cada una de las variables, y es así como se descartaron aquellas cuyo porcentaje de ausentes sea mayor al 85%. A su vez también se descartaron variables que se concentran en un solo valor. Se consideró de esta manera ya que no permiten una correcta segmentación de la población por la variable objetivo al no poder indicarnos un comportamiento relevante.

Posteriormente se consultó la definición de las variables para tener un mayor entendimiento de estas, es así como se descubrió que en varias de ellas se conocían los valores; posterior al robo de hipoteca. Estas variables se descartaron ya que no son preventivas al hecho de estudio de este trabajo.

2.4.1. Segmentación de la población

Una vez teniendo mayor conocimiento de la población se buscó dividirla en diferentes grupos de acuerdo a las características asociadas que tuvieran, ya que de esta manera se buscaba emplear modelos diferenciados para cada segmento de la población y así conseguir las variables que mejor explican la variable objetivo.

2.4.1.1. Metodología

Para esta segmentación se usó el análisis *cluster*, el cual es un conjunto de técnicas multivariantes[12] utilizadas para clasificar a un conjunto de individuos en grupos homogéneos de acuerdo a una característica, este es un método de aprendizaje no supervisado. Este algoritmo agrupa los datos basándose en la distancia entre cada uno y se basa en las similitudes de las variables de entrada.

Este sirve para realizar dos tareas fundamentales:

1. Análisis Taxonómico o de clasificación.
2. Cambio o reducción de la dimensión de los datos.

En estos las consideraciones más importantes son que cada grupo debe ser homogéneo, es decir, debe tener características muy similares, y los grupos deben ser lo más distintos posibles; de lo contrario no existe una posible segmentación de la población.

Dado lo anterior se pretende encontrar un conjunto de grupos a los cuales se les asignará algún criterio de homegeneidad.

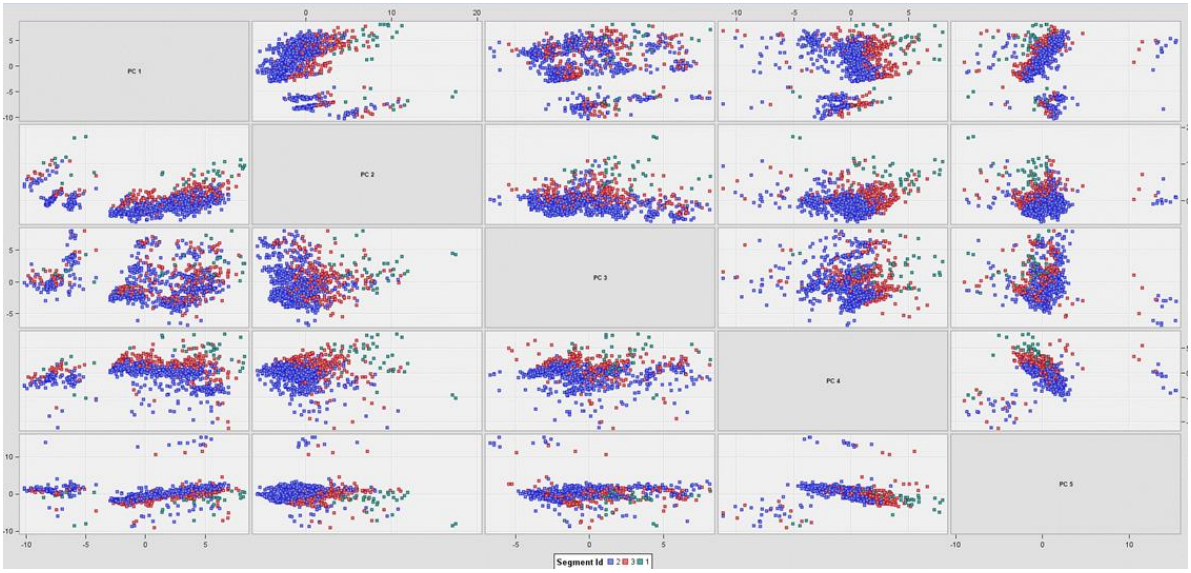
[12] Es un tipo de análisis simultáneo de más de una variable.

En una primera prueba se trabajó con métodos jerárquicos los cuales consisten en dividir el conjunto de datos en tantos grupos el algoritmo considere como adecuados. Dado que en esta prueba creaba demasiados grupos se recurrió a un método no jerarquizado el cual nos permitió definir el número de grupos que se desea obtener.

Una vez realizado este método nos dio por resultado una segmentación la cual nos indica que no existe una clara separación entre las observaciones, es decir, no se puede segmentar a la población por algún grupo de variables.

De forma gráfica, estas conclusiones se pueden observar en la figura 3.4 ya que muestra las diferentes dimensiones del *clúster* y no se pueden distinguir distintas poblaciones. Derivado de estos resultados se procede a realizar un solo modelo de *scorecard* para la población total.

Figura 3.4 Matriz de Clúster No Jerárquico



Fuente: Elaboración propia con SAS MINER

2.4.2. Modificación de Variables

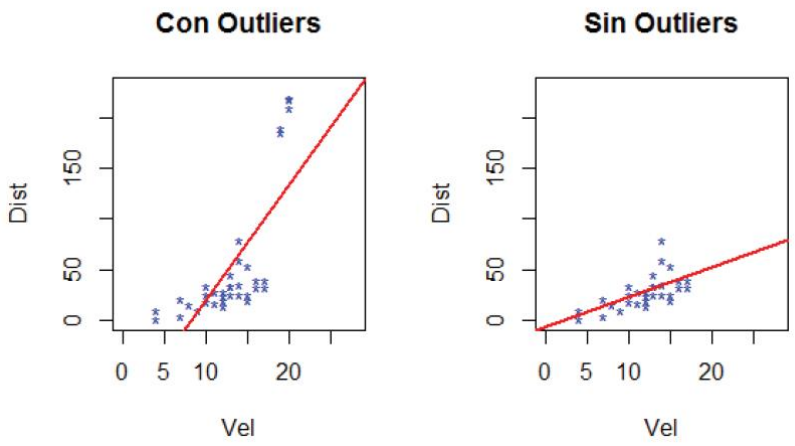
Para poder considerar variables cuya calidad de datos era baja se consideró aplicar dos criterios adicionales los cuales refieren a la modificación de valores atípicos (outliers) y a la modificación de valores ausentes (missings).

Un *outlier* es una observación anormal y extrema en una muestra estadística que puede afectar potencialmente a la estimación de los parámetros del mismo. Estos datos atípicos pesan más que los datos cercanos a la media (Freed, A linear programming approach to the discriminant problem, 1991).

La mayoría de las estadísticas paramétricas, como la media, la desviación estándar y las correlaciones, así como las medidas estadísticas de tendencia central y todas las estadísticas derivadas de estas son altamente sensibles a los outliers. De manera que si la información contiene una cantidad significativa de outliers es importante realizar un tratamiento de estos antes de entrar a la fase de entrenamiento del modelo.

En nuestro caso, se identificaron aquellas variables las cuales su valor en el percentil 99th Pctl estaba muy alejado de su máximo, se midió esta proporción en unidades y en porcentaje para poder sustituir un nuevo máximo que no se alejara de manera significante. En la población objeto de nuestro estudio solo 18 variables se modificaron bajo este argumento.

Figura 3.5 Influencia de valores atípicos en un modelo



Fuente: Universidad Politécnica de Madrid

La presencia de datos ausentes es considerada un problema en el campo de la medición debido a que disminuye el poder estadístico. En la actualidad se pueden contar con enfoques para el manejo de datos perdidos más eficientes y fáciles de implementar gracias a los avances tecnológicos, permitiendo así recuperar los valores perdidos y restablecer el poder estadístico.

En el caso de los valores ausentes de variables categóricas se creó una nueva categoría indicando la ausencia de estos datos, esto debido a que se comprobó que este valor tiene la posibilidad de no reflejarse en las fuentes de información.

Para las variables de intervalo no se hallaron valores ausentes por lo que no se realizó un tratamiento especial. Cabe destacar que una de las alternativas más comunes para sustituir valores ausentes en este tipo de variables es por el valor de la media de la variable.

2.4.2.1. Reagrupación de variables

Otro inconveniente que se encontró fue el gran número de niveles dentro de las variables categóricas. Las variables con mayor número de categorías son las siguientes:

- Estado civil vs número de dependientes
- Zona Territorial
- Zona de apertura del crédito
- Tipo de cliente vs ingresos

Para poder reagrupar estas variables y crear menos categorías se asociaron de acuerdo a la distribución de abandono que presenta cada una de las categorías, cuidando que entre estas categorías resultantes el *bad rate*[13] fuera lo más distinto posible y el número de observaciones fuera similar entre los grupos para poder crear categorías significativas.

Tabla 1.3 Resultados de Agrupación

Variable	# Categórias	# Categorias Agrupadas
Estado civil vs número de dependientes	10	4
Zona territorial	13	4
Zona de apertura del crédito	234	5
Tipo de cliente vs ingresos	15	3

Fuente: Elaboración propia

Dado los métodos descritos anteriormente, las pruebas realizadas sobre las variables, así como las modificaciones realizadas en ellas nos ha permitido prescindir de 120 variables que dadas las conclusiones ya descritas presentan inconsistencias con su información debido a

[13] Tasa de robo de hipoteca

que muchas de ellas tenían un gran porcentaje de ausentes o se trataban de variables con un solo valor, además se descartaron aquellas entradas cuya definición era una vez hecho el robo de hipoteca mismo que nos permitió mantener la mejor información en solo 42 variables.

2.5. Análisis Multivariado

2.5.1. Análisis de Correlación

El coeficiente de correlación expresa el grado de asociación entre dos variables y se utiliza principalmente para el análisis de datos en la etapa multivariada. Este coeficiente mide la fuerza y la dirección de la asociación de dos variables (x,y) y tiene valores entre -1 y +1.

En la actualidad se utilizan los coeficientes de Pearson y Spearman, que principalmente se diferencian en que el segundo funciona para series de datos en los que existan valores atípicos y por lo tanto su distribución no sea normal (Hand, 2002).

De acuerdo al valor de coeficiente podemos indicar lo siguiente:

- Si $\rho = 1$ – Perfecta correlación positiva
- Si $0 < \rho < 1$ – Correlación positiva, si una de las variables crece la otra también crecerá
- Si $\rho = 0$ – Correlación nula
- Si $-1 < \rho < 0$ – Correlación negativa, si una de las variables decrece la otra también decrecerá
- Si $\rho = -1$ – Perfecta correlación negativa

Estos coeficientes de correlación como lo mencionamos previamente nos indican la asociación entre dos variables lo que nos ayuda a reducir la dimensión del número de variables a considerar en el modelo de robo de hipoteca. En el Anexo 1 se pueden encontrar las matrices de correlación con ambos métodos.

2.5.2. V de Cramer

El coeficiente de Cramer o V de Cramer es otro de los coeficientes más utilizados para ver la asociación de variables categóricas ya que funciona como una medida de relación estadística basada en la prueba ji-cuadrada. Lo que se pretende con este coeficiente es hacer una corrección de la ji-cuadrada donde se pueda precisar la fuerza de asociación entre dos o más variables.

En este sentido el resultado del coeficiente varía entre cero y uno, por lo que su interpretación es la siguiente,

- V = 0 – Indica que no hay asociación
- 0 < V <= 0.2 – Existe asociación baja
- 0.2 < V <= 0.6 – Existe asociación moderada
- 0.6 < V < 1 – Existe asociación fuerte

Algunas de las consideraciones que debemos tomar en cuenta para su aplicación son el cálculo previo de la prueba ji-cuadrada y que las variables deben tener al menos dos categorías. Cabe destacar que solo trabaja con variables categóricas y que su resultado siempre será positivo por lo que en comparación con el coeficiente de Pearson o Spearman aquí no se puede hacer un análisis sobre la dirección de la relación entre variables.

2.5.3. Estadístico GINI

El estadístico GINI fue propuesto en la década de los 60's y es un aproximación que proviene de la curva de Lorenz, este estadístico es sumamente utilizado ya que su importancia radica en poder medir la desigualdad entre dos poblaciones, en este caso, clientes buenos y malos[14].

La curva de Lorenz está dada por las funciones de distribución H (x) y G (x) y es el subconjunto del producto cartesiano dado por:

$$L(H,G) = \{(u,v)|u = F(x) \wedge u \ = G(x); x \epsilon \mathbb{R}\} \quad (3)$$

Sea H y G las funciones de distribución teóricas asociadas a los clientes buenos y malos respectivamente, donde x es el corte. En caso de que la proporción de clientes buenos sea mayor a la proporción de clientes malos, la curva de Lorenz de H y G es cóncava hacia arriba. Si H(x) = G(x) esto tiene como consecuencia que u = v ϵ (0,1), por lo que mientras L se separe más de la recta u = v mayor será la diferencia entre H(x) y G(x).
Dado lo anterior el área A que se encuentra entre la identidad y la curva de Lorenz, es una medida de desigualdad entre las distribuciones H y G. Es decir, calcula el área entre la curva y la diagonal de la curva de Lorenz, mientras mayor sea el valor en absoluto, mayor será la asociación.

En caso de no tener certeza sobre las funciones de distribución H(x) y G(x) pero se tenga información sobre las muestras aleatorias de cada una de estas dos distribuciones empíricas de tamaño n_1 y n_2 respectivamente, se puede estimar la curva de Lorenz y por lo tanto el estadístico GINI.

[14] De acuerdo a la definición de *target* o variable objetivo.

Comenzaremos estableciendo un punto de corte xi con i =0,1,…,k y seguidamente se obtendrán los estimadores H y G en los puntos xi de la siguiente manera,

$$\hat{\mathrm{H}}(xi) = \frac{\#\ de\ elementos\ en\ la\ muestra\ 1\ \leq xi}{n_1} \tag{4}$$

$$\hat{\mathrm{G}}(xi) = \frac{\#\ de\ elementos\ en\ la\ muestra\ 2\ \leq xi}{n_2} \tag{5}$$

La estimación de la curva de Lorenz de H(x) y G(x) viene dada por la unión de los segmentos de recta que unen a los puntos. El área por debajo de la curva de Lorenz estimada para un intervalo tiene una forma de trapecio y se calcula de la siguiente manera,

$$Ai = \frac{(\hat{\mathrm{H}}_i - \hat{\mathrm{H}}_{i-1})((\hat{\mathrm{G}}_i - \hat{\mathrm{G}}_{i-1}))}{2} \tag{6}$$

La totalidad del área por debajo de la curva de Lorenz estimada es $(\hat{\mathrm{H}}(x_i), \hat{\mathrm{G}}(x_i))$,

$$A = \sum_{i=2}^{k} Ai \tag{7}$$

Resultando el cálculo del GINI estimado dada la siguiente expresión,

$$GINI = \frac{1/2 - A}{1/2} \tag{8}$$

Utilizando estas metodologías en las 42 variables restantes logramos reducir la dimensionalidad de variables a elegir para el uso del modelo de robo de hipoteca quedándonos con las 14 variables con mayor desempeño estadístico y que mejor nos permiten conocer las causas de la variable dependiente.

Capítulo 3

3. Aplicación del modelo y análisis de resultados.

Se propone un modelo supervisado el cual tiene como objetivo disminuir el robo de créditos hipotecarios, considerando el funcionamiento del crédito en el mercado financiero mexicano.

En el presente capítulo se busca obtener un modelo estadísticamente significativo, considerando metodologías que potencialicen las variables restantes. El modelo será capaz de asociar características a clientes que migran un crédito hipotecario, así mismo, una vez concluido el modelo se podrán crear estrategias en una institución financiera para crear mayor vinculación.

3.1. Análisis de Agrupaciones Binning

El análisis de agrupaciones también conocido como *Binning* es un paso comúnmente utilizado credit scoring para la transformación de variables continuas en un conjunto de grupos o bins. El objetivo de este proceso es colocar atributos (valores) con comportamientos similares en un mismo grupo con el fin de mejorar el poder predictivo del modelo. Es una técnica en la que se busca categorizar a las variables continuas (Hand, 2002).

Durante esta etapa del modelaje se buscó revisar la predictibilidad de las variables divididas por rangos, buscando crear categorías diferenciadas por su concentración de población mala, es decir, se pretendía componer categorías que fueran monótonas crecientes o decrecientes en relación a su *bad rate*. También se consideró mantener un porcentaje de la distribución de la población con el fin de poder reclasificar mediante estas nuevas variables la población de manera más eficiente.

La formación de estas agrupaciones nos permitirá introducir dos conceptos que se relacionan con el reagrupamiento de variables los cuales son el Weight of Evidence y el Information Value los cuales abordaremos en este mismo capítulo.

Tabla 3.4 Ejemplo de agrupación en la variable Último Pago

Rangos	Tasa de Malos	% Población
Último Pago < 5,000	0.69%	22.74%
5,000 <= Último Pago < 9,000	1.34%	25.97%
9,000 <= Último Pago < 14,700	2.17%	26.21%
14,700 <= Último Pago < 20,000	3.58%	11.48%
20,000 <= Último Pago	4.43%	13.60%

Fuente: Elaboración propia

En la tabla 3.4 se observa el *binning* realizado en la variable Último Pago, es aquí donde se procuró que la tasa de malos sea monótona creciente y que su vez el porcentaje de la población que vive en cada rango sea considerable y proporcional para cada uno.

En general, los diferentes algoritmos para reagrupar los datos deben considerar lo siguiente:

- Separar los valores missings
- Crear grupos con al menos un caso positivo y uno negativo
- Agrupar de manera que se maximice la diferencia entre los mismos grupos (casos positivos y negativos)

Es preciso puntualizar que estas agrupaciones deben considerar también un orden lógico y sentido operacional, el cual no debe de ser necesariamente lineal (Siddiqi, 2006). En la práctica la creación de grupos suficientes se debe a las simulaciones que se lleven a cabo para conseguir una mejor segmentación de la variable.

3.2. Weight of Evidence & Information Value

El peso de la evidencia o comúnmente conocido como *Weight of Evidence* (WOE) hace referencia a la contribución del riesgo relativo. Cabe destacar que el WOE describe la relación entre una variable predictiva y una variable de objeto binaria.

El *Weight of Evidence* (WOE) de un atributo o rango de una variable se define como el logaritmo de la razón de la proporción de buenos sobre la proporción de malos donde los valores negativos grandes corresponden a un elevado riesgo mientras que los valores positivos indican bajo riesgo.

Para realizar este cálculo previamente se deben realizar las agrupaciones *binning* con el fin de maximizar la concentración de malos y buenos.

El WOE se define matemáticamente como:

$$WOE = ln\left(\frac{Distribución\ de\ casos\ positivos}{Distribución\ de\ casos\ negativos}\right) \tag{9}$$

Las distribuciones mencionadas en este cálculo refieren a la proporción de casos positivos y negativos por rango o agrupación en relación con el total de casos respectivamente para cada característica.

Consideraremos lo siguiente:

$$b_i = Número\ de\ casos\ negativos\ por\ grupo$$

$$g_i = Número\ de\ casos\ positivos\ por\ grupo$$

$$i = Número\ de\ grupos\ en\ los\ que\ se\ segmentó\ la\ variable$$

Por lo que,

$$Distribución\ de\ casos\ negativos = \frac{b_1}{(b_1 + b_2 + \cdots + b_{k+1})} \tag{10}$$

$$Distribución\ de\ casos\ positivos = \frac{g_1}{(g_1 + g_2 + \cdots + g_{k+1})} \tag{11}$$

En la siguiente tabla (3.5) se ilustra el cálculo para los diferentes k+1 grupos creados de una x variable.

Tabla 3.5 Cálculo del WOE por grupo.

Grupo	Casos positivos	Casos negativos	WOE
1 …	g_1 …	b_1 …	$\ln\left(\frac{\frac{g_1}{g_1+g_2+\cdots+g_{k+1}}}{\frac{b_1}{b_1+b_2+\cdots+b_{k+1}}}\right)$
k	g_k	b_k	$\ln\left(\frac{\frac{g_k}{g_1+g_2+\cdots+g_{k+1}}}{\frac{b_k}{b_1+b_2+\cdots+b_{k+1}}}\right)$
$k+1$	g_{k+1}	b_{k+1}	$\ln\left(\frac{\frac{g_{k+1}}{g_1+g_2+\cdots+g_{k+1}}}{\frac{b_{k+1}}{b_1+b_2+\cdots+b_{k+1}}}\right)$

Fuente: Zeng

Otro estadístico importante es el *Information Value* (IV) el cual mide el poder predictivo de una característica y la diferencia entre ambas distribuciones. Los valores de este estadístico siempre serán positivos.

Considerando los cálculos previos, el IV se define matemáticamente como:

$$IV = \sum_{i=1}^{k+1} \left(\frac{g_{k+1}}{g_1 + g_2 + \cdots + g_{k+1}} - \frac{b_{k+1}}{b_1 + b_2 + \cdots + b_{k+1}} \right) x\ WOE \tag{12}$$

Al observar la fórmula, el IV describe la suma ponderada de los valores individuales del WOE donde se incorpora como pesos la diferencia absoluta de las distribuciones, cabe destacar que el WOE es quien capta la diferencia relativa de las mismas.

Esta técnica tiene una regla empírica que describe el poder predictivo del IV (Siddiqi, 2006) la cual se ilustra en la Figura 3.6.

Figura 3.6 Cálculo del WOE por grupo.

Valor de IV	Fuerza predictiva
< 0.02	Impredecible
$0.02 \leq IV < 0.1$	Débil
$0.1 \leq IV < 0.3$	Mediana
$0.3 \leq IV$	Fuerte
$0.5 \leq IV$	Uso controlado por sobre predicción

Fuente: Siddiqi 2006

El *Weigth Of Evidence* y el *Information Value* juegan dos papeles distintos en el análisis de datos, mientras que el WOE describe la relación entre una variable predictiva y una variable de objeto binaria, el IV mide la fuerza de la relación[15]. Derivado de esto, es importante considerarlas para la creación de modelos de credit scoring.

[15] https://www.listendata.com/2015/03/weight-of-evidence-woe-and-information.html

3.2.1. Aplicación de WOE e IV

Recapitulando, con la metodología *binning* dividimos las variables independientes continuas en n grupos o rangos a los cuales se les calculó el WOE y el IV para determinar el tipo de relación y comportamiento de las variables independientes con respecto a la variable de interés.

Considerando las 14 variables restantes se realizaron estas metodologías y se pueden observar sus resultados en la Tabla 3.6.

Tabla 3.6 Cálculo de WOE e Information Value para variables.

Variable	Grupos	Rangos	WOE	IV	GINI
Último importe de pago	1	Último importe de pago < 5,000	1.114	0.383	31.445
	2	5,000 <= Último importe de pago < 9,000	0.451		
	3	9,000 <= Último importe de pago < 14,700	-0.040		
	4	14,700 <= Último importe de pago < 20,000	-0.554		
	5	Último importe de pago >= 20,000	-0.776		
Tasa de interés	1	Tasa de interés < 10.25	0.312	0.062	11.190
	2	10.25 <= Tasa de interés < 12	-0.116		
	3	Tasa de interés >= 12	-0.516		
Canal de contacto con el cliente	1	Canal de contacto con el cliente = 0	0.199	0.113	13.279
	2	Canal de contacto con el cliente = 1	-0.573		
Crédito Hipotecario	1	Crédito Hipotecario = 1	0.208	0.366	18.188
	2	Crédito Hipotecario = 0	-1.808		
Seguro Patrimonial	1	Seguro Patrimonial = 0	-0.287	0.123	16.673
	2	Seguro Patrimonial = 1	0.433		
Cuenta de vista	1	Cuenta de vista = 1	0.183	0.090	13.696
	2	Cuenta de vista = 0	-0.499		
Zona de oficina de originación	1	Zona de oficina de originación in [1,2]	-0.191	0.106	16.605
	2	Zona de oficina de originación = 3	0.025		
	3	Zona de oficina de originación in [4,5]	0.707		
Segmento del Cliente	1	Segmento del Cliente = 1	-0.255	0.217	19.275
	2	Segmento del Cliente = 2	0.381		
	3	Segmento del Cliente = 3	1.208		
Edad del Cliente	1	Edad del Cliente < 50	-0.064	0.014	6.320
	2	Edad del Cliente >= 50	0.221		
Comisión de Apertura	1	Comisión de Apertura < 1	0.266	0.023	6.010
	2	Comisión de Apertura >= 1	-0.085		
Comisión de Autorización	1	Comisión de Autorización < 0.1	-0.190	0.037	10.219
	2	Comisión de Autorización >= 0.1	0.193		
Segmento de vivienda	1	Segmento de vivienda = Media	0.084	0.258	25.440
	2	Segmento de vivienda = Residencial	-0.391		
	3	Segmento de vivienda = Tradicional	0.879		
	4	Segmento de vivienda = Residencial Plus	-0.836		
	5	Segmento de vivienda = Popular	1.386		
	6	Segmento de vivienda = Económica	-1.100		
Máximo en los ult. 6 meses de Saldo Medio en Vista	1	Max Saldo Medio < 10,000	0.232	0.032	9.459
	2	Max Saldo Medio >= 10,000	-0.139		
Límite máximo de crédito en alguna TDC	1	Max Límite de TDC < 75,000	0.149	0.050	11.299
	2	75,000 <= Max Límite de TDC < 200,000	-0.113		
	3	Max Saldo Medio >= 200,000	-0.418		

Fuente: Elaboración propia

De acuerdo con el criterio de Siddiqui las variables más fuerte son el último importe de pago y la variable de crédito hipotecario[16], las cuales se buscará obtener en el modelo de retención.

3.3. Modelo de Retención de Clientes

En esta sección se presentará la aplicación de un modelo de regresión logística aplicado a la gestión de retención de clientes de crédito hipotecario.

3.3.1. Regresión Logística

Históricamente el modelo de regresión logística ha sido el método estadístico más empleado para determinar la probabilidad de una variable objetivo en los modelos de credit scoring. Una de sus principales características es que es de fácil interpretación por lo que su explicación es más accesible a diferentes perfiles (Jiménez, 2000).

Pertenece al grupo de métodos de regresión para variables categóricas, que permiten la modelación de fenómenos que se dividen en clases. En este modelo la variable dependiente presenta dos categorías que indican la ocurrencia y la no ocurrencia del acontecimiento definido por la variable objetivo, en nuestro caso, el robo o no del crédito hipotecario por otra entidad financiera identificándose con los valores uno y cero respectivamente. Cabe destacar que las variables explicativas pueden ser de tipo intervalo o categóricas.

El modelo de regresión logística expresa la variable dependiente en términos de probabilidad, esto mediante el uso de la función logística, por lo que es necesario definir la variable dependiente de la siguiente manera:

$$Y = ln\left(\frac{P}{1-P}\right) \tag{13}$$

[16] Esta variable refiere a si el cliente ha tenido un crédito hipotecario previo al actual.

Considerando la fórmula de la regresión incluyendo este término, la expresión de la variable dependiente queda de la siguiente forma:

$$ln\left(\frac{P}{1-P}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-1} x_{k-1} + \beta_k x_k \tag{14}$$

De esta expresión, procedemos a despejar la función logaritmo,

$$\frac{P}{1-P} = exp\{\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-1} x_{k-1} + \beta_k x_k\} \tag{15}$$

$$\rightarrow\ P(Y=1) = \frac{exp\{\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-1} x_{k-1} + \beta_k x_k\}}{1 + exp\{\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-1} x_{k-1} + \beta_k x_k\}} = \tag{16}$$

$$= \frac{1}{(1 + exp\{\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-1} x_{k-1} + \beta_k x_k\})} \frac{1}{exp\{-\alpha - \beta_1 x_1 - \beta_2 x_2 - \cdots - \beta_{k-1} x_{k-1} - \beta_k x_k\}}$$

Por lo que simplificando la expresión tenemos:

$$P(Y=1) = \frac{1}{1 + exp\{-\alpha - \beta_1 x_1 - \beta_2 x_2 - \cdots - \beta_{k-1} x_{k-1} - \beta_k x_k\}} = P \tag{17}$$

Donde:

Y = Variable dependiente

x_i = Variables independientes

α = Coeficiente intercepto

β_i =Coeficientes de la pendiente

Derivado de esta fórmula, la interpretación de los coeficientes considerando su signo nos indica la dirección de la relación con la variable dependiente, es decir, nos describe si se trata de una correspondencia directa o inversa. El valor de las β_i no tiene una interpretación directa ya que solo representan el efecto que un cambio en x_i tiene sobre el resultado final.

3.3.1.1. Aplicación de Regresión Logística

Se aplicó una regresión logística a las 14 variables restantes, utilizando el método *Stepwise* que consiste en encontrar un conjunto de variables independientes que influyan significativamente en la variable dependiente, esto se puede lograr probando una variable independiente a la vez e incluyéndola en el modelo de regresión (Hand, 2002).

Las seis variables resultantes de la regresión se pueden observar en la tabla 3.7, así mismo se observa su coeficiente Beta para cada una.

Tabla 3.7 Resultados de Regresión Logística

Variable	Grupos	Rangos	WOE	Beta	IV	GINI
Último importe de pago	1	Último importe de pago < 5,000	1.114	-0.835	0.383	31.445
	2	5,000 <= Último importe de pago < 9,000	0.451			
	3	9,000 <= Último importe de pago < 14,700	-0.040			
	4	14,700 <= Último importe de pago < 20,000	-0.554			
	5	Último importe de pago >= 20,000	-0.776			
Canal de contacto con el cliente	1	Canal de contacto con el cliente = 0	0.199	-1.195	0.113	13.279
	2	Canal de contacto con el cliente = 1	-0.573			
Crédito Hipotecario	1	Crédito Hipotecario = 1	0.208	-0.924	0.366	18.188
	2	Crédito Hipotecario = 0	-1.808			
Cuenta de vista	1	Cuenta de vista = 1	0.183	-0.749	0.090	13.696
	2	Cuenta de vista = 0	-0.499			
Edad del Cliente	1	Edad del Cliente < 50	-0.064	-1.422	0.014	6.320
	2	Edad del Cliente >= 50	0.221			
Máximo en los ult. 6 meses de Saldo Medio en Vista	1	Max Saldo Medio < 10,000	0.232	-0.429	0.032	9.459
	2	Max Saldo Medio >= 10,000	-0.139			

Fuente: Elaboración propia

Estas variables representarían a través de diferentes iteraciones computacionales las características que mejor nos ayudan a describir nuestra variable objetivo.

3.3.2. Scorecard

El *credit scoring* es un sistema de calificación que permite a las entidades determinar la posibilidad de que un cliente sea bueno o malo y por tanto establecer estrategias comerciales que se pueden llevar a cabo dependiendo de la variable objetivo.

Estas calificaciones provienen de convertir los resultados de la regresión logística en una escala de puntuación asociada a unos odds. Estos odds[17] se definen como momios que miden la posibilidad de ocurrencia de un evento. Otra característica del *scorecard* es que debe ordenar monótonamente el *bad rate*.

El método para llegar a estos puntajes es aplicando la siguiente formula:

[17] Se define como el cociente entre la probabilidad de ocurrencia y la probabilidad de no ocurrencia.

$$Scorecard = -\left(\ln\left(\frac{Distribución\ de\ buenos}{Distribución\ de\ malos}\right) * \beta_i + \frac{\alpha}{n}\right) * factor + \frac{offset}{n} \quad (18)$$

Donde;

β_i = Cociente de la regresión para cada variable

α = Intercepto de la regresión

$factor$ = Puntos para doblar los odds[18].

n = Número de variables

$offset$ = Base de puntuación[19]

Cabe destacar en el cálculo anterior que el logaritmo de la razón de distribuciones de buenos entre la distribución de malos corresponde al WOE de las agrupaciones de cada variable.

Otro punto importante a destacar es el valor del offset (Siddiqi, 2006) nos indica que la base para cada modelo debe ser de acuerdo al criterio de negocio de cada desarrollador y este no debe afectar el poder predictivo del modelo.

Estas puntuaciones se otorgarán a cada agrupación por variable, de este modo el puntaje final para cada cliente será la suma de puntos que obtuvo para cada una de las variables.

3.3.2.1. Aplicación del cálculo de Scorecard

Una vez obtenidos los resultados de la regresión logística, se puede obtener la tabla de puntajes para las agrupaciones de cada variable.

[18] Por convención se utiliza un factor de 20 / ln (2).
[19] Mínimo de calificación para todos los clientes.

Tabla 3.8 Valores para Scorecard.

Intercepto	Factor	Offset	Número Variables
-3.8392	28.85	720	6

Fuente: Elaboración propia

Para la obtención de puntajes consideramos el intercepto obtenido de la regresión logística, el factor de convención, un offset de 720 puntos y el número de variables resultantes de la regresión logística.

De esta forma se procedió al cálculo y se obtuvieron los puntos para las agrupaciones de cada variable (Tabla 3.9).

Tabla 3.9 Scorecard de variables

Variable	Categoría	Población	%	Bad Rate	WoE	IV	β	Score Card
1	Último importe de pago < 5,000	20,593	22.7%	0.7%	1.1144	0.383	-0.835	165
	5,000 <= Último importe de pago < 9,000	23,525	26.0%	1.3%	0.4512			149
	9,000 <= Último importe de pago < 14,700	23,744	26.2%	2.2%	-0.0395			138
	14,700 <= Último importe de pago < 20,000	10,401	11.5%	3.6%	-0.5542			125
	Último importe de pago >= 20,000	12,314	13.6%	4.4%	-0.7761			120
2	Canal de contacto con el cliente = 0	73,145	80.8%	1.7%	0.1995	0.113	-1.195	145
	Canal de contacto con el cliente = 1	17,432	19.2%	3.6%	-0.5732			119
3	Crédito Hipotecario = 1	87,012	96.1%	1.7%	0.2083	0.366	-0.924	144
	Crédito Hipotecario = 0	3,565	3.9%	11.5%	-1.8079			90
4	Cuenta de vista = 1	71,727	79.2%	1.7%	0.1828	0.090	-0.749	142
	Cuenta de vista = 0	18,850	20.8%	3.4%	-0.4987			128
5	Max Saldo Medio < 10,000	37,789	41.7%	1.7%	0.2319	0.032	-0.429	141
	Max Saldo Medio >= 10,000	52,788	58.3%	2.4%	-0.1391			137
6	Edad del Cliente < 50	67,960	75.0%	2.2%	-0.0642	0.014	-1.422	136
	Edad del Cliente >= 50	22,617	25.0%	1.7%	0.2208			148

Fuente: Elaboración propia

Cabe señalar que este tipo de puntuaciones pueden también utilizarse como un método de control ya que segmenta a la población de acuerdo a la variable objetivo y pueden ayudar a plantear estrategias más orientadas.

Es común que la muestra utilizada para el desarrollo del modelo sea de años anteriores a la fecha en que se empezará a emplear el scorecard, es por ello que durante este lapso de tiempo pudiera ocurrir algún desplazamiento de las características o algún cambio en las

variables por lo que se requiere aplicar una serie de pruebas para determinar que esto no tenga lugar; estas pruebas se enfocan principalmente en:

- Poder predictivo – Prueba ROC y Prueba KS
- Estabilidad – PSI

Mismos procesos que se describirán a continuación.

3.3.3. Pruebas de Diferencias de Dos Poblaciones

3.3.3.1. Prueba Kolmogorov-Smirnov

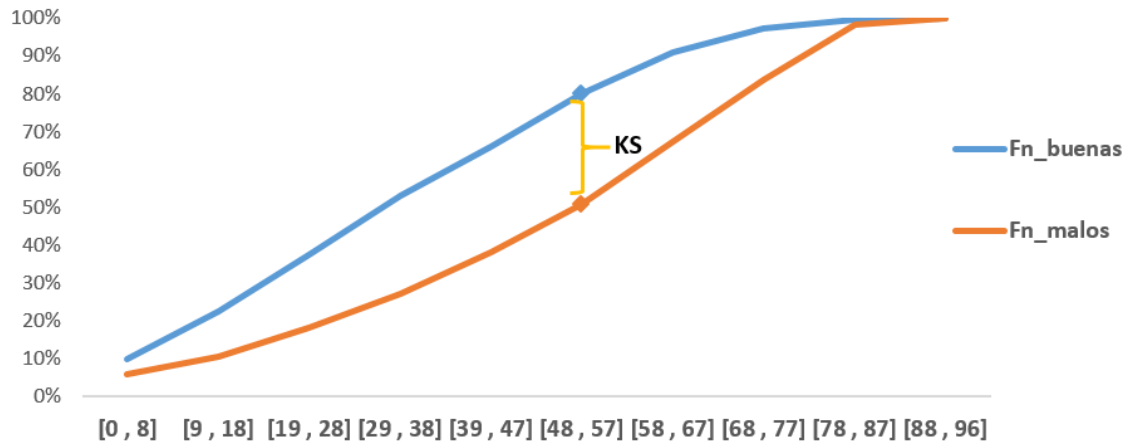
La prueba Kolmogorov – Smirnov (KS) es un contraste no paramétrico que tiene como objetivo determinar si la frecuencia de dos conjuntos de datos sigue la misma distribución alrededor de su media, en otras palabras, el KS es una prueba que se ajusta a la forma de los datos y se utiliza principalmente para comprobar si dos muestras distintas siguen la misma distribución.

La cual viene dada por la siguiente expresión,

$$KS = max_x|F_1(x) - F_2(x)| \tag{19}$$

Como se puede interpretar, esta fórmula nos ayuda a identificar la distancia más grande entre dos poblaciones muestras, en este caso las distribuciones corresponden a la población propensa a robo de hipoteca y la que no, en términos del score.

Figura 3.7 Gráfica KS



Fuente: Elaboración propia

La figura 3.7 muestra un ejemplo de la gráfica que podría formarse bajo las distribuciones acumuladas de buenos y malos, se señala también el KS como la máxima distancia entre ambas distribuciones.

3.3.3.2. Curva ROC

La curva de Característica Operativa del Receptor o bien llamada curva ROC (Receiver Operating Characteristic) constituye una de las formas más eficaces y utilizadas en la evaluación del rendimiento de un modelo orientado a la clasificación binaria.

Cabe destacar que es una representación bidimensional del rendimiento del modelo de clasificación. Esta se obtiene representando para cada uno de los posibles puntos de corte el porcentaje acumulado de casos positivos y de casos negativos.

La precisión del modelo se mide como el área que queda bajo la curva ROC, llamada AUC. El modelo “perfecto” tendrá un área igual a 1 y el peor un área igual a 0.5. Es importante mencionar que el AUC está relacionado con el índice GINI, ya que este es el doble del área entre la diagonal y la curva ROC.

Por lo que se tendría la siguiente expresión,

$$AUC = \frac{GINI + 1}{2} \tag{20}$$

Actualmente la convención para interpretar la curva ROC utiliza la siguiente escala:

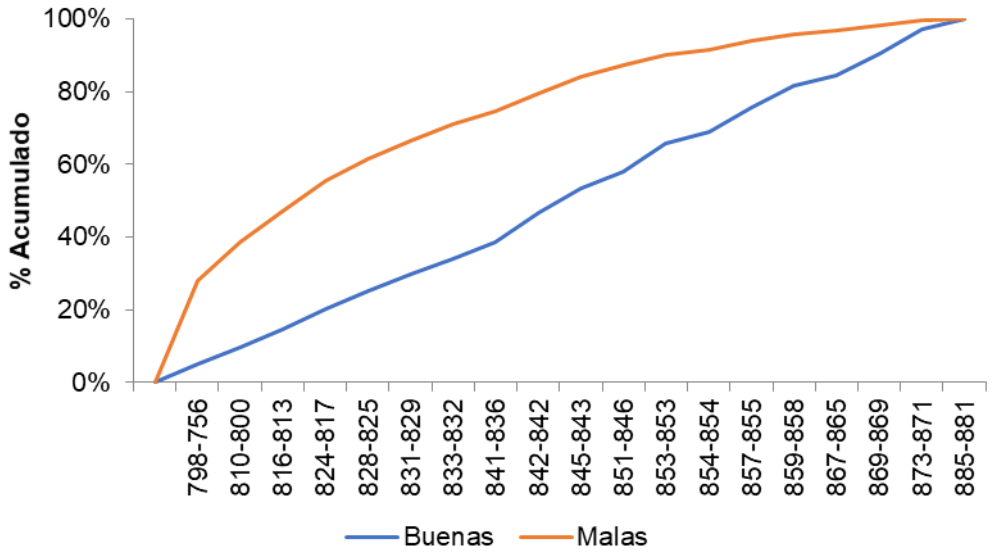
- ROC >= 1; la discriminación del modelo es perfecta.
- 0.8 <= ROC < 1; la discriminación del modelo es excelente.
- 0.7 <= ROC < 0.8; la discriminación del modelo es bueno.
- 0.5 < ROC < 0.7; la discriminación del modelo es pobre.
- 0.5 = ROC; la discriminación del modelo es nula.

Resultados de Evaluaciones

Aplicando los criterios descritos anteriormente podemos concluir que el modelo presenta una excelente separación entre buenos y malos, es decir, el KS de 37% nos indica la máxima distancia entre la población de clientes más propensos a robo de hipoteca contra los que no.

Además, al medir en conjunto la sensibilidad y la especificidad del modelo, el cual viene representado por la curva ROC, nos otorga un estadístico de 74% que de acuerdo a la interpretación anteriormente expuesta se trata de un modelo bueno.

Figura 3.8 Curva ROC



Fuente: Elaboración propia

Al observar los resultados de estas pruebas, ambas nos indican que es posible llevar a producción este modelo.

3.3.3.3. Population Stability Index

La prueba estadística PSI (*Population Stability Index*) sirve para medir la variación en la distribución de la población a lo largo del tiempo. Lo que realiza es una comparación de dos poblaciones mediante segmentos.

Su fórmula está dada por la siguiente expresión:

$$PSI = \sum \left((Población\ Actual - Población\ Esperada) * ln\left(\frac{Población\ Actual}{Población\ Esperada}\right) \right) \quad (21)$$

Donde los valores más pequeños indicarían que la población no presenta cambios significativos una vez hecho el desarrollo del modelo, ya que así se tendrían poblaciones estables, caso contrario si los valores fueran muy grandes significaría que no se trata de variables duraderas, es decir, la variable necesitaría una revisión a detalle y por lo tanto se excluiría del modelo.

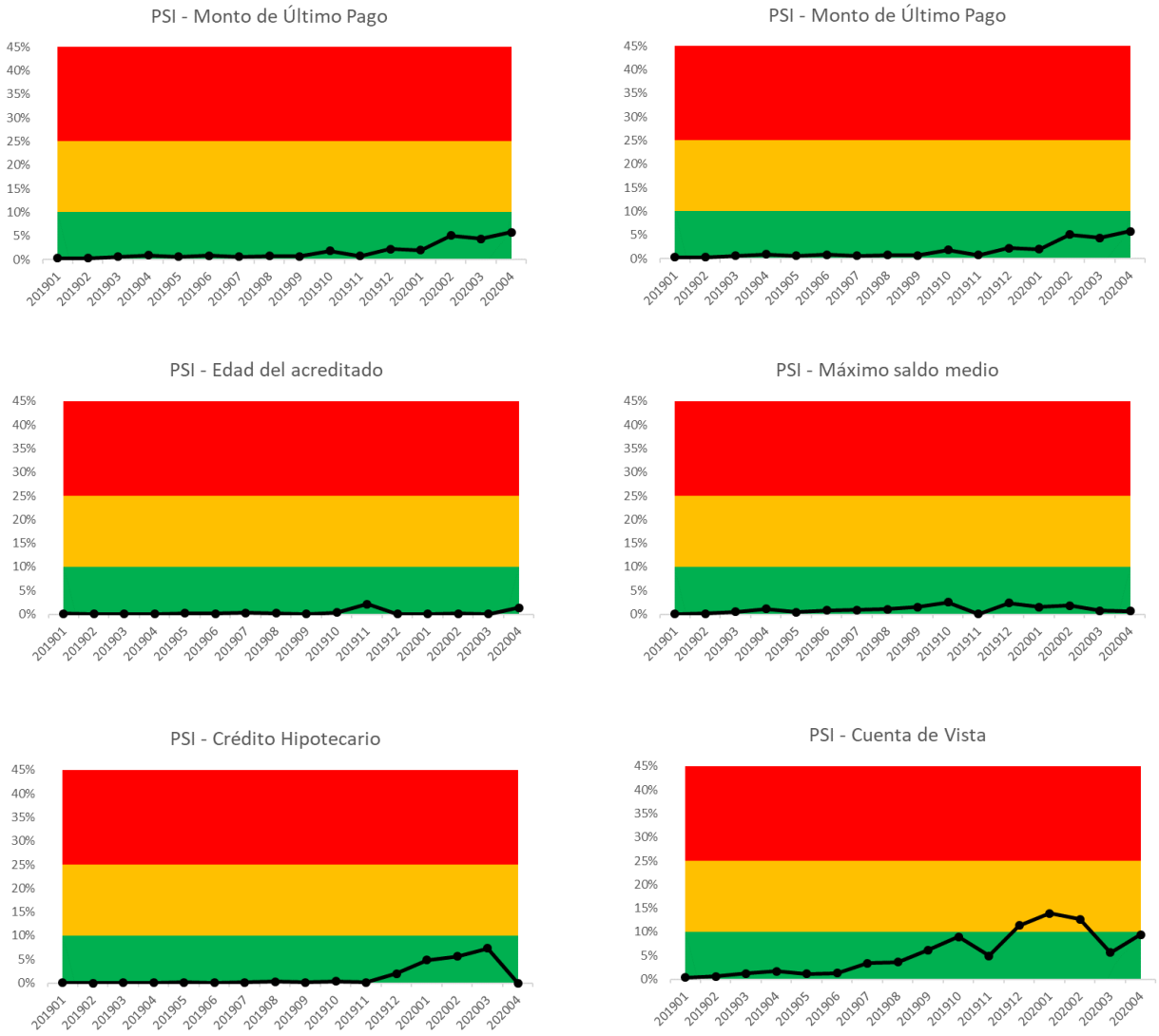
Actualmente la gráfica de esta prueba tiene 3 secciones donde se puede visualizar el tipo de cambio que tiene la variable que se está analizando. La sección en rojo nos indicaría cambios drásticos en la distribución de la variable por lo que nos daría pie a descartar inmediatamente la variable, la sección naranja nos señala un cambio moderado en la composición de la variable y en caso de agregarse al modelo se debería dar seguimiento especial a este tipo de variables; por último el escenario más favorable es cuando la variable no se desplaza fuera de la sección verde ya que nos indica que se trata de una variable cuya distribución permanece sin cambios importantes a través del tiempo.

Aplicación de la prueba PSI

Al aplicar este índice para las 6 variables que construyen el modelo nos arroja que 5 de ellas tienen una distribución que permanece con cambios poco importantes a través del tiempo; la variable restante tiene un cambio poco significativo y dado las anteriores pruebas se decidió conservar la variable por el poder predictivo que esta misma le aporta al modelo.

A continuación, se muestra el comportamiento de cada una de estas variables.

Figura 3.9 Pruebas de PSI



Fuente: Elaboración propia

4. Conclusiones

El modelo de retención de clientes fue realizado a través de una regresión logística en la que se obtuvo como resultado un modelo con 6 variables, así mismo se obtuvieron puntajes para una mejor interpretación de la clasificación de clientes. La selección de variables presentadas en esta tesis fue obtenida de un análisis univariado y multivariado empleando técnicas estadísticas en las que siempre se atendió la visión del negocio. Cabe destacar que la selección de variables consideró de forma preferencial la experiencia de los clientes en el sector hipotecario.

A través de las pruebas de KS y ROC se pudo comprobar la viabilidad del modelo ya que ambos métodos nos describen un modelo con alto poder predictivo. La estabilidad de las variables medidas a través del PSI nos otorga una visión inmediata de la posible estabilidad del modelo ya que al evaluarse con periodos fuera del desarrollo del mismo nos reitera su factibilidad.

Al menos en lo referente a los modelos de calificación crediticia suele utilizarse en la práctica la regresión logística con más frecuencia ya que considerablemente, obtenemos resultados más certeros y su cálculo no es complicado. Cuando se desarrolla este tipo de modelos dicotómicos se profundiza en el análisis estadístico y de negocio con lo que ayuda a una entidad financiera a la rápida toma de decisiones de forma eficiente.

Para finalizar, podemos afirmar que la propuesta de este modelo cumple con el objetivo de la tesis ya que logramos identificar las características de los clientes que son más propensos a migrar un crédito hipotecario a otra institución financiera, de esta manera, con base en este trabajo somos capaces de sostener que las técnicas de *machine learning* pueden proporcionar soluciones más eficientes a la estimación de calificaciones crediticias.

Por otra parte, la deserción de los clientes es un tema muy importante en la industria de servicios financieros. Este tipo de modelos es crucial en las campañas de marketing para crear estrategias de retención de clientes. A partir de estas estrategias, la entidad financiera

será responsable de medir la rentabilidad considerando el costo beneficio de la implementación.

Como se observa, la mayor probabilidad de deserción al crédito hipotecario son los clientes con menor tiempo con su crédito hipotecario, los clientes con experiencia crediticia y por último los clientes adversos al endeudamiento prolongado, de acuerdo con nuestra variable objetivo.

Con todo lo anterior, y en un contexto donde la gestión de clientes de crédito y la competencia por captar clientes buenos, se puede decir que las acciones que deriven de este tipo de estrategias otorgarán al acreditado una mayor gama de opciones que beneficien la originación o seguimiento de su crédito hipotecario, por ejemplo, opciones con tasa de interés o plazo.

Este trabajo se puede complementar con estudio de costo de oportunidad en función del uso de distintos modelos.

Las futuras líneas de investigación son: elaboración de técnicas que automaticen el seguimiento de los modelos con el fin de identificar preventivamente cambios de población que provoquen la errónea segmentación, también como complemento de este trabajo se puede realizar un estudio del costo de oportunidad en función del uso de distintos modelos.

Referencias

Alpaydin, E. (2020). *Introduction to machine learning. MIT press.*

Banco de México. (n.d.). *Divulgación: Sistema Financiero.*

Banco de México. (n.d.). *Servicios de Crédito*. Retrieved from http://www.anterior.banxico.org.mx/dyn/divulgacion/sistema-financiero/sistema-financiero.html#Serviciosdecredito

Banco de México. (n.d.). *Sistema Financiero*. Retrieved from http://educa.banxico.org.mx/banco_mexico_banca_central/sistema-financiero.html

Banco de México. (2005). *Definiciones Básicas de Riesgo. México.*

Banco de pagos Internacionales. (n.d.). *Comité de Supervisión Bancaria Basilea*. Retrieved from https://www.bis.org/publ/bcbs189_es.pdf

Bishop, C. M. (2006). *Pattern recognition and machine learning.* Springer.

Bourel, M. (2012). *Métodos de agregación de modelos y aplicaciones.*

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, 1145-1159.

Circular Unica de Bancos. (2007). *Diario Oficial de la Federación.* México.

Comisión Nacional Bancaria y de Valores. (n.d.). *Glosario de Términos.* Portafolio de Información.

Comisión Nacional Bancaria y de Valores. (n.d.). *Supervisión Bancaria Basilea*. Retrieved from https://www.cnbv.gob.mx/PrevencionDeLavadoDeDinero/Documents/Comit%C3%A9%20de%20Supervisi%C3%B3n%20Bancaria%20de%20Basilea.pdf

Crook, J. N. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, 1447-1465.

De Lara Haro, A. (2008). *Medición y control de riesgos financieros.* México: Limusa.

Elizondo, A., Altman, E., Gutiérrez, R., Gutiérrez, J., Mina, J., Segoviano, M., & Márquez, J. (2012). *Medición Integral del riesgo de crédito.* Ciudad de México: Limusa.

Freed, N. y. (1991). A linear programming approach to the discriminant problem. *Decision Sci*, 68-74.

Freed, N. y. (1991). Simple but powerful goal programming formulations for the discriminant problem. *European J*, 44-60.

Gilchrist, S., & Zakrajsek, E. (2005). *The importance of credit for macroeconomic activity: Identification through heterogeneity.*

Goldberg, D. E. (1988). *Genetic algorithms and machine learning.*

Hand, D. J. (2002). Superscorecards. *IMA Journal of Management Mathematics.*, 273-281.

Jiménez, M. V. (2000). La predicción del rendimiento académico: regresión lineal versus regresión logística. *Psicothema*, 248-525.

M., G. S. (2005). *Riesgo de crédito en México: Aplpicación del modelo Credit Metrics.* Universidad de las Américas, Puebla.

Marquez Diez-Canedo, J. (2009). *Una nueva visión del riesgo de crédito* . Limusa.

Menard, S. (2002). *Applied logistic regression analysis (Vol. 106).* Sage.

Michie, D. S. (1994). *Machine learning. Neural and Statistical Classification.*

Millán Solarte, J. C., & Caicedo Cerezo, E. (2018). Modelos para otorgamiento y seguimiento en la gestio´n de riesgo de cr´edito. *Revista de métodos cuantitativos para la economía y la empresa*, 23-41.

Molina, S. (2015). *El ciclo del crédito.* LID.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective. MIT press.*

Restrepo, L., & Julian González. (2007). From pearson to Spearman. *Revista Colombiana de Ciencias Pecuarias*, 183-192.

Siddiqi, N. (2006). *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring* . SAS Institute Inc.

Sierra Núñez, L. (2011). *La regulación internacional de Basilea y su evolución en el Sistema Financiero Mexicano*. UNAM.

Universidad Autónoma de Yucatán. (n.d.). *Sistema Financiero Mexicano*. Retrieved from https://www.contaduria.uady.mx/files/cuerpo-acad/caef/aief/sistema_financiero_mexicano.pdf

Ustáriz González, L. H. (2003). El comité de Basilea y la supervisión bancaria. *Pontificia Universidad Javeriana*, 431- 462.

Wod, I. J. (1985). Weight of evidence: A brief survey. *Bayesian statistics*, 249-270.

Zeng, G. (2014). A necessary condition for a good binning algorithm in credit scoring. *Applied Mathematical Sciences*, 3229-3242.

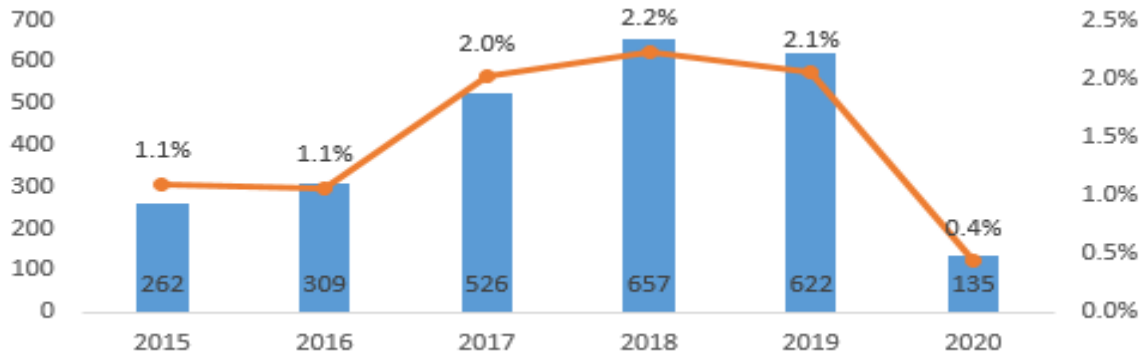
Apéndice A

Definición de Variable Objetivo

```
DATA MUESTRA;
        SET MUESTRA;
        IF ANIOS_CREDITO <= 4 AND  ANIOS_CREDITO >= 1 THEN
                DO;
                        IF ROBO_HIPOTECA = 1 THEN TARGET = 1;
                        ELSE TARGET = 0;
                        OUTPUT;
                END;
RUN;
```

Evolución de Fugas

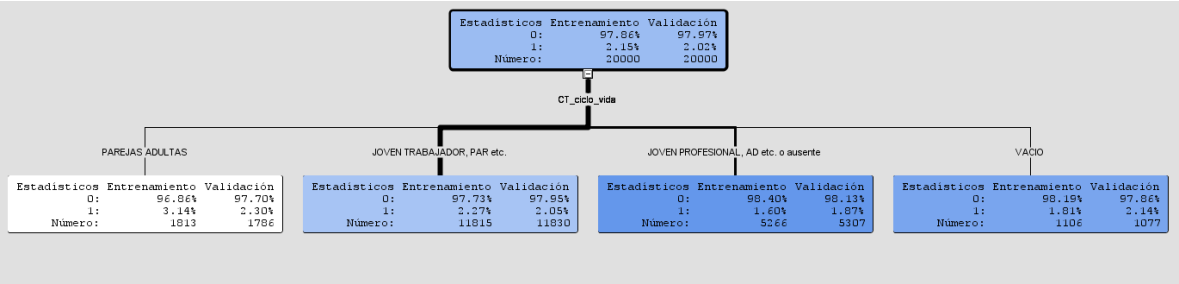
a) Distribución por cosecha: en las cosechas 2017 a 2019 se ha duplicado el número de casos



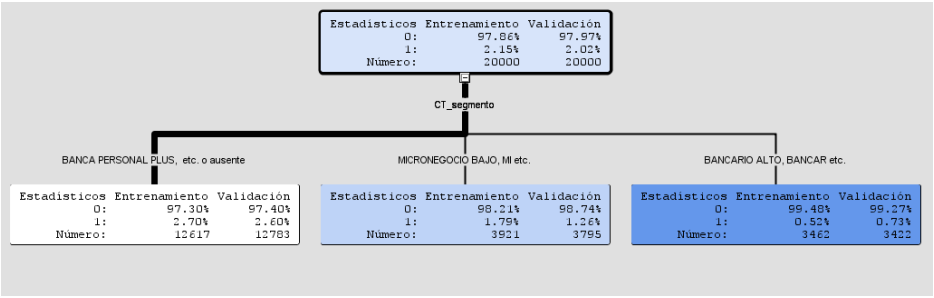
Apéndice B

Árboles de decision para Agrupación de Variables Categóricas

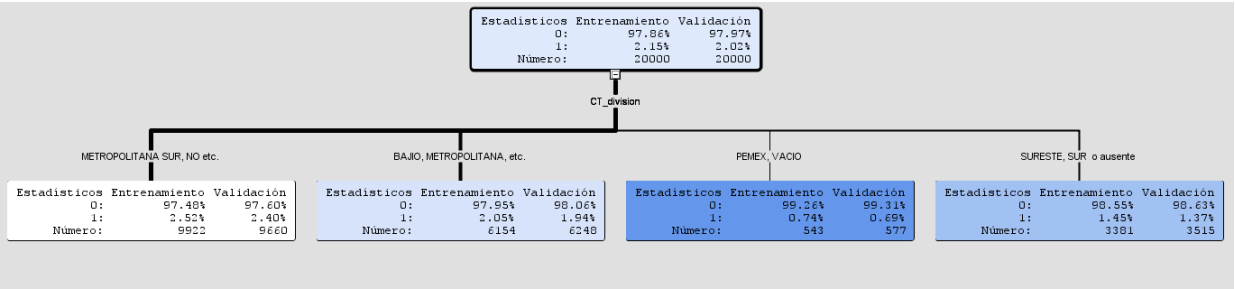
- Ciclo de vida del cliente



- Segmento del cliente



- División territorial



Análisis Univariado

Análisis Univariado de variables finales con tipo intervalo.

Variable	N	Tasa de Ausentes	Promedio	Desviación Estándar	Moda	Mediana	Mínimo	Pctl5	Pctl25	Pctl50	Pctl75	90th Pctl	Pctl95	Máximo
Comisión de Autorización	90,577	0%	0.1	0.1	-	0.2	-	-	-	0.2	0.2	0.3	0.3	0.5
Comisión de Apertura	90,577	0%	0.7	0.4	1.0	1.0	-	-	-	1.0	1.0	1.0	1.0	1.3
Edad del Cliente	90,577	0%	41.3	11.5	31.0	39.0	-	27.0	32.0	39.0	49.0	59.0	63.0	80.0
Tasa de interés	90,577	0%	10.4	1.0	10.3	10.3	1.0	9.4	10.0	10.3	10.9	11.2	12.2	14.0
Límite máximo de crédito en alguna TDC	90,577	0%	96,278.6	184,647.8	-	-	-	-	-	-	105,000.0	308,700.0	513,300.0	1,000,000.0
Máximo en los ult. 6 meses de Saldo Me	90,577	0%	146,233.8	460,512.3	-	16,836.8	-	-	2,136.8	16,836.8	78,199.4	286,524.4	648,119.3	3,500,295.8
Último importe de pago	90,577	0%	11,688.9	10,298.6	65,000.0	9,232.1	-	1,000.0	5,336.6	9,232.1	14,718.9	22,613.8	30,400.3	65,000.0

Prueba x^2 cuadrada

Se utiliza cuando se tiene una muestra de n individuos que se clasifican respecto a dos variables, preferentemente cualitativas (nominales, dicotómicas) y se desea conocer a partir de datos muestrales, si existe asociación de estas a nivel poblacional.

Hipótesis:

H_0: Existe poblacionalmente independencia entre las variables estudiadas (no existe asociación a nivel poblacional entre las variables estudiadas)

H_1: No existe poblacionalmente independencia (existe asociación a nivel poblacional entre las variables estudiadas)

Esta prueba se utiliza en la metodología V-Cramer.

Apéndice C

Matriz de correlación – Pearson

VARIABLES	GINI	INT_IM_ULT_PAGO	Variable 1	Variable 2	Variable 3	Variable 4	INT_max_IM_LIMITE	INT_TASA_DE_INTERES	INT_COMISION_DE_AUTORIZACION	Variable 5	INT_max_IM_SDO_MEDIO	INT_edad	INT_COMISION_POR_APERTURA	Variable 6	Variable 7	Variable 8	Variable 9	Variable 10
INT_IM_ULT_PAGO	31.45	1.00	0.86	0.85	0.86	0.78	0.29	-0.03	-0.04	0.27	0.31	0.16	-0.07	0.21	0.18	-0.18	-0.10	-0.07
Variable 1	29.07	0.86	1.00	0.87	0.99	0.89	0.31	0.02	-0.06	0.29	0.32	0.15	-0.03	0.19	0.22	-0.22	0.00	-0.09
Variable 2	28.03	0.85	0.87	1.00	0.88	0.77	0.25	-0.09	-0.04	0.27	0.27	0.09	-0.08	0.19	0.23	-0.15	0.06	-0.05
Variable 3	27.73	0.86	0.99	0.88	1.00	0.89	0.31	-0.09	-0.08	0.29	0.35	0.14	-0.05	0.20	0.22	-0.20	0.00	-0.07
Variable 4	26.83	0.78	0.89	0.77	0.89	1.00	0.32	-0.04	-0.07	0.29	0.36	0.18	-0.05	0.21	-0.12	-0.14	-0.04	-0.01
INT_max_IM_LIMITE	11.30	0.29	0.31	0.25	0.31	0.32	1.00	-0.02	-0.12	0.75	0.39	0.19	-0.01	0.49	-0.01	-0.15	-0.09	-0.04
INT_TASA_DE_INTERES	11.19	-0.03	0.02	-0.09	-0.09	-0.04	-0.02	1.00	0.20	0.02	-0.13	0.11	0.16	-0.04	-0.03	-0.17	-0.01	-0.14
INT_COMISION_DE_AUTORIZACION	10.22	-0.04	-0.06	-0.04	-0.08	-0.07	-0.12	0.20	1.00	-0.07	-0.08	-0.04	-0.35	0.00	-0.03	0.04	0.08	-0.03
Variable 5	9.61	0.27	0.29	0.27	0.29	0.29	0.75	0.02	-0.07	1.00	0.27	0.12	-0.01	0.43	0.02	-0.14	-0.05	-0.06
INT_max_IM_SDO_MEDIO	9.46	0.31	0.32	0.27	0.35	0.36	0.39	-0.13	-0.08	0.27	1.00	0.16	-0.02	0.30	-0.02	-0.12	-0.08	-0.04
INT_edad	6.32	0.16	0.15	0.09	0.14	0.18	0.19	0.11	-0.04	0.12	0.16	1.00	0.00	0.12	-0.08	-0.34	-0.32	-0.05
INT_COMISION_POR_APERTURA	6.01	-0.07	-0.03	-0.08	-0.05	-0.05	-0.01	0.16	-0.35	-0.01	-0.02	0.00	1.00	-0.30	0.03	-0.05	-0.05	-0.02
Variable 6	6.21	0.21	0.19	0.19	0.20	0.21	0.49	-0.04	0.00	0.43	0.30	0.12	-0.30	1.00	0.00	-0.11	-0.06	-0.04
Variable 7	5.65	0.18	0.22	0.23	0.22	-0.12	-0.01	-0.03	-0.03	0.02	-0.02	-0.08	0.03	0.00	1.00	-0.31	0.11	-0.26
Variable 8	5.59	-0.18	-0.22	-0.15	-0.20	-0.14	-0.15	-0.17	0.04	-0.14	-0.12	-0.34	-0.05	-0.11	-0.31	1.00	0.18	0.64
Variable 9	2.94	-0.10	0.00	0.06	0.00	-0.04	-0.09	-0.01	0.08	-0.05	-0.08	-0.32	-0.05	-0.06	0.11	0.18	1.00	0.09
Variable 10	2.51	-0.07	-0.09	-0.05	-0.07	-0.01	-0.04	-0.14	-0.03	-0.06	-0.04	-0.05	-0.02	-0.04	-0.26	0.64	0.09	1.00

Matriz de correlación – Spearman

VARIABLES	GINI	INT_IM_ULT_PAGO	Variable 1	Variable 2	Variable 3	Variable 4	INT_max_IM_LIMITE	INT_TASA_DE_INTERES	INT_COMISION_DE_AUTORIZACION	Variable 5	INT_max_IM_SDO_MEDIO	INT_edad	INT_COMISION_POR_APERTURA	Variable 6	Variable 7	Variable 8	Variable 9	Variable 10
INT_IM_ULT_PAGO	31.45	1.00	0.78	0.77	0.78	0.71	0.19	-0.08	-0.05	0.18	0.29	0.16	-0.09	0.20	0.25	-0.22	-0.04	-0.19
Variable 1	29.07	0.78	1.00	0.83	0.99	0.87	0.20	-0.02	-0.09	0.19	0.28	0.17	-0.04	0.18	0.34	-0.28	0.04	-0.25
Variable 2	28.03	0.77	0.83	1.00	0.84	0.72	0.14	-0.13	-0.06	0.15	0.18	0.08	-0.09	0.16	0.32	-0.16	0.13	-0.14
Variable 3	27.73	0.78	0.99	0.84	1.00	0.88	0.20	-0.13	-0.11	0.19	0.28	0.15	-0.06	0.18	0.33	-0.26	0.04	-0.23
Variable 4	26.83	0.71	0.87	0.72	0.88	1.00	0.22	-0.16	-0.11	0.20	0.30	0.21	-0.08	0.20	-0.06	-0.16	-0.02	-0.13
INT_max_IM_LIMITE	11.30	0.19	0.20	0.14	0.20	0.22	1.00	0.02	-0.14	0.93	0.54	0.16	-0.03	0.73	0.00	-0.19	-0.10	-0.17
INT_TASA_DE_INTERES	11.19	-0.08	-0.02	-0.13	-0.13	-0.16	0.02	1.00	0.16	0.03	-0.04	0.11	0.23	-0.04	0.08	-0.23	-0.03	-0.23
INT_COMISION_DE_AUTORIZACION	10.22	-0.05	-0.09	-0.06	-0.11	-0.11	-0.14	0.16	1.00	-0.12	-0.11	-0.04	-0.35	0.00	-0.02	0.03	0.08	0.00
Variable 5	9.61	0.18	0.19	0.15	0.19	0.20	0.93	0.03	-0.12	1.00	0.48	0.11	-0.04	0.71	0.02	-0.17	-0.07	-0.16
INT_max_IM_SDO_MEDIO	9.46	0.29	0.28	0.18	0.28	0.30	0.54	-0.04	-0.11	0.48	1.00	0.18	-0.07	0.65	0.00	-0.19	-0.11	-0.17
INT_edad	6.32	0.16	0.17	0.08	0.15	0.21	0.16	0.11	-0.04	0.11	0.18	1.00	0.00	0.12	-0.03	-0.37	-0.32	-0.31
INT_COMISION_POR_APERTURA	6.01	-0.09	-0.04	-0.09	-0.06	-0.08	-0.03	0.23	-0.35	-0.04	-0.07	0.00	1.00	-0.30	0.03	-0.03	-0.05	-0.01
Variable 6	6.21	0.20	0.18	0.16	0.18	0.20	0.73	-0.04	0.00	0.71	0.65	0.12	-0.30	1.00	0.00	-0.15	-0.06	-0.14
Variable 7	5.65	0.25	0.34	0.32	0.33	-0.06	0.00	0.08	-0.02	0.02	0.00	-0.03	0.03	0.00	1.00	-0.37	0.10	-0.34
Variable 8	5.59	-0.22	-0.28	-0.16	-0.26	-0.16	-0.19	-0.23	0.03	-0.17	-0.19	-0.37	-0.03	-0.15	-0.37	1.00	0.22	0.94
Variable 9	2.94	-0.04	0.04	0.13	0.04	-0.02	-0.10	-0.03	0.08	-0.07	-0.11	-0.32	-0.05	-0.06	0.10	0.22	1.00	0.20
Variable 10	2.51	-0.19	-0.25	-0.14	-0.23	-0.13	-0.17	-0.23	0.00	-0.16	-0.17	-0.31	-0.01	-0.14	-0.34	0.94	0.20	1.00

Matriz de correlación – V Cramer

VARIABLES	GINI	CT_SEGMENTO_VIVIENDA_SHF	Variable 1	Variable 2	CT_CRED_HIPO	Variable 3	CT_SEG_PYP	Variable 4	Variable 5	Variable 6	CT_VISTA_BBVA	Variable 7	Variable 8	Variable 9	Variable 10	Variable 11	CT_CANALES	Variable 12	Variable 13	Variable 14	Variable 15	Variable 16	Variable 17	Variable 18	Variable 19	Variable 20	Variable 21	Variable 22	Variable 23	Variable 24
CT_SEGMENTO_VIVIENDA_SH	25.44	1.00	0.78	0.50	0.08	0.08	0.08	0.05	0.07	0.09	0.07	0.07	0.07	0.07	0.07	0.07	0.18	0.12	0.05	0.08	0.09	0.13	0.12	0.12	0.09	0.11	0.07	0.21	0.10	0.14
Variable 1	24.53	0.78	1.00	0.49	0.08	0.08	0.08	0.04	0.07	0.09	0.07	0.07	0.07	0.07	0.07	0.07	0.18	0.13	0.05	0.07	0.08	0.13	0.12	0.12	0.09	0.11	0.07	0.21	0.10	0.14
Variable 2	19.28	0.50	0.49	1.00	0.04	0.04	0.04	0.04	0.02	0.03	0.09	0.09	0.09	0.09	0.09	0.09	0.13	0.08	0.04	0.09	0.07	0.16	0.17	0.17	0.16	0.15	0.05	0.12	0.06	0.16
CT_CRED_HIPO	18.19	0.08	0.08	0.04	1.00	1.00	0.20	0.01	0.17	0.16	0.39	0.39	0.40	0.40	0.40	0.40	0.10	0.09	0.12	0.03	0.03	0.21	0.18	0.18	0.18	0.16	0.02	0.07	0.03	0.17
Variable 3	18.19	0.08	0.08	0.04	1.00	1.00	0.20	0.01	0.17	0.16	0.39	0.39	0.40	0.40	0.40	0.40	0.10	0.09	0.12	0.03	0.03	0.21	0.18	0.18	0.18	0.16	0.02	0.07	0.03	0.17
CT_SEG_PYP	16.67	0.08	0.08	0.04	0.20	0.20	1.00	0.05	0.89	0.83	0.26	0.26	0.26	0.26	0.26	0.26	0.09	0.06	0.17	0.04	0.02	0.26	0.25	0.25	0.25	0.25	0.01	0.13	0.06	0.21
Variable 4	16.61	0.05	0.04	0.04	0.01	0.01	0.05	1.00	0.06	0.07	0.03	0.03	0.02	0.02	0.02	0.02	0.02	0.05	0.03	0.06	0.31	0.04	0.04	0.04	0.04	0.04	0.02	0.23	0.05	0.03
Variable 5	16.56	0.07	0.07	0.02	0.17	0.17	0.89	0.06	1.00	0.69	0.22	0.22	0.22	0.22	0.22	0.22	0.07	0.06	0.15	0.03	0.02	0.20	0.20	0.20	0.19	0.19	0.01	0.09	0.05	0.15
Variable 6	15.06	0.09	0.09	0.03	0.16	0.16	0.83	0.07	0.69	1.00	0.22	0.22	0.22	0.22	0.22	0.22	0.07	0.06	0.10	0.05	0.02	0.19	0.20	0.20	0.19	0.19	0.01	0.08	0.05	0.15
CT_VISTA_BBVA	13.70	0.07	0.07	0.09	0.39	0.39	0.26	0.03	0.22	0.22	1.00	1.00	0.94	0.94	0.94	0.93	0.24	0.04	0.31	0.06	0.03	0.52	0.40	0.40	0.39	0.40	0.01	0.13	0.04	0.40
Variable 7	13.70	0.07	0.07	0.09	0.39	0.39	0.26	0.03	0.22	0.22	1.00	1.00	0.94	0.94	0.94	0.93	0.24	0.04	0.31	0.06	0.03	0.52	0.40	0.40	0.39	0.40	0.01	0.13	0.04	0.40
Variable 8	13.42	0.07	0.07	0.09	0.40	0.40	0.26	0.02	0.22	0.22	0.94	0.94	1.00	1.00	1.00	0.99	0.25	0.04	0.30	0.06	0.03	0.53	0.45	0.45	0.44	0.39	0.01	0.13	0.05	0.41
Variable 9	13.39	0.07	0.07	0.09	0.40	0.40	0.26	0.02	0.22	0.22	0.94	0.94	1.00	1.00	1.00	0.99	0.25	0.04	0.30	0.06	0.03	0.53	0.45	0.45	0.44	0.39	0.01	0.13	0.05	0.41
Variable 10	13.39	0.07	0.07	0.09	0.40	0.40	0.26	0.02	0.22	0.22	0.94	0.94	1.00	1.00	1.00	0.99	0.25	0.04	0.30	0.06	0.03	0.53	0.45	0.45	0.44	0.39	0.01	0.13	0.05	0.41
Variable 11	13.37	0.07	0.07	0.09	0.40	0.40	0.26	0.02	0.22	0.22	0.93	0.93	0.99	0.99	0.99	1.00	0.23	0.04	0.30	0.06	0.03	0.52	0.45	0.45	0.44	0.39	0.01	0.13	0.05	0.41
CT_CANALES	13.28	0.18	0.18	0.13	0.10	0.10	0.09	0.02	0.07	0.07	0.24	0.24	0.25	0.25	0.25	0.23	1.00	0.08	0.08	0.21	0.08	0.28	0.26	0.26	0.24	0.24	0.09	0.16	0.06	0.27
Variable 12	10.68	0.12	0.13	0.08	0.09	0.09	0.06	0.05	0.06	0.06	0.04	0.04	0.04	0.04	0.04	0.04	0.08	1.00	0.04	0.07	0.11	0.06	0.06	0.06	0.05	0.06	0.17	0.10	0.06	0.06
Variable 13	9.48	0.05	0.05	0.04	0.12	0.12	0.17	0.03	0.15	0.10	0.31	0.31	0.30	0.30	0.30	0.30	0.08	0.04	1.00	0.11	0.03	0.35	0.29	0.29	0.30	0.29	0.03	0.09	0.04	0.27
Variable 14	9.11	0.08	0.07	0.09	0.03	0.03	0.04	0.06	0.03	0.05	0.06	0.06	0.06	0.06	0.06	0.06	0.21	0.07	0.11	1.00	0.06	0.10	0.08	0.08	0.08	0.09	0.04	0.08	0.04	0.11
Variable 15	8.99	0.09	0.08	0.07	0.03	0.03	0.02	0.31	0.02	0.02	0.03	0.03	0.03	0.03	0.03	0.03	0.08	0.11	0.03	0.06	1.00	0.06	0.06	0.06	0.05	0.06	0.03	0.46	0.05	0.08
Variable 16	8.77	0.13	0.13	0.16	0.21	0.21	0.26	0.04	0.20	0.19	0.52	0.52	0.53	0.53	0.53	0.52	0.28	0.06	0.35	0.10	0.06	1.00	0.71	0.71	0.70	0.73	0.01	0.26	0.09	0.79
Variable 17	7.67	0.12	0.12	0.17	0.18	0.18	0.25	0.04	0.20	0.20	0.40	0.40	0.45	0.45	0.45	0.45	0.26	0.06	0.29	0.08	0.06	0.71	1.00	1.00	0.98	0.85	0.02	0.27	0.09	0.46
Variable 18	7.67	0.12	0.12	0.17	0.18	0.18	0.25	0.04	0.20	0.20	0.40	0.40	0.45	0.45	0.45	0.45	0.26	0.06	0.29	0.08	0.06	0.71	1.00	1.00	0.98	0.85	0.02	0.27	0.09	0.46
Variable 19	7.66	0.09	0.09	0.16	0.18	0.18	0.25	0.04	0.19	0.19	0.39	0.39	0.44	0.44	0.44	0.44	0.24	0.05	0.30	0.08	0.05	0.70	0.98	0.98	1.00	0.83	0.01	0.25	0.07	0.45
Variable 20	6.64	0.11	0.11	0.15	0.16	0.16	0.25	0.04	0.19	0.19	0.40	0.40	0.39	0.39	0.39	0.39	0.24	0.06	0.29	0.09	0.06	0.73	0.85	0.85	0.83	1.00	0.02	0.26	0.07	0.43
Variable 21	6.43	0.07	0.07	0.05	0.02	0.02	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.09	0.17	0.03	0.04	0.03	0.01	0.02	0.02	0.01	0.02	1.00	0.03	0.02	0.02
Variable 22	5.88	0.21	0.21	0.12	0.07	0.07	0.13	0.23	0.09	0.08	0.13	0.13	0.13	0.13	0.13	0.13	0.16	0.10	0.09	0.08	0.46	0.26	0.27	0.27	0.25	0.26	0.03	1.00	0.21	0.21
Variable 23	5.74	0.10	0.10	0.06	0.03	0.03	0.06	0.05	0.05	0.05	0.04	0.04	0.05	0.05	0.05	0.05	0.06	0.06	0.04	0.04	0.05	0.09	0.09	0.09	0.07	0.07	0.02	0.21	1.00	0.08
Variable 24	5.42	0.14	0.14	0.16	0.17	0.17	0.21	0.03	0.15	0.15	0.40	0.40	0.41	0.41	0.41	0.41	0.27	0.06	0.27	0.11	0.08	0.79	0.46	0.46	0.45	0.43	0.02	0.21	0.08	1.00

Distribución de Tasa de malos en puntuaciones.