



UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

**EVALUACIÓN DEL COMPORTAMIENTO DE LOS CLIENTES DE UN
BANCO DADO LOS SHOCKS ECONÓMICOS DE LA PANDEMIA EN CHILE**

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN ECONOMÍA APLICADA

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERA CIVIL INDUSTRIAL

LISSETTE ALEJANDRA LUENGO MADRID

PROFESOR GUÍA:
JOSÉ MIGUEL CRUZ GONZÁLEZ

MIEMBROS DE LA COMISIÓN:
GUILLERMO SÁEZ SÁEZ
RICHARD WEBER HAAS

SANTIAGO DE CHILE
2023

**RESUMEN DE LA TESIS PARA OPTAR AL GRADO DE
MAGÍSTER EN ECONOMÍA APLICADA Y MEMORIA PARA
OPTAR AL TÍTULO DE INGENIERA CIVIL INDUSTRIAL
POR: LISSETTE ALEJANDRA LUENGO MADRID
FECHA: 2023
PROFESOR GUÍA: JOSÉ MIGUEL CRUZ GONZÁLEZ**

EVALUACIÓN DEL COMPORTAMIENTO DE LOS CLIENTES DE UN BANCO DADO LOS SHOCKS ECONÓMICOS DE LA PANDEMIA EN CHILE

Durante los últimos tres años Chile ha vivido una importante recesión económica que surge tras la crisis social y que se agrava con la llegada de la pandemia al país.

Como medida para frenar la baja liquidez y la contracción del mercado el gobierno impulsa una serie de medidas orientadas a inyectar recursos a los sectores más vulnerables del país, es decir a la clase media y baja. Adicionalmente, el Congreso impulsa el retiro de fondos desde el sistema de las administradoras de fondos de pensiones (AFPs).

Estas medidas impactaron en el mercado bancario. Aumentó la capacidad de pago de los usuarios y se redujo el nivel de riesgo ya que contaban con dinero para cancelar sus deudas, sin embargo, dicho comportamiento en muchos casos no se llevó a cabo. Hubo una proporción de personas a las que se denominarán Can't pay, que si cancelaron sus deudas y que previo a la inyección de recursos no poseían dinero para hacerlo. Mientras que hubo otras, las Won't pay que, bajo las mismas condiciones, decidieron no pagar los compromisos bancarios adquiridos, aún contando con el dinero.

Este trabajo busca establecer relaciones e identificar las variables que permitan predecir y/o explicar el comportamiento y las diferencias existentes entre clientes Can't pay, y los Won't pay.

Con este propósito se empleó como metodología la minería de datos, junto con herramientas de análisis que permitan comprender el incumplimiento de los clientes a través de diversas variables.

Como resultado obtenido a partir del modelo final del puntaje de comportamiento que este estudio presenta, se determinó la existencia de 14 variables que tienen incidencia en que un cliente pertenezca al segmento Can't pay y/o al Won't pay. Entre los más importante destacan la antigüedad, edad y sexo del cliente.

ABSTRACT

During the last three years, Chile has experienced a significant economic recession that arose after the social crisis and that worsened with the arrival of the pandemic in the country.

To curb low liquidity and market contraction, the government promotes a series of measures aimed at injecting resources into the country's most vulnerable sectors. The lower and middle class is targeted. Additionally, Congress pushed for the withdrawal of funds from the pension fund management system.

These measures had an impact on the banking market. The payment capacity of users increased, and the level of risk was reduced of debts, since they had money to pay off their debt. However, in many cases, this was not carried out. There was a proportion of people whom we will call "Can't pay", who did cancel their debts and who, prior to the injection of resources, did not have the money to do so. While there were others that "Won't pay", under the same conditions, decided not to pay the acquired bank commitments, even with the money.

This work seeks to establish relationships and identify the variables that allow predicting and/or explaining the behavior and differences between Can't pay customers and Won't pay customers.

For this purpose, data mining was used as a methodology for the analysis of information, together with analysis tools that allow understanding the default of clients through various variables.

As a result, obtained from the final model of the behavior score that this study presents, 14 variables are determined. These have an impact on whether a client belongs to the Can't pay and/or Won't pay segment. The most important factors are the client's seniority at the bank, age, and gender.

*A mis padres Alejandro y Andrea,
a mis hermanos Alejandra y Raúl,
a mis sobrinos Isabel, Paz y Gaspar,
por su apoyo, cariño y amor
por creer en mí, ¡lo logramos!*

AGRADECIMIENTOS

Quiero partir agradeciendo a mi familia, por siempre ser mi sustento, mi guarida y mi apoyo. Mi padre, quien creyó en mí y me insistió tantas veces para que diera lo mejor de mí, mi madre, quien estuvo ahí cuando ya no podía más para darme aliento y consejos, mi hermana, con quien muchas veces hice cábalas para pasar ramos, a mi hermano por haber sido el lector número uno de mis trabajos y por haberme dedicado palabras cuando lo necesitaba, a mis tres sobrinos gracias por alegrar mis días, por su amor y por ser quienes son, mi motivación del día a día, los amo. Gracias a la Chabe, mi tía, por haberme apoyado en esta etapa, sin ti no podría haberlo logrado; y a mi tía Sari por haberme ayudado cuando lo necesitaba. Agradezco a todos los integrantes de mi familia que alguna vez me dedicaron una palabra de aliento y cómo olvidar a la gente que partió en este trayecto universitario, gracias Leli, tío Lucho y tío Tato, quienes en vida siempre me apoyaron y creyeron en mí.

Gracias a mi segunda familia, los Dornhofer, a Michi y Andreas, sin ellos no hubiera estudiado ingeniería, gracias por siempre ser parte de mi vida, a Luis y Theo por ser mis hermanos pequeños y recibirme cada vez que los visitó.

Un agradecimiento especial a mi pololo Harm, quien ha tenido una paciencia especial en este proceso, me ha escuchado y acompañado siempre que lo necesito, junto con brindarme ayuda si puede. Gracias también a su familia, por su apoyo en este proceso, en especial a Marcel y Janne, que me ayudaron a pensar muchas veces cómo avanzar con la tesis junto a un café.

El paso por la universidad no es algo fácil y agradezco a todas las personas que se cruzaron en mi camino estos años, sin ellos la experiencia no hubiera sido lo que fue, a la archi, que sin ellos no hubiera sobrevivido los primeros años, en especial a las chiquillas, quienes han estado ahí desde el día uno. A mis amigos de Industrias, de los intercambios y de plan común. También gracias a mis amigos de Iquique.

Agradezco también a los grupos de los que fui parte en la universidad y a la gente con la que ahí compartí: el World Class, Redes Beauchef y el CEIN 2020.

Un gran agradecimiento a mi profesor guía, José Miguel el cual me dedicó tantas de sus horas para tener reuniones y me dió muchos consejos, gracias por confiar en mí. Gracias al equipo de Ripley, a Joshua e Ignacio, por siempre estar dispuestos a ayudarme. Gracias a mis profesores miembros de la comisión por haberme dado parte de su tiempo para este proceso.

Gracias a todas las personas que estuvieron en mi vida en esta etapa.

TABLA DE CONTENIDO

1.- Introducción	1
2.- Descripción del estudio	3
2.1 Exposición del tema	3
2.2 Objetivos	5
3.- Marco teórico	7
3.1 Proceso KDD	7
3.2 Riesgo de créditos	8
3.2.3. Puntaje de comportamiento	8
3.3 Regresión Logística	12
3.4 Árboles de decisión	13
3.4.1 Método CHAID	14
3.5 Índice de estabilidad de la población	14
3.6 Conceptos Relevantes	15
3.6.1 Deuda	15
3.6.2 Probabilidad de Default	15
3.6.3 Clientes Won't pay	15
3.6.4 Clientes Can't Pay	15
4.- Metodología	16
4.1 Datos	16
4.2 Selección de los Datos	16
4.2.1 Verificación de los datos	17
4.2.2 Composición de los clientes	17
4.3 Pre-procesamiento de Datos	18
4.4. Transformación de datos	19
4.5 Análisis descriptivo	19
4.5.1 Clientes buenos	22
4.5.2 Clientes malos	27
5. Discusión	31
5.1 Riesgo según características de los clientes	31
5.1.1 Variable de la edad de los clientes	32
5.1.2 Variable del saldo de los clientes	34

5.1.3 Variable antigüedad de los clientes en el banco	37
5.1.4 Variable del sexo de los clientes	39
6. Resultados	42
6.1 Base de datos	42
6.2 Variables seleccionadas	43
6.2.1 Variable 1: Tendencia de los días de mora de un cliente	43
6.2.2 Variable 2: Grupo socioeconómico de los clientes	44
6.2.3 Variable 3: Edad	45
6.3 Resultados del modelo	47
7. Conclusiones y recomendaciones	56
Bibliografía	59
Anexos	61
Anexo A: Índice PSI de los clientes buenos	61
Anexo B: Descripción de categorías de variable producto más riesgoso de la cartera del cliente	65
Anexo C: Índice PSI de los clientes malos	66
Anexo D: Gráficas de riesgo para la variable subsegmento de clientes y producto más riesgoso de la cartera	70
Anexo E: Análisis de resultados de variables del modelo	74
Anexo F: Resultados del modelo, puntaje de todos los datos	85
Anexo G: Resultados finales	98

ÍNDICE DE TABLAS

<i>Tabla N°1: “Segmentación de los clientes”</i>	<u>18</u>
<i>Tabla N°2: “Tabla de WOE e IV para variable 1”</i>	<u>44</u>
<i>Tabla N°3: “Tabla de WOE e IV para variable 2”</i>	<u>45</u>
<i>Tabla N°4: “Tabla de WOE e IV para variable 3”</i>	<u>46</u>
<i>Tabla N°5: “Tabla de IV para todas las variables”</i>	<u>48</u>
<i>Tabla N°6: “Resultado modelo de regresión”</i>	<u>49</u>
<i>Tabla N°7: “Tabla de puntaje de comportamientos”</i>	<u>50</u>

ÍNDICE DE ILUSTRACIONES

Gráfico N°1: “Distribución de clientes por clasificación...”	4
Gráfico N°2: “Distribución de grupos socioeconómica en Chile”	4
Gráfico N°3: “Comparación data del banco con la de CMF”	17
Gráfico N°4: “Distribución de clientes por sexo”	20
Gráfico N°5: “Distribución de clientes por edad”	20
Gráfico N°6: “Distribución de clientes por actividades que realiza”	21
Gráfico N°7: “Distribución de los clientes de la data según días...”	22
Gráfico N°8: “Composición de los clientes buenos según...”	23
Gráfico N°9: “Composición de los cliente buenos según edad”	23
Gráfico N°10: “Composición de los clientes buenos según saldo”	24
Gráfico N°11: “Composición de los clientes buenos según producto”	25
Gráfico N°12: “Composición de los clientes buenos según subsegmentos”	26
Gráfico N°13: “Composición de los clientes buenos según su sexo”	26
Gráfico N°14: “Composición promedio de clientes malos según antigüedad”	27
Gráfico N°15: “Composición promedio de clientes malos según edad”	28
Gráfico N°16: “Composición de clientes malos según saldo adeudado”	28
Gráfico N°17: “Composición de los clientes malos según producto...”	29
Gráfico N°18 “Composición de clientes malos según segmentos de clientes”	30
Gráfico N°19: “Composición promedio de los clientes malos según el sexo”	30
Gráfico N°20: “PD a través del tiempo en categoría de edades”	32
Gráfico N°21: “Tasa de incumplimiento a 12 meses según edades”	33
Gráfico N°22: “Pérdida porcentual del segmento con diferentes...”	34
Gráfico N°23: “PD a través del tiempo de saldo”	35
Gráfico N°24: “Porcentaje de incumplimiento a 12 meses de saldo”	36
Gráfico N°25: “Pérdida porcentual del segmento con diferentes deudas...”	36
Gráfico N°26: “PD a través del tiempo de antigüedad”	37
Gráfico N°27: “Porcentaje de incumplimiento a 12 meses de antigüedad”	38
Gráfico N°28: “Pérdida porcentual del segmento según la ...”	39
Gráfico N°29: “PD a través del tiempo para la variable sexo”	40
Gráfico N°30: “Porcentaje de incumplimiento a 12 meses según sexo”	40
Gráfico N°31: “Pérdida porcentual dependiendo del sexo que posea”	41
Gráfico N°32:” WOE de la tendencia de los días de mora de un cliente”	43
Gráfico N°33:” WOE de los grupos socioeconómicos de los clientes”	44
Gráfico N°34: “WOE de edad de los clientes”	45
Gráfico N°35: “Resultados del modelo - Distribución y probabilidad”	49
Gráfico N°36: “Resultados del modelo - KS y ROC”	50
Gráfico N°37: “Grupos socioeconómicos de los clientes”	51
Gráfico N°38: “Producto más riesgoso que posee el cliente”	52
Gráfico N°39: “Sexo de los clientes”	52
Gráfico N°40: “Actividad que realiza el cliente”	53

Gráfico N°41: "Número de instituciones en las que el cliente tiene..."	54
Gráfico N°42: "Antigüedad en años del cliente en el banco"	54
Gráfico N°43: "Edad de los clientes"	55

CAPÍTULO 1

1.- Introducción

El 18 de octubre del año 2019 en Chile tuvo lugar un estallido social sin precedentes. Trajo consigo una marcada caída en la tasa de crecimiento de la economía, pasando del 4,0% en 2018 al 0,8%. Esta caída se produjo a causa del menor dinamismo de la demanda interna y externa, junto a la caída del consumo (Banco Central de Chile, 2019). A su vez, el año 2020 se asentó en el mundo la pandemia del COVID-19, cuyo impacto afectó el día a día de la mayor parte de los habitantes del planeta.

Un estudio de la CEPAL determinó que Chile vivió una de las peores crisis económicas de su historia y lo atribuyó principalmente a los factores de: cierre de fronteras, cuarentenas prolongadas y a la suspensión de ciertas actividades económicas. Todas, iniciativas impulsadas con objeto de frenar el avance de contagios pero que causaron una notoria disminución en la demanda doméstica, causando que los niveles de producción cayeran a un 5,8% y el índice de desempleo aumentará a 10,9% (CEPAL, 2021).

A raíz de esto, el Gobierno decidió implementar diversos programas de ayuda orientados al segmento más vulnerable del país y a la clase media. Se llevó a cabo la entrega de bonos a pequeñas y medianas empresas (Pymes), como, asimismo, a personas naturales. Se entregaron cajas de alimentos y se creó un Ingreso Familiar de Emergencia (IFE) dirigidos a las familias inscritas en el Registro Social de Hogares y que pertenezcan al 90% más vulnerable de la población. Este último, se materializó mediante una transferencia monetaria directa y transitoria que a junio del año 2021 ya alcanzaba los 14,5 millones de beneficiarios, es decir, un 96% del total de personas que se encuentran en el Registro Social de Hogares, según lo indicado en el informe de finanzas del Gobierno (Dirección de Presupuesto - Gobierno de Chile, 2021).

Además, el Congreso impulsó una ley, donde se permitía el retiro excepcional, desde los fondos de las Administradoras de Fondos de Pensiones (AFP). Se realizaron en total tres retiros con una cifra máxima que ascendía al 10% del total de fondos de cada persona (Chile Atiende, 2022).

Dichas medidas generaron una mayor disponibilidad de liquidez, disminución en la adquisición de créditos de consumo a comienzos del año 2021 y acotados índices de morosidad (Central).

En cuanto al endeudamiento, en junio del año 2021 el nivel de deuda representativa de los deudores bancarios en Chile fue de \$1,87 millones, cifra que disminuyó en comparación al año anterior. (CMF, 2021).

El presente estudio se focalizará en datos de personas que adquirieron créditos previos al año 2020 en un banco de Retail. Se emplearán herramientas de análisis económicos preestablecidas, contemplando como factor de amenaza para cada usuario su riesgo crediticio.

Los clientes a considerar con categoría de malos serán los registros en donde los días de deuda superen los 90 días, ya que estos según lo estipulado por el Compendio de Normas Contables son considerados en estado de incumplimiento (Superintendencia de Bancos e Instituciones Financieras - SBIF., 2014).

Luego, se distinguirán dos grupos más. Por un lado, los que quieren pagar sus deudas, pero no pueden hacerlo por problemas económicos (*Can't pay*) y un segundo grupo compuesto por individuos que escogen no hacerlo, pese a tener la disponibilidad de dinero (*Won't pay*). A raíz de esto último, surgen las principales interrogantes que aborda este trabajo. ¿Cómo es posible identificar a los clientes *Won't pay* y *Can't pay* que mostraron diferencias en su comportamiento de pago posterior al shock económico vivido? ¿A qué se debe la diferencia de comportamiento entre estos individuos que bajo el análisis de riesgo previo eran iguales?

CAPÍTULO 2

2.- Descripción del estudio

2.1 Exposición del tema

En la presente tesis se da a conocer el trabajo realizado en conjunto con Ripley Chile S.A, bajo el contexto de riesgo crediticio en los clientes del banco.

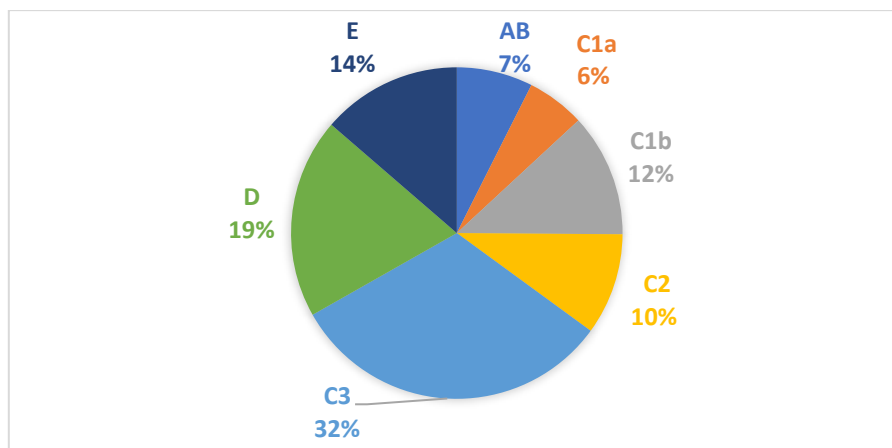
Ripley en Chile posee importante presencia en el rubro de Retail, asimismo en el sector Inmobiliario y en el Bancario. Es en este último en el que se enmarca esta tesis. Banco Ripley otorga servicios bancarios y financieros orientados a satisfacer las necesidades de personas pertenecientes en su mayoría, al segmento de ingresos medios de la población chilena.

Banco Ripley ofrece créditos de consumo a sus clientes, los que principalmente se otorgan a través de Tarjeta Ripley, la cual puede ser usada como medio de pago y permite a la vez, el acceso a otros diversos beneficios.

Es importante señalar que Banco Ripley es regulado por la CMF (Comisión para el Mercado Financiero) y se acoge a las Normas de la NIIF (Normas Internacionales de Información Financiera). En Chile, la Ley general de bancos en su Artículo 15, faculta a la CMF para impartir normas contables de aplicación general a las entidades sujetas a su fiscalización.

Para el desarrollo de esta tesis se considerará que la totalidad de los clientes de Ripley vivieron un shock positivo en sus ingresos con las medidas paliativas que otorgó el gobierno. Esto, dado que la cartera del banco es la siguiente según los grupos socioeconómicos de los clientes:

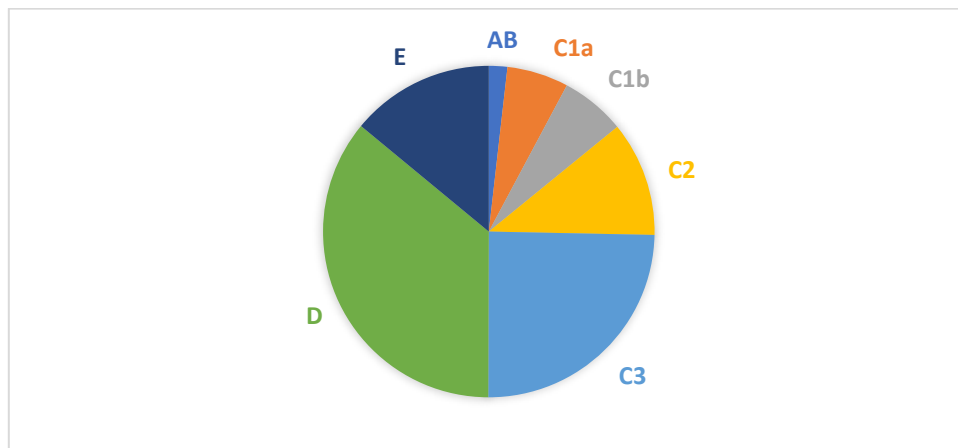
Gráfico N°1: “Distribución de clientes por clasificación de segmento socioeconómico”



Fuente: Elaboración propia

En el gráfico N°1, se observa que cerca de un 65% de los clientes pertenecen a los segmentos E, D y C3. Según el preinforme de ingreso familiar de emergencia, un 73,8% pertenece al tramo del 40% de hogares de mayor vulnerabilidad que recibió al menos una ayuda económica (Ministerio de desarrollo social y familia, 2020. Comparando esta información con las distribuciones, según la Asociación de investigaciones de mercado y opinión pública (AIM) de los grupos sociales en Chile, se tiene lo siguiente:

Gráfico N°2: “Distribución de grupos socioeconómica en Chile”



Fuente: Elaboración propia

El gráfico N°2 en el que se aprecia que el 40% de hogares de mayor vulnerabilidad estarían ubicados en el sector E y D de la población, donde pertenece gran parte de los clientes del banco Ripley pertenece. Junto con esto, es importante señalar que el bono clase media, tuvo impacto en los sectores C3, C2, C1b, por lo que es posible decir que la mayoría de los clientes de banco Ripley obtuvo al menos un shock económico positivo en pandemia. Esto, permite dar un sustento a la hipótesis de esta tesis.

Es de crucial importancia indicar cuáles serán los sucesos que se considerarán para el análisis. Como primer punto se considerará el estallido social ocurrido el 18 de octubre de 2019, momento en el que empezaron a mostrarse los primeros impactos económicos de la crisis. El segundo punto a considerar serán las ayudas económicas dispuestas por el gobierno, en específico las que corresponden al retiro del 10%, procesos que se llevaron a cabo el 30 de julio de 2020, el 10 de diciembre de 2020 y el 28 de abril de 2021 respectivamente. Por último, se considera el primer IFE, como periodo de inicio de ayudas económicas con fecha 01 de mayo de 2020.

2.2 Objetivos

El objetivo principal de esta tesis es evaluar si existe alguna razón que justifique las diferencias en el comportamiento entre individuos que presentan el mismo riesgo crediticio, que no están al día con sus pagos y que son, por ende, considerados morosos. En específico, se distinguirán dos grupos de clientes morosos, aquellos que no han podido pagar por deficiencias económicas y los que no tienen incentivos para hacer el pago.

Como parte de los objetivos específicos se encuentran; analizar el comportamiento de pago en los individuos que poseen productos de créditos en un banco de Retail, identificar variables que expliquen una diferencia de comportamiento al momento de pagar las cuotas, además de identificar segmentos relevantes, donde existan comportamientos diferenciados de riesgo.

2.3 Alcances

Esta tesis está principalmente enfocada en el análisis de los datos de deuda de los clientes de la empresa, a través de la información recaudada por el Banco sobre los clientes activos que presentan al cierre del mes.

El trabajo contempla, por una parte, la identificación de variables relevantes para establecer diferencias en el comportamiento. Por otra parte, establece un modelo de puntaje de comportamiento con las respectivas utilidades que podría brindar a la industria, al sector bancario.

Es así como el resultado final de esta tesis busca responder las interrogantes planteadas, junto con presentar un modelo útil para cuantificar las diferencias entre los tipos de clientes.

CAPÍTULO 3

3.- Marco teórico

3.1 Proceso KDD

El proceso KDD, es un proceso no trivial, que alude sus siglas al término en inglés Knowledge Discovery in Databases, donde se busca el descubrimiento de información útil dentro de una base de datos, identificando patrones que sean potencialmente útiles. Su principal etapa es la minería de datos, donde se aplican algoritmos específicos para modelar los datos. Está junto a las otras etapas son descritas a continuación:

a.- Selección de datos: En esta etapa del proceso se seleccionan los datos desde las diferentes fuentes de información que se encuentran disponibles. Existen tres tipos de fuentes: internas, que corresponden a bases de datos propias de la entidad producto de los procesos que lleva a cabo en ella; externas, que corresponden a datos provenientes de otras entidades o del entorno; y las variables generadas, que corresponden a indicadores definidos en función de los datos de ambas fuentes antes mencionadas. Esta etapa garantiza la efectividad de los modelos, ya que se requiere poseer información de calidad para que el modelo sea representativo del fenómeno de estudio.

b.- Pre-procesamiento de Datos: Por lo general, los datos poseen mucho ruido e inconsistencia, los que deben ser tratados para evitar los posibles efectos en el modelo final. En esta etapa, se desarrolla este proceso, analizando las variables que se tienen, para ver si todas son explicativas del fenómeno de estudio. Se preocupa del tratamiento de datos faltantes y en blanco, datos inconsistentes y fuera de rango mediante diversas estrategias.

c.- Transformación de datos: En esta etapa, se realizan todas las transformaciones a los datos que permitan cumplir con los requerimientos establecidos en el modelo, en específico a las variables con las que se decidió trabajar. Los métodos comúnmente utilizados son normalización de variables, creación de variables binarias para representar variables categóricas y discretización de variables continuas.

d. Minería de Datos: En esta etapa, se ajustan los modelos estadísticos con el fin de obtener patrones válidos, nuevos, potencialmente útiles y comprensibles. Lo primero es seleccionar los atributos, dejando fuera aquellas variables que no poseen poder discriminante desde el punto de vista estadístico. El paso siguiente es dividir la base de datos en dos subconjuntos; el primero, es la base de entrenamiento que se utiliza para calibrar el modelo y el segundo, es la base de testeo que se utiliza para validar el modelo y probar su capacidad.

e. Interpretación y evaluación de resultados: Esta etapa final, consiste en analizar los resultados obtenidos con el fin de determinar si el modelo utilizado logra explicar de manera satisfactoria el fenómeno en estudio.

3.2 Riesgo de créditos

La palabra crédito tiene su origen en el latín *creditum*, que significa “haber confiado”, ya que la persona que presta confía en que el otro retribuirá lo prestado.

El riesgo de crédito nació como solución a las necesidades de las entidades crediticias ante la amenaza de quedar sin solvencia, dándoles una predicción certera a la probabilidad de que clientes fallaran en sus pagos y no cumplieran con los términos preestablecidos entre las partes, es decir, que tuvieran morosidad con la entidad (Siddiqi, 2006).

La analogía que se conoce de este método se realiza con una fotografía, en donde se intenta hacer coincidir una imagen del consumidor de su información económica y financiera en la solicitud, con una fotografía de su posible estado después de un período fijo de préstamo. Esto se realiza a través de un análisis histórico de los consumidores y así se predice el comportamiento futuro (Thomas, 2009).

Para el autor Siddiqi, el mejor método que se puede utilizar para predecir la falta de pago de la morosidad es la scorecards, este método incluye una metodología de puntuación que evalúa el riesgo de deserción de los clientes. Consta de un grupo de características del tipo demográficas, de relación existente e historial crediticio, lo que se acerca a ser un perfil de riesgo completo para el individuo, que, mediante la aplicación de técnicas estadísticas, determina la predicción que brinda una probabilidad de que el cliente sea buen o mal pagador. De esta manera se decide si se otorga o no el crédito a un solicitante (Siddiqi, 2006).

A continuación, se describe el proceso de puntaje de comportamiento según el autor Thomas.

3.2.3. Puntaje de comportamiento

El nombre *puntaje de comportamiento* viene de la traducción del inglés de “*Behavioural Scoring*” y su objetivo es generar un puntaje de clasificación según las variables que estén presentes en el método. En este procedimiento la información del comportamiento de los clientes debe estar disponible y se debe tener en cuenta el tiempo en el que se considerarán las características de comportamiento. Por lo general, se usa un año como el periodo de rendimiento, es decir el período que se aplica el modelo en primera instancia y un año posterior para evaluar los resultados, observándose así, si el

contraste de resultados con el año anterior sigue un comportamiento lógico. Por este motivo, se requerirán datos de al menos dos años.

Con este fin, los clientes son clasificados en buenos y malos clientes. Los buenos hacen referencia a las personas que cumplen con sus obligaciones a tiempo, mientras que los malos son los clientes que incumplen o hacen los pagos con atrasos. Según Banco Ripley, esta clasificación de malos clientes, corresponde a aquellos que han acumulado más de 90 días de mora.

Se dice que la probabilidad de ser bueno o malo depende de las características del individuo, denominadas como X , donde $X = (X_1, X_2, \dots, X_m)$, es decir, X abarca todas las características que se tienen del cliente. Estas probabilidades se pueden expresar como:

$$Pr\{\text{buenos} | \text{data } x\} = p(B|x); \quad Pr\{\text{malos} | \text{data } x\} = p(M|x)$$

Ecuación N°1: "Probabilidad de buenos y malos"

donde,

$$p(B|x) + p(M|x) = 1$$

Ecuación N°2: "Suma de probabilidad de buenos y malos"

Para obtener las probabilidades de ser un buen o mal cliente, se puede usar la probabilidad de ocurrencia de un evento, conocido en inglés como odds, esto es: tomar la posibilidad de que ocurra el evento y dividirlo por la posibilidad de que no ocurra.

En este caso, el evento puede ser un resultado bueno o malo, por lo que para calcular la probabilidad de ser bueno se realiza la división de buenos por malos y así inversamente si se quiere obtener la probabilidad de ser malo, lo que se expresa como:

$$o(B|x) \triangleq \frac{p(B|x)}{p(M|x)} \text{ y } o(M|x) \triangleq \frac{p(M|x)}{p(B|x)} = \frac{1}{o(B|x)}$$

Ecuación N°3: "Probabilidad de ocurrencia del evento buenos y malos"

El resultado, se expresa como la razón de posibles clientes buenos/malos versus los malos/buenos respectivamente.

Para obtener los scores de comportamiento se realizan los siguientes pasos que se describen en los subtítulos a continuación:

3.2.3.1 Muestra de datos

Lo primero es considerar la información que se tiene de los clientes y cuántos son los registros con los que se cuenta. Si la base de datos es de una escala de un millón lo recomendable es sacar un muestreo de ella.

Para sacar los muestreos existen dos formas, la primera es una forma aleatoria, lo que generará la misma proporción de buenos y malos de la base de datos completa. Mientras que la segunda es muestreo estratificado, donde lo que se intenta es aumentar la selección de malos en el muestreo, para que sea una proporción similar a la de los buenos.

3.2.3.2 Validación de los datos y limpieza

En este paso el objetivo principal es corroborar que las entradas de información sean válidas. Es decir, que los valores de las variables sean valores reales y lógicos según cual sea el caso, de no ser así se debe limpiar la base de datos.

3.2.3.3 Segmentación

Posterior se debe tomar la decisión, si se desea, segmentar la población y construir diferentes scores de comportamiento para ellos, buscando que la segmentación aumente el poder predictivo.

3.2.3.4 Datos de entrenamiento y testeo

Con este paso la idea es generar una división de la base de datos en una base de entrenamiento o “training data” cuyo fin es usar estos datos para entrenar el modelo y una de prueba o “testing data” donde los datos serán usados para comprobar que el modelo funcione. Por lo general la proporción de la base de datos total que es usada para crear ambas divisiones es 80:20 correspondientemente.

3.2.3.5 Elección de variables

Para este proceso, se considera primeramente el análisis univariado y multivariado, el que se detalla más adelante. Con estos análisis, se buscan las variables más discriminantes, por lo que la cantidad de variables que son ser llevadas al modelo se reduce.

Para realizar este procedimiento de manera más sencilla existen softwares que han desarrollado funciones o paquetes que realizan los puntajes de comportamiento. En ellos, se deben completar los pasos hasta el punto anterior, para luego utilizar la función, que generará los pasos siguientes.

En el software *Rstudio*, que es el empleado para este estudio, la herramienta a utilizar es el paquete *scorecard*.

3.2.3.6 Clusterización de una variable

Luego de tener elegidas las variables, ahora es importante discernir si se desea hacer una clusterización manual de ella o a través del algoritmo. Por lo general, las variables que son categóricas se pueden agregar de forma manual, ya sea según una agrupación de ellas o según el orden original.

3.2.3.7 Chi-cuadrado y valor de la información de variables

En esta etapa se busca seleccionar cuales de las variables que se están ingresando al modelo tienen implicancia en el comportamiento de la variable dependiente.

Con chi-cuadrado se buscan las variables con más diferenciación entre el ratio de buenos y malos, por lo que se buscan los valores más altos.

Para obtener el chi-cuadrado, se realiza un test de hipótesis, tomando la hipótesis de que el ratio de buenos y malos es el mismo para todos los grupos, si el valor obtenido es bajo el punto de corte, la hipótesis nula se rechaza.

Por otro lado, para buscar el valor de la información (VI), que mide el poder predictivo de las variables, se tiene que utilizar la siguiente ecuación en base a un intervalo i sobre los N intervalos que presenta cada variable analizada:

$$VI = \sum_{i=1}^N \left(\frac{Buenos_i}{Total\ de\ buenos} - \frac{Malos_i}{Total\ de\ malos} \right) \ln \left(\frac{Buenos_i}{Total\ de\ buenos} \div \frac{Malos_i}{Total\ de\ malos} \right)$$

Ecuación N°4: “Valor de la información”

3.2.3.8 Pesos de la evidencia

Una vez más se quiere determinar la importancia de las variables en el modelo, para mejorar la predicción que tiene, a través de los pesos de la evidencia. Son conocidos como WOE por sus siglas en inglés, proveniente de “*Weight of Evidence*”.

Mide la fuerza de cada atributo, o atributos agrupados, al separar cuentas buenas y malas. Se buscará las puntuaciones más altas, ya que cuanto mayor sea la puntuación de WOE, más fuerte es la relación entre la categoría y la variable objetivo.

El WOE está definido como:

$$WOE = \log \left(\frac{Buenos_i}{Total\ de\ buenos} \div \frac{Malos_i}{Total\ de\ malos} \right)$$

Ecuación N°5: “Pesos de la evidencia”

Una vez seguidos estos siete pasos, las variables con las que se trabajarán ya están seleccionadas, por lo tanto, se puede pasar al paso final de creación de puntaje a través de un modelo. Por lo general, para predecir variables categóricas, la técnica escogida para obtener el puntaje de comportamiento es la regresión logística.

3.3 Regresión Logística

Es un modelo que utiliza un conjunto de características para predecir la probabilidad de un resultado específico. Se utiliza comúnmente para desarrollar scorecards, ya que la variable a pronosticar es categórica. En este caso estudia la relación entre las características del individuo, como variables explicativas y si este es o no un buen cliente pagador, variable objetivo. El resultado obtenido de la regresión será la probabilidad de pertenecer al grupo de los buenos o al grupo de los malos pagadores (Siddiq, 2006).

Se utiliza este tipo de regresiones, ya que hay variables dicotómicas, y se puede crear una variable ficticia de la siguiente estructura:

$$x_i = \{1 \text{ Cuando ocurre el evento en } i ; 0 \text{ Cuando no ocurre el evento en } i$$

Ecuación N°6: "Variable binaria"

Donde el subíndice i , representa cada observación que se posee en la base de datos.

La ecuación de la regresión logística está constituida por una función de transformación Logit. La función Logit convierte la relación no lineal entre las variables predictoras y la variable de respuesta en una relación lineal que puede ser modelada por una regresión logística. En este caso esta desarrollada según lo siguiente:

$$\text{Logit}(P_i) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

Ecuación N°7: "Ecuación modelo Logit"

Donde:

p = probabilidad de ocurrencia del evento

x = variables de entrada

β_0 = intersección de la línea de regresión

β_k = parámetros estimados

Para finalizar se debe asignar puntuaciones a cada variable agrupada y la suma de todos los puntajes para cada atributo proporciona el puntaje final.

Es importante señalar que la tendencia y la diferencia entre pesos de evidencia en los atributos agrupados afectará los puntos asignados utilizando este enfoque. La fórmula que permite obtener los puntajes es la siguiente:

$$Puntaje = \sum_{j,i=1}^{k,n} \left(- \left(WOE_j * \beta_i + \frac{a}{n} \right) * factor + \frac{offset}{n} \right)$$

Ecuación N°8: “Cálculo de puntaje”

Donde:

WOE = Peso de la evidencia para cada atributo agrupado

β = Coeficiente de regresión para cada característica

a = Término de intersección de la regresión logística

n = Número de características

k = Número de atributos en cada característica

Factor: Ponderador que tiene relación con la probabilidad de ocurrencia del evento

Offset: Factor para ajustar el valor dentro de la escala

El offset y factor son utilizados para poder ajustar los datos dentro de una escala de puntajes, de tal manera que se pueda distinguir entre buenos y malos en un rango. Por lo general los rangos se encuentran entre 300 a 850 puntos.

Una vez obtenido los puntajes por cada variable, ahora es posible obtener el puntaje para cada individuo y ver como estos se distribuyen en los datos.

3.4 Árboles de decisión

Los árboles de decisión son una técnica de clasificación que parte de un nodo, y es de gran ayuda cuando se presenta una gran cantidad de datos que incluye un gran número de variables. Es un proceso de iteración automática, por lo general basado en un algoritmo recursivo de clasificación, en donde las ramas de los árboles son representaciones de conjunto de decisiones y de cada decisión se genera un nuevo conjunto de decisiones, formando así grupos homogéneos respecto a la variable que se desea discriminar. Este proceso se repite hasta que se alcanza un punto final, donde no es posible realizar más iteraciones que cumplan con los criterios (Pineda, 2009).

3.4.1 Método CHAID

CHAID, es una abreviatura a la frase en inglés *Chi-square Automated Interaction Detector* y es un tipo de árbol de decisión, el cual se utilizará para realizar esta tesis. Es un análisis basado en una variable crítica que puede poseer dos o más categorías y que busca determinar el segmento que posee esta variable.

Se hace una diferenciación entre la variable dependiente e independiente, las que deben ser cuidadosamente elegidas previo a la implementación. Los subconjuntos son elaborados a partir de la variable independiente y va iterando sobre el número de éstas, haciendo que la variable dependiente esté mejor explicada por las variables independientes resultantes. En este método hay dos tipos de análisis, el univariado y multivariado, los que son descritos a continuación (Vanesa Berlanga Silvente, 2013).

3.4.1.1 Análisis univariado

Este análisis permite analizar la relación de una variable dependiente con otra independiente. El proceso consiste en elegir variables que presentan un mayor poder explicativos de la variable dependiente, obteniendo como resultado el valor Chi-cuadrado de cada una de estas interacciones y va generando nodos hasta llegar al punto más preciso de relación, seleccionando en las iteraciones los valores más altos (Flora M Díaz-Pérez, 2006).

3.4.1.2 Análisis multivariado

Este análisis permite analizar simultáneamente conjunto de datos y variables, en función de un objetivo. Las variables que son analizadas, por lo general provienen de la última etapa del análisis univariado, la principal diferencia con este análisis es la cantidad de variables independientes que se toman simultáneamente al realizar la comparación y que estas pueden interactuar para analizar el comportamiento conjunto de ellas y cómo se relacionan entre sí.

3.5 Índice de estabilidad de la población

El índice de estabilidad de la población, conocido como PSI, es una métrica que mide cuánto ha cambiado una variable en cuanto a su distribución entre dos muestras a lo largo del tiempo, su fórmula es la siguiente:

$$PSI = \sum \left((\% \text{ Muestra 1} - \% \text{ Muestra 2}) \times \ln \left(\frac{\% \text{ Muestra 1}}{\% \text{ Muestra 2}} \right) \right)$$

Ecuación N°9: "Índice de estabilidad de la población"

Donde el valor se evalúa según lo siguiente:

- <0.1: Cambio muy leve
- 0.1-0.2: Algún cambio menor
- > 0.2: Cambio significativo

3.6 Conceptos Relevantes

Dentro de los términos que se suele hablar en el riesgo crediticio son las deudas que pueden tener los clientes y la probabilidad de incumplimiento que estos tienen, los siguientes términos serán descritos según la definición de Siddiqi.

3.6.1 Deuda

Para este estudio, se considera que deuda corresponde a todo préstamo de dinero que la institución bancaria otorgó a una persona natural que lo solicitó, para fines de consumo.

3.6.2 Probabilidad de Default

El default de una persona hace mención de cuando un deudor no ha pagado sus deudas a los acreedores, por lo que en este caso se evalúa la probabilidad de que un solicitante o cliente se atrase 90 días seguidos en sus pagos durante los próximos 12 meses.

3.6.3 Clientes Won't pay

Se definirá como clientes *Won't pay* a aquellos que, teniendo las condiciones económicas para realizar el pago de sus deudas, prefieren no pagar, es decir, no quieren pagar. Este grupo será identificado como el grupo que previo al shock económico contaban con deuda impaga, y que posterior al mismo, continuaron sin pagarlas.

3.6.4 Clientes Can't Pay

Se definirá como clientes *Can't pay* a aquellos que, dada sus condiciones económicas no pueden realizar el pago de sus deudas, por no tener los recursos necesarios para aquello. Este grupo de clientes se identificará ya que previo al shock económico presentaban deuda y no la estaban pagando, pero con las ayudas económicas pudieron realizar este pago.

CAPÍTULO 4

4.- Metodología

4.1 Datos

Para este estudio, se cuenta con tres tipos de bases de datos, las que fueron proporcionadas por la empresa Ripley S.A. Estas bases de datos se pueden unir a través del ID encriptado de un usuario. La primera base de datos tiene la información demográfica de las personas, en ella se puede identificar la edad, sexo, ocupación de una determinada persona, es decir de un determinado ID. Esto sirve para indagar la existencia de una relación explicativa entre las variables demográficas y el tipo de cliente *Won't Pay* o *Can't pay*.

La segunda base de datos da cuenta de todas las deudas que puede tener el cliente, tanto en la entidad financiera con la que se trabaja como en otras entidades financieras, los datos son recopilados por la CMF. es mensual y dado el tiempo considerado para este análisis, se tomaron 48 meses.

Por último, se cuenta con la base de datos de la entidad financiera con la que se trabaja, en esta se puede identificar si el cliente posee deudas, cuánto le debe al banco, su probabilidad de default, su tipo de segmentación según diferentes categorizaciones, además de otras variables que pueden ser importante para el estudio.

4.2 Selección de los Datos

Relevante para el estudio y para el análisis en sí es discriminar entre las bases de datos existentes qué información será relevante considerar y cuáles serán las variables cuya calidad informativa es valiosa para ser considerada.

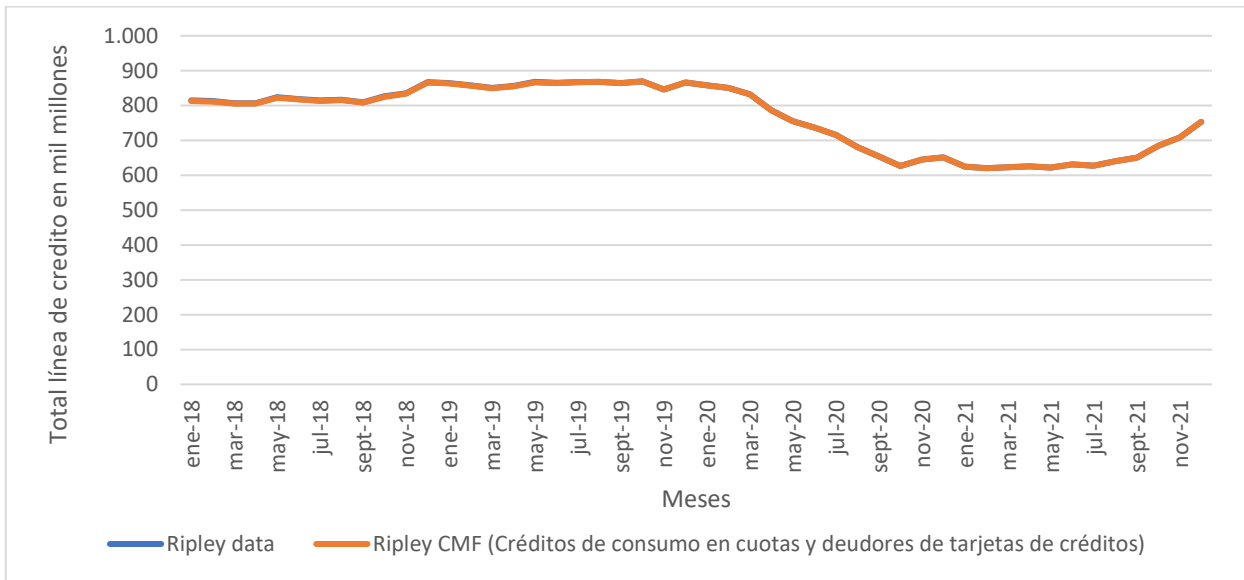
Otro elemento significativo es la fidelidad de los datos, por lo que se realizará una contrastación de los datos del banco con los de la CMF.

El trabajo se desarrolló con todas las bases de clientes, para ver si todos estos clientes eran relevantes para análisis posteriores. Esto se hizo a través de análisis de los subsegmentos que poseen los clientes, lo que hizo disminuir el número de clientes a analizar mes a mes.

4.2.1 Verificación de los datos

Para poder comenzar los análisis, lo primero que se debe realizar es establecer que se cuenta con una base fidedigna de datos. Dado que se trabaja con bancos, las cifras deben ser informadas en su totalidad a la CMF, por lo que se realizará la comparación de los datos que se encuentra en el sitio oficial de la CMF con los que brindó el banco, con el fin de observar si el total en ambos coincide.

Gráfico N°3: Comparación datos del banco con la de CMF”



Fuente: Elaboración propia

Observando el gráfico N°3, es posible identificar que ambas líneas realizan las mismas trayectorias. Existe un margen minúsculo que en promedio alcanza el 0.01%, en la mayoría del transcurso del gráfico se hace imposible diferenciar ambas líneas, reflejando que la información que se posee es fiable.

4.2.2 Composición de los clientes

La principal base de datos del banco está compuesta por 48 meses, está será usada para poder seleccionar los clientes activos del banco. Tomando el número de registros de ID diferentes dentro de estos 48 meses el total es de 3.520.134.

Para entender quiénes son estos clientes, se debe saber que el banco le asigna a cada cliente una segmentación, dependiendo de distintas características de ellos.

A continuación, se muestra la tabla resumen que especifica los grandes 5 segmentos junto a sus características de estos grupos.

Tabla N°1:” Segmentación de los clientes”

Segmento 1: Nuevos con solo compras y antigüedad <= 6 meses	Segmento 2: Antiguos con solo saldo compras (> 6 meses)	Segmento 3: Normalizados: Renegociación o Repactación	Segmento 4: Productos Financieros	Segmento 5: Clientes con solo saldo contingente
1A: Bancarizados Fuertes y al día Banco Ripley	2A: Clean (<= 6 días de vencimiento con el banco en los últimos 3 meses.)	3A: Refinanciado 1-30 días de mora	4A: Refinanciados sin días de mora	Prime: Clientes que no han tenido saldo vigente en los últimos 11 meses
1B: No bancarizados (máximo de 50.000 de deuda en el sistema) o morosos	2B: Dirty (> 6 días de vencimiento con el banco en los últimos 3 meses.)	3B: Repactados o renegociados	4B: SAV-Consumo >= \$2MM	Fija: Clientes que tuvieron saldo vigente en los últimos 11 meses
			4C: SAV-Consumo < \$2M M o Avance	

Fuente: Elaborada por Banco Ripley

En la tabla N°1, en la última columna se observa los clientes del Segmento 5, quienes son los que tienen cuenta en el banco, pero no están activos. Este grupo será excluido del análisis dado que el fin de la tesis es analizar los clientes activos.

Al eliminar de la muestra las observaciones que pertenecen al Segmento 5, el total desciende a 2.457.668 observaciones, las que serán utilizados para realizar los análisis correspondientes. Es importante señalar, que esta cantidad de observaciones no es una suma constante mensual, sino que son formados por la entrada de nuevos registros a lo largo del tiempo.

4.3 Pre-procesamiento de Datos

Debido a la gran cantidad de datos con que se está trabajando, es importante establecer ciertas métricas antes de emplearlos para el análisis. Se deben estimar:

- **Consistencia:** Los datos deben estar entre los rangos establecidos y al crear nuevas variables o métricas deben ser hechas bajo los consensos establecidos por el banco. En el caso de la métrica fecha de nacimiento hay que considerar que una persona no puede haber nacido en 1920 o años menores.
- **Transformaciones realizadas:** Las transformaciones deben ser hechas con el fin de ayudar al proceso de análisis y de manera certera, para que no brinden información errónea. De igual forma, es importante tener todas las bases de datos en la misma unidad de medida para los términos que lo requieran.

4.4. Transformación de datos

En esta etapa se crean variables de interés para el análisis como lo es, la edad de los clientes, que será calculada con la variable fecha de nacimiento. Será la diferencia entre 01.08.2022 y la fecha que entrega la base, fijando como resultado de interés, las edades a considerar entre 18 y 95 años de edad. Esta variable se usará fija para los clientes en los 4 años de las bases de datos.

A la vez, se crean variables relacionadas con las deudas de la persona, las que serán comentadas más adelante al momento de hablar sobre el modelo.

Por otra parte, las bases de datos son normalizadas, en específico la de deuda y del banco, ya que no están expresadas en la misma unidad.

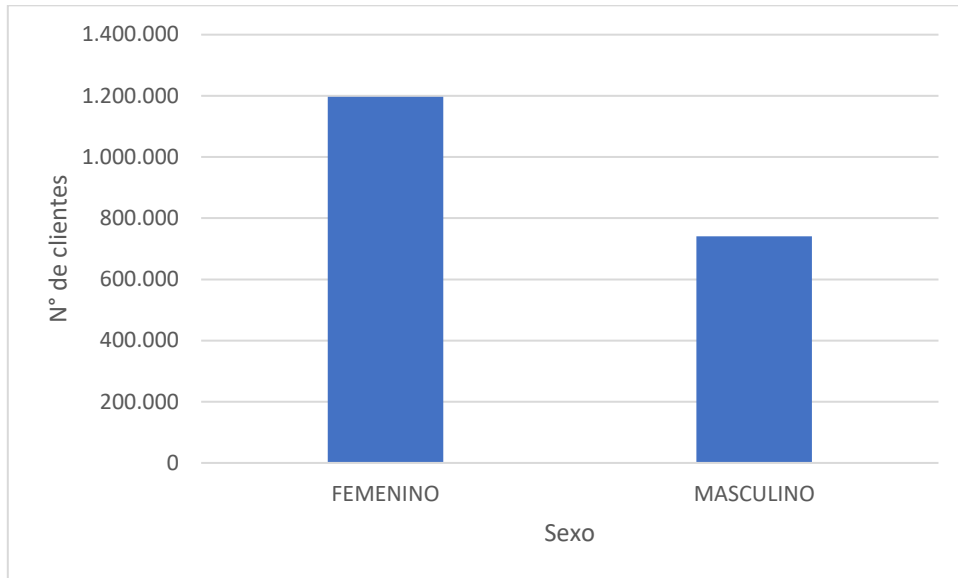
También con la ayuda de árboles de decisión se crearán categorías para facilitar los análisis posteriores y de esta manera posibilitar la segmentación de los clientes del banco.

4.5 Análisis descriptivo

Posterior a la construcción y selección de variables, se procede a realizar el análisis descriptivo de ellas, con el fin de poder entender quiénes son los clientes del banco y, en específico, qué diferencia a un cliente bueno de uno malo.

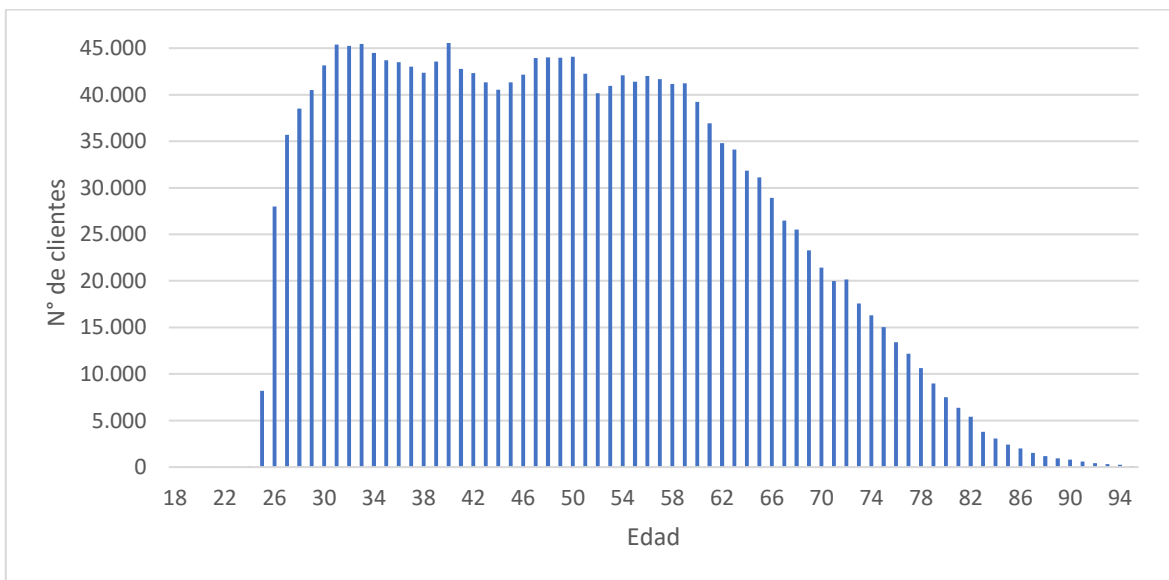
Para caracterizar a los clientes se puede hacer uso de la base de datos demográficos, donde obtenemos la variable sexo, fecha de nacimiento y actividad. Si se toma la variable sexo, se ve una preponderancia según el gráfico N°4 de clientes del género femenino. En el gráfico N°5 se observa la edad de los clientes, y se puede observar que la mayor densidad está entre el rango de 29 a 59 años. Por último, en la variable actividad, cuyo comportamiento es observable en el gráfico N°6, predominan los registros dependientes.

Gráfico N°4:” Distribución de clientes por sexo”



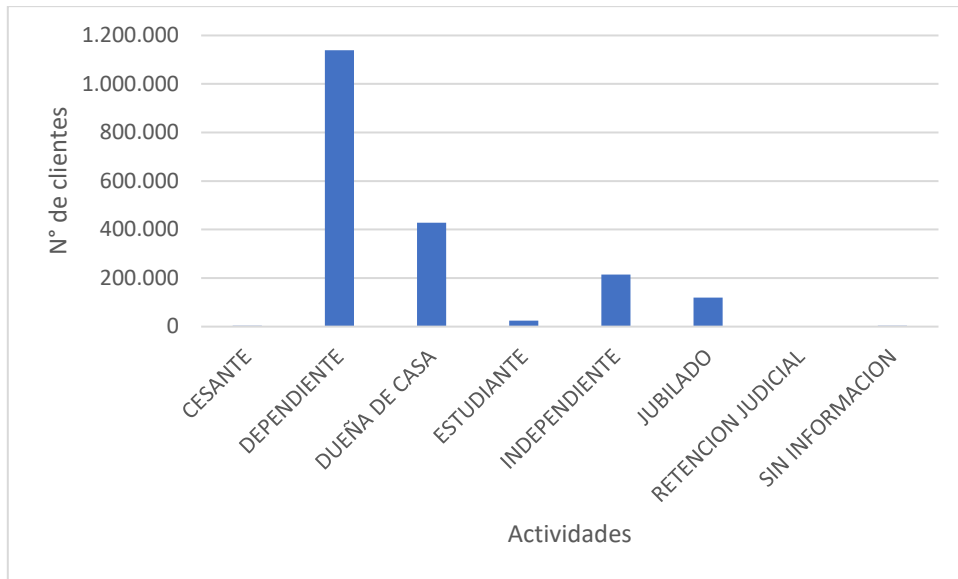
Fuente: Elaboración propia

Gráfico N°5:” Distribución de clientes por edad”



Fuente: Elaboración propia

Gráfico N°6:” Distribución de clientes por actividades que realiza”



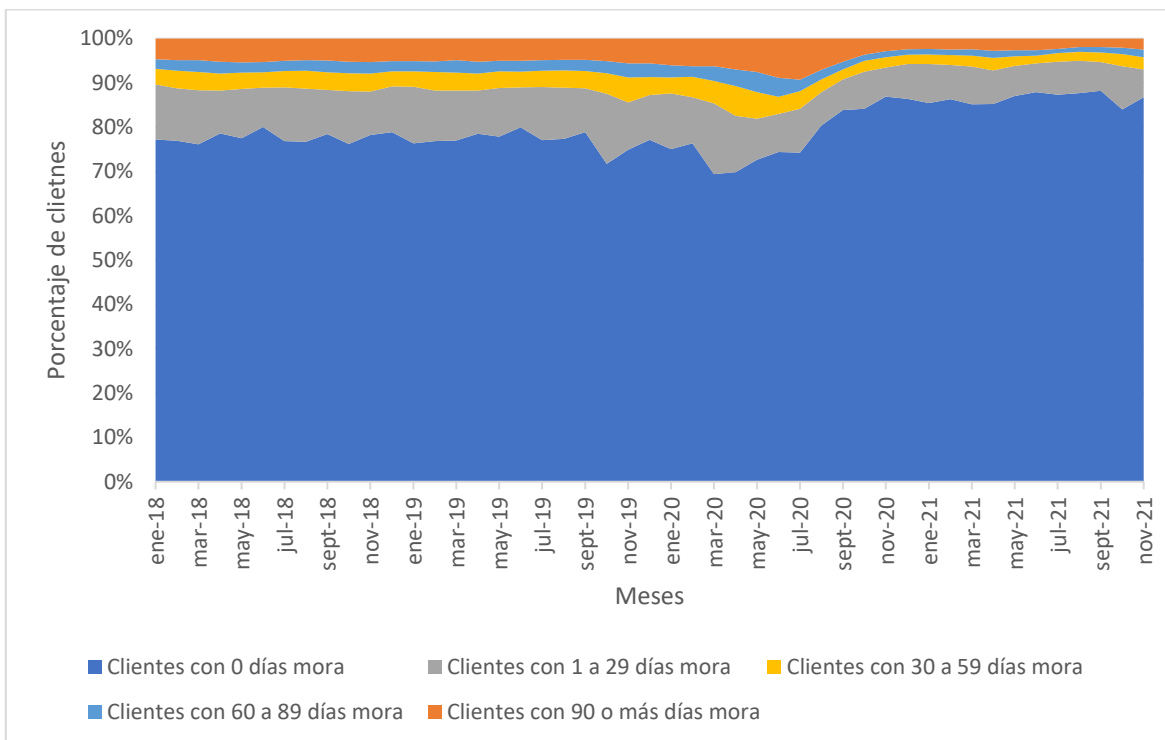
Fuente: Elaboración propia

Para generar una categorización del tipo de clientes, se utilizan los días de mora que estos poseen, variable que se encuentra en la base de datos del banco y expresa cuántos días acumulan sin haber pagado sus obligaciones.

Si se analiza la base de datos total, se pueden considerar los clientes que no tienen mora. Luego, se considerarán aquellos que presentan menos de un mes de mora, posteriormente, quienes poseen entre un mes y menos de dos meses, consecutivamente, quienes acumulan entre dos meses y menos de tres, para finalmente, dar cuenta de quienes evidencian 90 o más días de mora. Este segmento corresponderá a los incumplidores.

En el gráfico N°7, se puede observar que la mayoría de los clientes corresponde al grupo de 0 días mora. Estos viven una baja en el mes de marzo del 2020, para luego presentar alza desde julio del 2020 hasta el mes de noviembre del 2020, mientras que los clientes que presentan registro de mora siguen un comportamiento similar, teniendo su punto máximo de baja en junio del 2020 y el punto máximo en noviembre 2020. Observando finalmente, que el número de clientes sin mora aumentó con relación al tiempo y el de clientes que presentaban moras, principalmente aquellos con más de un mes, disminuyó.

Gráfico N°7:” Distribución de los clientes de la base de datos según días mora”



Fuente: Elaboración propia

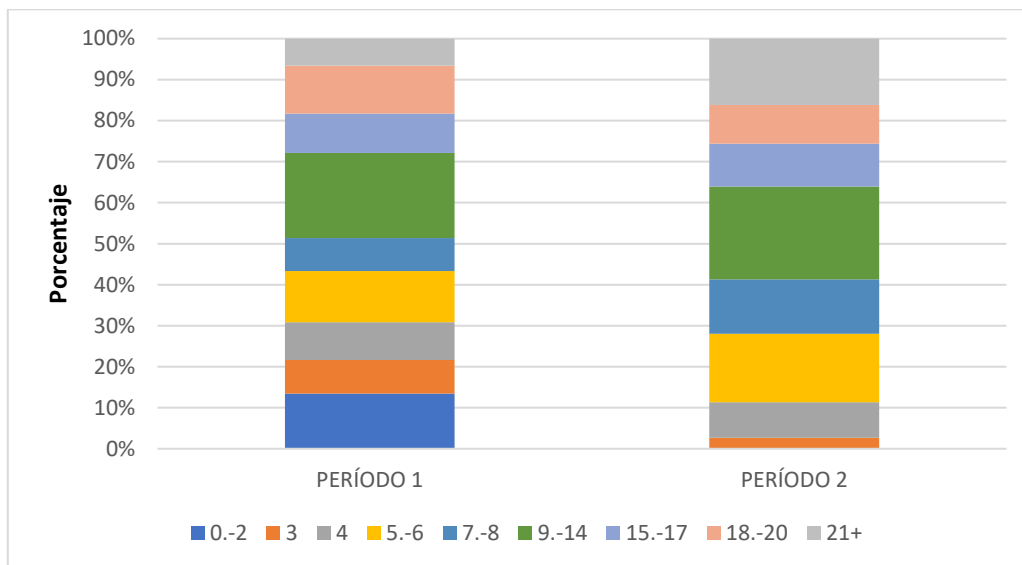
El banco considera a las personas con más de 90 días de mora cómo incumplidores, estos conformarán los malos clientes de la base de datos, mientras que los buenos clientes serán considerados como los que tienen entre 0 a 89 días de mora.

Es importante saber quiénes son estos clientes y qué características comparten entre ellos. Es así como el siguiente paso es buscar características que los identifique en dos períodos de tiempo, el primero, previo las ayudas económicas, y el segundo, posterior a éstas. Es decir, a malos y buenos pagadores desde 01.2018 al 08.2020, nombrado período 1, y entre 09.2020 y el 12.2021 nombrado como período 2.

4.5.1 Clientes buenos

A continuación, serán descritos los buenos clientes tanto por variables demográficas como por las del banco. Esta sección se complementa con el Anexo 1, donde se pueden encontrar los PSI.

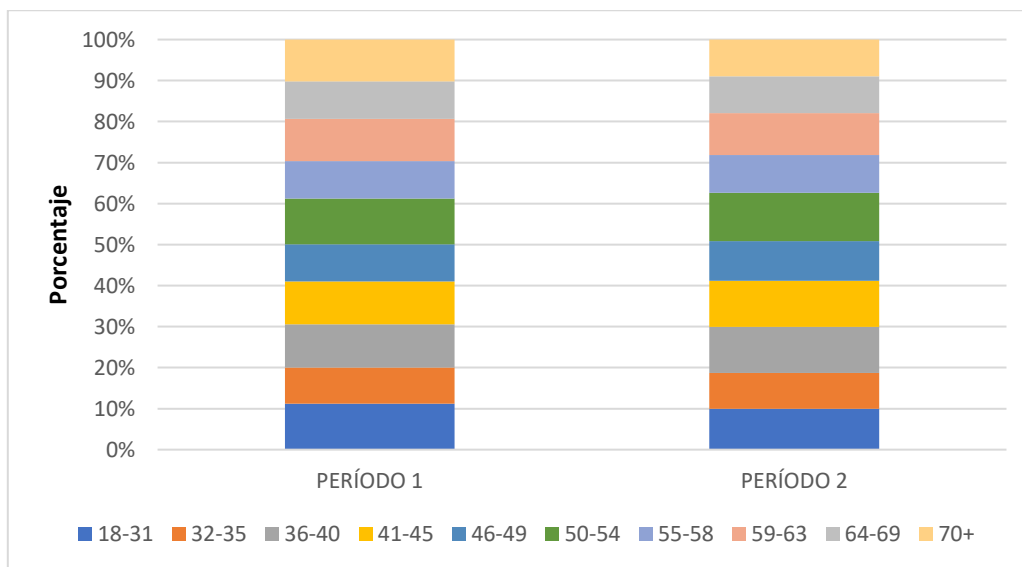
Gráfico N°8: “Composición de los clientes buenos según antigüedad en años”



Fuente: Elaboración propia

Los clientes buenos se caracterizan por tener un mayor porcentaje de personas que poseen entre 9 a 14 años de antigüedad en el banco. Los mayores cambios entre el período 1 y 2 se asocian al porcentaje de gente que posee entre 0 a 2 años de antigüedad.

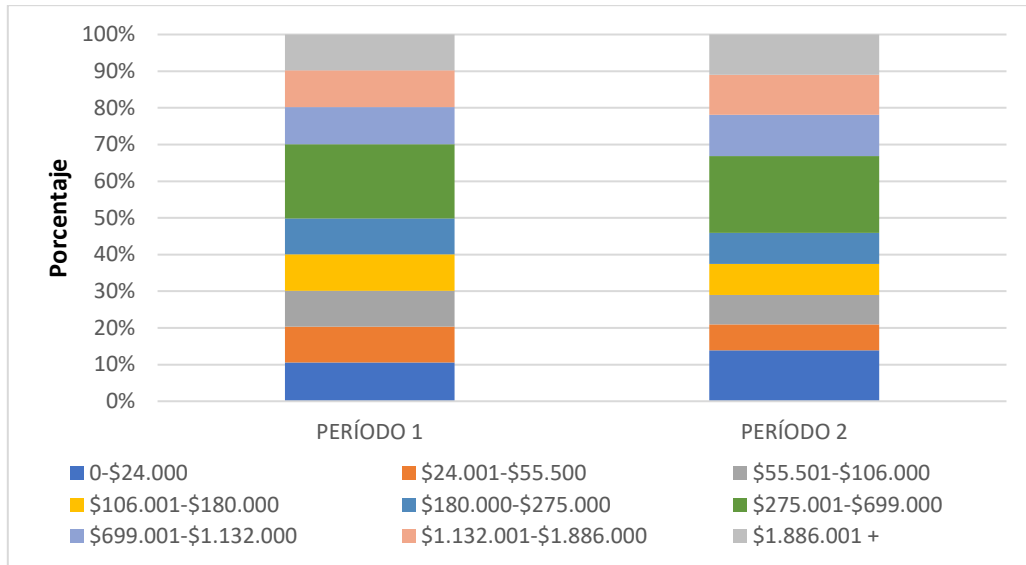
Gráfico N°9: “Composición de los cliente buenos según edad”



Fuente: Elaboración propia

En cuanto a la edad la distribución entre los segmentos etarios es pareja. Las que más variaron entre un período y otro, con un PSI mayor a 0,1, son los correspondientes al menor rango etario (18-31) junto a los de mayor rango etario (70+).

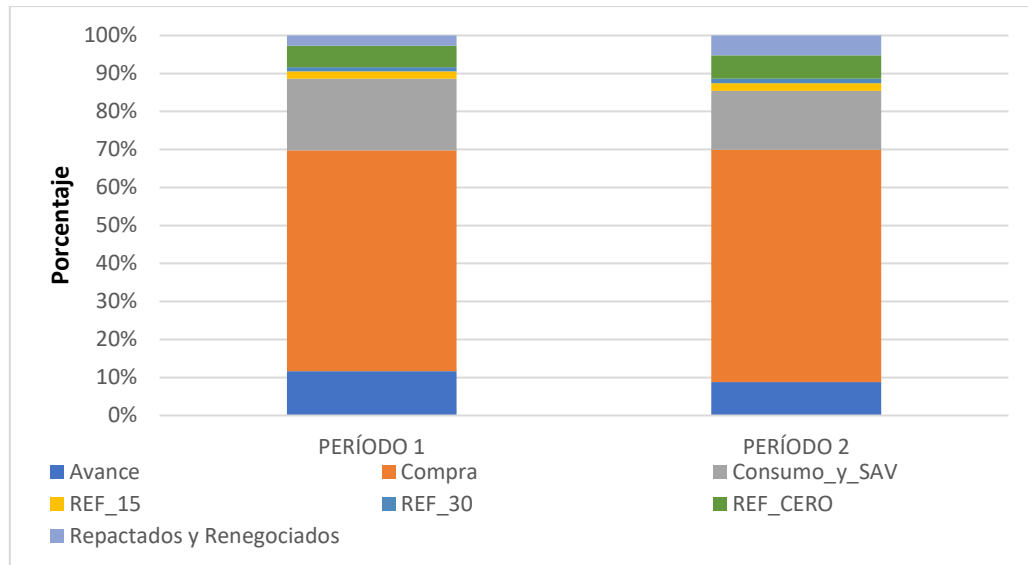
Gráfico N°10: “Composición de los clientes buenos según monto de deuda”



Fuente: Elaboración propia

Referente a los montos adeudados, el mayor porcentaje corresponde al segmento de entre \$275.000-\$699.000, representando cerca del 50% de los segmentos que se encuentran bajo los \$275.000. El mayor cambio entre ambos períodos corresponde a los rangos entre 0-\$24.000 y entre \$24.001-\$55.500, dándose un aumento del primer segmento y una disminución del segundo.

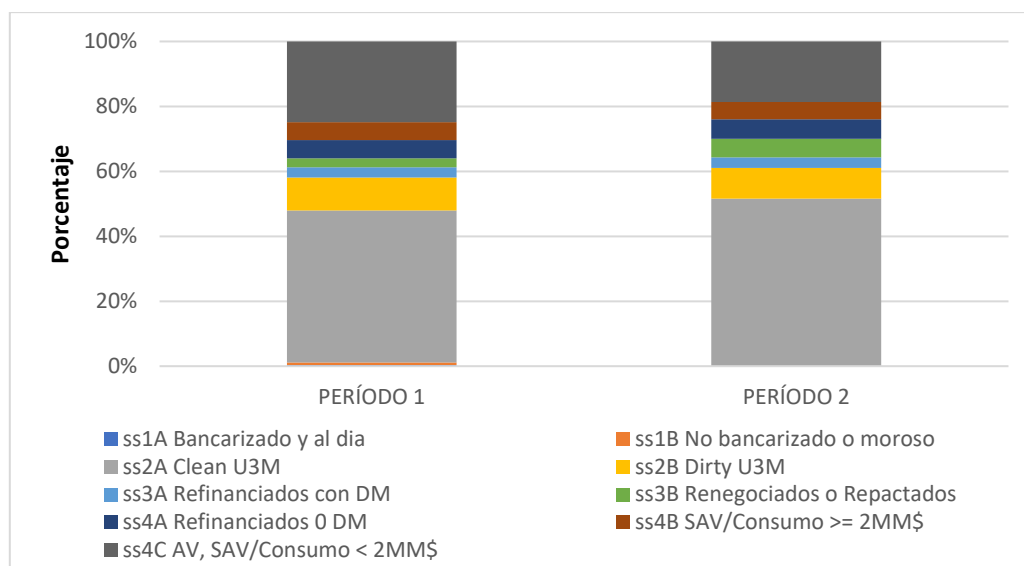
Gráfico N°11: “Composición de los clientes buenos según su producto más riesgoso en la cartera”



Fuente: Elaboración propia

Los clientes buenos, en su mayoría, presentan como producto más riesgoso la compra. Por su parte, la categoría que tiene mayor variación, dado el PSI, corresponde a los repactados y renegociados, avances y consumo_y_SAV. Esta variable presenta sus categorías explicadas en el Anexo 2.

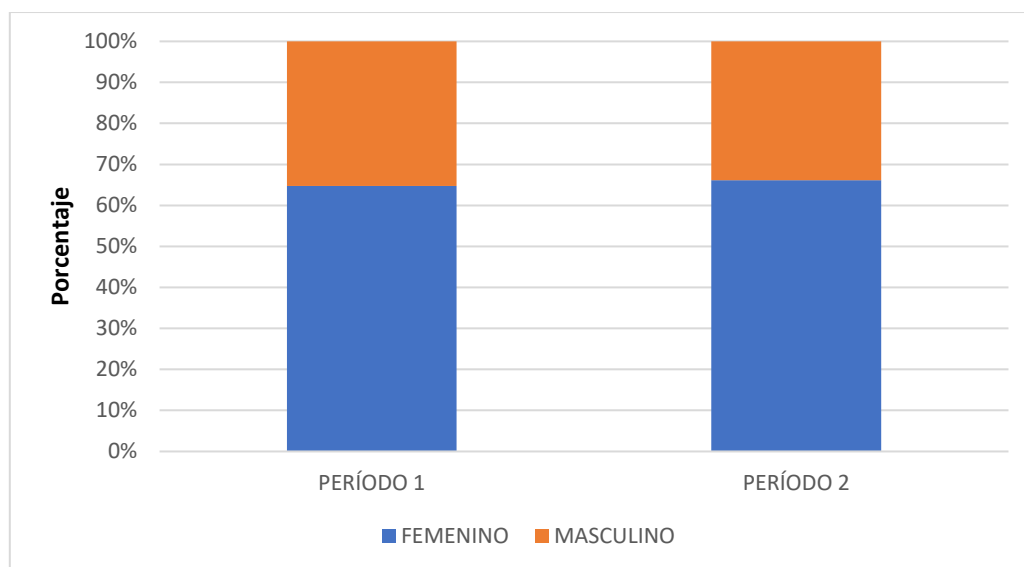
Gráfico N°12: “Composición de los clientes buenos según los subsegmentos a los que pertenecen”



Fuente: Elaboración propia

Los segmentos que predominan en los buenos clientes son los SS2A y SS4C. Los que tienen más variación son SS1B, SS3B y SS4C, donde el primero disminuye, el segundo aumenta y el tercero igualmente disminuye.

Gráfico N°13: “Composición de los clientes buenos según su sexo”



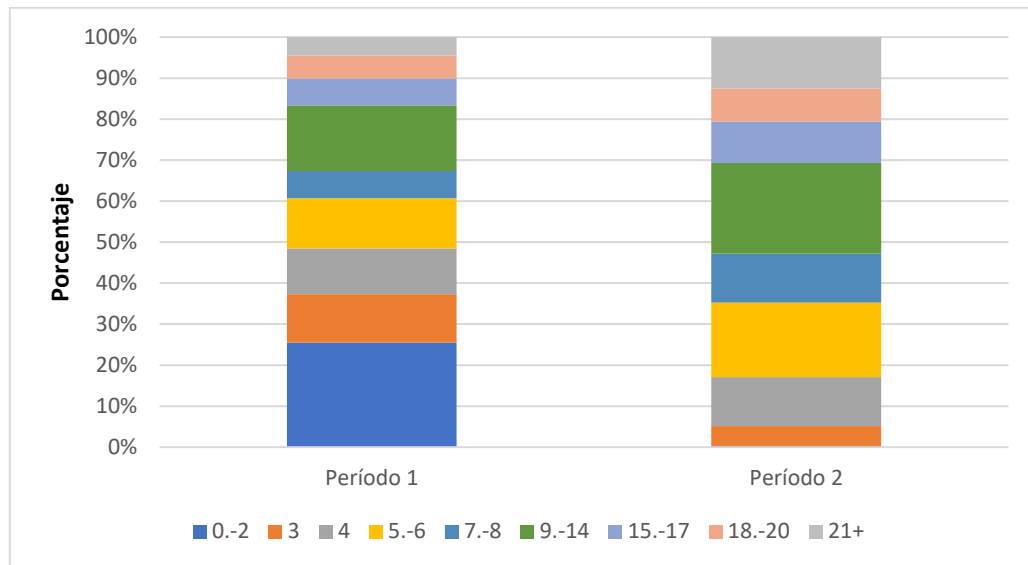
Fuente: Elaboración propia

En la variable sexo se exhibe, que el mayor porcentaje de buenos clientes corresponde al sexo femenino y esta variable no presenta una variación significativa entre período 1 y 2, lo que también pasa con la variable actividad, donde tienen mayor preponderancia los dependientes.

4.5.2 Clientes malos

De manera similar a la anterior, serán descritos los clientes malos. Los índices de PSI están por categoría en el Anexo 3.

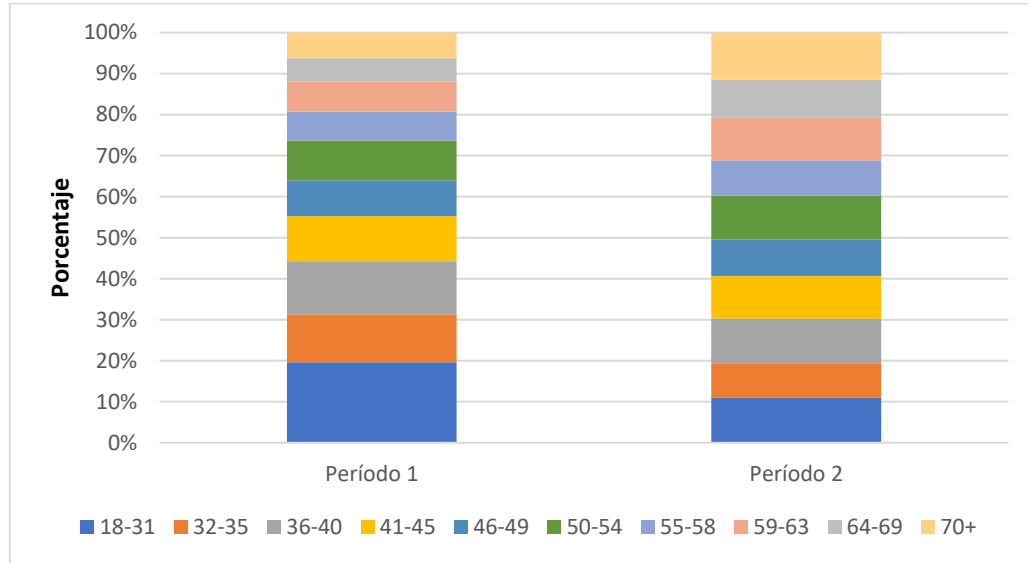
Gráfico N°14: “Composición promedio de clientes malos según antigüedad”



Fuente: Elaboración propia

Los clientes malos están representados en un mayor porcentaje por clientes que poseen entre 0 a 2 años de antigüedad. Los mayores cambios entre el período 1 y 2 se reflejaron en este grupo e igualmente en los que poseen menos de 3 años de antigüedad. Esta razón explica que clientes con menos tiempo en el banco dejaron de pertenecer a la categoría de malos.

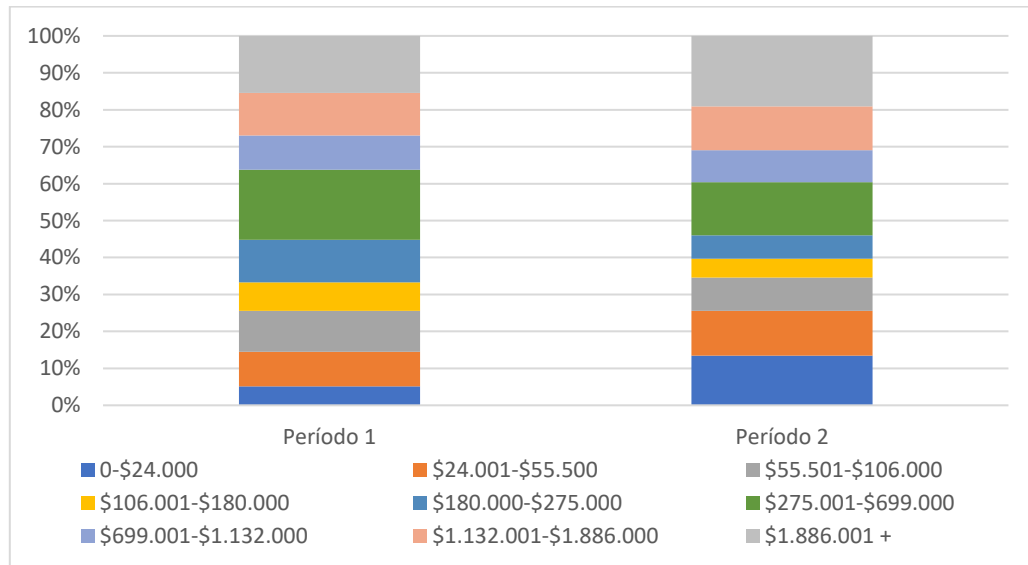
Gráfico N°15: “Composición promedio de clientes malos según edad”



Fuente: Elaboración propia

En cuanto a la edad y la distribución entre los segmentos etarios, el mayor porcentaje de clientes malos se encuentra en el rango de 18-31 años. Dicho rango junto a los de 70+ fueron los que más variaron entre periodos con un PSI mayor a 0.1.

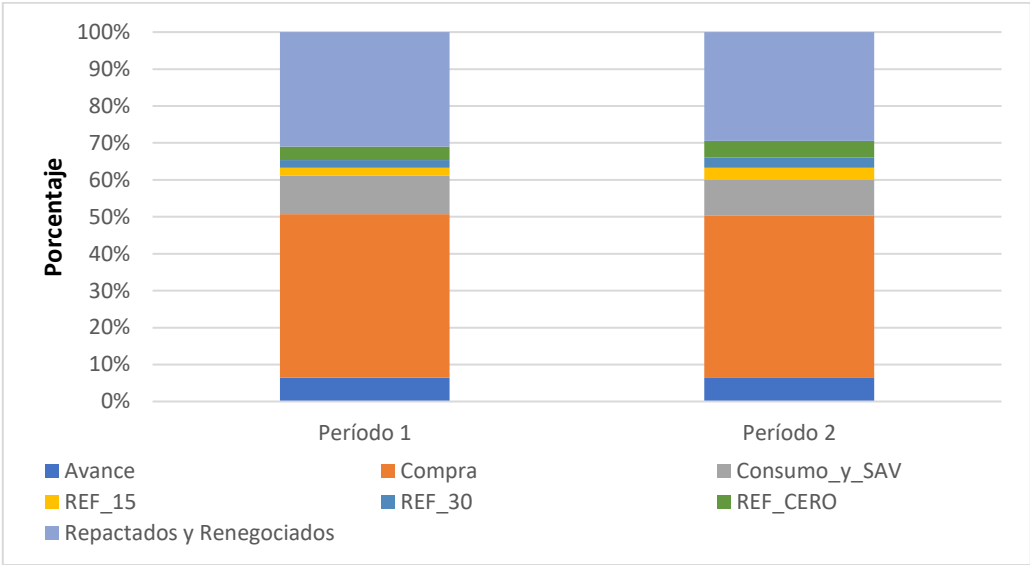
Gráfico N°16: “Composición de clientes malos según saldo adeudado”



Fuente: Elaboración propia

Respecto a los montos adeudados, estos presentan un mayor porcentaje en el segmento mayor a \$1.886.001. El mayor cambio visto entre ambos períodos corresponde a los montos situados entre los 0-\$24.000 que evidenciaron un incremento. Por su parte el grupo entre \$180.000-\$275.000 reflejó una disminución.

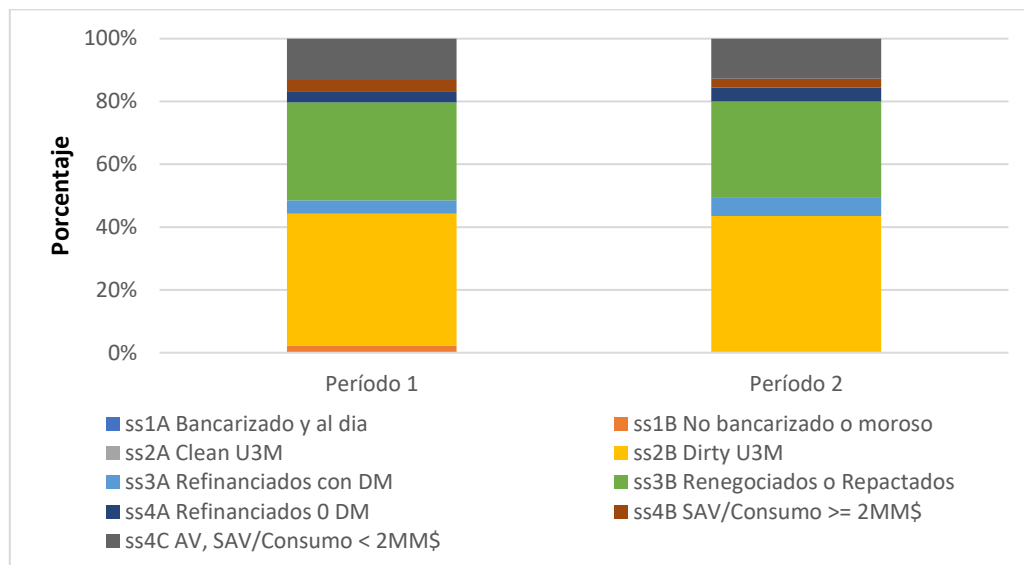
Gráfico N°17: “Composición de los clientes malos según producto más riesgoso de la cartera”



Fuente: Elaboración propia

Los clientes malos, en su mayoría presentan como producto más riesgoso la compra, los repactados y renegociados, mientras que el que tiene una mayor variación, es el REF_15.

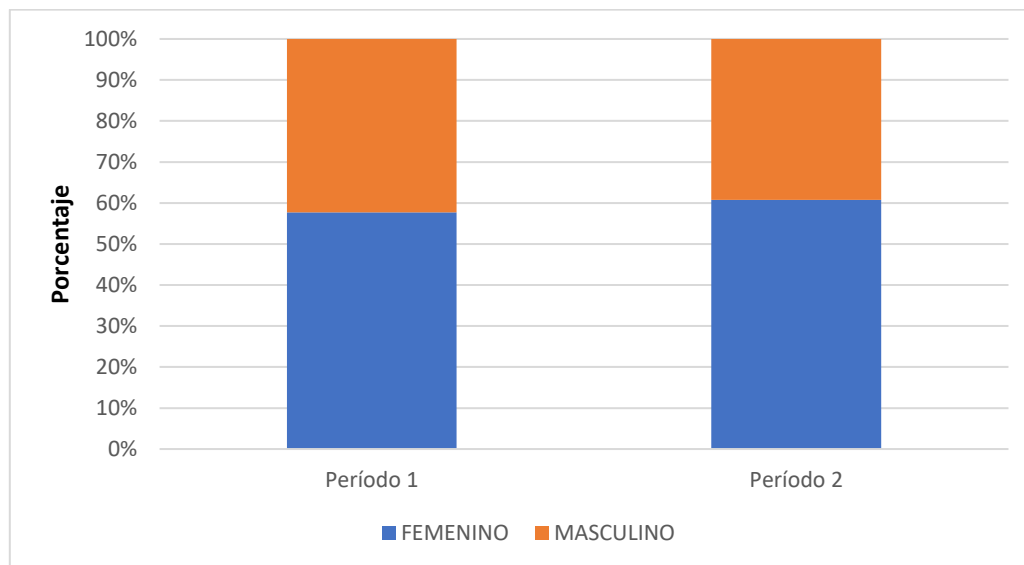
Gráfico N°18 “Composición de clientes malos según segmentos de clientes”



Fuente: Elaboración propia

En la variable de los segmentos, los que predominan en los malos son los SS2B y SS3B y el que tiene más variación es SS1B disminuyendo su porcentaje.

Gráfico N°19: “Composición promedio de los clientes malos según el sexo”



Fuente: Elaboración propia

Por último, en cuanto a sexo se aprecia, que el mayor porcentaje de malos clientes corresponde al sexo femenino. Dicha variable no presentó una gran variación entre el período 1 y 2. En cambio, el grupo de clientes masculinos sí presentó una variación mayor entre periodos.

CAPÍTULO 5

5. Discusión

5.1 Riesgo según características de los clientes

Obteniendo las variables de interés y las que más discriminan el comportamiento de los individuos, se procedió a realizar clusterización de las variables a través de árboles de decisión, con el fin de poder graficar el riesgo en función de diferentes cluster de individuos dentro de la variable en estudio.

Estos análisis se realizaron considerando las fechas que ocurrieron los shocks, especialmente cuando sucedió la crisis social, primer IFE, primer retiro y último retiro. Las cuatro fechas de cortes que se considerarán para los análisis serán:

- Octubre, 2019
- Mayo, 2020
- Agosto, 2020
- Mayo, 2021

Estas fechas serán graficadas con los colores morado, rosado, azul y rojo respectivamente, perpendiculares al eje de las abscisas.

Es importante señalar, que el riesgo se puede analizar de tres maneras distintas, la primera es tomando la Probabilidad de incumplimiento (PD) de los clientes que entrega la base de datos del banco, esta variable es calculada a través de un modelo interno que ellos poseen. En segundo lugar, se puede ver el riesgo con una variable similar a la PD, que es el incumplimiento a 12 meses, esta variable entrega 1 si la persona incumplió en 12 meses más y 0 si no, por lo que se considerará como tasa de incumplimiento del segmento i a:

$$Tasa\ de\ incumplimiento_i = \frac{Total\ de\ individuos\ en\ incumplimiento}{Total\ de\ individuos\ del\ segmento\ i}$$

Ecuación N°10: "Tasa de incumplimiento"

La tercera forma de analizar el riesgo es a través de la pérdida porcentual que tuvo el segmento, la cual se mide dado el saldo que adeuda la gente y la PD, según la siguiente fórmula:

$$\text{Pérdida esperada del segmento} = \frac{\sum_{i=1}^n PD_i \times Saldo_i}{\sum_{i=1}^n Saldo_i}$$

Ecuación N°11: “Pérdida esperada del segmento”

A continuación, se muestran los tres tipos de análisis para diferentes variables presentes en la base de datos.

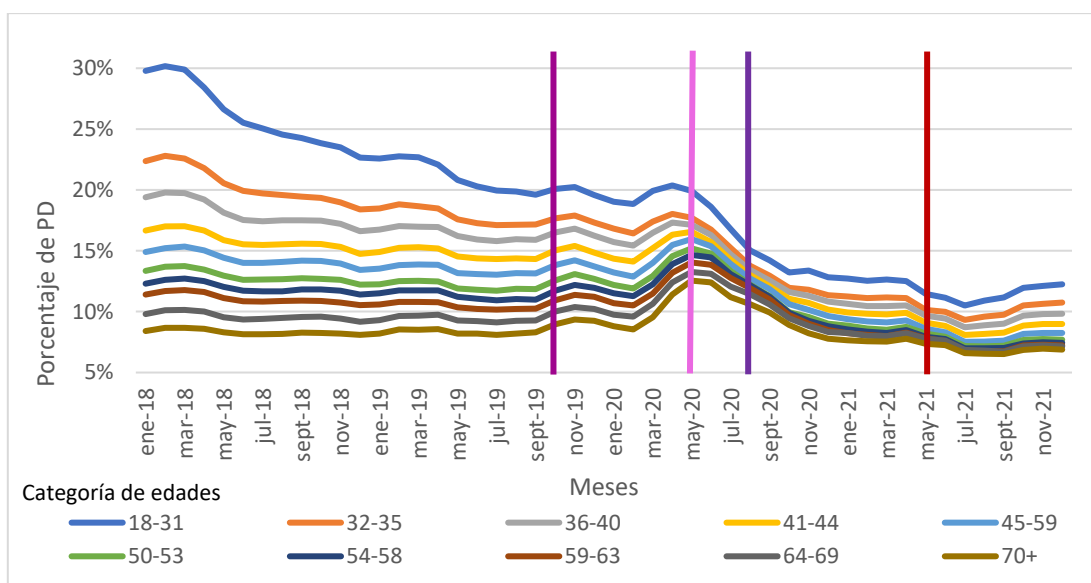
5.1.1 Variable de la edad de los clientes

Para la variable edad el árbol de decisión arrojó 10 categorías, las cuales se graficaron según los tres tipos de análisis de riesgo.

El primer gráfico de esta sección, número 20, que se puede observar, el de PD a lo largo del tiempo, es decir la probabilidad de default de los clientes, según las categorías de edad. De donde se pueden extraer ciertas características: Los PD antes de octubre de 2019 presentaban una diferencia clara entre los distintos grupos etarios y tienen una tendencia a la baja. Mientras que en marzo del 2020 se observa una subida de PD y que las líneas empiezan a juntarse. Posterior a mayo 2021 se empiezan a separar las líneas nuevamente.

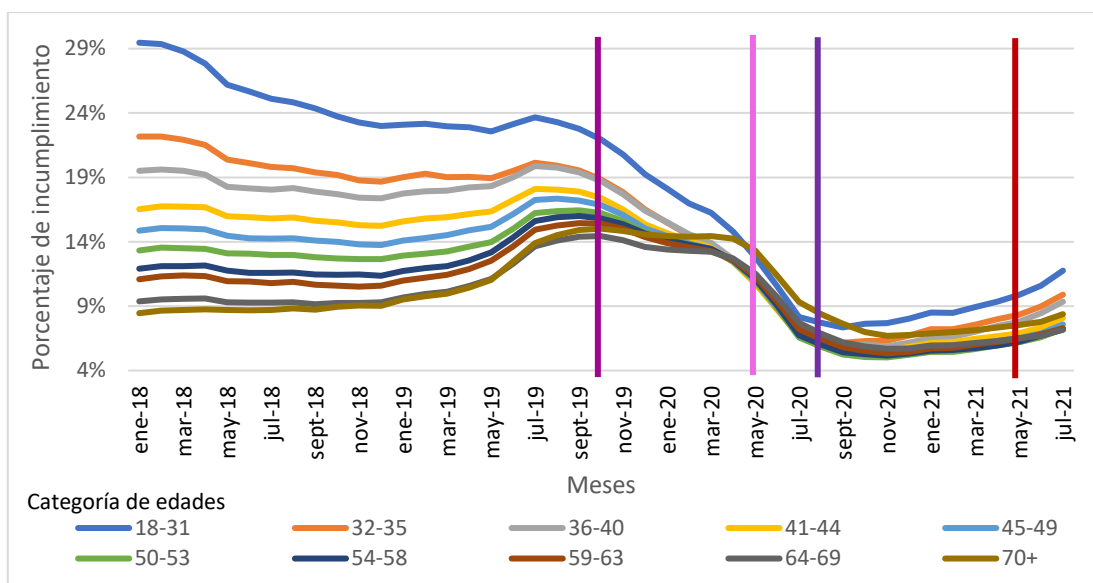
Esto quiere decir que en el período catalogado como de ayudas económicas, los grupos etarios disminuyeron su riesgo y alguno de estos grupos se comportaron iguales entre sí, por lo que los grupos comenzaron a ser más homogéneos en su riesgo.

Gráfico N°20: “PD a través del tiempo según categorías de edades”



Fuente: Elaboración propia

Gráfico N°21: “Tasa de incumplimiento a 12 meses según categoría de edades”



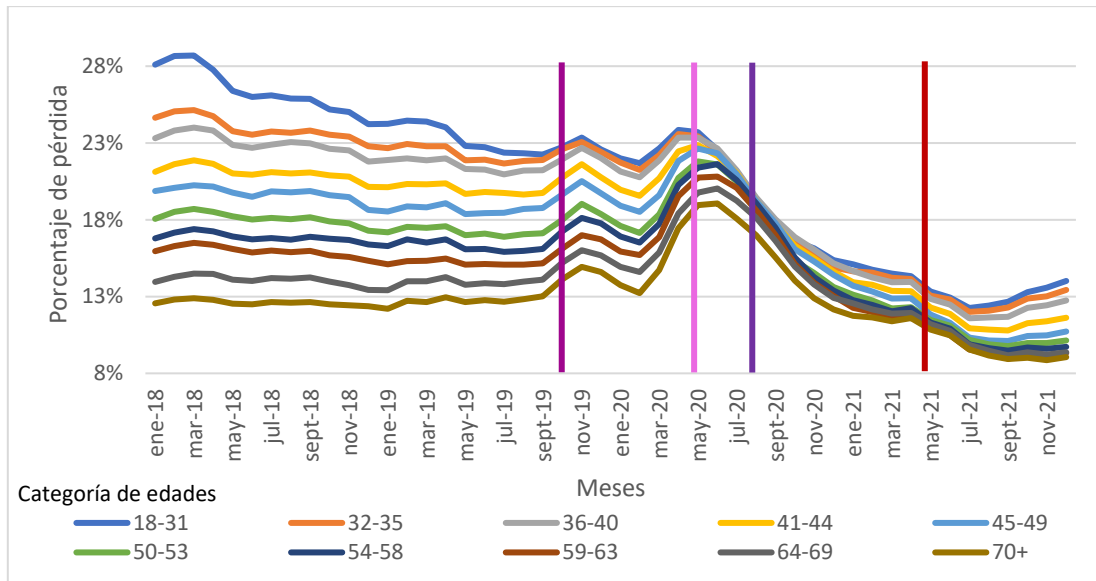
Fuente: Elaboración propia

El gráfico N°21 de edad corresponde al de la tasa de incumplimiento mensual. Entre los elementos más significativos se reflejan: En el período antes de octubre de 2019, es posible diferenciar la trayectoria de las edades con claridad, teniendo una cúspide en julio 2019, la que se atribuye a políticas implementadas en el banco de cómo considerar el incumplimiento. Posterior a octubre 2019, la tasa de incumplimiento una cae, donde la pendiente es más abrupta desde mayo 2020, tiempo en el que comenzaron las ayudas económicas, hasta el primer retiro de agosto 2020, donde se mantienen más constante las trayectorias.

Un grupo etario a destacar son los pertenecientes al grupo 70+, dado que no siguieron la misma trayectoria que los demás, si no que aumentaron su nivel de incumplimiento posterior a octubre 2019 y vivieron la caída de éste desde mayo de 2020.

Por otro lado, tanto el gráfico de PD como el de tasa de incumplimiento son comparables. Es así como se observa que en el período previo a octubre de 2019 ambos gráficos son similares en las trayectorias de los grupos etarios, pero posterior a esto tienden a diferenciarse bastante, esto se debe a que modelo que tiene el banco para predecir la PD está captando las caídas después y de forma más suave de lo que realmente está ocurriendo.

Gráfico N°22: “Pérdida esperada del segmento con diferentes categorizaciones de edad”



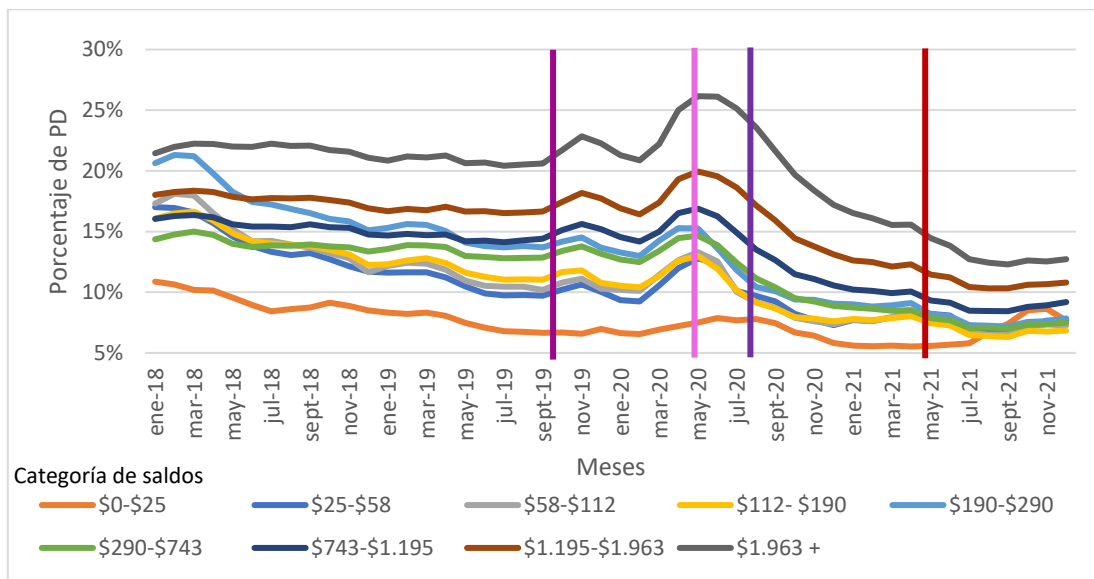
Fuente: Elaboración propia

En el gráfico N°22, se observa la pérdida esperada según los segmentos etarios, donde también ocurre que las líneas se juntan a medida que hay más solvencia en el mercado y posterior a esto se vuelven a separar.

5.1.2 Variable del saldo de los clientes

Para la variable saldo, que representa el monto que adeuda una persona, se realizaron los mismos análisis anteriores, obteniendo 9 clusterizaciones. Al observar el gráfico N°23 de PD los saldos empiezan con una trayectoria más compacta, la que se abre posterior a las crisis para luego volverse a comprimir. Un grupo a destacar es el de los montos bajos, entre \$0-\$25.000 dado que no siguen la trayectoria general del grupo.

Gráfico N°23: “PD a través del tiempo de saldo”

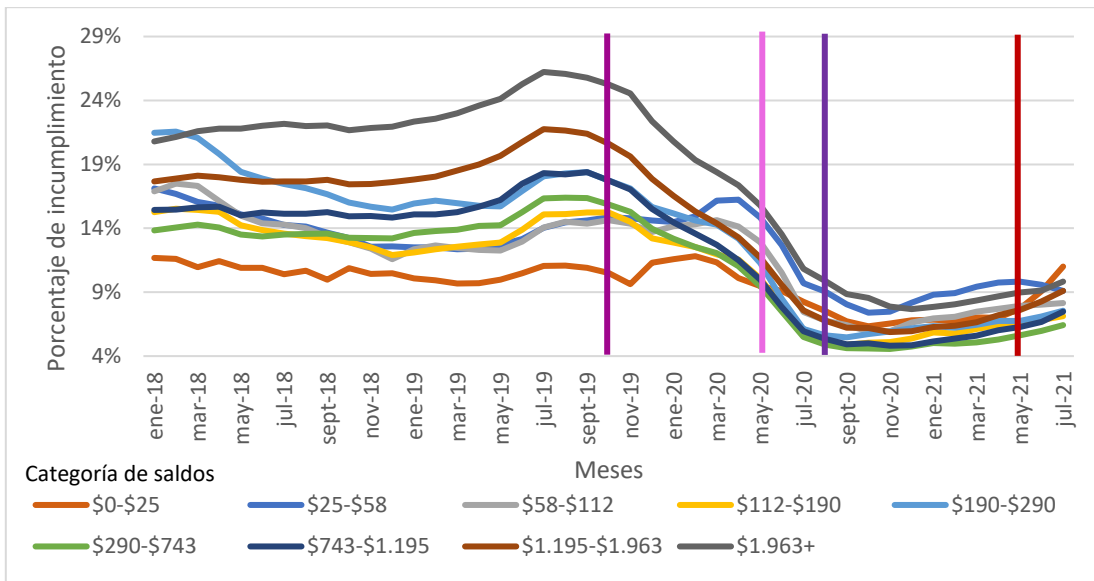


Fuente: Elaboración propia

En el análisis del incumplimiento para saldo, se observa que existen dos tipos de trayectorias en el gráfico, las menores a \$290.000 y las mayores a este valor.

Para los valores menores, se podrían agrupar tres segmentos que están teniendo la misma trayectoria hasta octubre de 2019, conformando el rango desde \$25.000-\$190.000. Al llegar al punto de la crisis las tres curvas se expanden separando sus trayectorias, mientras que el rango menor a \$25.000 vive una disminución del porcentaje de incumplimiento más bajo que el resto. Por otro lado, los valores mayores tienen la misma trayectoria, aumentando sus diferencias en el tiempo de crisis y teniendo pendientes más pronunciadas que el grupo de menor saldo.

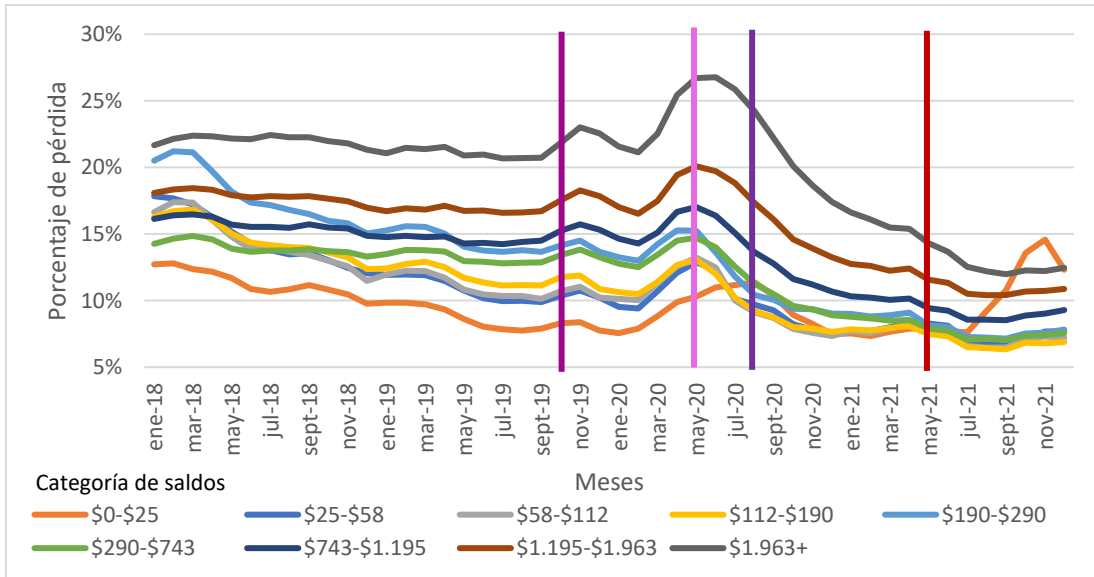
Gráfico N°24: “Porcentaje de incumplimiento a 12 meses de saldo”



Fuente: Elaboración propia

Comparando ambos gráficos, se puede ver que la caída del incumplimiento es mayor que en la PD, lo que nuevamente determina que el modelo no está captando la verdadera disminución de riesgo que se vive.

Gráfico N°25: “Pérdida porcentual del segmento con diferentes categorizaciones de deuda del cliente”



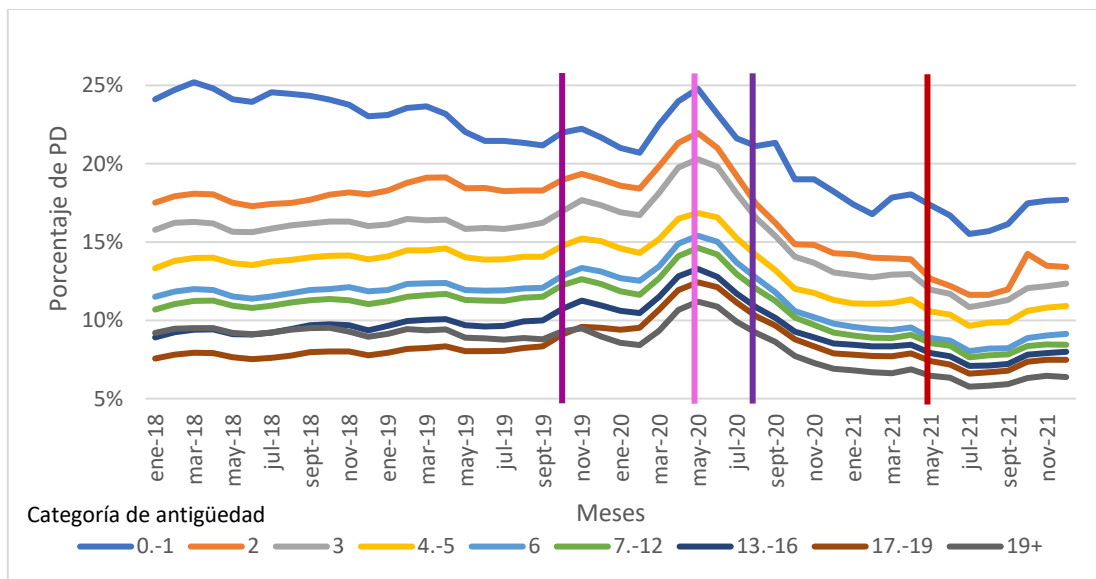
Fuente: Elaboración propia

Por último, en el gráfico de pérdida esperada, el gráfico N°24 se asimila al del incumplimiento y los segmentos se comportan de la misma manera que él, por lo que acá se observa que las diferencias entre los grupos se acentúan en el período de crisis.

5.1.3 Variable antigüedad de los clientes en el banco

En cuanto a la variable antigüedad, las categorías creadas son 9. Lo que ocurre es similar a la edad, en el gráfico N°25 de PD, las trayectorias de las distintas 9 categorías se pueden observar claramente, pero luego estas se unen y posteriormente se vuelven a separar. El grupo a destacar es el que lleva un año o menos en el banco, ya que ellos presentan un riesgo superior y su pendiente de caída es menor en comparación al resto.

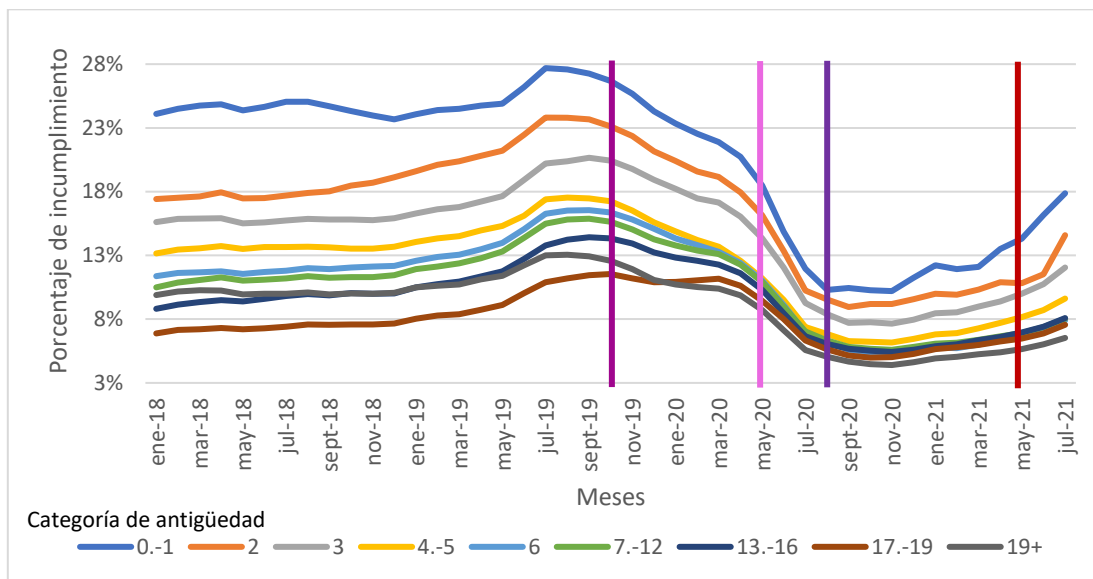
Gráfico N°26: “PD a través del tiempo de antigüedad”



Fuente: Elaboración propia

Si se observa el porcentaje de incumplimiento, gráfico N°26, se ve que hay dos grupos de comportamientos, los que llevan tres años o menos en el banco y los que llevan más de 4 años. En el momento de crisis los de menor años en el banco viven una caída de incumplimiento similar, mientras que los que tienen mayor estadía en el banco juntan sus las líneas desde mayo 2020 hasta aproximadamente mayo 2021. Es decir, que dado las ayudas económicas estos grupos tienden a ser más homogéneos en su comportamiento.

Gráfico N°27: “Porcentaje de incumplimiento a 12 meses de antigüedad”

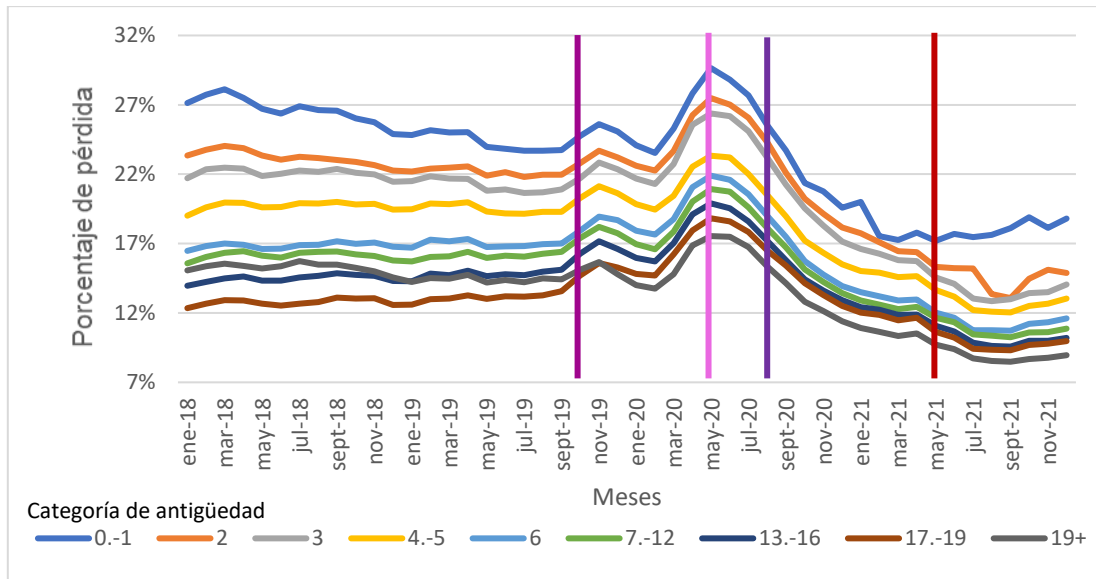


Fuente: Elaboración propia

Al comparar ambos gráficos, igual que en los casos anteriores de las variables edad y saldo, la caída del gráfico de incumplimiento en período de crisis es mayor que la del gráfico de PD.

En cuanto a la pérdida porcentual esperada de la antigüedad, antes de octubre del 2019 los grupos eran diferenciables unos de otros y luego se comenzaron a juntar las líneas, para ya al final tener una leve separación nuevamente. Vale destacar, que el grupo con un año o menos de estadía en el banco, presenta un comportamiento de fluctuaciones entre octubre del 2019 y mayo del 2021, donde tiene pequeños períodos de subida que no viven los otros segmentos.

Gráfico N°28: “Pérdida porcentual del segmento según la antigüedad del cliente en el banco”

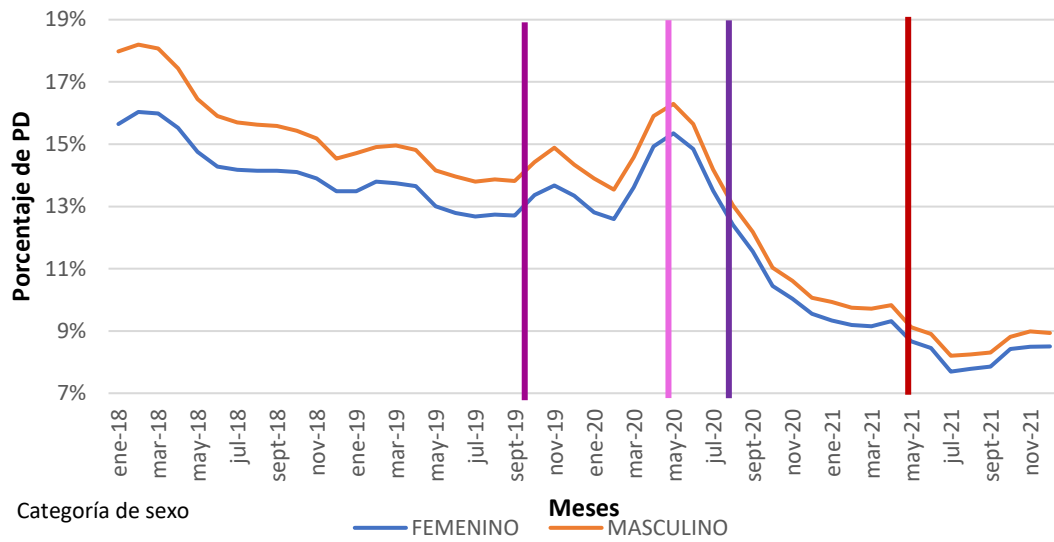


Fuente: Elaboración propia

5.1.4 Variable del sexo de los clientes

En cuanto a la variable sexo, que presenta la categorización de femenino y masculino, el comportamiento de estos gráficos no es similar al resto. Por un lado, se tiene el gráfico N°28 de PD, donde se observa que hasta octubre 2019 ambas líneas presentan las mismas fluctuaciones con tendencia a la baja, para posteriormente aumentar hasta mayo del año 2020, donde vive una caída, dando como resultado que ambas líneas ahora no tengan tanta separación entre ella, por lo que tanto hombres como mujeres se empezaron a comportar de manera más similar.

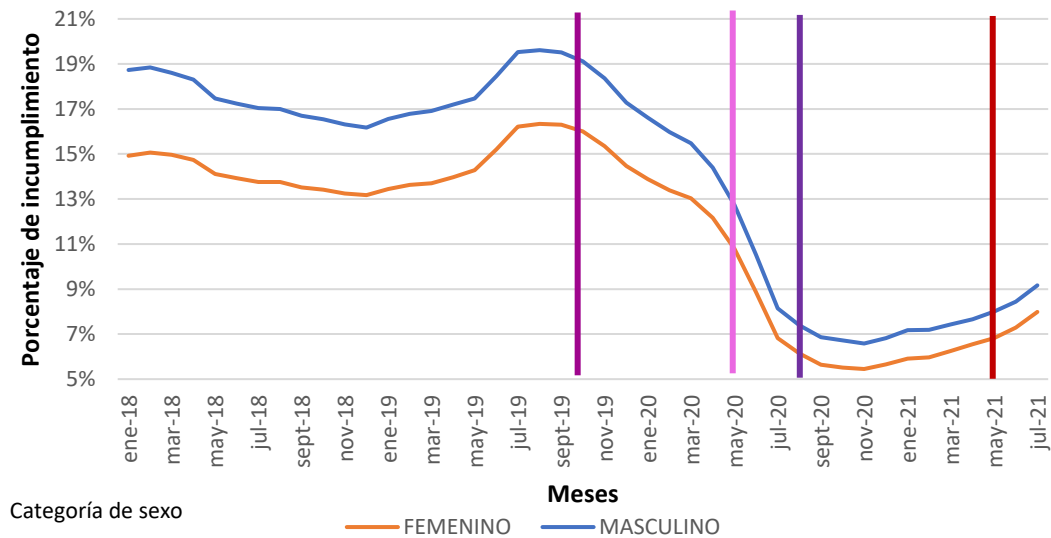
Gráfico N°29: “PD a través del tiempo para la variable sexo”



Fuente: Elaboración propia

Para el incumplimiento, se observa que la pendiente de caída es mayor entre período de crisis que la PD y posterior a las ayudas económicas las líneas de ambos son más unidas.

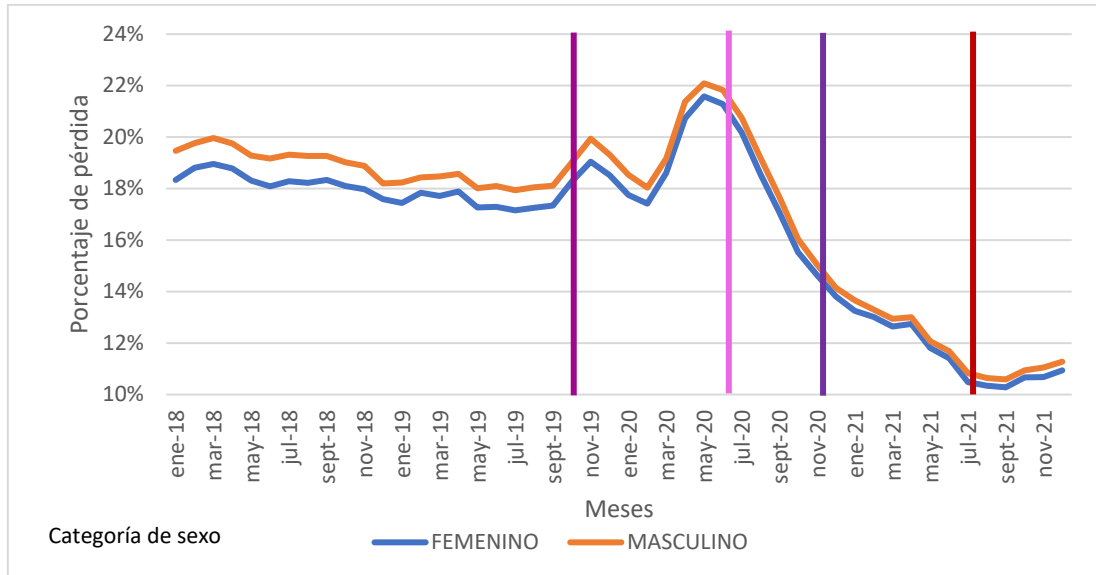
Gráfico N°30: “Porcentaje de incumplimiento a 12 meses según el sexo”



Fuente: Elaboración propia

Con respecto a la pérdida porcentual, esta tiene mayor similitud entre hombres y mujeres que en los otros gráficos anteriores, aminorando su diferencia posterior a la crisis. Es importante recalcar que tanto hombres como mujeres viven las mismas fluctuaciones a lo largo del gráfico.

Gráfico N°31: “Pérdida porcentual dependiendo del sexo que posea”



Fuente: Elaboración propia

En cuanto a las variables segmentación y la de producto más riesgoso, gráficos incluidos en el Anexo 4, ambas son creadas por el banco y como categorías tienen a gente que incumple. Debido a esto los gráficos tienen una tendencia a poseer valores tanto de PD como de incumplimiento altos. Sin embargo, igualmente es posible apreciar una baja del riesgo crediticio durante los períodos de ayuda económica.

La gran diferencia que se observa entre ambos gráficos de PD e incumplimiento son que las bajas son suaves en la PD, ya que ambas variables están predeterminadas previamente a tener un rango de PD, por lo que no varía mucho el comportamiento, en cambio, el incumplimiento si ve reflejado una baja significativa. Los grupos que se ven más afectados en estos cambios son los que están constituido por gente que incumple.

A través de los datos observados en los gráficos es posible concluir que los grupos se acercan entre sí durante el período de crisis. Esto, debido a la inyección de liquidez que aminora el riesgo dado que los clientes poseen los recursos para cumplir con sus compromisos económicos. Al término de estos períodos los grupos se vuelven a separar.

CAPÍTULO 6

6. Resultados

A partir de los resultados obtenidos en un primer análisis, se establecen los cimientos para establecer un modelo de puntaje de comportamiento que refleje el actuar de los clientes. Dicho modelo se describe en tres partes; primero, se dará a conocer la base de datos que se utilizó para su implementación, segundo, se detallarán las variables que fueron seleccionadas para ser parte del modelo y, en tercer lugar, se darán a conocer los resultados obtenidos, con sus respectivos puntajes por variables seleccionadas.

6.1 Base de datos

En la base de datos final, se tomó el mes de octubre del 2019 como fecha de corte para seleccionar los clientes que fueran activos y buenos, es decir, dando seguimiento a lo estipulado por el banco; que no estuvieran en incumplimiento ni pertenecieran al segmento 5. Al aplicar dicho criterio, la base de datos que en primer momento contaba con 2.824.606 registros pasó a tener 1.142.997.

Con la base de datos filtrada se tomó una ventana de tiempo catalogada como crisis, correspondiente al periodo existente entre noviembre del 2019 hasta mayo del 2020, momento en el que se materializa el pago de la primera de las ayudas económicas. Durante ese período de tiempo, se seleccionó a los clientes que incumplieran al menos una vez en la ventana de tiempo, quitando a aquellos que luego de haber incumplido fueron castigados por el banco en este mismo período. Las personas castigadas por el banco corresponden aquellas que tienen más de 180 días de mora.

Una vez obtenida esta base de datos, los registros con los que se cuenta son 78.880 y a estos clientes se les asigna una nueva variable, la de castigado o no castigado post período de ayudas. Esta variable es añadida, dado que los clientes que siguieron sin pagar posterior al período de ayuda son los clientes que no quisieron pagar, por lo tanto, los clientes Won't pay y los clientes que se pudieron recuperar, pagando sus deudas son los Can't pay.

6.2 Variables seleccionadas

Las variables seleccionadas fueron las que previo a un análisis univariado y multivariado se decidieron incluir en el posible modelo y obtener su WOE e IV, quedando las con mayor poder predictivo finalmente en él. Esto, con la ayuda de la función *scorecard* del software *Rstudio*. Estas variables serán mostradas con sus aperturas y su gráfica de WOE, otorgando nombres de lo que representa, pero no del nombre real de la variable por temas de confidencialidad.

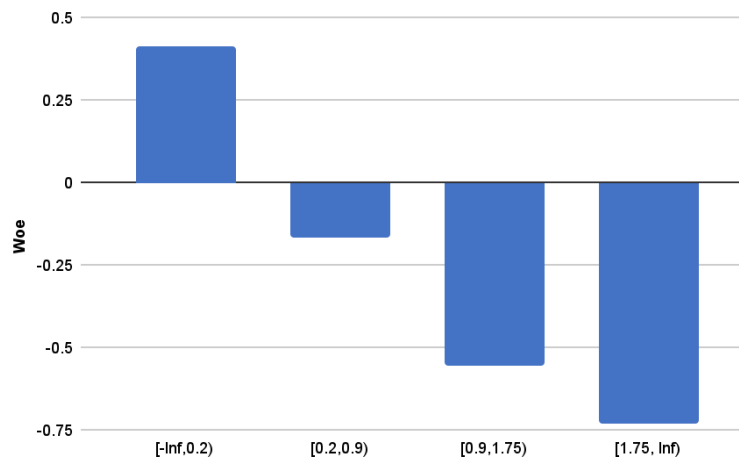
6.2.1 Variable 1: Tendencia de los días de mora de un cliente

Tabla N°2:” Tabla de WOE e IV para variable 1”

Variable 1	WOE	Contribución IV
$[-\text{Inf}, 0.2)$	0.413	0.082
$[0.2, 0.9)$	-0.165	0.007
$[0.9, 1.75)$	-0.555	0.054
$[1.75, \text{Inf})$	-0.730	0.046
Total IV		0.188

Fuente: Elaboración propia

Gráfico N°32:” WOE de la tendencia de los días de mora de un cliente”



Fuente: Elaboración propia

En el gráfico N°32, es posible observar un comportamiento lineal de los WOE, que es lo que se espera obtener en estos gráficos para poder considerar la variable en el modelo. Igualmente, en la tabla N°2 se puede ver que aquellos que tienen menores valores aportan más información al modelo.

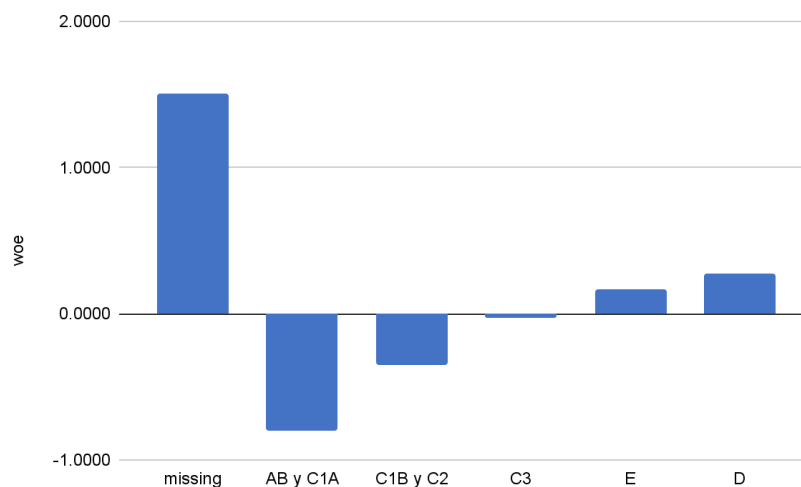
6.2.2 Variable 2: Grupo socioeconómico de los clientes

Tabla N°3:” Tabla de WOE e IV para variable 2”

Variable 2	WOE	Contribución IV
NA	1.503	0.019
AB y C1A	-0.805	0.035
C1B y C2	-0.355	0.021
C3	-0.029	0.000
E	0.168	0.005
D	0.278	0.019
Total IV		0.100

Fuente: Elaboración propia

Gráfico N°33:” WOE de los grupos socioeconómicos de los clientes”



Fuente: Elaboración propia

Al observar el gráfico N°33 se observa que el comportamiento no es lineal, dado que existe datos faltantes (missing), pero como es una variable sin nombre y creada por el modelo, está podría agregarse y aparecer al final de la secuencia, por lo que se considerará la trayectoria desde el segundo grupo correspondiente a AB y C1A hasta D, donde se cumple la trayectoria lineal que se necesita. En cuanto al grupo que mayor IV aporta dada la tabla N°3 es el grupo de AB y C1A.

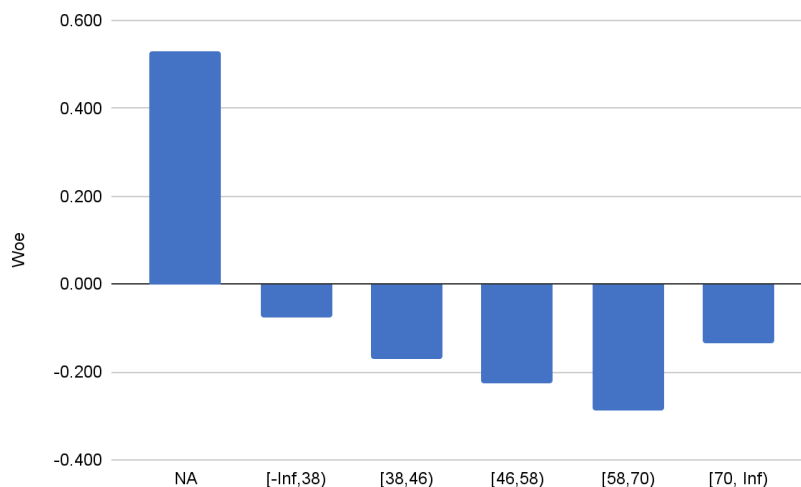
6.2.3 Variable 3: Edad

Tabla N°4: “Tabla de WOE e IV para variable 3”

Variable 3	WOE	Contribución IV
missing	0.529	0.070
[-Inf,38)	-0.077	0.001
[38,46)	-0.170	0.004
[46,58)	-0.226	0.010
[58,70)	-0.288	0.012
[70, Inf)	-0.134	0.001
Total IV		0.099

Fuente: Elaboración propia

Gráfico N°34: “WOE de edad de los clientes”



Fuente: Elaboración propia

La variable edad, la cual presenta un comportamiento lineal descendente en el gráfico N°34 hasta la edad de los 70, donde vuelve a bajar, este comportamiento no es el esperado para considerar la variable, sin embargo, se sabe que a los 70 años las personas son jubiladas, por lo que sus ingresos disminuyen y si presentan un comportamiento más riesgoso que en el rango anterior, por lo que se explica esta subida de WOE. En cuanto a la categoría que presenta mayor influencia en el IV de la variable, se puede observar en la tabla N°4 que corresponde a la que está compuesta por valores vacíos, es decir a gente que no declaró su edad. Pese a que esta variable presenta un comportamiento anómalo, será incluido en el modelo, ya que este comportamiento tiene relación con lo que verdaderamente ocurre.

Por último, los gráficos de las otras once variables están incluidas en el Anexo 5, junto con sus respectivos nombres. Todas las gráficas del anexo presentan un comportamiento similar al de la variable 1. Es decir, su tendencia es lineal, tanto ascendente como descendente. Lo que significa, que estas variables son aptas para el modelo.

Para finalizar con esta sección, se presentan los IV totales en orden, según el valor de información que tienen las distintas variables seleccionadas:

Tabla N°5: “Tabla de IV para todas las variables”

Número de variable IV total	
Variable 6	0.198
Variable 1	0.188
Variable 12	0.140
Variable 7	0.102
Variable 4	0.100
Variable 2	0.100
Variable 3	0.099
Variable 8	0.095
Variable 9	0.092
Variable 14	0.091
Variable 13	0.090
Variable 10	0.077
Variable 11	0.052
Variable 5	0.048

Fuente: Elaboración propia

Es importante señalar, que el objetivo del análisis del valor WOE era observar si presentaba una consistencia con los valores obtenidos, a través de gráficas, en donde se buscaba que tuvieran una tendencia lineal, ya sea ascendente o descendente. Este valor es relevante, ya que a posterior permite asignar los puntajes de comportamiento para

cada tramo de las variables, por lo que se busca que este sea asignado de forma coherente.

6.3 Resultados del modelo

El modelo final que se mostrará a continuación corresponde a la regresión logística con sus respectivos betas.

Tabla N°6: “Resultado modelo de regresión”

Variables	Beta	Std. Error	Significancia
Intercepto	-0.11709	0.008	0.000
Variable 1	0.28927	0.075	0.000
Variable 2	0.59261	0.028	0.000
Variable 3	-0.72211	0.109	0.000
Variable 4	0.35013	0.028	0.000
Variable 5	0.42337	0.044	0.000
Variable 6	0.59582	0.073	0.000
Variable 7	1.06357	0.079	0.000
Variable 8	-0.36862	0.110	0.000
Variables 9	0.30375	0.042	0.000
Variable 10	0.32099	0.042	0.000
Variable 11	0.54148	0.057	0.000
Variable 12	0.50577	0.037	0.000
Variable 13	0.09482	0.047	0.045
Variable 14	-0.11465	0.050	0.024

Fuente: Elaboración propia

Igualmente, se muestra el resultado obtenido en el puntaje de comportamiento asociado a las variables malo como castigado y bueno como no castigado, junto con la distribución de puntajes de la base en general que se encuentra en ANEXO 6.

Tabla N°7: “Tabla de puntaje de comportamientos”

Variable	Separaciones	Puntos
Basepoints		396
Variable 1	[-Inf,0.2)	-9
	[0.2,0.9)	3
	[0.9,1.75)	12
	[1.75, Inf)	15
Variable 2	[-Inf,380)	-10
	[380,390)	-1
	[390,420)	3
	[420, Inf)	8
Variable 3	Missing	-64
	AB%,%C1a	34
	C1b%,%C2	15
	C3	1
	E	-7
	D	-12
Variable 4	missing	-3
	REF_15%,%Repactados y Renegociados%, %REF_30%,%REF_CERO	10
	Consumo_y_SAV	5
	Avance%,%Compra	-5
Variable 5	[-Inf,2)	-19
	[2,18)	-2
	[18,30)	7
	[30, Inf)	27
Variable 6	FEMENINO	20
	MASCULINO	4
	SIN INFORMACION	-41
Variable 7	ESTUDIANTE, RETENCION JUDICIAL Y DEPENDIENTE	-6
	CESANTE, DUEÑA DE CASA, INDEPENDIENTE Y JUBILADO	-4
	SIN INFORMACION	14

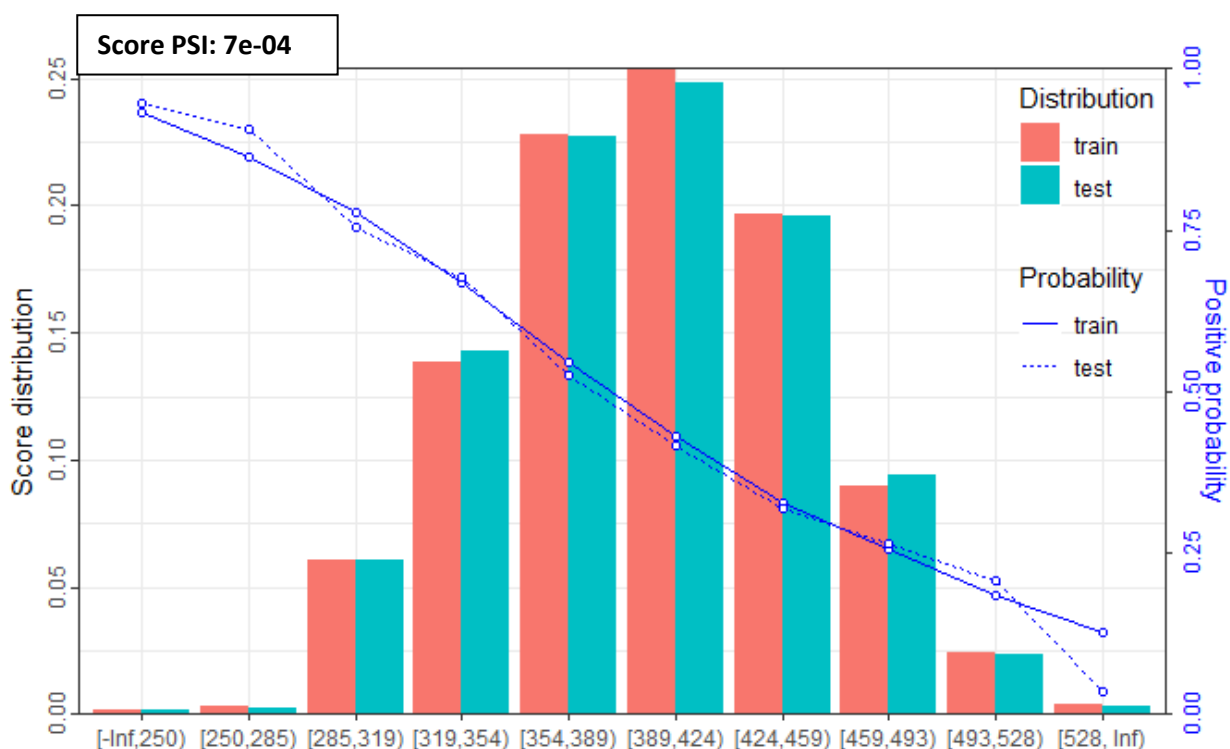
Variable	Separaciones	Puntos
Variable 8	[-Inf,1.000)	-8
	[1.000,4.000)	0
	[4.000,22.500)	4
Variable 9	[22.500, Inf)	13
	[-Inf,2)	-10
	[2,3)	-2
Variable 10	[3,4)	4
	[4, Inf)	7
	[-Inf,200)	-7
	[200,400)	-1
Variable 11	[400,900)	3
	[900,3.500)	9
	[3.500, Inf)	22
	[-Inf,2)	-23
	[2,4)	-6
Variable 12	[4,16)	7
	[16, Inf)	17
	[-Inf,400)	-2
	[400,2.000)	1
	[2.000,7.800)	2
Variable 13	[7.800, Inf)	4
	[-Inf,500)	3
	[500,1.000)	1
	[1.000,4.500)	-1
Variable 14	[4.500, Inf)	-3
	NA	28
	[-Inf,38)	-4
	[38,46)	-9
	[46,58)	-12
	[58,70)	-15
	[70, Inf)	-7

Fuente: Elaboración propia

El resultado obtenido también está complementado por dos gráficos, el primero el gráfico N°34. En él se puede observar que la proporción entre la base de entrenamiento y testeo es similar, que es lo que requiere el modelo, tanto en la probabilidad como en la distribución. A la vez de conocer su índice de estabilidad de la población (PSI), que en este caso es de 0.0007. Cuando los valores de PSI son menos que 0.1 implican que el modelo fue bien ejecutado y hay cambios insignificantes si se quiere mejorar, por lo que no requiere ninguna acción de mejora.

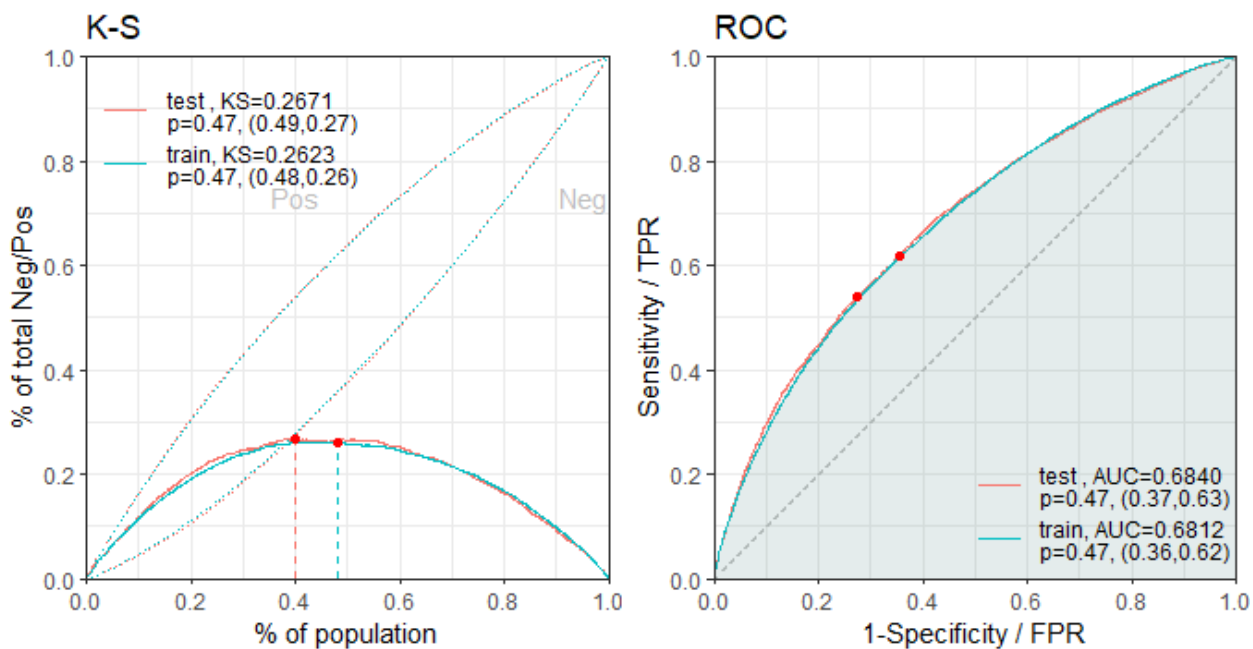
En el segundo gráfico N°35, se muestra el indicador Kolmogórov-Smirnov (KS) obtenido, que mide la bondad de ajuste entre dos distribuciones de probabilidad, en este caso alcanza el 26,7% y el gráfico de Característica Operativa del Receptor (ROC), que muestra el área bajo la Curva (AUC) de un modelo. El AUC mide la sensibilidad de la representación de un modelo para una clasificación binaria, el que alcanzó un 0,68.

Gráfico N°35: “Resultados del modelo - Distribución y probabilidad”



Fuente: Rstudio - elaboración propia

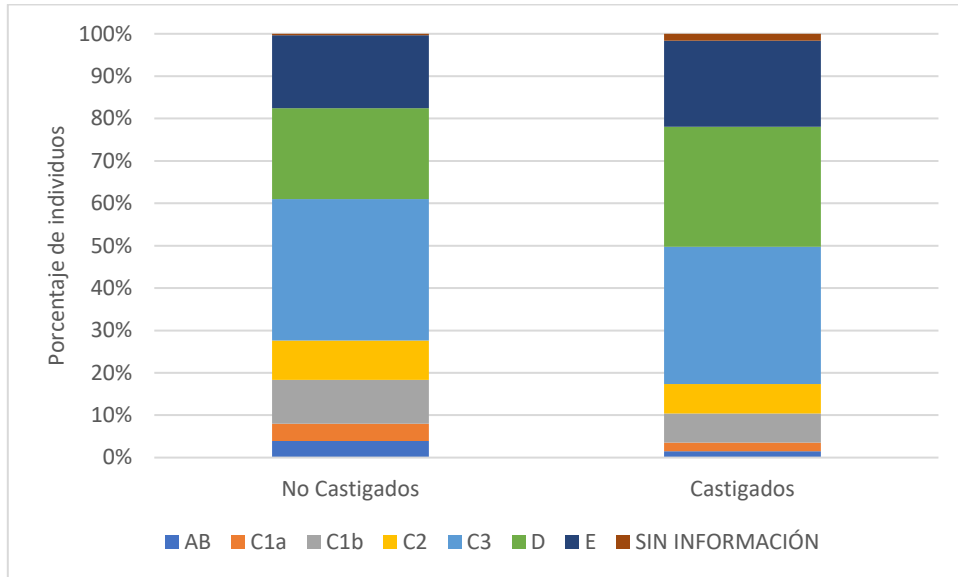
Gráfico N°36: “Resultados del modelo - KS y ROC”



Fuente: Rstudio - elaboración propia

Otro resultado, es la obtención de las variables finales que sirven para discriminar a un cliente que será castigado de uno que no. Se realizan los gráficos correspondientes a las variables, junto con su distribución de las categorías en los clientes castigados y no castigados. Estos se muestran a continuación:

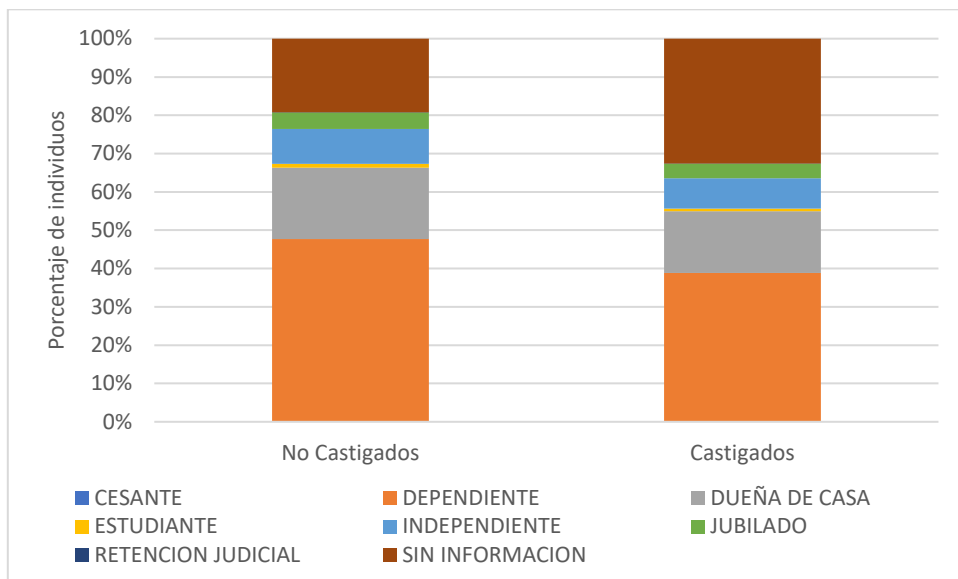
Gráfico N°37: “Grupos socioeconómicos de los clientes”



Fuente: Elaboración propia

En cuanto a la distribución de clientes según los grupos socioeconómicos en el grupo de no castigados y castigados predomina el grupo C3, pero donde existen mayores diferencias entre ambos grupos son los grupos etarios E y D.

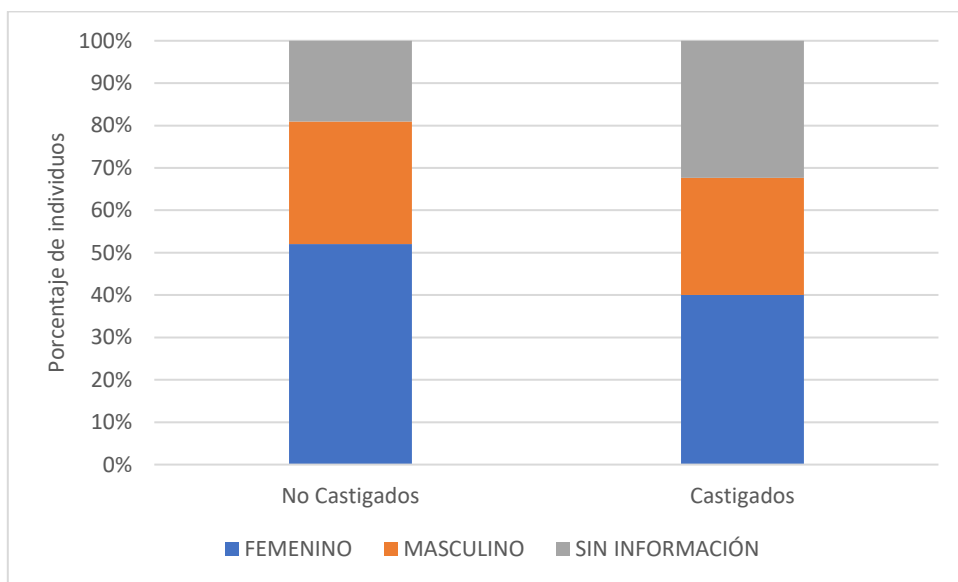
Gráfico N°38: “Producto más riesgoso que posee el cliente”



Fuente: Elaboración propia

Si se observa la composición del grupo de no castigados y castigados, domina el segmento de compra, sin embargo, esa preponderancia es aún mayor en los castigados.

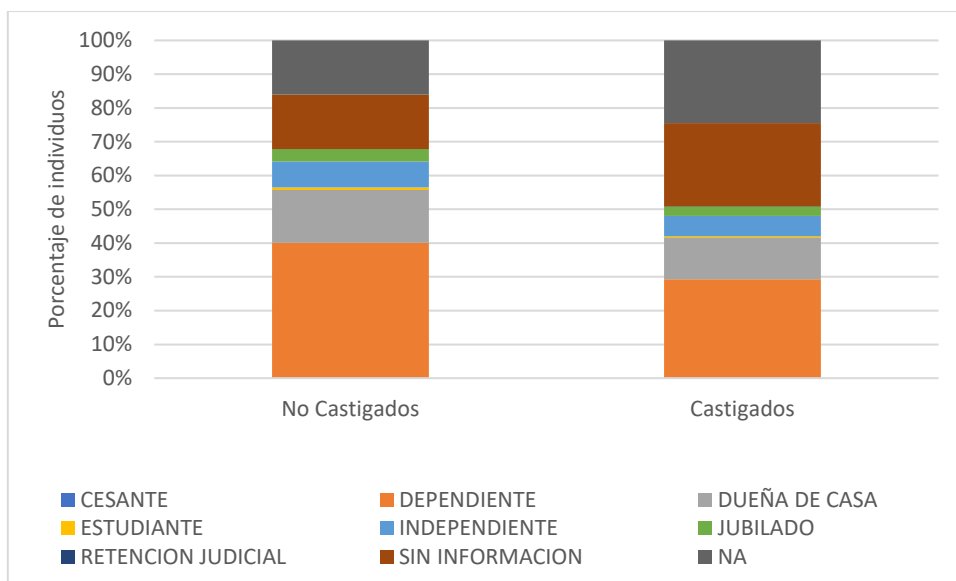
Gráfico N°39: “Sexo de los clientes”



Fuente: Elaboración propia

En este gráfico se puede apreciar la preponderancia del género femenino tanto en no castigados como castigados, sin embargo, se puede acotar que el grupo sin información también tiene un gran porcentaje en castigados.

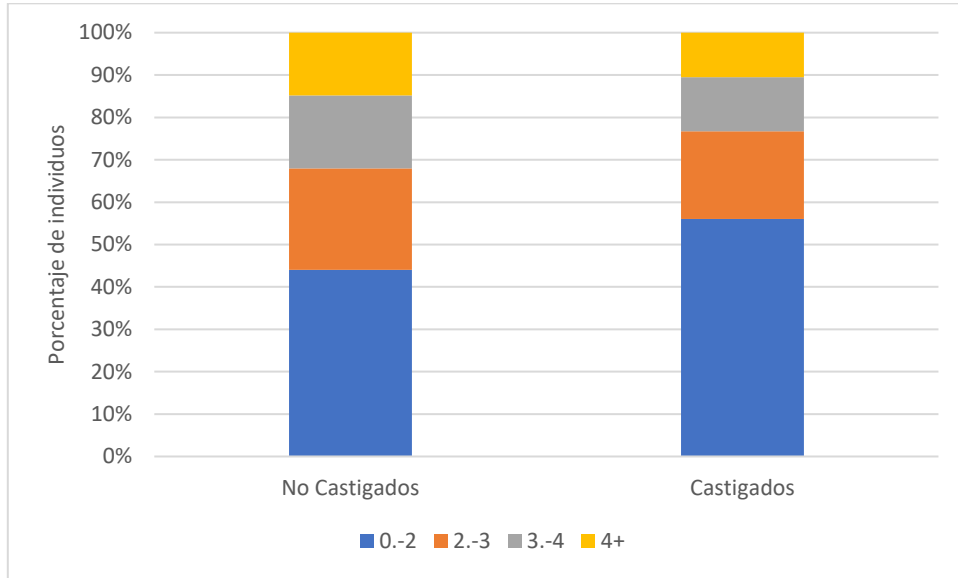
Gráfico N°40: “Actividad que realiza el cliente”



Fuente: Elaboración propia

El grupo de no castigados y castigados está compuesto en su mayoría por personas que trabajan de forma dependiente. Mientras que la mayor diferencia entre grupos corresponde a las personas sin información.

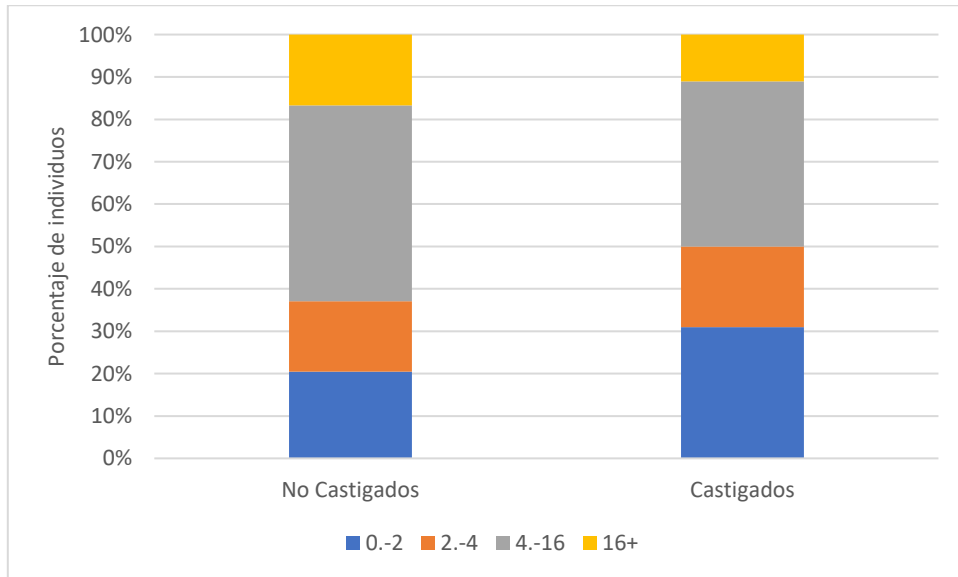
Gráfico N°41: “Número de instituciones en las que el cliente tiene crédito”



Fuente: Elaboración propia

Para los grupos de no castigados y castigados, la mayoría de ellos presenta clientes con 1 o 2 instituciones en las que tiene créditos, aumentando el porcentaje en los clientes que están castigados.

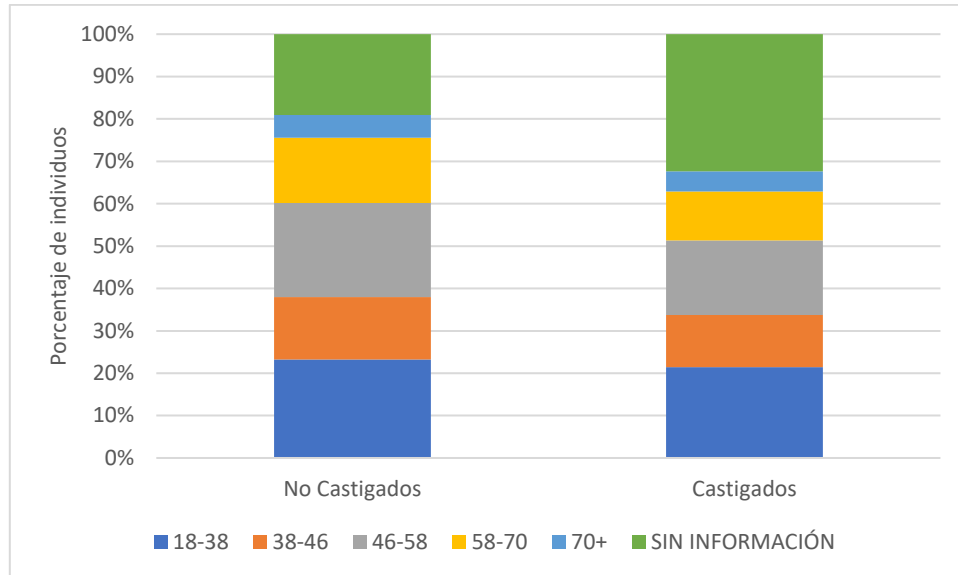
Gráfico N°42: “Antigüedad en años del cliente en el banco”



Fuente: Elaboración propia

En cuanto a la variable antigüedad y su relación con los clientes castigados y no castigados, se puede observar que los clientes castigados en su mayoría presentan una antigüedad menor a 4 años, mientras los no castigados poseen en su mayoría una antigüedad por sobre los 4 años.

Gráfico N°43: “Edad de los clientes”



Fuente: Elaboración propia

Para la variable edad de los clientes, su relación con los no castigados y castigados se ve la preponderancia de los jóvenes y de los que poseen valores no disponibles (NA), siendo la mayor cantidad de castigados atribuidos a esta categorización.

Las demás variables se pueden encontrar en Anexo 7, junto a su descripción. Estas variables, son dependientes de las conductas de los individuos, por lo que estos pueden cambiar sus conductas para poder mejorar su puntaje. Mientras, que las anteriormente nombradas son en su mayoría de carácter demográfico.

CAPÍTULO 7

7. Conclusiones y recomendaciones

A través de este estudio se puede observar un análisis descriptivo de clientes que son considerados como buenos y malos por la entidad bancaria. Se permitió diferenciar el comportamiento de estos a través de dos períodos, uno pre crisis y otro post crisis.

Al exponer los hallazgos sobre el comportamiento de los datos en estos grupos, se puede aludir que, en el caso de buenos clientes, la variable antigüedad refleja diferencias respecto del grupo de malos. Esta variable apunta a que la mayor parte de clientes buenos (Can't pay) poseen datos de antigüedad mayor a 4 años. Proporción que disminuye en el grupo de malos clientes, en donde, solo la mitad de sus integrantes cumpliría esta condición.

Por su parte, para los clientes malos, las variables edad de los clientes y saldo de deuda mostrarían resultados relevantes.

En el caso de la primera, la conformación del grupo etario mostraría que casi un tercio de este está compuesto por personas cuya edad oscila entre los 50 y 55 años. Por su parte, para el grupo de los buenos clientes, solo 1/5 se encontraría en dicho rango de edad.

En el grupo de clientes malos, la variable monto disponible para crédito evidenció que el grupo estaba compuesto por un 56 % de personas con capacidad de endeudamiento sobre a \$200.000, mientras que en los clientes buenos esta cifra aumenta al 69%.

Los resultados finales de la tesis, eventualmente, podrían aportar a las entidades bancarias a contar con más información predictiva para discriminar si el comportamiento de un cliente será bueno o malo, en otras palabras, si se comportaran como can't pay o won't pay.

En el modelo se pueden destacar el comportamiento que las 14 variables tienen en estos clientes. Las primeras a destacar son las relacionadas con los días de mora que tiene el cliente, estas variables son evaluadas en tendencia, el total y como variable sola. Las tres presentan un alto poder predictivo en el modelo. Sin embargo, esta relación es conocida y tiene que ver con un comportamiento que lleva a cabo el cliente, más que una característica de él. Por tal, las variables que más interés le generan al banco son las variables demográficas, que tiene relación a las características de una persona.

Entre las características demográficas que destacaron, se tiene a la variable que señala la antigüedad del individuo en el banco, el sexo del individuo, el grupo socioeconómico del cliente y por último la edad que este tiene. Por lo que un cliente a

más años de antigüedad, con sexo femenino y perteneciente a los grupos más altos socioeconómicos (AB1 y C1A) posee mejor puntaje de comportamiento, por ende, es mejor cliente para el banco. En el caso de la edad, surgió la disyuntiva de comportamiento que el comportamiento no es lineal y dado un cierto tope de edad, que en este caso fue 70 años vuelve a tener comportamiento similar a los de 38-46 años. Esta relación es común que ocurra, ya que incluye a la gente jubilada, quienes tienen menor ingreso que en su vida laboral activa.

En cuanto al análisis de riesgo, en todas las variables la crisis trajo consigo un período de alza en el riesgo de los clientes, sin embargo, las ayudas económicas aportaron a que hubiera más liquidez en la población, lo que se reflejó en los clientes del Banco Ripley. Esto hizo que el riesgo se aminorara en el periodo que abarca desde la asignación del primer IFE hasta la entrega del último retiro. Por consiguiente, independiente de la variable que describa a la persona su riesgo bajó.

Es importante apreciar que los valores omitidos (NA) tienen un efecto considerable al momento de discriminar el comportamiento de los clientes. La alta cifra de NA se puede deber a campañas realizadas por el banco y su meta de llegar a cierta cantidad de clientes, por lo que no solicitan estos datos y por ende estos quedan guardados como NA. Por lo tanto, descubrir que estos valores marcan una diferencia, podría mejorar la capacidad de predicción de modelos del banco, en cuanto a cómo pagarán estos clientes.

En cuanto a los resultados de la precisión del modelo, los valores de KS presentan un valor de 26%, que es una métrica baja para el banco, sin embargo, el AUC supera lo mínimo que debiese tener con un 0.68, por lo que se pueden considerar como válidos los resultados obtenidos y utilizar la métrica encontrada.

Dado los resultados anteriores, tanto de riesgo como caracterización del tiempo de cliente se puede corroborar lo planteado por C. Bravo en su estudio (Bravo Román, C, 2012). Los clientes jóvenes no se preocupan de su reputación y tienen más probabilidad de tomar responsabilidades que no pueden cumplir en la posterioridad, por lo que tienden a tener altos niveles de riesgo y tener gran presencia en el grupo de malos.

La misma situación ocurre con clientes que llevan poco tiempo en el banco, estos no se preocupan de su reputación, porque no tienen intenciones de permanecer por largo tiempo en el banco, en cambio, los que llevan más tiempo se dedican a ser buenos clientes para mantener su reputación y así poder optar a mejores créditos.

La riqueza de este trabajo radica en el hecho de haber analizado un amplio universo de variables que incorporan información proveniente de diversas fuentes. Es por esto que se puede hacer uso de los resultados obtenidos para futuros análisis del comportamiento de los clientes y para poder mejorar su poder predictivo. Por otro lado, entrega un análisis del modelo con una nueva clasificación de clientes, lo que le permite a la empresa analizar posibilidades de mejora para el modelo con el que ellos trabajan.

Como trabajo futuro sería interesante poder agregar variables demográficas que incluyan los ingresos de los clientes mensualmente en el modelo y enfocar los análisis en mejorar el KS obtenido del modelo.

Interesante sería aplicar este estudio en una entidad bancaria dependientes del estado para identificar si existen diferencias en el comportamiento de clientes respecto de la entidad privada. Esto dado que el umbral de clientes entre los distintos bancos varía y las variables de interés que nos muestran los modelos podrían diferir.

Bibliografía

- AIM. (s.f.). Asociación de Investigación de Mercado y Opinión Pública. Obtenido de <https://aimchile.cl/gse-chile/>
- Bravo Román, C. (2012). Métodos para estimar riesgo crediticio en base a minería de datos y teoría de juegos. Disponible en <https://repositorio.uchile.cl/handle/2250/115063>
- Central, B. https://www.bcentral.cl/documents/33528/3245688/IEF_2021_semestre2.pdf/2cf22294-5e9a-9ff0-52e2-ae0ed4e1d2e6. Obtenido de Informe de Estabilidad Financiera Segundo Semestre 2021.
- CEPAL. (febrero de 2021). CEPAL ORG. Obtenido de https://repositorio.cepal.org/bitstream/handle/11362/46504/10/PO2020_Chile_en.pdf
- Chile Atiende. (2022). Obtenido de <https://www.chileatiende.gob.cl/fichas/81027-retiros-del-10-de-los-fondos-de-afp>
- Dirección de Presupuesto - Gobierno de Chile. (28 de septiembre de 2021). DIPRES. Obtenido de https://www.dipres.gob.cl/598/articles-244739_Informe_PDF.pdf
- Flora M Díaz-Pérez, M. B.-C. (2006). *CHAID algorithm as an appropriate analytical method for tourism market segmentation*. Journal of Destination Marketing & Management, Volume 5, Issue 3.
- Ministerio de desarrollo social y familia. (2020). *Preinforme ingreso familiar de emergencia*.
- Pineda, S. C. (2009). *Comparación de Árboles de Regresión y clasificación y regresión logística*. Colombia.
- Siddiq, N. (2006). *Credit Risk Scorecards - Developing and Implementing Intelligent Credit Scoring*. Hoboken, New Jersey: John Wiley & Sons, Inc.
- Superintendencia de Bancos e Instituciones Financieras (SBIF). (diciembre de 2014). *Compendio de Normas Contables, Capítulo B-1*. Obtenido de http://www.sbif.cl/sbifweb3/internet/archivos/norma_10635_3.pdf
- Thomas, L. C. (2009). *Consumer Credit Models: Pricing, Profit, and Portfolios*. Oxford New York: Oxford university press.

Vanesa Berlanga Silvente, M. J. (2013). Cómo aplicar árboles de decisión en SPSS.
Revista d'Innovació i Recerca en Educació.

Anexos

Anexo A: Índice PSI de los clientes buenos

Tabla N°8: “Tabla de PSI para clientes buenos según variable antigüedad en años”

CATEGORÍA	PSI
0.-2	59.81
3	6.48
4	0.03
5.-6	1.20
7.-8	2.64
9.-14	0.15
15.-17	0.07
18.-20	0.47
21+	8.63
TOTAL PSI	79.49

Fuente: Elaboración propia

Tabla N°9: “Tabla de PSI para clientes buenos según variable edad”

CATEGORÍA	PSI
18-31	0.146
32-35	0.000
36-40	0.044
41-45	0.060
46-49	0.041
50-54	0.024
55-58	0.001
59-63	0.000
64-69	0.004
70+	0.164
TOTAL PSI	0.485

Fuente: Elaboración propia

Tabla N°10: “Tabla de PSI para clientes buenos según variable monto de deuda”

CATEGORÍA	PSI
0-\$24.000	0.886
\$24.001-\$55.500	0.868
\$55.501-\$106.000	0.353
\$106.001-\$180.000	0.242
\$180.000-\$275.000	0.181
\$275.001-\$699.000	0.026
\$699.001-\$1.132.000	0.130
\$1.132.001-\$1.886.000	0.073
\$1.886.001 +	0.132
TOTAL PSI	2.890

Fuente: Elaboración propia

Tabla N°11: “Tabla de PSI para clientes buenos según variable producto más riesgoso en cartera”

CATEGORÍAS	PSI
Avance	0.817
Compra	0.152
Consumo_y_SAV	0.630
REF_15	0.004
REF_30	0.000
REF_CERO	0.054
Repactados y Renegociados	1.635
TOTAL PSI	3.292

Fuente: Elaboración propia

Tabla N°12: “Tabla de PSI para clientes buenos según variable antigüedad en años”

CATEGORÍAS	PSI
ss1A Bancarizado y al día	1.211
ss1B No bancarizado o moroso	3.129
ss2A Clean U3M	0.473
ss2B Dirty U3M	0.048
ss3A Refinanciados con DM	0.003
ss3B Renegociados o Repactados	2.082
ss4A Refinanciados 0 DM	0.040
ss4B SAV/Consumo >= 2MM\$	0.012
ss4C AV, SAV/Consumo < 2MM\$	1.786
TOTAL PSI	8.784

Fuente: Elaboración propia

Tabla N°13: “Tabla de PSI para clientes buenos según sexo”

CATEGORÍAS PSI	
FEMENINO	0.031
MASCULINO	0.059
TOTAL PSI	0.091

Fuente: Elaboración propia

Anexo B: Descripción de categorías de variable producto más riesgoso de la cartera del cliente

- Refinanciamiento: Es una campaña que ofrece Banco Ripley a todos los clientes que poseen un crédito de consumo vigente (no renegociado) en el Banco, con el fin de modificar las condiciones iniciales de su crédito vigente.

- Ref_Cero: Es un refinanciamiento de clientes al día.

- Ref_15: Es un refinanciamiento de clientes con 1 a 15 días de mora.

- Ref_30: Es un refinanciamiento de clientes con 16 a 30 días de mora.

- Renegociación (car) o Repactados (Banco): Es una campaña que ofrece Banco Ripley a todos los clientes con operaciones vigentes en Banco Ripley al momento de la evaluación. Su objetivo es adecuar el plan de pago a condiciones viables para el cliente y aceptadas por el Banco.

- SAV: Producto de libre disposición para clientes. Estos giros de dinero se realizan desde los terminales financieros de las Tiendas Ripley y la red de sucursales de Banco Ripley.

- Avance: Producto de libre disposición para clientes. Estos giros de dinero se realizan desde los terminales financieros de las Tiendas Ripley y la red de sucursales de Banco Ripley.

Anexo C: Índice PSI de los clientes malos

Tabla N°14: “Tabla de PSI para clientes malos según variable antigüedad en años”

CATEGORÍA	PSI
0.-2	110.721
3	6.297
4	0.062
5.-6	2.287
7.-8	3.214
9.-14	1.930
15.-17	1.514
18.-20	0.938
21+	8.177
TOTAL PSI	135.141

Fuente: Elaboración propia

Tabla N°15: “Tabla de PSI para clientes malos según variable edad”

Categorías	PSI
18-31	5.05
32-35	1.06
36-40	0.39
41-45	0.03
46-49	0.01
50-54	0.08
55-58	0.31
59-63	1.02
64-69	1.86
70+	3.20
TOTAL PSI	12.99

Fuente: Elaboración propia

Tabla N°16: “Tabla de PSI para clientes malos según variable saldo de deuda”

Categorías	PSI
0-\$24.000	8.14
\$24.001-\$55.500	0.71
\$55.501-\$106.000	0.45
\$106.001-\$180.000	1.00
\$180.000-\$275.000	3.22
\$275.001-\$699.000	1.26
\$699.001-\$1.132.000	0.05
1.132.001-\$1.886.000	0.01
\$1.886.001 +	0.75
TOTAL PSI	15.59

Fuente: Elaboración propia

Tabla N°17: “Tabla de PSI para clientes malos según producto más riesgoso”

Categorías	PSI
Avance	0.00
Compra	0.01
Consumo_y_SAV	0.04
REF_15	0.45
REF_30	0.19
REF_CERO	0.20
Repactados y Renegociados	0.08
TOTAL PSI	0.97

Fuente: Elaboración propia

Tabla N°18: “Tabla de PSI para clientes malos según segmento de categorización”

Categorías	PSI
ss1A Bancarizado y al día	0.00
ss1B No bancarizado o moroso	8.13
ss2A Clean U3M	0.00
ss2B Dirty U3M	0.05
ss3A Refinanciados con DM	0.61
ss3B Renegociados o Repactados	0.02
ss4A Refinanciados 0 DM	0.16
ss4B SAV/Consumo >= 2MM\$	0.16
ss4C AV, SAV/Consumo < 2MM\$	0.01
TOTAL PSI	9.12

Fuente: Elaboración propia

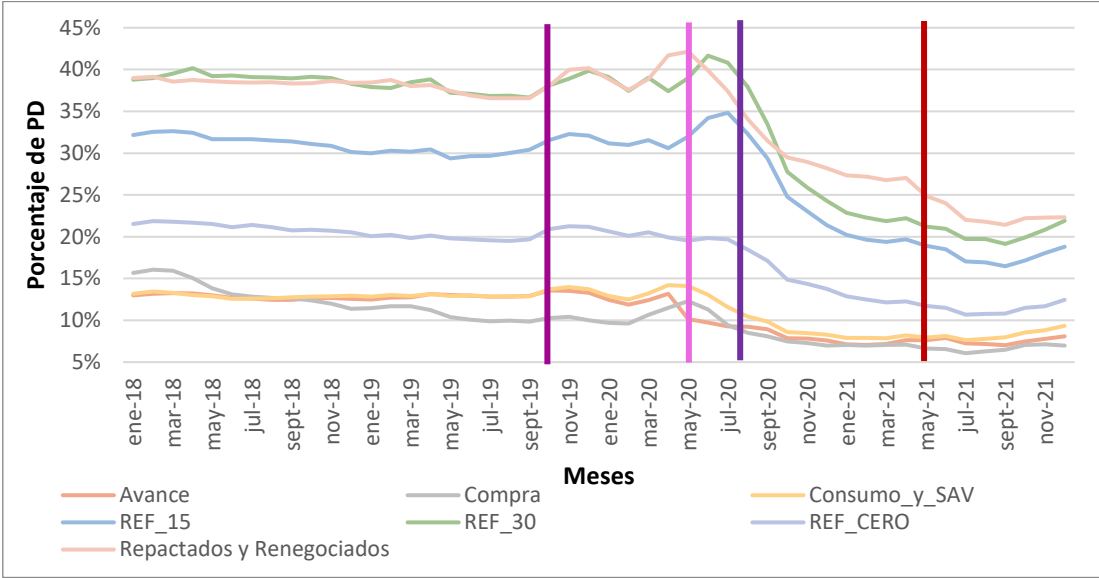
Tabla N°19: “Tabla de PSI para clientes malos según sexo”

CATEGORÍA	PSI
FEMENINO	0.156
MASCULINO	0.226
TOTAL PSI	0.381

Fuente: Elaboración propia

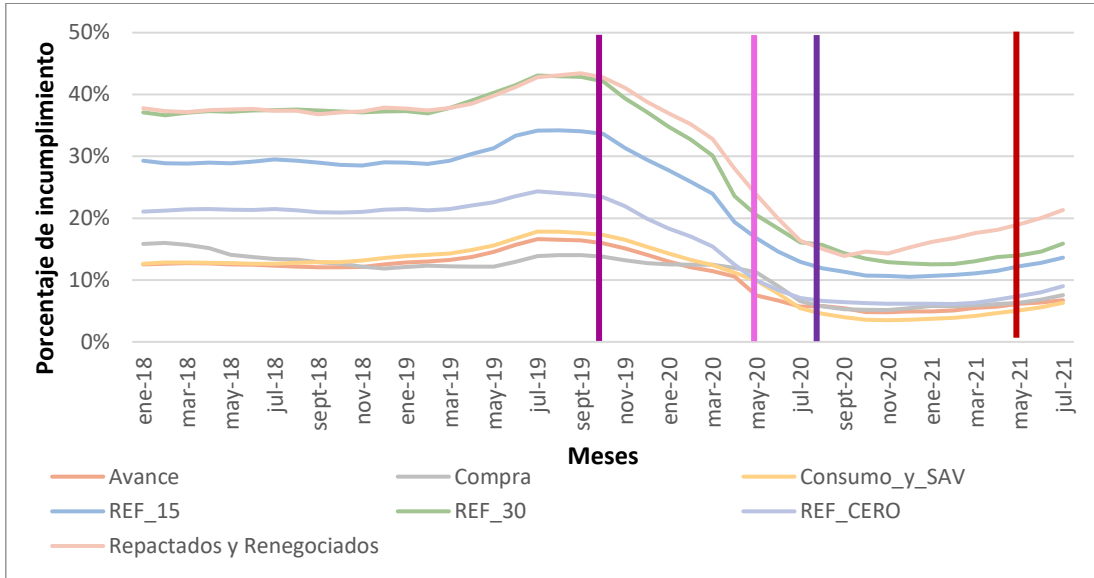
Anexo D: Gráficas de riesgo para la variable subsegmento de clientes y producto más riesgoso de la cartera

Gráfico N°44: “PD a través del tiempo”



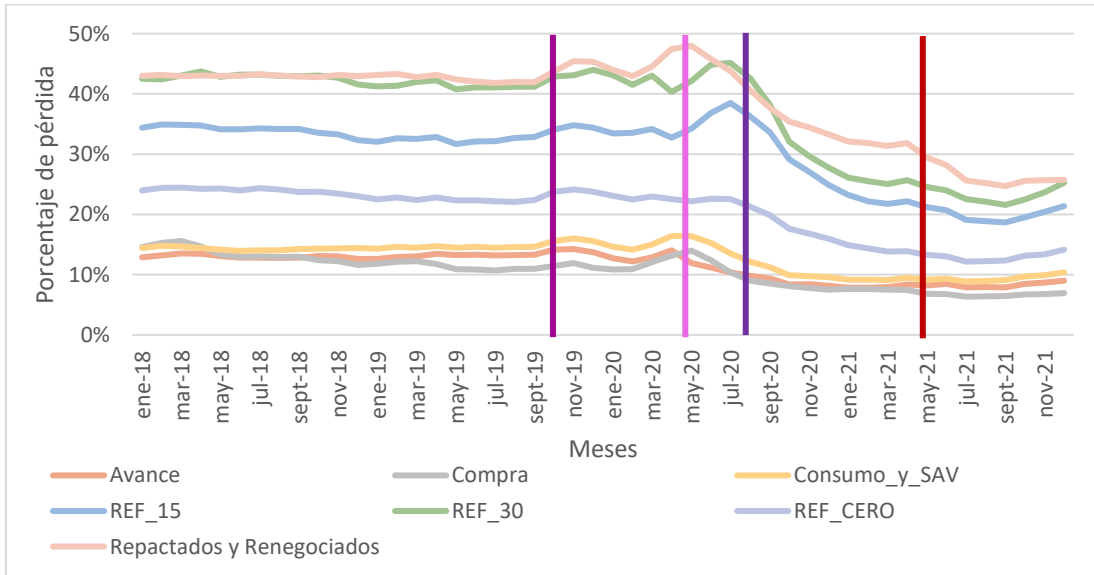
Fuente: Elaboración propia

Gráfico N°45: “Porcentaje de incumplimiento a través del tiempo”



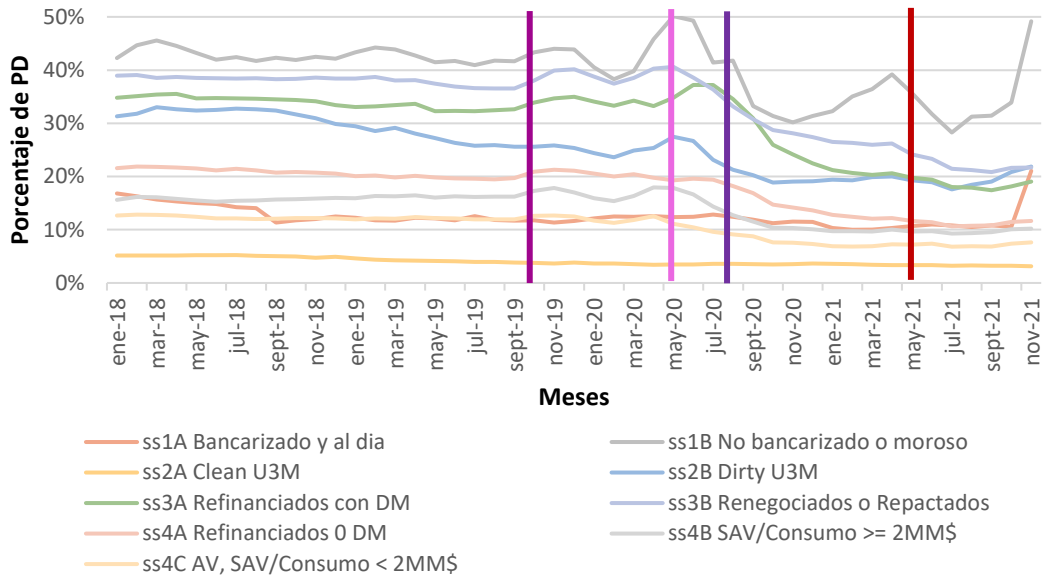
Fuente: Elaboración propia

Gráfico N°46: “Pérdida porcentual dado el producto más riesgoso que tiene en cartera el cliente”



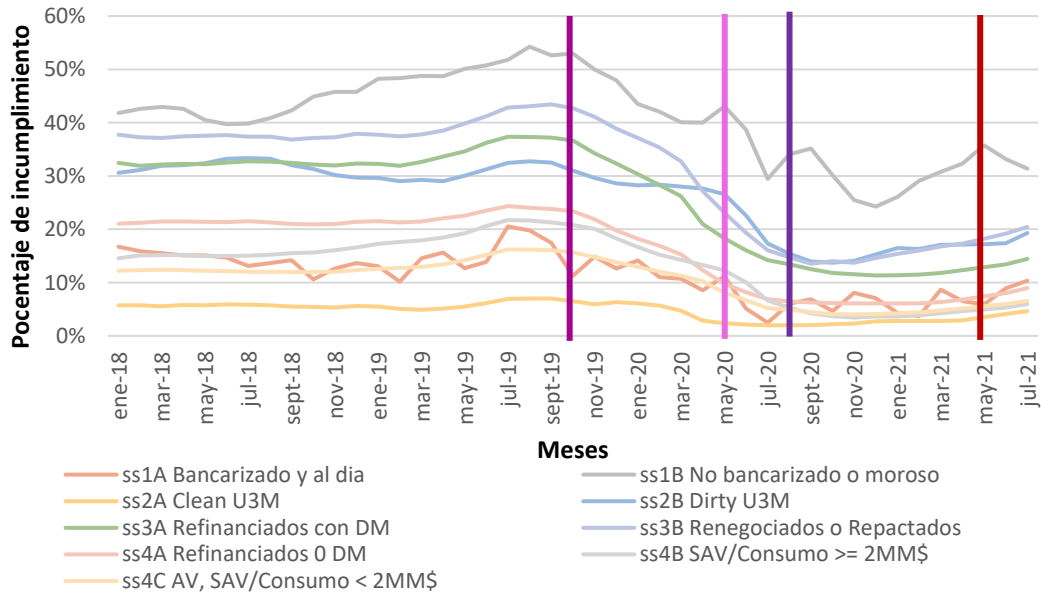
Fuente: Elaboración propia

Gráfico N°47: “PD a través del tiempo”



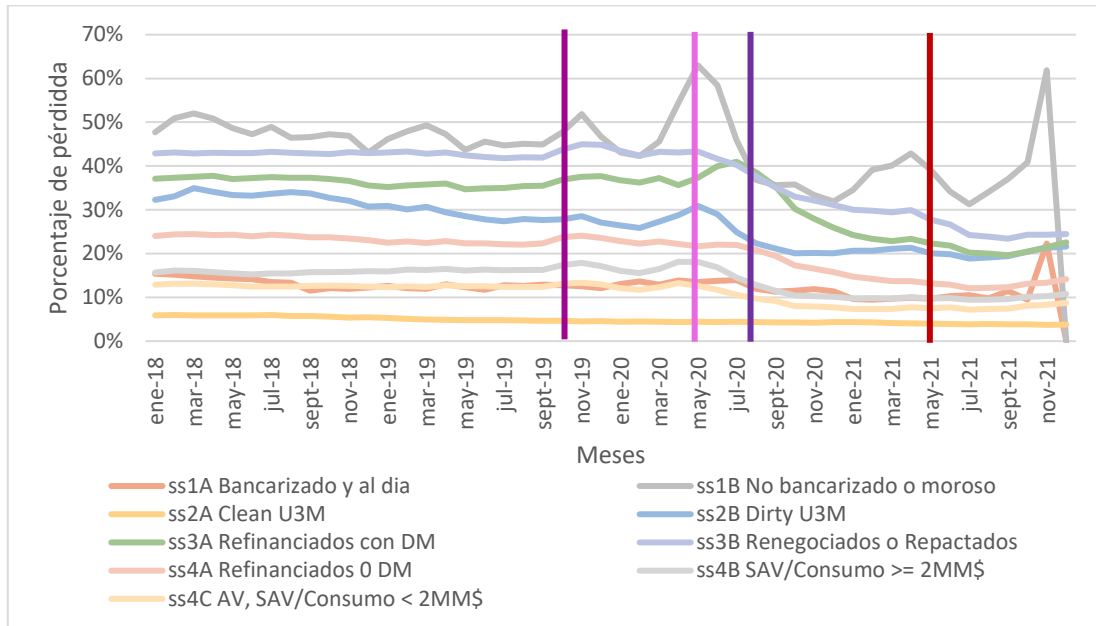
Fuente: Elaboración propia

Gráfico N°48: “Porcentaje de incumplimiento a 12 meses”



Fuente: Elaboración propia

Gráfico N°49: “Perdida porcentual dado el segmento de cliente al que pertenece”



Fuente: Elaboración propia

Anexo E: Análisis de resultados de variables del modelo

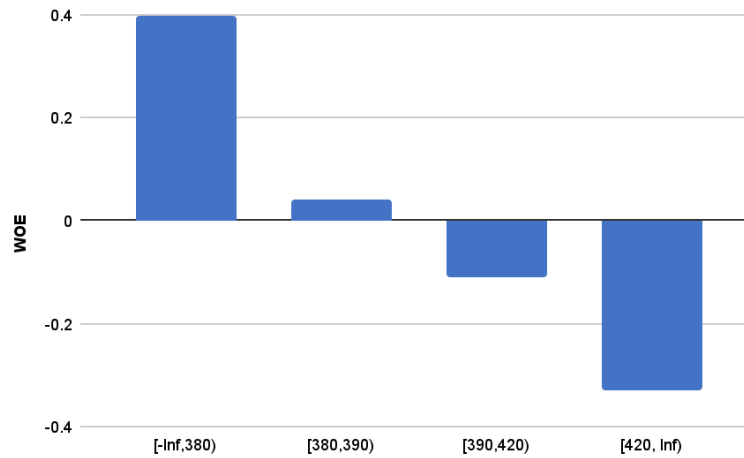
Variable 4: Total días mora en un año

Tabla N°20: “Tabla de WOE e IV para variable 4”

Variable 4	WOE	Contribución IV
[-Inf,380)	0.399	0.055
[380,390)	0.042	0.000
[390,420)	-0.109	0.001
[420, Inf)	-0.329	0.043
Total IV		0.100

Fuente: Elaboración propia

Gráfico N°50: “WOE del total días de mora en un año”



Fuente: Elaboración propia

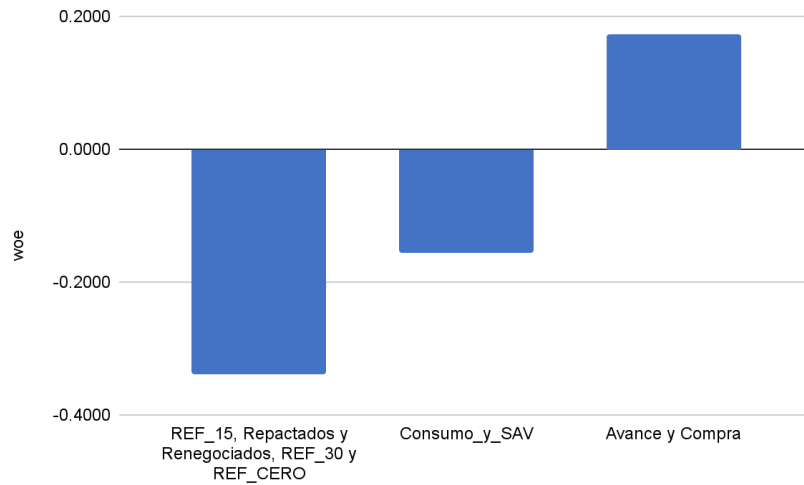
Variable 5: Producto más riesgoso que posee el cliente

Tabla N°21: “Tabla de WOE e IV para variable 5”

Variable 5	WOE	Contribución IV
REF_15, Repactados y Renegociados, REF_30 y REF_CERO	-0.338	0.026
Consumo_y_SAV	-0.156	0.004
Avance y Compra	0.173	0.018
Total IV		0.048

Fuente: Elaboración propia

Gráfico N°51:” WOE del producto más riesgoso que posee el cliente”



Fuente: Elaboración propia

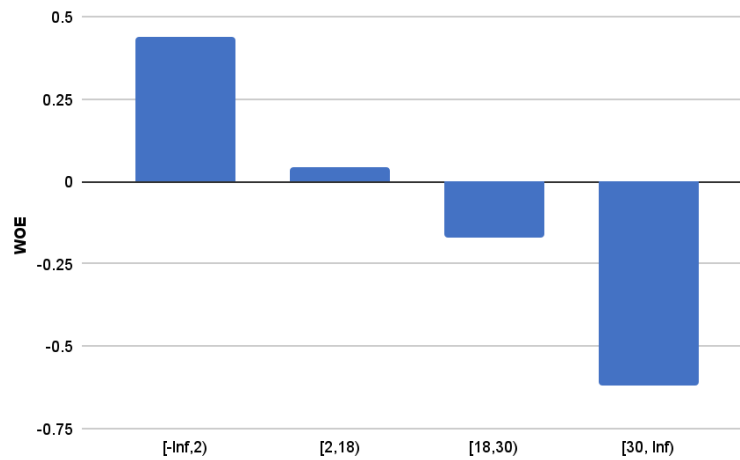
Variable 6: Días de mora del cliente

Tabla N°22: “Tabla de WOE e IV para variable 6”

Variable 6	WOE	Contribución IV
[-Inf,2)	0.436	0.086
[2,18)	0.043	0.000
[18,30)	-0.172	0.005
[30, Inf)	-0.621	0.107
Total IV		0.198

Fuente: Elaboración propia

Gráfico N°52:” WOE de los días de mora del cliente”



Fuente: Elaboración propia

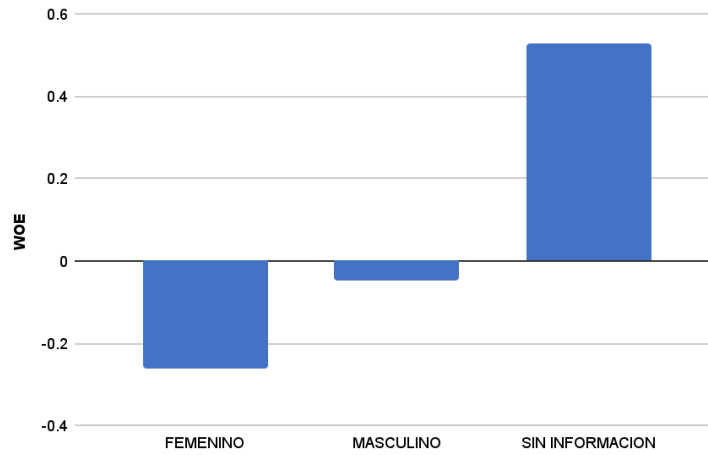
Variable 7: Sexo de los clientes

Tabla N°23: “Tabla de WOE e IV para variable 7”

Variable 7	WOE	Contribución IV
FEMENINO	-0.262	0.031
MASCULINO	-0.046	0.001
SIN INFORMACIÓN	0.529	0.070
Total IV		0.102

Fuente: Elaboración propia

Gráfico N°53:” WOE del sexo de clientes”



Fuente: Elaboración propia

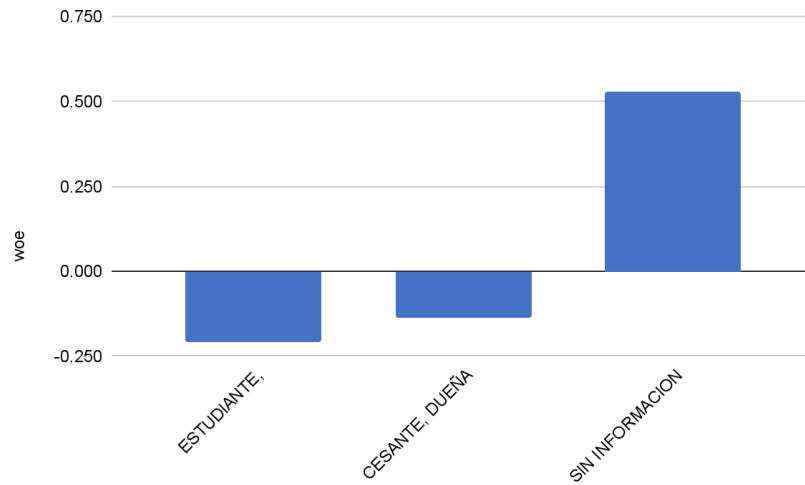
Variable 8: Actividad que realiza el cliente

Tabla N°24: “Tabla de WOE e IV para variable 8”

Variable 8	WOE	Contribución IV
ESTUDIANTE,RETENCIÓN JUDICIAL Y DEPENDIENTE	-0.210	0.019
CESANTE, DUEÑA DE CASA,INDEPENDIENTE Y JUBILADO	-0.138	0.006
SIN INFORMACIÓN	0.527	0.070
Total IV		0.095

Fuente: Elaboración propia

Gráfico N°54:” WOE de actividad que realiza el cliente”



Fuente: Elaboración propia

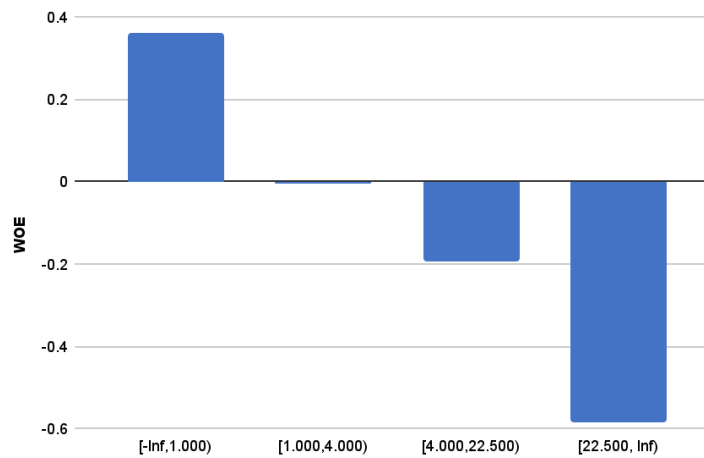
Variable 9: Deuda vigente en las instituciones financieras

Tabla N°25: “Tabla de WOE e IV para variable 9”

Variable 9	WOE	Contribución IV
[-Inf,1.000)	0.359	0.044
[1.000,4.000)	-0.006	0.000
[4.000,22.500)	-0.193	0.011
[22.500, Inf)	-0.585	0.036
Total IV		0.092

Fuente: Elaboración propia

Gráfico N°55:” WOE de deuda vigente en todas las instituciones”



Fuente: Elaboración propia

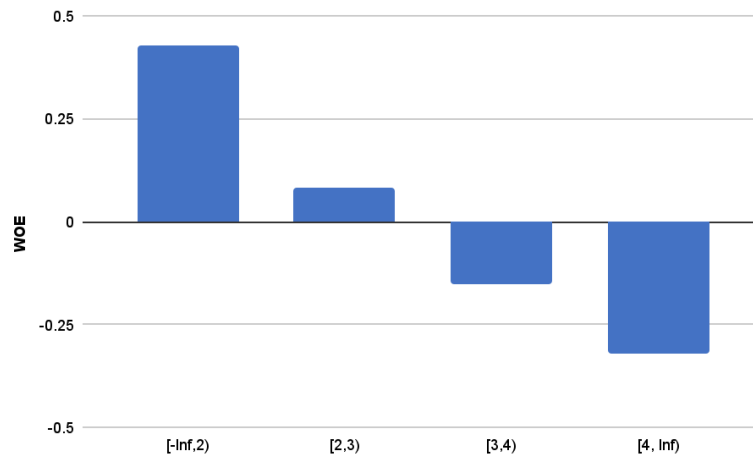
Variable 10: Número de instituciones en las que el cliente tiene créditos

Tabla N°26: “Tabla de WOE e IV para variable 10”

Variable 10	WOE	Contribución IV
$[-\text{Inf}, 2)$	0.429	0.043
$[2, 3)$	0.082	0.002
$[3, 4)$	-0.153	0.005
$[4, \text{Inf})$	-0.319	0.028
Total IV		0.077

Fuente: Elaboración propia

Gráfico N°56: “WOE del número de instituciones en las que el cliente tiene créditos”



Fuente: Elaboración propia

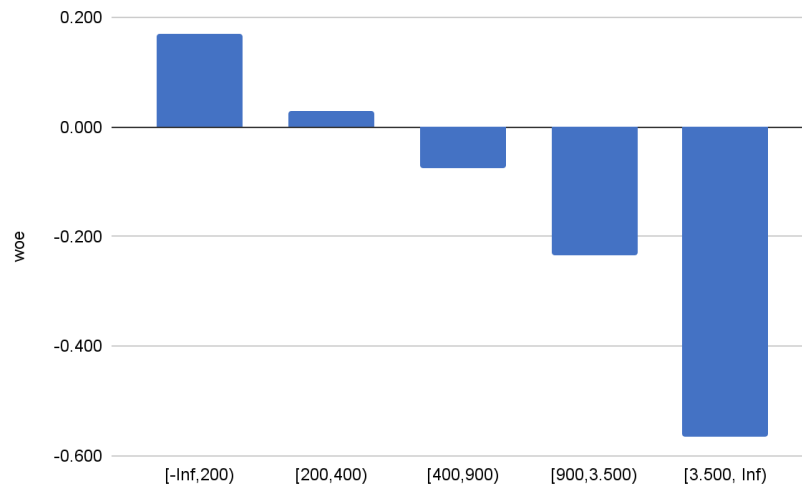
Variable 11: Monto disponible

Tabla N°27: “Tabla de WOE e IV para variable 11”

Variable 11	WOE	Contribución IV
[-Inf,200)	0.170	0.016
[200,400)	0.029	0.000
[400,900)	-0.074	0.001
[900,3.500)	-0.234	0.009
[3.500, Inf)	-0.566	0.026
Total IV		0.052

Fuente: Elaboración propia

Gráfico N°57: “WOE del monto disponible para optar a crédito que tiene el cliente”



Fuente: Elaboración propia

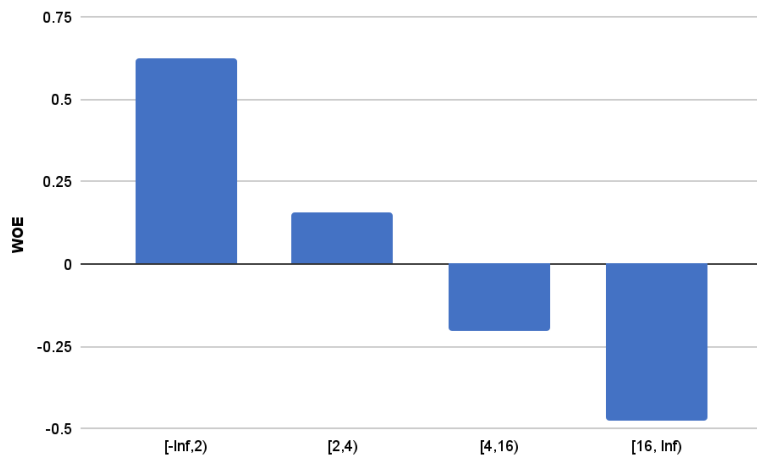
Variable 12: Antigüedad en años

Tabla N°28 “Tabla de WOE e IV para variable 12”

Variable 12	WOE	Contribución IV
$[-\text{Inf}, 2)$	0.624	0.082
$[2, 4)$	0.155	0.005
$[4, 16)$	-0.205	0.018
$[16, \text{Inf})$	-0.477	0.036
Total IV		0.140

Fuente: Elaboración propia

Gráfico N°58: “WOE de la antigüedad en años del cliente en el banco”



Fuente: Elaboración propia

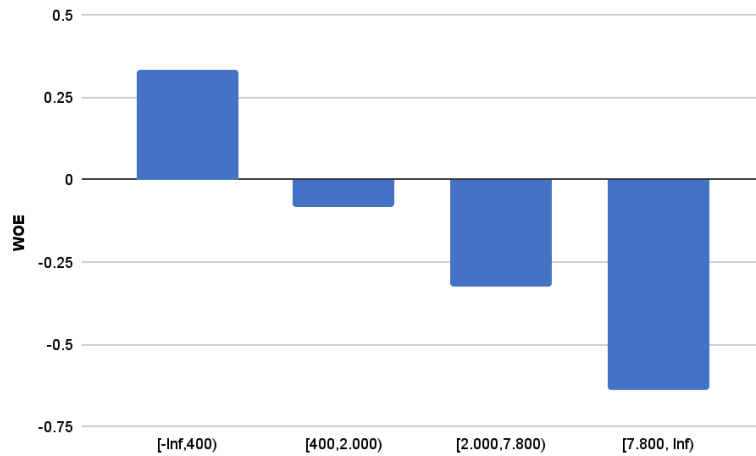
Variable 13: Monto máximo de crédito que el cliente puede optar en un año

Tabla N°29: “Tabla de WOE e IV para variable 13”

Variable 13	WOE	Contribución IV
[-Inf,400)	0.331	0.043
[400,2.000)	-0.085	0.003
[2.000,7.800)	-0.325	0.020
[7.800, Inf)	-0.638	0.025
Total IV		0.090

Fuente: Elaboración propia

Gráfico N°59: “WOE del monto máximo que tiene el cliente para optar a crédito en 12 meses antes”



Fuente: Elaboración propia

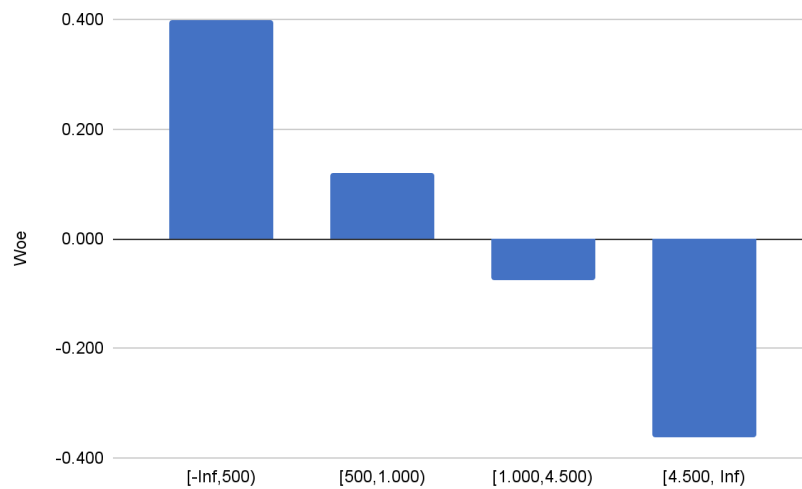
Variable 14: Valor máximo de crédito de consumo en 12 meses antes

Tabla N°30: “Tabla de WOE e IV para variable 14”

Variable 14	WOE	Contribución IV
[-Inf,500)	0.399	0.048
[500,1.000)	0.120	0.001
[1.000,4.500)	-0.077	0.002
[4.500, Inf)	-0.362	0.040
Total IV		0.091

Fuente: Elaboración propia

Gráfico N°60: “WOE del valor máximo de crédito de consumo”



Fuente: Elaboración propia

Anexo F: Resultados del modelo, puntaje de todos los datos
Tabla N°31: “Tabla de puntaje de los datos”

PUNTAJE	NO CASTIGADO	CASTIGADO	TOTAL
243	9	166	175
246	0	1	1
248	0	1	1
249	0	4	4
251	3	15	18
252	3	18	21
253	0	2	2
255	0	2	2
256	1	17	18
257	1	5	6
258	0	2	2
259	0	1	1
260	0	11	11
261	2	13	15
262	0	2	2
264	0	3	3
265	2	5	7
266	0	2	2
267	3	7	10
268	0	7	7
269	0	8	8
270	0	1	1
271	1	3	4

PUNTAJE	NO CASTIGADO	CASTIGADO	TOTAL
272	5	20	25
273	2	4	6
274	0	2	2
275	0	2	2
276	1	5	6
277	0	6	6
278	2	7	9
279	0	2	2
280	2	3	5
281	4	19	23
282	0	4	4
283	0	7	7
284	0	1	1
285	2	11	13
286	1	3	4
287	0	5	5
288	1	9	10
289	0	2	2
290	1	16	17
291	1	1	2
292	2	7	9
293	0	6	6
294	2	7	9
295	87	513	600
296	1	7	8

PUNTAJE	NO CASTIGADO	CASTIGADO	TOTAL
297	1	4	5
298	3	3	6
299	4	12	16
300	173	635	808
301	5	37	42
302	1	4	5
303	45	192	237
304	16	78	94
305	3	17	20
306	24	41	65
307	22	68	90
308	242	860	1102
309	59	155	214
310	19	38	57
311	10	28	38
312	46	147	193
313	53	185	238
314	61	169	230
315	25	81	106
316	147	372	519
317	79	209	288
318	85	217	302
319	34	84	118
320	104	265	369
321	98	227	325

PUNTAJE	NO CASTIGADO	CASTIGADO	TOTAL
322	115	256	371
323	60	198	258
324	55	152	207
325	102	247	349
326	113	317	430
327	69	151	220
328	77	215	292
329	109	265	374
330	113	248	361
331	111	245	356
332	54	178	232
333	127	247	374
334	130	292	422
335	85	192	277
336	129	254	383
337	120	255	375
338	170	338	508
339	130	259	389
340	97	208	305
341	153	280	433
342	125	281	406
343	148	237	385
344	151	243	394
345	122	232	354
346	207	392	599

PUNTAJE	NO CASTIGADO	CASTIGADO	TOTAL
347	189	295	484
348	132	219	351
349	139	303	442
350	163	273	436
351	190	331	521
352	180	288	468
353	167	270	437
354	229	360	589
355	204	325	529
356	191	285	476
357	188	305	493
358	213	333	546
359	248	347	595
360	227	328	555
361	204	281	485
362	222	310	532
363	239	331	570
364	259	335	594
365	226	284	510
366	239	301	540
367	259	317	576
368	272	277	549
369	288	310	598
370	240	277	517
371	288	345	633

PUNTAJE	NO CASTIGADO	CASTIGADO	TOTAL
372	240	328	568
373	266	334	600
374	292	321	613
375	281	334	615
376	325	342	667
377	285	292	577
378	294	300	594
379	305	321	626
380	320	328	648
381	315	314	629
382	300	290	590
383	409	353	762
384	331	321	652
385	333	288	621
386	287	278	565
387	307	282	589
388	470	340	810
389	329	309	638
390	326	300	626
391	314	299	613
392	350	285	635
393	329	307	636
394	311	278	589
395	328	284	612
396	528	373	901

PUNTAJE	NO CASTIGADO	CASTIGADO	TOTAL
397	372	273	645
398	333	318	651
399	333	246	579
400	312	322	634
401	368	279	647
402	316	288	604
403	383	306	689
404	393	259	652
405	392	276	668
406	347	301	648
407	351	243	594
408	422	261	683
409	329	218	547
410	413	265	678
411	417	292	709
412	347	233	580
413	385	247	632
414	349	241	590
415	350	228	578
416	385	289	674
417	363	204	567
418	315	230	545
419	396	283	679
420	346	209	555
421	396	231	627

PUNTAJE	NO CASTIGADO	CASTIGADO	TOTAL
422	389	211	600
423	370	206	576
424	336	207	543
425	375	212	587
426	370	221	591
427	403	255	658
428	350	189	539
429	334	208	542
430	390	209	599
431	344	181	525
432	353	184	537
433	374	193	567
434	343	175	518
435	364	189	553
436	335	154	489
437	365	182	547
438	366	173	539
439	312	174	486
440	327	180	507
441	336	173	509
442	296	157	453
443	345	127	472
444	319	140	459
445	283	133	416
446	340	138	478

PUNTAJE	NO CASTIGADO	CASTIGADO	TOTAL
447	306	131	437
448	331	122	453
449	356	126	482
450	290	117	407
451	294	133	427
452	280	116	396
453	304	90	394
454	279	129	408
455	275	112	387
456	270	124	394
457	259	96	355
458	265	98	363
459	246	107	353
460	277	110	387
461	247	77	324
462	256	106	362
463	241	80	321
464	225	76	301
465	223	90	313
466	210	77	287
467	219	72	291
468	180	80	260
469	195	72	267
470	192	67	259
471	172	52	224

PUNTAJE	NO CASTIGADO	CASTIGADO	TOTAL
472	178	58	236
473	162	45	207
474	154	59	213
475	158	57	215
476	154	60	214
477	109	46	155
478	142	46	188
479	147	58	205
480	150	44	194
481	135	31	166
482	119	31	150
483	111	31	142
484	102	37	139
485	124	24	148
486	117	25	142
487	124	27	151
488	110	22	132
489	105	33	138
490	94	14	108
491	82	20	102
492	107	22	129
493	74	23	97
494	80	20	100
495	68	19	87
496	66	22	88

PUNTAJE	NO CASTIGADO	CASTIGADO	TOTAL
497	69	16	85
498	75	11	86
499	69	16	85
500	63	20	83
501	58	24	82
502	71	18	89
503	38	7	45
504	56	14	70
505	51	12	63
506	39	10	49
507	36	18	54
508	55	9	64
509	36	12	48
510	51	8	59
511	51	6	57
512	61	4	65
513	38	11	49
514	43	7	50
515	43	5	48
516	37	6	43
517	24	5	29
518	27	11	38
519	22	3	25
520	35	5	40
521	23	3	26

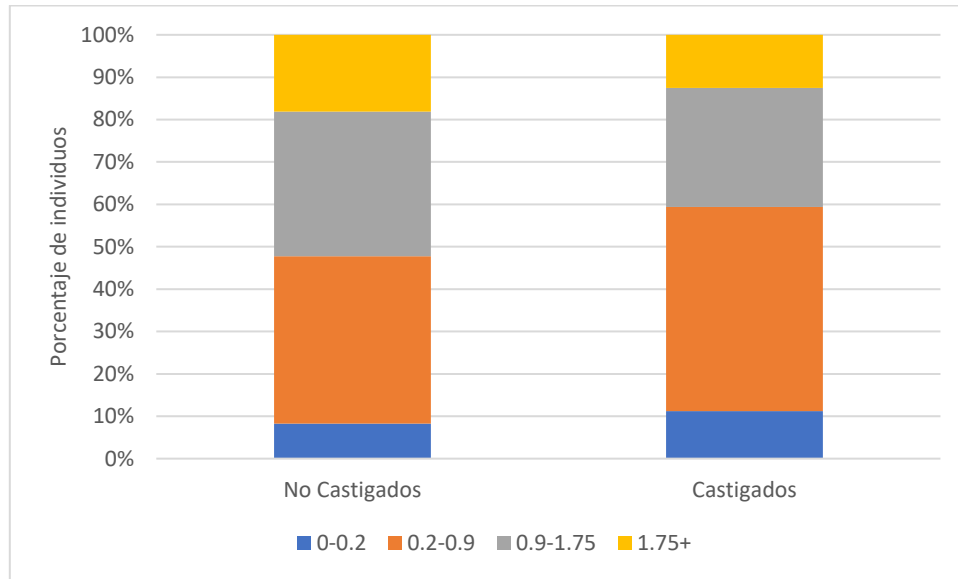
PUNTAJE	NO CASTIGADO	CASTIGADO	TOTAL
522	19	2	21
523	22	2	24
524	20	6	26
525	19	6	25
526	17	0	17
527	16	3	19
528	22	4	26
529	13	0	13
530	9	4	13
531	16	2	18
532	7	1	8
533	8	3	11
534	8	1	9
535	8	2	10
536	16	2	18
537	17	1	18
538	4	0	4
539	12	4	16
540	3	1	4
541	2	0	2
542	3	0	3
544	2	0	2
546	1	0	1
547	3	0	3
549	3	0	3

PUNTAJE	NO CASTIGADO	CASTIGADO	TOTAL
550	2	0	2
552	2	0	2
555	1	0	1

Anexo G: Resultados finales

Variable: Tendencia de los días de mora de un cliente

Gráfico N°61: “Tendencia de los días de mora de un cliente”

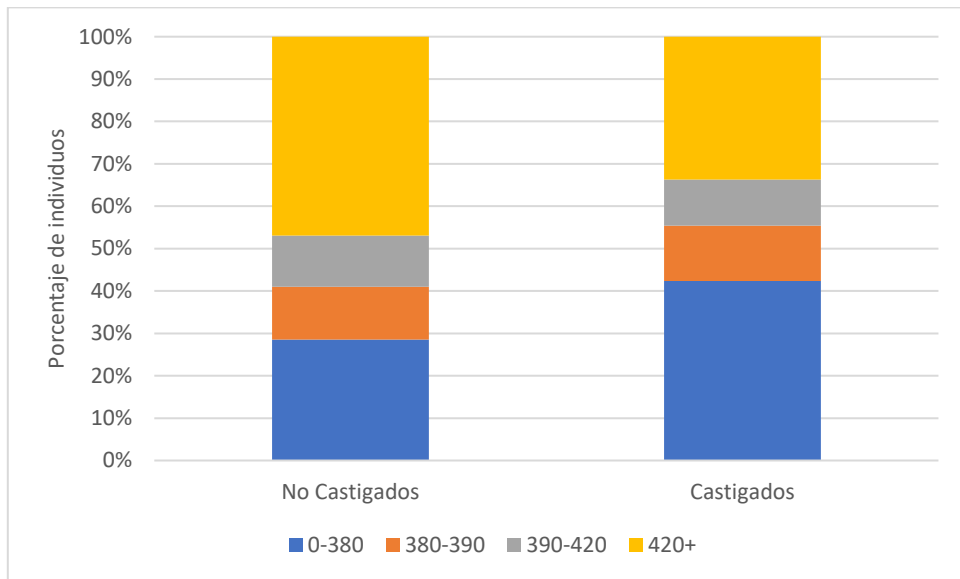


Fuente: Elaboración propia

Al observar la composición de la base de datos trabajada, entre castigados y no castigados, se puede observar que la variable tendencia de los días de mora de un cliente genera diferencias en ambos grupos. En su mayoría ambos presentan un alto porcentaje de personas que poseen tendencia de días mora de 0.2-0.9, siendo el porcentaje aún mayor en gente castigados.

Variable: Suma de los días mora en un año de un cliente

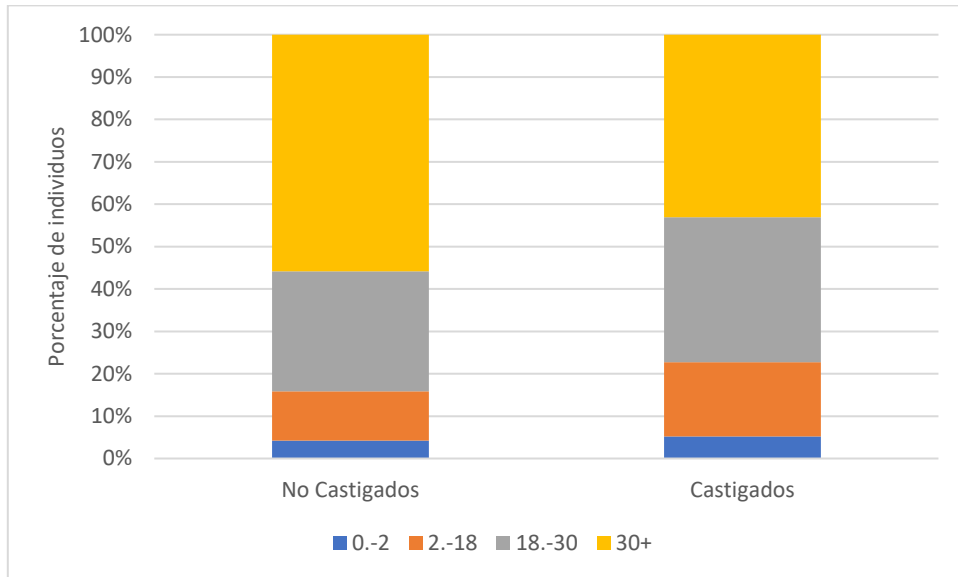
Gráfico N°62: “Suma de los días mora en un año de un cliente”



En cuanto a la composición de los clientes castigados y no castigados, dada la variable días de mora en un año, los no castigados tienen los valores más altos de la suma, mientras que los no castigados en su mayoría son entre 0-380.

Variable: Días de mora del cliente

Gráfico N°63: “Días de mora del cliente”

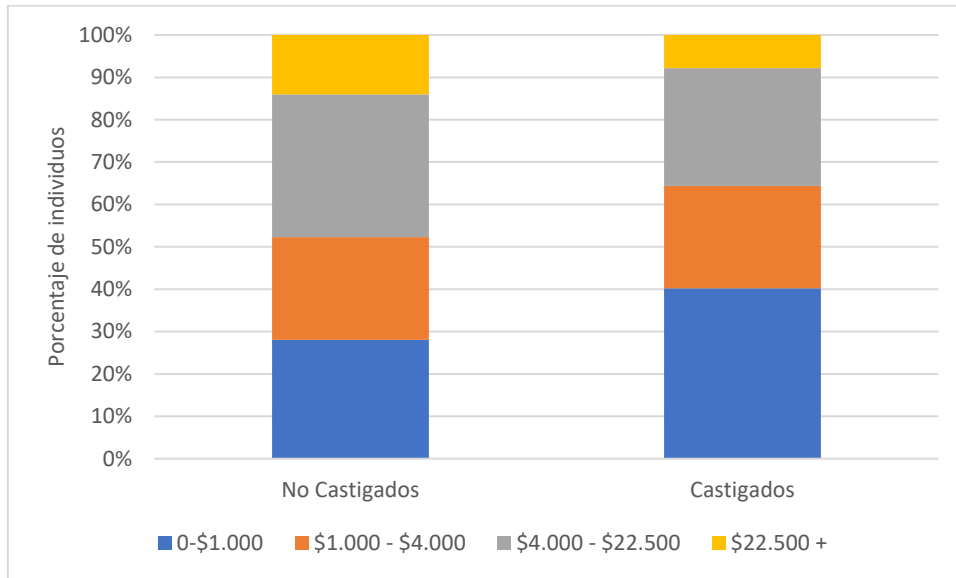


Fuente: Elaboración propia

La composición de los no castigados y castigados se ve mayoritariamente compuesta por personas con más de 30 días de mora, siendo este segmento el que mayor cambio presentan entre ambos grupos.

Variable: Deuda vigente en todas las instituciones

Gráfico N°64: “Deuda vigente en todas las instituciones”

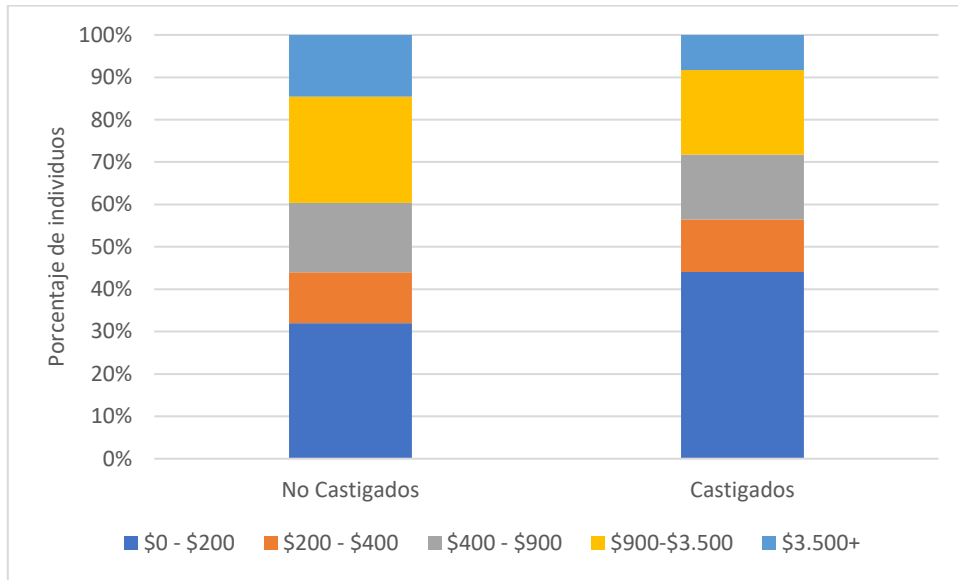


Fuente: Elaboración propia

Para la composición final de los clientes castigados y no castigados la mayoría de los no castigados están con deuda vigente de \$4.000.000-\$22.500.000, mientras que la otra gran composición de los castigados corresponde a gente que tiene deuda entre \$0-\$1.000.000.

Variable: Monto disponible para optar a crédito que tiene el cliente

Gráfico N°65: “Monto disponible para optar a crédito que tiene el cliente”

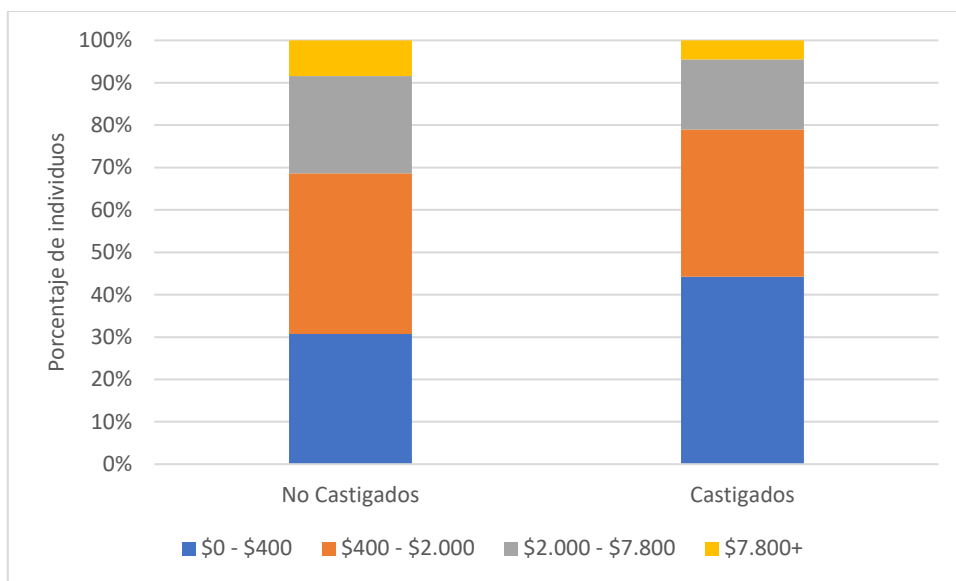


Fuente: Elaboración propia

El monto que les queda aún disponible para crédito de los clientes castigados y no castigados en su mayoría está compuesto por clientes de entre \$0-\$200.000, siendo este porcentaje mayor en el grupo de castigados.

Variable: Monto máximo que tiene el cliente para optar a crédito en 12 meses antes

Gráfico N°66: “Monto máximo que tiene el cliente para optar a crédito en 12 meses antes”

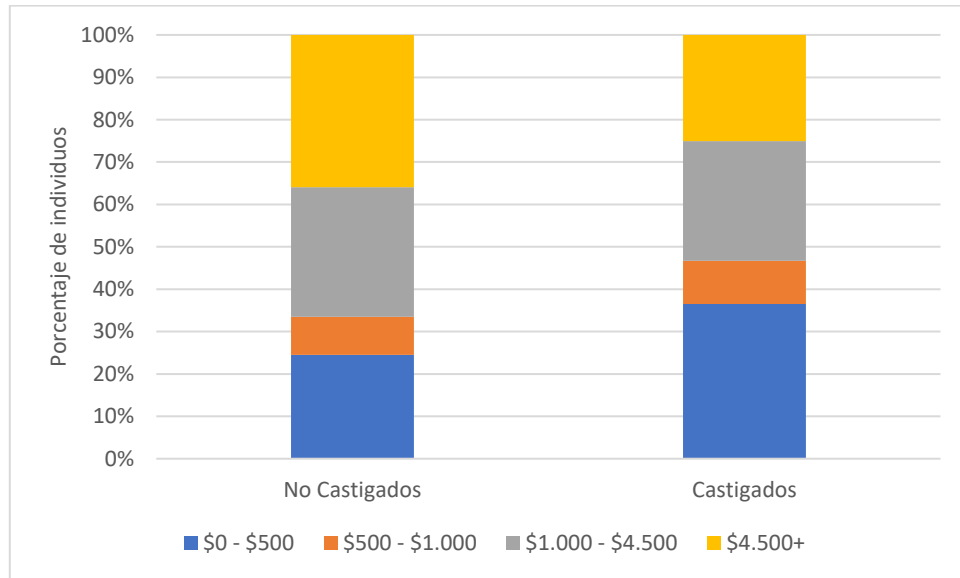


Fuente: Elaboración propia

La variable diferencia a los grupos de no castigados y castigados, teniendo los castigados un valor inferior en el monto que pueden adquirir de deuda en comparación a los no castigados.

Variable: Valor máximo del crédito de consumo en 12 meses antes

Gráfico N°67: “Valor máximo del crédito de consumo en 12 meses antes”



Fuente: Elaboración propia

Para la variable valor máximo de crédito de consumo que tiene un cliente en 12 meses antes de octubre 2019 con respecto a los clientes que son no castigados y castigados, se puede observar en los primeros una preponderancia de los valores de créditos de consumo máximo entre \$1.000.000 a más de \$4.500.000, mientras que los castigados presentan un valor cercano al 50% de créditos de consumo entre \$0-\$1.000.000.