UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

Ventaja de la Segmentación de una Cartera de Clientes en la Construcción de un Credit Score

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

Actuaria

PRESENTA:

Itzel Anai Flores Terrones

TUTOR

Act. José Fernando Soriano Flores

2016

Ciudad Universitaria CDMX

1.Datos del alumno.
Flores
Terrones
Itzel Anai
55 34 08 23 02
Universidad Nacional Autónoma de México
Facultad de Ciencias
Actuaría
303047638

2.Datos del tutor
Act.
José Fernando
Soriano
Flores

3.Datos del sinodal 1
Dra.
Lizbeth
Naranjo
Torres

4.Datos del sinodal 2
Act.
Alejandra
Juárez
Torres

5.Datos del sinodal 3
Act.
Ricardo
Villegas
Azcorra

6.Datos del sinodal 4
Act.
Felipe
Zamora
Ramos

7.Datos del trabajo escrito
Ventaja de la Segmentación de una Cartera de Clientes en la Construcción de un Credit Score
126p.
2016

Agredecimientos

Gracias Dios por mi vida y darme la oportunidad de recrearme todos y cada uno de mis días.

A Juan Flores y María Guadalupe por ser mis padres, y hacer de mi la profesionista que presenta esta tesis.
Se cuanto significa para ustedes este trabajo...
Toda la vida les estaré agradecida.

A mi hija Paula, por ser mi inspiración y mi nuevo mundo. Eres la hija perfecta para mi. Te amo.

A ti Gerardo por la paciencia y ese último empujón que necesitaba para terminar éste ciclo. Te adoro, amor.

A mis compañeros de vida Marcos, Nayely, Leticia y Alejandro por su apoyo y cariño, no hubiera podido lograrlo sin ustedes. Son lo máximo.

A mi asesor Fernando Soriano por su apoyo y atención para que llevar a cabo esta tesis. No hubiera podido terminar sin tu guía.

A mis 4 sinodales por aceptar ser parte de este proyecto.

Y todos los maestros, amigos y compañeros que de alguna forma contribuyeron en mi vida académica, pues sin ellos no hubiera sido tan hermoso este camino como lo fue.

A la Universidad Nacional Autónoma de México por abrirme las puertas de tus aulas y crearme un mejor futuro.

Gracias.

Índice general

Capítulo 1

Introducción

En la actualidad para todo individuo es de suma importancia la administración del dinero. Los principales puntos para su administración resultan cuando se obtiene el dinero, ya sea trabajando como empleado o de forma independiente, el siguiente punto es dónde tenerlo, para qué obtener una ganancia, puede ser en un fondo de inversión o en un seguro, y finalmente dónde gastarlo, como un negocio o en necesidades personales o familiares. Por ello es esencial el estudio de los temas relacionados con el dinero, éste trabajo abordará el tema crédito como punto central.

Cuando existe la necesidad de adquirir un bien o satisfacer una carencia familiar o personal, normalmente se encuentran dos posturas. Cuando se tiene el capital y se puede realizar la adquisición del bien o servicio y cuando no se tiene dinero y se toma la decisión de recurrir al crédito. Regularmente para solicitar un préstamo se acude a una institución bancaria o financiera.

Estas instituciones encargadas de otorgar créditos cuentan internamente con una estructura operacional para dar atención a la originación del préstamo y darle seguimiento al mismo. Estas sociedades cuentan con medidas de control de riesgo, que son las directrices encargadas de regular las condiciones en las que un crédito puede ser autorizado.

A lo largo de éste trabajo se desarrollarán temas como el crédito y posteriormente abordaremos como tema principal el indicador score de crédito, con la peculiaridad de que será para una cartera ya existente dentro de la institución, es decir, para aquel solicitante de crédito que ya cuenta con un comportamiento de pago dentro de una institución y requiera realizar un mantenimiento de cuentas como incremento o decremento de línea de crédito, ofrecimiento de una línea de crédito adicional, ser sujeto a una campaña comercial o cualquier modificación a su crédito previamente pactado.

Para generar un *Behavior Score* (Score de Comportamiento) es necesario conocer diferentes conceptos acerca del crédito de consumo, por lo que a lo largo de ésta introducción se irán desarrollando temas, como la definición del crédito que se encuentra en el siguiente apartado.

1.1. El crédito

¿Qué es el crédito?

La palabra crédito proviene del latín *credititus* que es una forma sustantiva del verbo *credere* (creer), cuyo significado a su vez es cosa confiada. Se le denomina crédito a una suma de dinero que se le debe a alguna entidad. Generalmente se estipula una fecha límite de devolución y ésta debe realizarse con intereses.

El crédito es un financiamiento utilizado comúnmente por las familias y empresas permitiéndoles adquirir bienes o servicios con el préstamo de una cantidad de dinero por una persona o entidad con el compromiso de devolverlo en un futuro con un interés adicional en un plazo y forma de pago acordado entre ambas partes. Se puede decir que el crédito brinda la facilidad de liquidez para comprar, hacer pagos, o en su caso alguna inversión.

Dadas las necesidades ya sea de un bien o servicio que se pueden tener, existe un tipo de crédito diferente para cada una de ellas, entre los más comunes se pueden encontrar:

- Crédito bancario
- Crédito hipotecario
- Crédito de producción o comercial
- Crédito de consumo (comúnmente tarjeta de crédito)

El listado superior indica que existe un crédito apropiado para cada una de las necesidades de vivienda, consumo y/o crecimiento de un negocio. Por lo que actualmente no es necesario contar con el capital liquido para realizar el gasto o inversión, se puede decir que el crédito pretende desplazar al dinero expresado en moneda líquida; con ello se muestra la pretensión de que el dinero desaparezca como principal medio de cambio.

Si bien cada tipo de crédito es importante e impacta de una u otra forma en la economía mexicana, en algunos casos brinda la oportunidad, por ejemplo, de obtener un inmueble como es el caso de crédito hipotecario o adquirir artículos de consumo básico. Sin embargo, para efectos de éste trabajo se considerará el crédito de consumo como uno de los temas principales a desarrollar.

En los siguientes incisos se tocarán temas relacionados con el crédito de consumo. Se comenzará por explicar que es un crédito de consumo.

1.2. ¿Qué es un crédito de consumo?

El crédito de consumo es un tipo de préstamo que permite financiar y satisfacer la adquisición de bienes como un coche, muebles para la casa, viajes o festejos familiares, siempre y cuando los pagos sean no comerciales o empresariales, es decir, que busquen

fondear un negocio.

Estos prestamos son realizados por entidades financieras que regularmente revisan condiciones básicas del solicitante de crédito que debe demostrar, como tener capacidad de pago para hacer frente a dicho crédito así como una cultura de pago en términos de riesgo.

Actualmente éste tipo de crédito es muy fácil de obtener, pueden solicitarse en cadenas comerciales y bancos por medio de sus tarjetas de crédito. En algunos casos se pueden obtener sin necesidad de entregar demasiada información o forzosamente entregar un historial crediticio, sólo basta con presentar un comprobante de domicilio y una identificación oficial para que otorguen el crédito.

Ya que actualmente es un medio común de obtención de bienes, es necesario conocer el impacto del crédito en México, en el siguiente apartado se hablará de éste tema.

1.3. Importancia del crédito de consumo

Si bien México no ha alcanzado niveles de compra con crédito comparados con Estados Unidos o España, el crédito de consumo toma un papel importante dentro de la economía Mexicana, ya que los especialistas hablan de un crecimiento gracias a la reducción de tasas de interés. En las últimas décadas el crédito ha tenido un crecimiento acelerado y volátil, en momentos se habla de un crecimiento del 20 % y en otros decrece, ya que se trata de una cantidad pequeña de dinero para un préstamo inmediato. La información económica habla de que 40 % de las compras o ventas que se realizan en el país se realizan por medio de un crédito de consumo [3].

Los datos muestran que el crecimiento del porcentaje se ha dado sin generar un endeudamiento excesivo en las familias mexicanas. Ya que se han implementado mejoras en las medidas de control de riesgo. Una de las medidas aplicadas es cuando se observa un deterioro en la cuenta, las instituciones prestatarias optan por frenar el nuevo crédito o incremento hasta que la cuenta muestre estabilidad.

Los otorgantes de crédito que identifican oportunidades en el crecimiento del mercado están dispuestos a asumir el riesgo que conlleve, con la finalidad de incrementar su cartera y por supuesto sus ingresos.

1.4. Características de un crédito de consumo

Al solicitar un crédito es importante conocer las características del crédito que la institución financiera está ofreciendo, ya que de ellas debe depender si el cliente toma o no el crédito. Algunas de las propiedades para el otorgamiento del crédito de consumo son:

- Es concertado por un profesional en otorgamiento de créditos.
- La relación es con un usuario, que requiere de ese monto para atender una necesidad personal o familiar ajena a su actividad profesional o no.

- El contrato del crédito se realiza por escrito y el cliente se lleva una copia. Éste debe contener la firma de ambas partes, de lo contrario es nulo.

Un tema delicado dentro de la adquisición de un crédito son los datos financieros que se reúnen en la operación, estos deben de estar especificados en el contrato. Muchas de las veces el cliente no tiene cuidado al considerar esta información para aceptar un crédito, lo que puede generar un riesgo para la institución financiera, ya que el usuario por falta de administración financiera puede representar una pérdida. Los siguientes puntos son la información básica que debe de estipularse en el contrato.

- **Tasa anual** corresponde a la cantidad prestada y consiste en el costo total del préstamo expresado en porcentaje, teniendo en cuenta comisiones, gastos administrativos, etc, también conocido como CAT.
- **Tipo de interés** puede ser nominal, en el caso de no establecer tasa anual, con sus gastos correspondientes.
- **Importe del crédito** el importe de dinero que la institución está dispuesta a financiar.
- **Número de cuotas a pagar** consiste en conocer el número de pagos en que se finiquitará el pago del préstamo.
- **El importe de cada cuota** estas pueden ser todas iguales, crecientes o decrecientes, según se acuerde.
- **Periodicidad de pago** consiste en establecer si los pagos o pago único se harán de forma semanal, mensual, bimestral, etc.
- Constitución de **garantías, penalizaciones** y otros gastos que se pueden aplicar.
- Por su parte, no deben faltar los **derechos de ambas partes** durante todo el proceso de otorgamiento y pago del préstamo, condiciones generales, y delimitar que este crédito no tiene un fin concreto, sino que el cliente lo puede usar para lo que desee, sin tener que informarlo a la entidad prestataria del dinero.

1.5. El ciclo de crédito en la administración de riesgo

Si bien la administración de riesgo varía entre instituciones que se dedican a otorgar créditos. Desde el punto de vista del riesgo, se detectan tres puntos principales a lo largo de la vida de un crédito, figura 1.1.

Figura 1.1: Ciclo de Crédito.

- **La Originación** es el primer paso dentro del ciclo de un crédito, en éste punto se revisan todas y cada una de las directrices de aceptación de clientes basadas en políticas de crédito, políticas de aprobación y políticas de asignación de atributos como lo son: monto de la línea de crédito, tasa de interés, etc. Dichas políticas se basan entre otras cosas en modelos de score denominados Score de Adquisición o *Acquisition Score*. La revisión de los lineamientos de aceptación tiene como finalidad filtrar un cliente potencial de un cliente que pueda representar una pérdida para la institución, y no otorgarle un crédito.

- **La Administración** de la cuenta consiste en gestionar los créditos ya aprobados modificando sus atributos tales como línea de crédito, tasa de interés basado en estrategias de riesgo construidas a partir del comportamiento de pago que tiene el cliente con la propia institución del crédito que le fue otorgado. Para ésto se utiliza principalmente el Score de Comportamiento o *Behavior Score*. Mismo que es el objeto de estudio en este trabajo.

- **La Cobranza** en este punto del ciclo del crédito, busca que un cliente cumpla con sus obligaciones de pago mediante estrategias de cobro segmentadas que permitan la mayor recuperación del crédito con el menor costo para la institución. En dicho cobro y como principal herramienta se usan modelos como *Collection Score*, o Score de Cobranza.

1.6. Administración de riesgo de crédito al consumo

Uno de los principales problemas bancarios está directamente relacionado con normas o metodologías débiles para otorgar un préstamo. La débil administración del riesgo y/o una falta de atención a un siempre cambiante entorno económico podrían causar un deterioro en el crédito de las contrapartes del banco o prestador.

La manera sencilla de describir el riesgo de crédito es como la posibilidad de que un prestatario o contraparte no pueda cumplir con sus obligaciones de acuerdo a los términos acordados en un contrato.

El objetivo de la administración del riesgo de crédito es maximizar la tasa de rendimiento ajustada por el riesgo del banco, manteniendo la exposición al riesgo de crédito dentro de límites aceptables. Ya que un banco administra el riesgo de su cartera de forma global y también de forma individual, es decir, por cliente.

Es importante considerar que la administración del riesgo varía entre bancos y prestamistas dependiendo de la índole, complejidad y apetito de riesgo a la que cada institución se encuentre alineada. Los bancos deben estar conscientes de la necesidad de identificar, medir, monitorear, controlar y cuantificar el nivel de riesgo de crédito y determinar si pueden hacer frente a estos riegos.

Por ello se han aplicado prácticas matemáticas basadas en modelos paramétricos y no paramétricos para poder administrar el riesgo de crédito. Una de ellas son los modelos de puntuación generalmente conocidos como *Modelos Scoring* mediante los cuales es posible calificar a un prospecto de cliente infiriendo así la probabilidad de incumplimiento de pago.

En la literatura se mencionan algunos puntos que debe de abarcar la administración del riesgo:

- Establecer un entorno apropiado para el riesgo de crédito.
- Operar bajo un proceso sano para otorgar créditos.
- Mantener un proceso adecuado para administrar, medir y monitorear el crédito.
- Garantizar controles adecuados del riesgo de crédito.

Como se mencionó en éste inciso una de las herramientas más conocidas para la administración de riesgos son los *Modelos de Scoring*. Por lo que es importante conocer su importancia, misma que se desarrolla a continuación.

1.7. Importancia de un modelo de scoring

En la historia de los modelos de scoring se habla de que en el año de 1936 Fisher [1] introduce la teoría del análisis de discriminante, el cual es un método estadístico utilizado para reconocer los patrones dentro de una población con el objetivo de encontrar combinaciones lineales de las características que separen en dos o más clases a una población. Mediante el desarrollo de una forma más amplia dicha idea se aplicó dentro del sector financiero para discriminar entre un buen y un mal pagador.

Para el año de 1958 Bill Fair y Earl Isaac se encargaron de desarrollar criterios analíticos. Uno de los modelos más usados como herramienta de modelo de riesgo es mejor cono-

[1] Fisher, un estadístico, genetista y biólogo nacido en Londres Inglaterra, una de sus principales aportaciones fue el avance en el análisis de varianza (ANOVA), este análisis fue desarrollado para revisar la gran cantidad de información que recopilaba de sus experimentos.

cido como FICO[2] Score (*Fair Isaac Corporation Score*). FICO fue la primera propuesta general de un *credit score*, transformo las bases por las cuales un prestamista otorgaría un crédito a un modelo analítico predictivo.

Los modelos de *credit scoring* mostraron su utilidad e importancia cuando en los años 60's surgieron nuevos instrumentos financieros como las tarjetas de crédito. Para el autor Mayer (1963) este tipo de modelos tenían mayor poder predictivo que el juicio de un experto [13].

El uso de modelos de *credit scoring* no sólo impacta en el nivel predictivo, sino también en los tiempos de ejecución de los procesos para el otorgamiento de un crédito. Ya que se puede deliberar el otorgamiento casi al instante, y así incrementar el consumo inmediato. En los créditos hipotecarios, se han reducido los tiempos de deliberación a horas y no a semanas como antes se acostumbraba.

Los modelos de *credit scoring* permiten automatizar y parametrizar las variables que se consideran por medio de software estadísticos como SAS, R o SPSS desarrollados con herramientas que permiten la construcción de un *credit score's*, de tal forma que se disminuye la necesidad de la intervención humana en la evaluación del crédito. Los bancos y las empresas financieras no son los únicos beneficiados con el uso de modelos de *scoring*, también los clientes del sector financiero. Pues reducen los errores al otorgar o no otorgar el crédito al momento de solicitarlo. La decisión de aprobación o rechazo es más objetiva, ya que se incorpora toda una metodología la cual integra variables y factores de riesgos basados en la elaboración del modelo.

Se puede decir que los modelos de *scoring* son una pieza fundamental en la administración de riesgo de crédito al consumo, pues ayudan a identificar de una manera confiable aquellos clientes con buena probabilidad de incumplimiento de pago. Es importante considerar que como todo modelo estadístico-matemático no es posible contar con un 100 % de certeza en el cálculo de dicha probabilidad, y es evidente que si existiera un modelo así las tasas de pérdida financiera derivada de los créditos otorgados sería cero, pues se tendría un modelo que predice con exactitud qué clientes van o no van a cumplir con sus compromisos crediticios y de ésta forma se podría saber a quién otorgar un crédito o no.

Es así, que en el presente trabajo se presenta al menos una forma con la que se puede pulir la predictibilidad de los modelos *scoring* de tal forma que se pueda ver una gran diferencia del desempeño de un modelo *scoring* realizado mediante la forma tradicional versus el método mostrado en el presente trabajo.

Esto permitiría sin duda alguna a todas aquellas instituciones que se dedican a otorgar crédito al consumo, ya sean reguladas o no reguladas, a tener una mejor gestión de riesgo, y por ende a controlar el costo de crédito para mejorar sus niveles de rentabilidad y sustentabilidad de la institución.

[2]En 1956 el ingeniero Bill Fair y el matemático Earl Isaac fundaron FICO bajo el principio. "Si los datos son usados de forma inteligente pueden mejorar la toma de decisiones del negocio."

Como ya se comentó en el presente trabajo, se propone una forma con la cual podemos mejorar el desempeño de los modelos *scoring*. Para ello en el capítulo 2 se hablará de cómo se construye tradicionalmente un modelo *scoring* desde la preparación de los datos, definición de la ventana de desarrollo y desempeño hasta la medición de predictibilidad de éste. En el capítulo 3, se hablará de todas aquellas herramientas matemáticas y estadísticas que se emplearán en todo el trabajo. En el capítulo 4, se construye un modelo *scoring*con datos reales de una institución bajo la metodología tradicional, y se construye un modelo *scoring* con la misma población pero con un cambio en la metodología. En el capítulo 5, se determinarán las conclusiones, revisando la diferencia de predictibilidad de los modelos, y evidenciando las mejoras que tuvo bajo el método de segmentación de población versus el modelo tradicional de *scoring* utilizando indicadores de estadística no paramétrica.

Capítulo 2

Construcción de un modelo tradicional

2.1. Modelo de *Scoring*

Un *Credit Scoring* o algunas veces llamado *Score-Cards*, es una calificación de crédito que la gran mayoría de prestamistas como bancos, tiendas de departamento, ventas de carros, casas y compañías de celulares utilizan con el fin de determinar qué tan riesgoso es prestar dinero a un solicitante. La mayoría de las veces se determina por medio de un algoritmo estadístico numérico basado generalmente en la información financiera que ha tenido el usuario en la institución como cliente de la cartera. Para la aplicación de un *credit score* es necesario clasificar en clases a los clientes, generalmente se utiliza una clasificación binaria para diferenciarlos, determinado por un cliente "bueno", "indeterminado" o "malo".

Gracias al apoyo de estos algoritmos los prestamistas han podido autorizar más préstamos y con una línea de crédito ajustada al tipo de riesgo del cliente, ya que la información de la puntuación del score resulta ser más precisa para la toma de decisión del crédito. Permite identificar a los clientes con un buen futuro, aunque el cliente presente problemas en su pasado crediticio. Pues regularmente los prestamistas enfocan diferentes productos para diferentes tipos de clientes.

Partiendo de este hecho es importante saber cómo se determina la clasificación, y a qué tipo de cliente pertenece cada uno. Por lo que en el siguiente apartado, se hablará de las reglas bajo las cuales regularmente se lleva a cabo esta clasificación.

2.2. Clasificación de buenos y malos

Como se mencionó la clasificación regularmente se hace de forma binaria, para la asignación de un cliente "bueno" o "malo" se cuenta con diferentes fuentes de información como la histórica de la cartera en la institución, datos de cobranza, cartera en mora, y de ser posible también con la información que comparten otros bancos acerca del cliente. A partir de esta información histórica, los prestamistas regularmente consideran un buen cliente cuando, por ejemplo:

a) El cliente se encuentra con número de meses vencidos cero (current).

b) Cuando realiza al menos su pago mínimo correspondiente.

Existe también el caso, cuando los clientes se encuentran en el intermedio, es decir que no se puede determinar si son "buenos" o "malos". Puede ser, por que tienen algún adeudo de capital e intereses o no se cuenta con el tiempo suficiente para determinar su estatus.

Un mal cliente se determina regularmente cuando éste causa pérdidas a la compañía, es decir, cuando sobrepasa los meses de tolerancia en mora establecidos por la institución. Regularmente las instituciones de préstamo consideran 5 meses de retraso en su pago para catalogarlo como pérdida [9].

Por lo general, la evaluación mediante un *credit score* se ve reflejada en la asignación de alguna medida que permite comparar y ordenar a los solicitantes en función de su riesgo, y de esta forma puede ser cuantificado. Generalmente los modelos de *credit score* asignan un valor, un puntaje, una calificación, clasificación o rating. Por ejemplo, podemos encontrar un score entre los números 300 y 850, un cliente se considera más solvente entre más alto sea su score.

Por ejemplo, un prestatario con un *credit score* con menos de 600 puntos no será sujeto de un crédito hipotecario con una tasa promedio del mercado. Tendrá que acudir a un prestamista con una mayor tasa de interés.

El *credit score* forma parte de un rol importe dentro de la toma de decisiones de un prestamista, ya sea para otorgarle un crédito o para ampliarle la línea de crédito, y determina los términos en que se otorgan los mismos.

2.3. Métodos de evaluación de *credit score*

Existe una gran variedad de métodos para construir un *credit score* como: análisis discriminante, regresión lineal, regresión logística, modelos probit, métodos no paramétricos de suavizado, métodos de programación matemática, modelos basados en cadenas de Markov, algoritmos de particionamiento (árboles de decisión), sistemas expertos, algoritmos genéticos, redes neuronales, entre otros y el juicio humano.

La literatura sugiere que todos los métodos de *credit scoring* arrojan resultados similares, por lo que la diferencia entre utilizar un método u otro recae en el tipo de información que contamos.

En general existen (no son los únicos) tres tipos de credit score, basados en los pilares fundamentales del ciclo de vida de un crédito:

- Score de Adquisición (*Acquisition Score*)
- Score de Comportamiento (*Behavior Score*)
- Score de Cobranza (*Collection Score*)

2.3.1. Score de Adquisición

Un soporte fundamental para la decisión de otorgar o negar un nuevo crédito, está basado en la información socio-demográfica y económica disponible al momento en que el cliente realiza la aplicación para el crédito mediante la información existente en la solicitud de crédito, así como su reporte de crédito obtenido a través de las sociedades de información crediticia.

La apertura de un crédito es un paso importante dentro de una institución financiera, pues determina el apetito y las directrices de riesgo que la organización está dispuesta a asumir, y previene futuros incumplimientos de pagos por parte de los clientes, es decir, pérdidas asociadas por "malos" clientes.

Cuando un cliente es aprobado se detonan diferentes procesos dentro de la institución, como áreas operativas, financieras y jurídicas, por mencionar algunas. Toda apertura de un crédito conlleva un gasto para la organización, y en caso de que los clientes "malos" sean clasificados como "buenos" generan un gasto no recuperable a la institución financiera.

2.3.2. Score de Comportamiento

Una vez que el cliente ha sido sujeto de un crédito, es necesario un seguimiento del mismo para ello se genera un *Behavior Score*. Este score se ha implementado dentro de diferentes instituciones financieras para la aplicación de campañas de incrementos o decrementos de líneas de crédito, tasas preferenciales, promociones para incentivar sus compras, meses sin intereses o bloqueo preventivo de líneas de crédito.

Para la construcción de éste score usualmente se utilizan variables socio-demográficas y financieras. Pero con mayor importancia también se considera la información histórica propia de la institución del comportamiento del cliente, como saldo promedio, monto de compras, número de meses con crédito aperturado, número de meses en incumplimiento de pago a la institución financiera, etc. El score de comportamiento como parte de la administración del riesgo, apoya en la generación de ingresos generosos sin que la cartera se deslice y caiga en incumplimiento y como consecuencia en pérdida para la empresa.

2.3.3. Score de Cobranza

Este tipo de score es el que cierra el ciclo de vida de un crédito cuando la cuenta se encuentra en incumplimiento de pago, por lo que en este caso se calcula la probabilidad de que el cliente se ponga al corriente dado que en la actualidad éste se encuentra en incumplimiento. Es un modelo diseñado para la recuperación de cobro de deudas, planteado para integrarse en sus operaciones para aumentar recursos recaudados y optimizar los recursos. Combina datos de crédito frescos con información crítica a nivel de cuenta para ayudar a su puntuación con precisión; brinda mayor información que la variable por sí misma.

Se construyen modelos de recuperación específicos para su tipo de deuda considerando la antigüedad de la deuda y la industria al ser más predictivo que los modelos agrupados

genéricos.

2.4. Construcción de un Score de Crédito

Con la descripción de cada uno de los tipos de score's, se aborda de forma particular el tema de score de Comportamiento. Para la construcción de un score de Crédito se pueden definir dos categorías de juicio, el criterio va en función del tipo de información que usaremos para su construcción. Al elegir el tipo de score es necesario estar consciente del tipo de juicio por el que optaremos al elaborarlo, pues es determinante en la toma de decisiones de la institución. Los dos diferentes tipos de juicio son:

- **Score Experto**

  Se conocen así los score's que se construyen cuando no se tiene suficiente información dentro de la institución, se utiliza información de comportamiento de pagos de instituciones similares, se basan principalmente en la experiencia de los desarrolladores y/o analistas. Este tipo de score es ideal para instituciones que apenas comienzan a otorgar crédito.

- **Score Estadístico**

  A diferencia del Score Experto, estos tipos de score's se construyen con información de comportamiento de pago de la propia institución, de tal forma que estos tipos de score siempre muestran un mejor desempeño para medir el riesgo de un cliente, ya que integra metodología estadística para su determinación.

Cuando se pretende construir un score es importante considerar si se hará de forma estadística o por experiencia ya que impacta directamente en el otorgamiento o no de una línea de crédito.

Para efectos de esta tesis se considerará un score de tipo estadístico. Se centrará en asignar un *credit score* considerando información histórica del cliente de la propia institución e información socio-demográfica y financiera.

No existe sólo una forma para la generación de un score de crédito, la gran diversidad de metodologías para el desarrollo de un *credit score* da lugar al siguiente apartado y al Capítulo 3 de la presente tesis.

2.5. Metodología de construcción de un *Credit Score*

Existen diferentes metodologías para la construcción de un *credit scoring*, ya que su forma de calcularse va en función de la información con la que se cuenta. En este trabajo se realiza una propuesta para el cálculo del *scoring* para una cartera de cliente de un banco. El desarrollo de cada uno de los pasos para la generación de score es importante para tener un buen resultado general.

A partir de este punto se presenta la metodología básica que generalmente las instituciones financieras siguen para la construcción de un score de comportamiento, los pasos clave son los siguientes:

1. Seleccionar la ventana de desarrollo y desempeño del modelo.
2. Obtención, validación y limpieza de la información con la que se va a construir el modelo.
3. Selección de información para entrenamiento del modelo y selección de información para validación del modelo.
4. Construcción de la variable objetivo (target).
5. Construcción de variables.
6. Análisis univariado de las variables en estudio.
7. Regresión logística.
8. Escalamiento numérico.
9. Construcción de Score Card.
10. Construcción de indicadores para medir el desempeño del modelo.
11. Validación del modelo.

Para tener claridad en cada uno de los pasos que se tienen que llevar a cabo a continuación se hace una descripción de cada uno de los puntos.

2.5.1. Seleccionar la ventana de desarrollo y desempeño del modelo

La ventana de desarrollo consiste en seleccionar el periodo medido en meses en el cual se revisará el comportamiento de pago de un cliente hasta el momento (t_0) en el que se pretende calcular su probabilidad de incumplimiento de pago.

El otro periodo es una ventana de desempeño, consiste en seleccionar el periodo medido en meses (t+n) en el cual se quiere pronosticar el incumplimiento de pago de un cliente, generalmente son 12 meses.

2.5.2. Obtención, validación y limpieza de la información para la construcción del modelo

Una vez definida la ventana de desarrollo y de desempeño, es necesario recopilar la información con la que se va a desarrollar el modelo, de tal forma que cumpla con los periodos de tiempo establecidos en el paso anterior. Dentro de las instituciones financieras, en muchas ocasiones las áreas operativas son las encargadas de suministrar la información de acuerdo a los requerimientos que se establezcan.

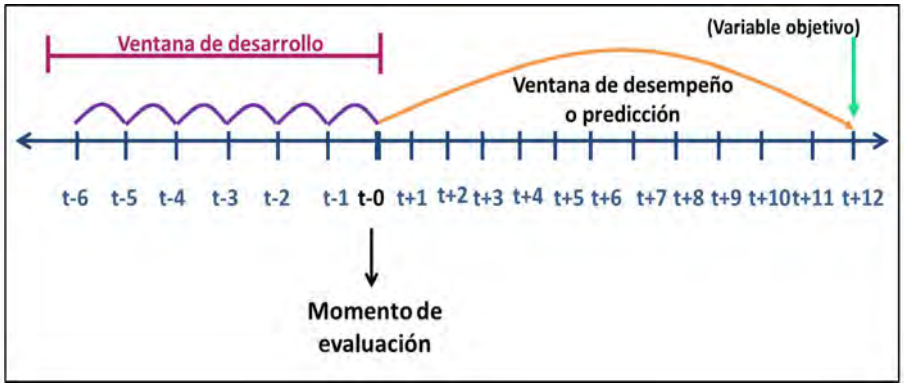


Figura 2.1: Ventana de desarrollo y desempeño.

Una vez que se cuenta con la información necesaria, se verifica la consistencia de la información. Regularmente se realiza la revisión de outliers o valores atípicos, basado en un análisis de distribución de las variables considerando medidas de tendencia central (media, moda y mediana) y medidas de dispersión o también llamadas medias de variabilidad (rango, varianza y desviación estándar).

Otro aspecto que se verifica es la existencia de valores perdidos o faltantes (*missings*), ya que estos valores pueden impactar de forma negativa en el nivel predictivo del score. Cabe mencionar que modelos como redes neuronales y regresión logística no te permiten trabajar con ellos generando error, por el contrario los árboles de decisión son una buena elección para trabajar con ellos, pues les asocia una clasificación.

Existen diferente formas para tratar los valores faltantes. Una de ellas es la imputación de valores.

- Imputación Simple

  Consiste en asignar un valor por cada valor faltante basándose en el valor de la propia variable o de otras variables, generando una base de datos completa. Los valores faltantes de una variable se sustituyen mediante la media de las unidades observadas en esa variable.

- Imputación Múltiple

  Consiste en asignar a cada valor faltante varios valores (m), generando (m) conjuntos de datos completos. En cada conjunto de datos completo se estiman los parámetros de interés y posteriormente se combinan los resultados obtenidos.

De las dos formas de imputar datos que se describieron, normalmente se recurre a la imputación simple.

2.5.3. Selección de información para entrenamiento del modelo y selección de información para validación del modelo

Cuando la información ya fue seleccionada y pasó por un proceso de limpieza que da la confianza para poder trabajar con los datos, se separa el conjunto de datos en dos

subconjuntos, uno de entrenamiento y otro de validación. Generalmente se considera un 70 % de la población para el conjunto de entrenamiento y un 30 % para el de validación. Sin embargo, los porcentajes quedan sujetos a la decisión del analista.

Esta partición tiene como finalidad minimizar los efectos de las diferencias de datos y comprender mejor el comportamiento del modelo, ya que el modelo se desarrolla sobre el conjunto de entrenamiento y posteriormente se validará sobre el conjunto de validación. La literatura habla de algunas herramientas de muestreo para realizar la partición como:

- Muestreo Aleatorio Simple
- Muestreo Estratificado
- Muestreo Sistemático
- Muestreo por Conglomerado

Regularmente la partición se realiza mediante una muestreo aleatorio simple. El cual consiste en seleccionar un tamaño de muestra n de una población de tamaño N de tal manera que cada muestra posible de tamaño n tenga la misma probabilidad de ser seleccionada. A la muestra así obtenida se le denomina muestra aleatoria simple.

2.5.4. Construcción de la variable objetivo (*target*) mediante matrices de transición

Para una mejor comprensión de la construcción de una variable *target* es necesario conocer algunas de las características de los *buckets* o también llamadas canastas, que es la clasificación que se le asigna a un cliente dependiendo su periodo de incumplimiento de su pago.

Se dice que un cliente es *current* cuando realiza sus pagos puntuales entre su fecha de corte de su crédito o su fecha límite de pago. Si el cliente decide pagar sólo el mínimo, el monto neto de su deuda se cargará al siguiente periodo. Cuando se presenta este caso, el cliente sigue conservando el estado de *current*, ya que no camino a *bucket* 1.

Cuando el cliente no realiza su pago, y cae en su primer periodo de mora, se le conoce como *bucket 1* con referencia de tener de 1 a 29 días de atraso, y si continua sin realizar su pago, avanza al siguiente *bucket 2* y así sucesivamente hasta el *bucket 7*. Es importante mencionar que cada institución financiera delimita el *bucket* a partir del cual se le considera pérdida (“*Write Off*”).

Bucket	Días de atraso de pago
Bucket 0	Al corriente con sus pagos
Bucket 1	De 1 a 29 días de atraso
Bucket 2	De 30 a 59 días de atraso
Bucket 3	De 60 a 89 días de atraso
Bucket 4	De 90 a 119 días de atraso
Bucket 5	De 120 a 149 días de atraso

Cuadro 2.1: Buckets.

Para la definición de una variable objetivo conviene conocer la probabilidad de que un cliente pase de un *bucket* (canasta) a otro, ya sea que avance o retroceda a través del tiempo y para ello son útiles las matrices de transición. Es importante el tiempo permisible en que el cliente bueno, es decir, que está al corriente en pagos, pase a un cliente indeterminado o un cliente con mora, es decir, malo.

Para definir la variable objetivo, hacemos uso de las matrices de transición, es decir, nos preguntamos cuál es la probabilidad de que un cliente estando en un cierto estado de morosidad migre al siguiente estado de morosidad en un tiempo determinado. Cuando dicha probabilidad supere el 50 % podremos decir que hemos encontrado la variable objetivo. Para ello se utilizan métodos de Cadenas de Markov.

Una Cadena de Markov mide la probabilidad de que ocurra un evento que depende de un evento inmediato anterior. Una peculiaridad de esta herramienta analítica es que se dice que tienen memoria ya que recuerdan el último evento y esto condiciona la probabilidades de eventos futuros. Esta dependencia entre evento distingue a las cadenas de Markov de las series de eventos independientes, como tirar unos dados o una moneda al aire.

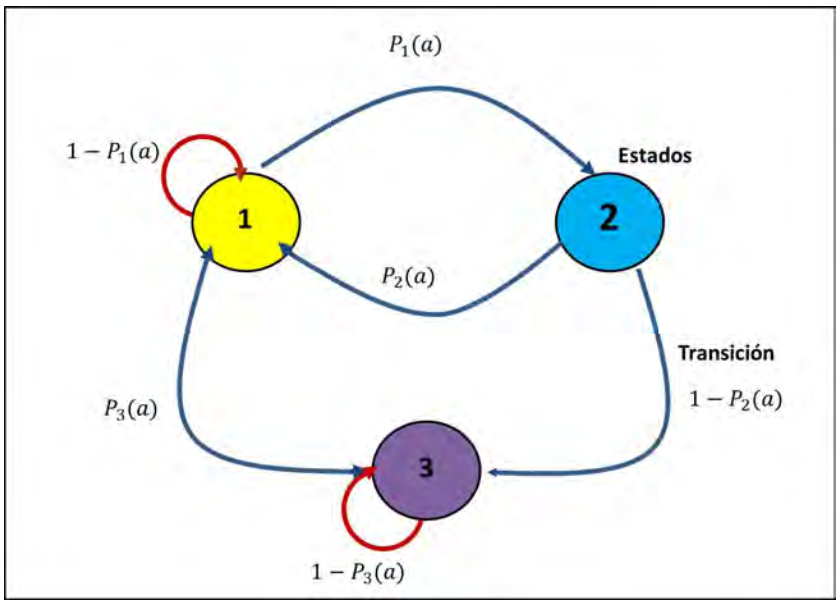


Figura 2.2: Flujo de Cadena de Markov.

Regularmente cuando se pretende generar un score para una cartera ya existente dentro de la institución, la variable objetivo se construye en función de si el cliente presentó o no presentó un comportamiento en particular, como puede ser mora en los últimos 90 días. Esta clasificación de clientes permite considerarla como variable target.

2.5.5. Construcción de variables

Como se mencionó anteriormente para realizar un *credit score* se cuenta con datos socio-demográficos e información propia de la institución. Sin embargo, para tener un mejor conocimiento del comportamiento de clientes resulta de gran atribución para el modelo generar variables nuevas a partir de las ya conocidas, a estas variables también es necesario aplicarles un análisis de relación con la variable objetivo, es decir, un análisis univariado. Come ejemplo tenemos:

PORCENTAJE DE UTILIZACIÓN = Saldo / límite de crédito.

2.5.6. Análisis univariado de las variables en estudio

Por medio del análisis univariado se inspecciona la causa efecto de la variable independiente sobre una única variable dependiente. Se busca encontrar la relación y efecto de una variación en la variable independiente y la dependiente.

Este análisis busca determinar la existencia o no relación entre dos variables, para la cual se realizan diferentes análisis estadísticos. La forma en que se relacionan dos variables se le denomina asociación entre dos variables. El interés se centra principalmente en cómo se distribuye la variable dependiente en función del comportamiento de la variable independiente.

Se considera variable dependiente a aquella que el investigador quiere explicar, que los valores que ésta tome depende de los valores que tomen las otras variables. A las variables que influyen en el cambio de valor de la variable dependiente se les denomina independientes.

Cuando se realiza la evaluación de la relación de cada una de las variables con la variable objetivo sólo consideramos las variables que pueden explicar el incumplimiento de pago de un cliente, de tal forma que dichas variables deben mostrar un comportamiento monótono creciente o decreciente para poder considerar la variable como potencial para ser incluida en el modelo final.

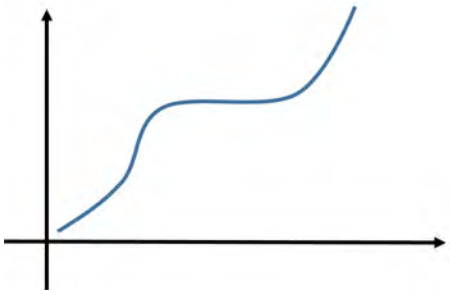


Figura 2.3: Función monótona creciente.

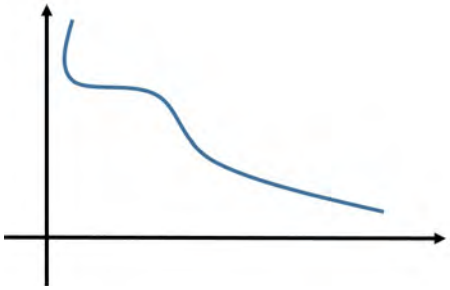


Figura 2.4: Función monótona decreciente.

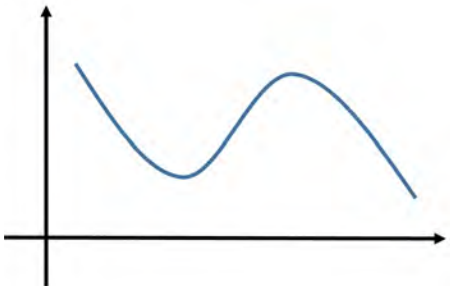


Figura 2.5: Función no decreciente.

2.5.7. Modelos predictivos

Existen modelos de clasificación supervisada (Hand 1997)[1], en estos modelos la variable objetivo es de tipo clase, y también existen los modelos donde no contamos con la variable objetivo. Un modelo predictivo asigna a cada caso un score que mide la propensión de que un caso pertenezca a una clase en particular. Cuando se presentan sólo dos tipos de casos, se dice que el modelo es binario y generalmente representan la ocurrencia o presencia/ausencia de un evento.

Cuando se habla de modelos predictivos también se considera la generalización. La *generalización* consiste en la habilidad de predecir el resultado en los nuevos casos. La estructura de un modelo predictivo es informal y tiene muchas aristas por estudiarse. Se dice que la validación de un modelo es empírica, si un modelo tiene una buena generalización, se puede considerar que el modelo funciona.

A continuación se describen los dos tipos de modelación:

Modelación No Supervisada

Se puede describir a la modelación no supervisada como el aprendizaje sin un conocimiento previo de clasificación, no se conoce la clasificación de "bueno" o "malo" que le corresponde, es decir, no tenemos la variable objetivo. La modelación no supervisada no sirve cuando la información de la base de datos es redundante, por ello uno de los retos de éste tipo de modelaje es discriminar la información redundante.

Algunos de los tipos de modelaje *no supervisado* que podemos encontrar son [2]:

- **Clusterización**

  En la clusterización se particiona a la población de tal forma que las observaciones de cada grupo sean lo más similares posibles entre ellas y lo más diferentes entre las observaciones de los otros grupos que se forman.

- **Prototipo**

  En este tipo de modelaje se forman categorías estereotipadas (Ejemplares).

- **Mapeo de Características**

  En el Mapaeo de Características se hace un mapa topológico de las entradas para que los patrones de entrada similares desencadenen unidades de salida cercanas. Es decir, se definen ciertas condiciones que deben de cumplir los agrupamientos y se buscan los grupos que las cumplan.

[1] David Hand es investigador senior y profesor emérito de Matemáticas en el *Imperial College de Londres*, donde antes tenía la cátedra de estadística. Ha publicado 300 artículos científicos y 28 libros, incluyendo los principios de la minería de datos, generación de información, medición teoría y práctica, la improbabilidad principio y el bienestar de las naciones. En 2002 fue galardonado con la medalla *Guy Royal Statistical Society* y en 2012 él y su grupo de investigación ganó el premio *Credit Collections and Risk Award* por su contribución a la industria del crédito. En 2013 fue nombrado OBE por sus servicios a la investigación y la innovación.

- **Detección de Novedad**

  Para realizar este modelo se puede realizar la siguiente pregunta: ¿Qué tan similar es un nuevo patrón de patrones típicos observados en el pasado?

- **Componentes Principales**

  En este modelo se crean nuevos sistemas de ejes con las variables. Lo que se busca es reducir las variables *n* alineados a las variables ortogonales *m*, se dice que se reduce la dimensionalidad del modelo.

Modelos Supervisados

Por otro lado encontramos a los Modelos Supervisados, estos modelos tienen la particularidad de contar con la variable objetivo, es decir que sabemos la clasificación del cliente o sujeto de evaluación.

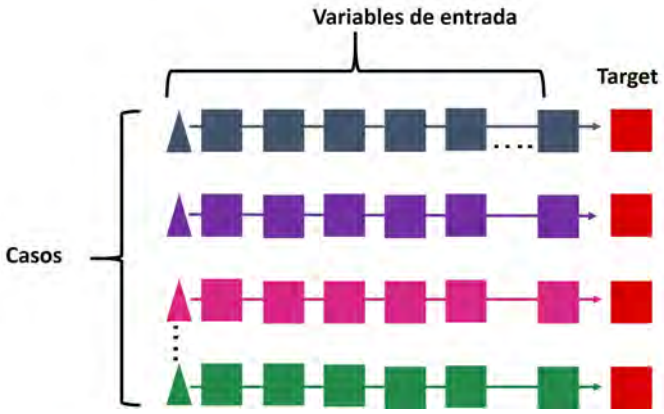


Figura 2.6: Variable Objetivo.

Los modelos que comúnmente se utilizan para éste tipo de problemas son:

- **Árboles de Decisión**

  Representan la segmentación de los datos aplicando un serie de simples reglas. Estas reglas asignan una observación a un segmento basado en su valor de entrada. Cada una de las reglas es empleada con orden, es decir una después de la otra, dan como resultado una segmentación jerárquica de un segmento con segmentos. A esta escala de jerarquías se le denomina árbol.

  Dos ventajas importantes que podemos mencionar acerca de los árboles de decisión es que se pueden interpretar con reglas lógicas a diferencia de red neuronal, que se mencionará más adelante y puede trabajar con valores nulos.

  Cada parte del árbol tiene un nombre, al segmento de árbol se le llama nodo. Al primer nodo del árbol se le conoce como nodo raíz y contiene la información completa de la base de datos. A los nodos con nodos sucesores se les denomina ramas, y a los nodos intermedios se les denomina nodos hijos.

Al nodo final de la rama se le conoce como hoja.

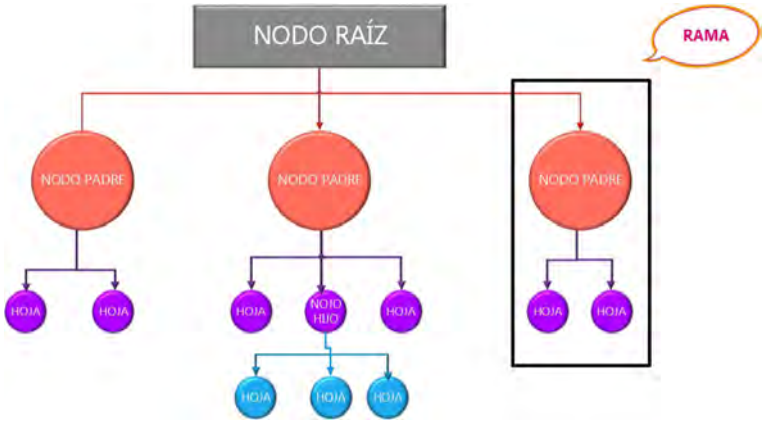


Figura 2.7: Árbol de decisión.

- **Regresión**

  La técnica de la regresión consiste en descomponer la variable Y para cada caso en función de las variables explicativas y en el componente específico de cada caso, al cual se le denomina error.

  Los objetivos de la regresión son explicar la variabilidad de la variable respuesta por medio de las variables explicativas y predecir el valor de la variable respuesta dando valores en las variables explicativas.

  Se puede mencionar que los modelos de regresión con variable respuesta binaria pueden ser modelos logísticos o probit, entre otros. Dado que la variable target para este score es binaria, para saber qué variables son las que mejor explican el incumplimiento de pago del cliente se utiliza este tipo de modelo para calcular el score en el caso práctico.

- **Red Neuronal**

  Las redes neuronales fueron originalmente desarrolladas por los investigadores que trataban de imitar la neurofisiología del cerebro humano. Por medio de la combinación de elementos computacionales en un sistema altamente intercomunicado los científicos esperaban producir fenómenos complejos como la inteligencia. En los últimos años los científicos han incorporado metodología estadística y análisis numérico en sus redes.

  Las redes neuronales son especialmente útiles para los problemas de predicción donde:

  - No hay una fórmula matemática específica para ellas, se sabe que relaciona las entradas a las salidas (tiene un enfoque de caja negra).
  - La predicción es más importante que la explicación.

- Existe una gran cantidad de datos de entrenamiento.

Una vez que se explica la herramienta de modelaje supervisado o no supervisado, continuamos con el siguiente paso del *credit score*, el escalamiento numérico.

2.5.8. Escalamiento numérico

Una vez obtenido los coeficientes de las variables que explican el incumplimiento de pago del cliente mediante la regresión logística, es necesario escalar dicha probabilidad a una escala de mejor comprensión. Generalmente una puntuación que va desde los 400 a los 800 puntos. Es decir, se establece el rango de puntaje que tomará el score. La escala depende del tipo de negocio y de las reglas del negocio, pues éstas varían entre instituciones financieras.

2.5.9. Construcción de un *Score Card*

Un Score Card se puede describir como una tabla que contiene un puntaje asignado a cada una de las características de cada una de las variables utilizadas. El puntaje determina la probabilidad de pago de su crédito cuando se le otorgue una tarjeta de crédito.

Los valores para cada uno de los atributos se obtienen mediante una translación y un cambio de escala de los estimadores de la regresión logística.

A continuación se muestra un ejemplo:

Se considera un score card simple con cinco variables o atributos:

Edad, Estado civil, antigüedad en el empleo, sexo y nivel de estudios. Se cuenta con un caso de un individuo con 37 años de edad, soltero, con 5 años de antigüedad con empleo, de sexo masculino y profesionista. De acuerdo a la tabla 2.2 tendrá un puntaje (score) de: 41 = 10 + 0 + 4 - 10 + 37.

Característica	**Atributos**	*Score*
Edad	Menor a 24 años	-40
	4-30 años	-28
	31-40 años	10
	Mayor a 40 años	-30
Estado Civil	Casado	-12
	Soltero	0
	Otros	-60
Antigüedad Empleo	0-1 años	-5
	2-5 años	4
	6-10 años	10
	Mayor de 10 años	15
Sexo	Masculino	-10
	Femenino	8
Nivel de Estudios	Superior	-15
	Medio	3
	Básica	20
	Profesionista	37

Cuadro 2.2: Ejemplo Score Card [1].

Para efectos de éste trabajo no se llevará a acabo la constitución de la puntuación por *score card*. Ya que principalmente se busca conocer el impacto de la Segmentación de una cartera de clientes en el cálculo del *credit score*.

2.5.10. Construcción de indicadores para medir el desempeño del modelo

Ya que se utilizó el conjunto de entrenamiento para generar el modelo, se realiza una prueba de predictivilidad con el conjunto de datos. Ya que en los datos del conjunto de pruebas se conoce los valores de la variable que deseamos predecir, resulta más sencillo determinar si el modelo tiene o no buena predicción.

Normalmente, la exactitud de la predicción de un modelo se cuantifica mediante la mejora respecto al modelo de predicción o la exactitud de la clasificación. Sobre este tema se hablará más adelante.

Existen dos tipos de medidas para evaluar el desempeño de los score de crédito:

- Las que comparan la distribución de score de los clientes en *clase=0* con los clientes en *clase=1*. Estas no tienen punto de corte (break point).
- Las que reconocen que lo esencial es la realización de una acción con el cliente. Teniendo en cuenta que se realizan acciones diferentes para los que son predichos en la *clase=0* y para los que son predichos en la *clase=1*. Éstas si tienen en cuenta el punto de corte (break point).

Las medidas de Tipo I tienen un enfoque teórico y también estadístico y la de Tipo II son más de uso práctico.

Si bien el criterio número II es más poderoso, ya que refleja mejor el uso de *credit score* y sus reglas de clasificación. No sólo requiere el uso del *credit score* si no también de un punto de corte. Pero muchas de las veces el punto de corte no es muy claro. De hecho casi siempre varía con el transcurso del tiempo, junto con las condiciones económicas.

Esto significa que se presenta la situación en que se quiere evaluar la efectividad del Score de Comportamiento si tiene un punto de corte explícito. Entonces, los criterios Tipo I son apropiados para este caso.

Los criterios Tipos I son:

- Coeficiente Gini.
- Estadístico de Kolmogorov-Smirnov.
- Diferencia de Medias.
- Divergence (Fair Isaac).

2.5.11. Validación del modelo

Es el último paso, y consiste en correr el modelo sobre la muestra de validación verificando que los indicadores del paso anterior sean consistentes, de ésta manera estaremos seguros que el modelo funcionará para poblaciones futuras.

En el siguiente capítulo se expondrán las bases estadísticas sobre las cuales se estará posteriormente trabajando el caso práctico.

Capítulo 3

Métodos estadísticos

A lo largo de este capítulo se desarrollarán los conceptos estadísticos necesarios para generar un *credit score* en el Capítulo 4. Se profundizará en la modelación supervisada, principalmente en la regresión logística y en los árboles de decisión, así como en sus medidas de desempeño.

Se comenzará haciendo alusión a la *base de datos* necesaria para ajustar un modelo predictivo. Esta base, consiste en un grupo de casos o también llamadas observaciones. A cada caso se le asocia un vector de variables de entrada, es decir variables explicativas, una *variable objetivo* o también llamada de respuesta. El modelo busca encontrar el vector de variables de entrada (variables independientes), es decir, del grupo de variables explicativas que nos lleven a la variable objetivo.

La *variable objetivo* (target) es la variable de salida que se debe predecir, por lo que se define como la variable dependiente del modelo. Una vez que se obtiene el modelo, éste se aplica para cada caso de la base de datos.

Regularmente se pueden encontrar casos cuando una base tiene un gran número de variables de entrada, y también tiene diferentes escalas de medida. Por lo que es necesario conocer los diferentes tipos de escalas en que se puede clasificar la variables:

- **Intervalo** como montos, sueldos y estaturas.
- **Binarias o nominales** como tipos de profesión o nombres.
- **Variables ordinales** como grado escolar. Es común que las variables nominales y ordinales con muchos niveles estén presentes en los modelos de regresión.

Cuando realizamos un modelo es necesario hablar de la *dimensionalidad*. Se refiere al número de variables de entrada (grados de libertad) que contiene una base de datos. Los modelos predictivos consideran un gran número de variables de entrada, lo que genera que su gran dimensionalidad limíte la habilidad de explorar y modelar la relación entre variables. Se dice que la complejidad de la base de datos aumenta rápidamente con el

incremento de la dimensionalidad (Breiman 1984)[1].

Una de las soluciones para disminuir la complejidad de la base es la reducción de dimensionalidad, para ello es necesario discriminar la variables redundantes sin dejar fuera las variables importantes del modelo.

En el modelo predictivo, el evento de interés suele ser raro. Por lo general, más datos conduce a mejores modelos, pero tener un número cada vez mayor en casos de no acontecimiento puede tener incluso efecto negativos en la efectividad de predicción del modelo. Cuando el evento que se quiere predecir es raro, el tamaño efectivo de la muestra para la construcción de un modelo de predicción fiable es cerca de 3 veces el número de casos de eventos que el tamaño nominal del conjunto de datos (Harrell 1997)[2]. El aparentemente conjunto de datos masivo podría tener el potencial predictivo de un conjunto más pequeño.

Una estrategia generalizada para predecir eventos raros es construir un modelo sobre una muestra desproporcionada, sobre la cual se representen eventos e igual no eventos. Tal análisis introduce sesgos que necesitan ser corregidos de manera que los resultados sean aplicables a la población.

Hay que considerar que un modelo predictivo es un problema de tipo multivariable. Muchos métodos de modelado clásicos (incluyendo regresión logística estándar) se desarrollaron para las entradas con los efectos que tienen una tasa constante de cambiar y no dependen de otros insumos.

Un error común que se comete es el sobreajuste de los datos; es decir, utilizar un modelo demasiado complejo. Un modelo complejo podría ser demasiado sensible a peculiaridades en los datos de la muestra establecidos y no generalizar así a los nuevos datos. Sin embargo, el uso de modelos demasiado simple puede conducir a una falta de ajuste.
Una de las herramientas estadísticas que se estarán utilizando en el Capítulo 3, son los árboles de decisión. A continuación se muestran sus bases estadísticas y sus características.

[1]Breiman, su investigación en los últimos años se centró en el análisis multivariante de cómputo intensivo, especialmente el uso de métodos no lineales para el reconocimiento de patrones y la predicción en espacios de dimensión de altura. Fue coautor de “Los árboles de clasificación y regresión” y desarrolló árboles de decisión como una alternativa computacional eficiente las redes neuronales. Él fue el autor de los libros como “La probabilidad y los procesos estocásticos con miras a aplicaciones”, “Estadísticas con miras a aplicaciones y Probabilidad”

[2]Harrell recibió un doctorado en Bioestadística de la Universidad de Carolina del Norte en 1979. Fue coeditor en jefe de la revista Servicios de Salud y Resultados Metodología de la Investigación de 1998-2005. Es editor asociado de Estadística en Medicina. Ha dedicado su carrera al estudio de los resultados de los pacientes en general y específicamente para el desarrollo de modelos y modelos de pronóstico y diagnóstico precisos para muchas otras respuestas de los pacientes. Gran parte de su trabajo se ha aplicado a los servicios de salud y los resultados de investigación, evaluación de tecnología, bases de datos observacionales y ensayos clínicos.

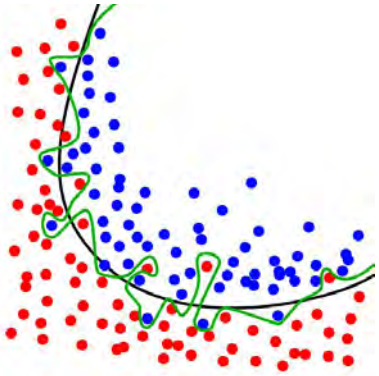


Figura 3.1: Sobreajuste de un modelo.

3.1. Árboles de decisión

Como se comentó en el capítulo anterior, los árboles de decisión también se encuentran dentro del grupo de modelos supervisados. Se dice que es un árbol de decisión porque puede ser representado como la estructura de un árbol. Un árbol de decisión suele ser leído de arriba para abajo. Cada uno de los nodos internos representa una división basada en los valores de una variable de entrada. Las variables de entrada pueden aparecer en cualquier división del árbol. Los casos que contengan estas divisiones irán filtrándose por las ramas hasta llegar a las hojas, la cual es el nodo final de la rama. Las hojas son la representación de la variable objetivo. Todos los casos que caigan dentro de una misma hoja tienen el mismo valor predictivo. Las variables de entrada pueden tener múltiples particiones en diferentes rangos.

Cuando nos encontramos con una variable objetivo categórica, al árbol de decisión se le llama árbol de clasificación. Las hojas nos dan la clase precedida así como la probabilidad de pertenecer a esa clase.

Cuando la variable objetivo es continua, el árbol que se genera es de regresión. Las hojas arrojan la probabilidad de que un caso este contenido en esa hoja. Las reglas que arrojan los arboles son de tipo Booleana. Las reglas Booleanas dan a todo enunciado uno de los dos valores Verdadero o Falso.

Un ejemplo de ellas es:

Si la variable evaluada pertenece a la región de las variables de entrada, por lo tanto el valor predictivo= valor de la región.

Para conocer más acerca de los árboles de decisión, a continuación se describen algunas de sus características básicas:

3.1.1. Poda de un árbol

Otro aspecto importante del árbol de decisión es el control de su crecimiento, a esta acción se le denomina poda (*pruning*). Por medio de la poda se puede elegir cuántos niveles y/o en cuántas hojas se va a dividir el árbol. Un árbol puede crecer hasta que todos su nodos sean puros. Cuando el árbol tiene el 100 % de exactitud de la base, se alcanza la máxima expansión con una variabilidad residual igual a cero.

Pueden aplicarse dos tipos de poda a un árbol de arriba para abajo y de abajo para arriba. Cuando se aplica una poda de arriba para abajo es análogo a aplicar una regresión "hacia adelante", es decir, tipo *forward*, para elegir la variable. Y cuando se realiza un poda de abajo para arriba es análogo a realizar una regresión "hacia atrás", de tipo backward, con las variables. Ya que en esta tesis se utilizará una poda hacia adelante (*forward*), a continuación se mencionan algunas puntos particulares de esta técnica:

- Limita la profundidad de árbol.
- Limita el monto de la fragmentación.
- Tiene significancia estadística.

3.2. Criterios de particionamiento

Una vez que ya se determinó el crecimiento del árbol, es importante determinar con qué criterio se realizarán las particiones de las reglas. Para el caso de los árboles de clasificación es común que se aplique el criterio de Gini, Entropía, y Ji-Cuadrada.

Para poder entender estos criterios es necesario hablar de la pureza del nodo. Entre los modelos que utilizan el nivel de pureza para realizar la partición es el índice de Gini y Entropía, pues ellos tratan de reducir la impureza realizando particiones.

3.2.1. Índice de Gini

El índice de Gini ayuda a conocer el nivel de concentración de los datos que tiene la distribución de la curva que componen la serie de observaciones de la población. En la economía regularmente se utiliza para medir la desigualdad en los ingresos o desigualdad de la riqueza, en general, se puede utilizar para medir cualquier distribución desigual.

El índice de Gini tiene valores entre 0 y 1 donde:

$$\begin{cases} 0 & \text{Concentración mínima (la muestra está uniformemente repartida a lo largo de todo su rango)} \\ 1 & \text{Concentración máxima (un sólo valor dela población acumula el 100\% de los resultados)[8]} \end{cases}$$

$$G = \frac{\sum(p_i - q_i)}{\sum p_i}$$

donde:

$$p_i = \frac{n_1 + n_2 + n_3 + ... + n_i}{n_n} \cdot 100 \qquad i = 1, ..., n \quad X_i \ Valores$$

$$q_i = \frac{X_1 \cdot n_1 + X_2 \cdot n_2 + ... + X_i \cdot n_i}{X_1 \cdot n_1 + X_2 \cdot n_2 + ... + X_n \cdot n_n}$$

En el siguiente ejemplo se puede observar la aplicación de este índice.

Se tienen la siguiente información de sueldos de una empresa.

Sueldos (en miles)	**Empleados**
15,000	20
25,000	17
32,000	13
37,000	9
40,000	7
50,000	5
90,000	2

El siguiente paso para realizar el cálculo es obtener sus frecuencias relativas y absolutas, para obtener el índice de Gini.

X_i	n_i	$\sum_{j=1}^{i} n_i$	p_i	$X_i \cdot n_i$	$\sum_{j=1}^{i} X_i \cdot n_i$	q_i	$p_i - q_i$
15,000	20	20	27.40	300,000	300,000	13.74	13.66
25,000	17	37	50.68	425,000	725,000	33.20	17.49
32,000	13	50	68.49	416,000	1,141,000	52.24	16.25
37,000	9	59	80.82	333,000	1,474,000	67.49	13.33
40,000	7	66	90.41	280,000	1,754,000	80.31	10.1
50,000	5	71	97.26	250,000	2,004,000	91.76	5.5
90,000	2	73	100.00	180,000	2,184,000	100.00	0.00

$\sum p_i$ (entre 1 y n-1)= 415.07 $\sum(p_i - q_i)$(entre 1 y n-1)= 76.33

$$G = 0.18$$

Un coeficiente de Gini de 0.18 indica que el nivel de concentración del sueldo no es excesivamente alto, es decir, que el sueldo tiene una buena distribución.

3.2.2. Entropía

Como se mencionó con anterioridad, la entropía, es una medida de variabilidad para los datos categóricos e impureza del nodo. Se puede describir como el promedio de la rareza y mide la incertidumbre del resultado. Es la medida de la aleatoriedad de la distribución de los registros en un subconjunto de registros con respecto al atributo de clase.

El criterio de entropía es equivalente a usar la estadística de prueba cociente o razón de probabilidad de la Ji-Cuadrada para la asociación entre las ramas y las categorías objetivo. Si una hoja es completamente pura, entonces las clases en la hoja se pueden describir como que todas caen en la misma clase.

La entropía de un nodo del árbol de decisión en particular, es la suma sobre todas las clases representadas en el nodo, de las proporciones de los registros que pertenecen a una clase particular multiplicado por la base dos de logaritmo de la proporción (en realidad esta suma se multiplica por lo general por -1 con el fin de obtener un número positivo).

Se tiene el caso cuando la variable objetivo es binaria.

$$H = -[p_1 log_2(p_1) + p_2 log_2(p_2)]$$
$$H = -1 * [0.5 * log_2(0.5 + 0.5 * log_2(0.5)]$$
$$H = 1$$

3.2.3. Ji-Cuadrada

La prueba de Ji-Cuadrada se rige bajo la distribución de su mismo nombre, esta prueba contrasta las frecuencias observadas con las frecuencias esperadas de acuerdo con la hipótesis nula.

Esta prueba se aplica cuando existen r categorías de la propiedad A y c categorías de la propiedad B, cada uno de los experimentos han sido clasificados en una de las $r\ x\ c$ categorías cruzadas. En un tabla $r\ x\ c$ el contenido de la celda *(i,j)* denotado por X_{ij} es la cantidad de experimentos que tuvieron la clasificación A_i y B_j.

Se establecen las hipótesis:

H_0: Todas las proporciones de la población son iguales.
H_1: Al menos un par de las proporciones de la población son distintas.

Para realizar esta prueba es necesario calcular las frecuencias esperadas y posteriormente su estadístico de prueba:

$$x_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

El estadístico de prueba se compara con el valor crítico de la x_α de la tabla de Ji-Cuadrada correspondiente con el grado de significancia que se haya establecido al inicio del problema. Y los grados de libertad que se reflejen en el tamaño de la tabla. Los grados de libertad de la columna son el número de filas (categorías) menos 1, o bien, $r-1$. Los grados de libertad de cada fila es igual al número de columnas (muestras) menos 1, o bien, $c-1$:

$$gl = (r-1)(c-1).$$

Si el valor estadístico de la prueba es menor que el valor tabular, la hipótesis nula H_0 es aceptada, caso contrario, H_0 es rechazada.

Para ilustrar mejor esta prueba a continuación se muestra un ejemplo numérico:

Se tiene una empresa que se dedica a la venta de cigarros a hombres y mujeres, tiene tres productos diferentes. A 177 personas fumadoras del producto se les preguntó, ¿qué tipo de cigarro preferían? con la finalidad de saber si la preferencia del cigarro es independiente del sexo. Los resultados se muestran en la tabla 3.1.

	Malvoro Rojos	**Malvoro Mentolados**	**Malvoro Light**	**Total**
Hombres	10	40	10	60
Mujeres	57	45	15	117
Total	67	85	25	177

Cuadro 3.1: Ejemplo Ji-Cuadrada.

Con los datos anteriores se formuló la siguiente hipótesis:

H_0: Las variables sexo y tipo de cigarro son independientes.
H_1: Las variables sexo y tipo de cigarro no son independientes.

Es necesario calcular la frecuencia esperada para cada una de las celdas de la tabla, en el siguiente ejemplo se calcula para una celda.

La frecuencia esperada para las 14 mujeres que eligieron Malvoro Mentolados es:

$$\frac{67 \cdot 60}{177} = 22.71$$

La tabla 3.2 muestra todas las frecuencias esperadas obtenidas:

	Malvoro Rojos	**Malvoro Mentolados**	**Malvoro Light**
Hombres	22.71	28.81	8.47
Mujeres	44.29	56.19	16.53

Cuadro 3.2: Frecuencias Esperadas.

Posteriormente se sustituye en la formula del valor del estadístico de prueba, y se tiene:

$$\frac{(10-22.71)^2}{22.71}+\frac{((40-28.81)^2}{28.81}+\frac{(10-8.47)^2}{8.47}+\frac{(57-44.29)^2}{44.29}+\frac{(45-56.19)^2}{56.19}+\frac{(15-16.53)^2}{16.53}=17.75$$

Es necesario comparar el 17.75 obtenido con el valor crítico de la tabla de Ji-Cuadrada:

$$x^2_{\alpha,(r-1)(c-1)}=5.991$$

Como el estadístico $x_0^2 > 5.991$ *por lo tanto se rechaza* H_0*. Existe una gran discrepancia entre lo observado y lo esperado. Entonces la preferencia del tipo de cigarro es diferente para los dos sexos.*

La prueba de Ji-Cuadrada se puede utilizar para juzgar el valor de la división. Evalúa si la distribución de las columnas (de los nodos hijo) son iguales a las filas (proporción de las clases). La prueba estadística mide la diferencia entre las celdas y lo que se esperaría que las ramas y la variable objetivo (columnas y filas) fueran independientes.

La significación estadística de la prueba no está monótonamente relacionada con el tamaño de la prueba estadística de la Ji-Cuadrada. Los grados de libertad de la prueba son:

$$(r-1)(B-1)$$

Donde r (niveles objetivo) y B (ramas) son las dimensiones de la tabla.

El valor esperado de una prueba estadística con v grados de libertad es igual V. Por consiguiente árboles más grandes con mayor cantidad de ramas, naturalmente tienen una Ji-Cuadrada más grandes.

3.3. Regresión logística

La regresión logística es un modelo estadístico en el que se busca explicar y predecir la relación entre una variable categórica binaria llamada dependiente en función de variables independientes que puedes ser de tipo cuantitativas o cualitativas.

Es importante mencionar que la variable dependiente debe ser de tipo dicótoma, es decir, que se pueda clasificar como 0 ó 1. Las regresiones logísticas permiten:

- Cuantificar la importancia de la relación existente entre cada una de las variables independientes y la variable dependiente, lo que lleva implícito también clarificar la existencia de interacción y confusión entre variables independientes respecto a la variable dependiente (es decir, conocer la *odds ratio* o también conocidos como cociente de momios para cada una de ellas).

- Clasificar individuos dentro de las categorías (presente/ausente) de la variable dependiente, según la probabilidad que tenga de pertenecer a una de ellas dada la presencia de determinadas variables independientes.

El modelo debe ser el más reducido que explique los datos, y que además sea congruente e interpretable. Hay que considerar que entre mayor cantidad de variables contenga, mayor será el error estándar asociado.

La regresión logística tiene como propósito:

- Predecir la probabilidad de que a alguien le ocurra cierto evento: por ejemplo, "estar desempleado" =1 o "no estarlo" = 0; "ser pobre" = 1 o "no ser pobre" = 0; "graduarse como actuario" =1 o "no graduarse" =0.
- Determinar qué variables pesan más para aumentar o disminuir la probabilidad de que a alguien le suceda el evento en cuestión.

En cuestión de variables deben incluirse todas aquellas variables que se consideren importantes para el modelo, esto se puede verificar con un análisis univariado, para demostrar su dependencia o independencia con la variable objetivo.

3.3.1. Métodos de selección

Una vez que se dispone del modelo inicial se debe proceder hasta encontrar el modelo más reducido posible. Para ello existen métodos de selección como "hacia adelante" o por eliminación "hacia atrás", o la selección de variables por mejores subconjuntos de covariables.

Algunas de las características de método **hacia adelante** son:

1. Se inicia con un modelo vacío.
2. Se ajusta un modelo y se calcula el valor p de incluir cada variable por separado.
3. Se selecciona el modelo con la variable más significativa.
4. Se ajusta un modelo con la(s) variable(s) seleccionada(s) y se calcula el valor p de añadir cada variable no seleccionada por separado.
5. Se selecciona el modelo con la más significativa.
6. Se repite el paso 4-5 hasta que no queden variables significativas para incluir.

Del método **hacia atrás** son:

1. Se inicia con un modelo con TODAS las variables candidatas.
2. Se eliminan, una a una, cada variable y se calcula la pérdida de ajuste al eliminar.
3. Se selecciona para eliminar la menos significativa.

4. Se repite el paso 2-3 hasta que todas las variables incluidas sean significativas y no pueda eliminarse ninguna sin que se pierda ajuste.

Y de la selección de variables por mejores subconjuntos de covariables o también llamado **stepwise**:

- Se combinan los métodos adelante y atrás.
- Puede empezarse por el modelo vacío o por el completo, pero en cada paso se exploran las variables incluidas, por si deben salir y las no seleccionadas, por si deben entrar.
- No todos los métodos llegan a la misma solución necesariamente.

3.3.2. Ecuación de la regresión logística

En este punto se presenta una de las formas más comunes de encontrar la ecuación para desarrollar un modelo de regresión logística univariante múltiple en la literatura:

$$y = \frac{exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k)}{1 + exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k)}$$

donde:

X es el conjunto de k variables $\{x_1, x_2, ...x_k\}$ que forman parte del modelo.

β_0 es la ordenada al origen
$\beta_i = \{\beta_1, \beta_2, \beta_3, ..., \beta k\}$ los coeficientes de las variables
$exp(\beta_0)$ y $exp(\beta_i)$ se denominan *odds ratio* o cociente de momios[7].

La regresión logística es utilizada regularmente en las ciencias médicas y sociales. En su mayoría cuando las variables que se analizan son cualitativas. En economía se utiliza, cuando se requiere saber si una empresa es rentable o no es rentable, si un cliente es de bajo o alto riesgo. Una de las pruebas médicas que utiliza una regresión logística saber si intervención de laparoscopía para tratar una hernia ofrece un menor riesgo de complicaciones postoperatorias que otra técnica tradicional. La variable objetivo es padecer o no complicaciones (si o no), se establecen como variables independientes el tipo de operación.

3.4. Medidas de desempeño

Como se mencionó con anterioridad para cualquier tipo de indicador es importante conocer su rendimiento, con estas medidas tendremos mayor certeza de la capacidad de predictiva de un modelo. De igual forma estas medidas de desempeño nos ayudan a comparar una primer versión del modelo versus la calibración del mismo después de un tiempo de uso.

A continuación se muestran algunas pruebas estadísticas que regularmente se utilizan para la medición del *credit score*.

3.4.1. Kolmogorov-Smirnov

Es una prueba no paramétrica de igualdad de distribuciones de probabilidad, que se puede usar para comparar una muestra con una distribución de probabilidad de referencia, o para comparar dos muestras. En la prueba se compara la distribución de frecuencia acumulada de la distribución teórica $F_0(x)$ con la distribución de frecuencia acumulada de la muestra $S_n(x)$.

La función de distribución empírica es:

$$S_n(x) = \frac{\text{Num de observaciones}}{n}$$
$$S_n(x) \text{ es la proporción de observaciones de la muestra.}$$

Si se tienen n observaciones distintas y ordenadas de manera ascendente, $x_{(1)}, x_{(2)}, ..., x_{(n)}$

$$S_n(x) = \begin{cases} 0 & \text{si } x < x_{(1)} \\ \frac{i}{n} & \text{si } x_{(i)} \leq x < x_{(i+1)} \\ 1 & \text{si } x > x_{(n)} \end{cases} \tag{3.1}$$

$$i = 1, 2, .., n$$

Las hipótesis se plantean de la siguiente forma:

$$H_0 : F_x = F_0(x) \text{ para toda x.}$$
$$H_1 : F_x \neq F_0 \text{ para al menos una x}$$

El estadístico que se utiliza de contraste es:

$$T = sup[|F_0(x_i) - S_n(x_i)|, |F_0(x_i) - S_n(x_i - 1)|] \tag{3.2}$$

Se rechaza H_0 con significancia α si T excede el cuantil $1 - \alpha$ de la tabla de cuantiles de la prueba de Kolmogrorov Smirnov.

Para tener una mejor idea de cómo se realiza la prueba de KS, a continuación se presenta un ejemplo:

Se tienen 9 clientes de crédito de tarjeta de crédito, en la tabla se muestra su porcentaje de utilización en el mes de julio. Se establece:

$H_0 =$ la función de distribución acumulada de las transacciones es de forma:

$$F_o(x) = \begin{cases} 0 & \text{si } x < x_{(1)} \\ x^2(3 - 2x) & \text{si } x_{(i)} \leq x < x_{(i+1)} \\ 1 & \text{si } x > x_{(n)} \end{cases} \tag{3.3}$$

$$i = 1, 2, .., n$$

H_A : la función de distribución acumulada de las transacciones no es de la forma $x^2(3-2x)$

De igual forma en el cuadro 3.3 se observan los cálculos de $F_0(x_i) - S_n(x_i)$ *y* $F_0(x_i) - S_n(x_i - 1)$ *para cada cliente.*

i	x	x'	$S_n(x_i)$	$F_0(x_i) = x^2 \cdot (3-2x)$	$F_0(x_i) - S_n(x_i)$	$F_0(x_i) - S_n(x_i - 1)$
1	40	0.4	0.11	0.35	0.24	0.35
2	50	0.5	0.22	0.50	0.28	0.39
3	52	0.52	0.33	0.53	0.20	0.31
4	58	0.58	0.44	0.62	0.17	0.29
5	60	0.6	0.56	0.65	0.09	0.20
6	65	0.65	0.67	0.72	0.05	0.16
7	80	0.8	0.78	0.90	0.12	0.23
8	82	0.82	0.89	0.91	0.03	0.14
9	95	0.95	1.00	0.99	0.01	0.10

Cuadro 3.3: Ejemplo Prueba KS.

En el cuadro anterior se puede ver el valor de la prueba $T = 0.39$*, ya que es el máximo entre* $F_0(x_i) - S_n(x_i)$ *y* $F_0(x_i) - S_n(x_i - 1)$ *de los 9 clientes. Mientras que en la tabla de cuantiles de la prueba para dos colas de Kolmogorov Smirnov considerando un* $\alpha = 0.05$ *el valor crítico* $= 0.43$*.*

Como $0.39 < 0.43$ *entonces No se rechaza* H_0*, por lo que no existe evidencia suficiente para decir que la utilización de los clientes no se distribuye de forma* $x^2(3-2x)$*.*

En el caso de un *credit score* $B(S)$ y $G(S)$ son las funciones de distribución acumuladas de los score's de las poblaciones de buenos y malos.

$$D = max|B(S) - G(S)|$$

Se determina el punto en el que estas dos distribuciones muestran su mayor separación. Dicho de otro modo, estima la máxima diferencia entre las funciones de distribución acumulada de los *credit score* de los clientes clasificados como "0" y los score de los clientes clasificados como "1".

Esta medida es equivalente a escoger un punto de corte que minimice la proporción de ceros mal clasificados y la proporción de unos mal clasificados, usando el punto de corte, el cual es una función de los datos. Esto es, potencialmente muy engañoso, puesto que el punto de corte se debe elegir en base a los costos relativos de la mala clasificación.

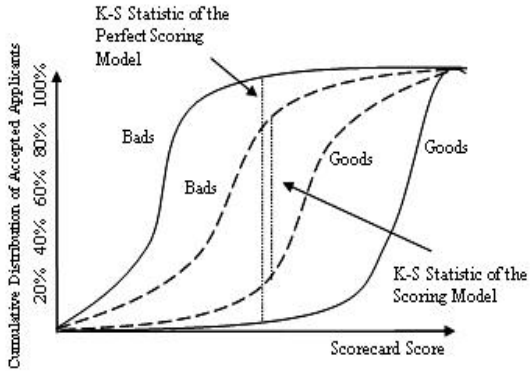


Figura 3.2: Buenos y Malos.

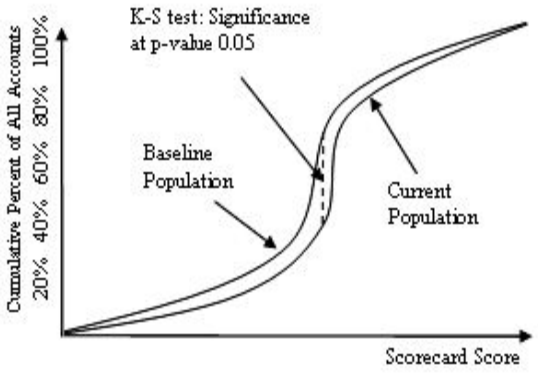


Figura 3.3: Prueba de Kolmogorov-Smirnov.

3.4.2. Índice de Gini

Como se mencionó, el índice de Gini mide la desigualdad entre valores de una distribución de frecuencias. Se aplica en economía para saber la distribución de la riqueza, comparación de importación/exportaciones y para saber la distribución del uso de internet por comunidades, entre otros. Toma valores entre 0 y 1.

$$\begin{cases} 0 & \text{Corresponde a la perfecta igualdad.} \\ 1 & \text{Corresponde con la perfecta desigualdad.} \end{cases}$$

El cálculo de coeficientes del índice se puede realizar por medio de la curva de Lorenz. Esta curva es un gráfico que regularmente se utiliza para representar la distribución relativa de una variable en un dominio determinado. En la Fig.3.4 la línea de 45° representa la igualdad perfecta y corresponde a una distribución perfecta de los ingresos entre la población. La línea azul de la figura representa la desigualdad perfecta, que está dada por:

$$\begin{cases} y = 0 & \text{para } x < 100 \\ y = 100 & \text{para } x = 100 \end{cases}$$

A partir de la curva de Lorenz el índice de Gini se calcula como el cociente entre el área comprendida entre la diagonal y la curva de Lorenz (área A en el gráfico) sobre el área bajo la diagonal (área A+B).

Para tener más claro el índice se presenta un ejemplo:

En la tabla se tiene el promedio de ingresos de 7 distritos y el número de empleados por distrito. Se requiere saber si la distribución del sueldo está concentrada o se distribuye entre los diferentes distritos. También se muestra la forma analítica de calcular el índice de Gini, calculando las frecuencias relativas y absolutas con la formula $G = \dfrac{\sum(p_i - q_i)}{\sum p_i}$, *que se describió en el capítulo anterior.*

Figura 3.4: Curva de Lorenz.

X_i	n_i	$\sum_{j=1}^{i} n_i$	p_i	$X_i \cdot n_i$	$\sum_{j=1}^{i} X_i \cdot n_i$	q_i	$p_i - q_i$
10,000	15	15	18.75	150,000	150,000	8.38	10.37
12,000	13	28	35	156,000	306,000	17.09	17.91
5,000	10	38	47.5	50,000	356,000	19.89	27.61
35,000	9	47	58.75	315,000	671,000	39.49	21.26
60,000	5	52	65	300,000	971,000	54.25	10.75
27,000	27	79	98.75	729,000	1,700,000	94.97	3.78
90,000	1	80	100.00	90,000	1,790,000	100.00	0.00

Obteniendo un índice de Gini$= 0{,}28$*, con un coeficiente de éste nivel no se puede decir que el ingreso se encuentra concentrado en algún distrito en particular, si no que se encuentra distribuido entre los 7 diferentes distritos.*

3.4.3. Divergencia

Es una prueba paramétrica, que asume normalidad de las dos distribuciones. Es la medida más sencilla para medir la separación entre dos distribuciones. Cuando se encuentra con una pequeña divergencia la distribución de una población con la otra es parecida y no se puede diferenciar entre un grupo y otro.

La divergencia se puede aplicar a cualquier variable continua, incluyendo el score. Sin embargo, tiene en contra que no dice nada de la forma de las distribuciones.

Se puede resumir como el cuadrado de la distancia entre las medias muestrales de las distribuciones de ceros y unos, referida a la dispersión promedio de las dos distribuciones:

$$DIV = \frac{2(\bar{x}_{sc.goods} - \bar{x}_{sc.bads})^2}{\sigma^2_{goods} - \sigma^2_{bads}}$$

La mayoría de los *credit score* tienen una divergencia entre 0.5 y 3.0. Y puede tomar valores $[0, \infty]$.

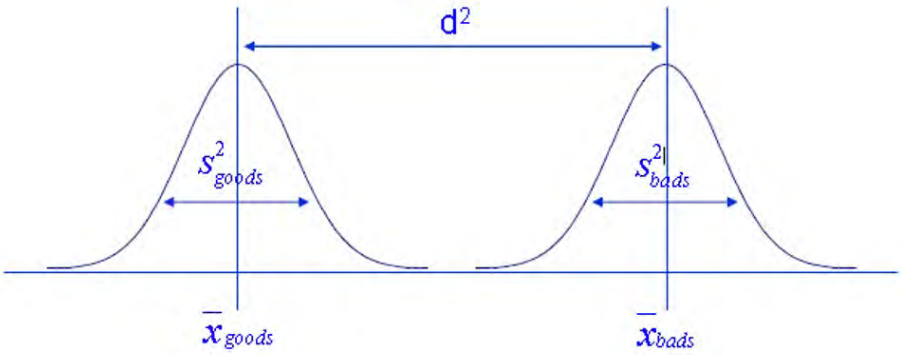


Figura 3.5: Prueba de divergencia.

A continuación se muestra un ejemplo para la divergencia:

Se tiene clasificada una cartera de clientes entre "buenos" y "malos". Se busca calcular la divergencia entre estas dos distribuciones. Para esto se calculó la media y las varianza de ambas distribuciones.

Rango Score	Clientes	Clientes Buenos	Clientes Malos	Clientes Buenos $\bar{x}$	Media Clientes Malos	σ^2_{goods}	σ^2_{bads}
50-200	2,169	2,161	15	270,125	1,875	177,727,333.54	3,889,327.62
200-250	4,227	3,205	35	721,125	7,875	111,812,779.54	5,860,671.20
250-300	5,050	3,970	80	1,091,750	22,000	74,274,433.56	10,322,189.50
300-350	6,360	6,216	150	2,020,200	48,750	46,811,875.26	14,341,048.34
350-400	13,305	11,982	375	4,493,250	14,0625	16,209,376.59	25,194,978.41
400-450	7,858	6,264	594	2,662,200	252,450	1,094,654.25	25,997,140.18
450-500	3,810	3,161	649	1,501,475	308,275	12,633,555.64	16,449,456.28
500-550	2,674	2,121	553	1,113,525	290,325	27,188,332.26	6,594,784.68
550-600	3,519	1,832	687	1,053,400	395,025	48,805,544.64	2,407,996.65
600-650	4,995	1,946	1,049	1,216,250	655,625	88,470,071.26	88,860.69
650-700	3,322	1,872	1,450	1,263,600	978,750	129,700,521.11	2,413,278.60
700-750	3,188	1,545	1,643	1,120,125	1,191,175	151,574,401.97	13,544,810.22
750-MÁS	2,757	1,454	2,303	1,126,850	1,784,825	191823823.33	45,653,682.60
Total	63,234	47,729	9,583	411.78	634.20	22,588.50	18,027.57

Con esta información se obtuvo una Divergencia=21.69. Como el valor de 21.69 mayor que 0.95, esto indica la separación entre las dos distribuciones de "buenos" y "malos".

Capítulo 4

Construcción del Score de comportamiento para una cartera

En los capítulos anteriores del presente trabajo se han abordado temas necesarios para la construcción de un Score de Comportamiento. Como se mencionó en la introducción del trabajo, en éste capítulo 4 se describirá el caso práctico que se llevó a cabo. Este caso se trabajó con el total de la población con la que se contaba, que contempla 352,777 diferentes clientes, y buscó comparar los resultados de la construcción de un score de comportamiento sin segmentación versus un score con una segmentación propuesta para la cartera. Para ello, en el siguiente punto se describirá la ventana de tiempo que se consideró para el cálculo de ambos score's.

4.1. Ventana de tiempo

Como se mencionó en el capítulo anterior es fundamental tener en claro el periodo de tiempo para el que se desarrollará el score. Para este caso se cuenta con una muestra de 6 meses de observación, durante éste periodo se registraron los movimientos de las variables demográficas y financieras del cliente. Como objetivo se busca conocer su *Behavior Score* después de 12 meses de acuerdo a su comportamiento. En la figura. 4.1 se ilustra el tiempo considerado para éste score.

Se puede observar que las variables como revolvencia, saldo, utilización, edad y saldo máximo se consideran dentro de la ventana de desarrollo, con el objetivo de proyectar y obtener un Score de Comportamiento después de 12 meses. También son de suma importancia considerar las variables que se encuentran al momento de la valuación t_0.

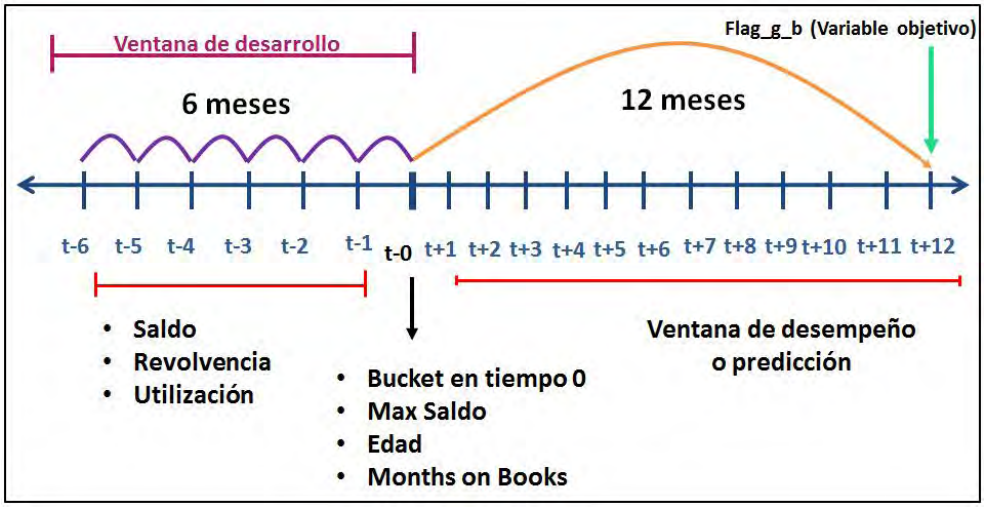


Figura 4.1: Ventana de tiempo.

4.2. Descripción de la información

En este apartado se describirá la información que se tuvo para realizar el caso práctico. Gracias a las facilidades de una entidad financiera, se obtuvo una muestra aleatoria de un universo de clientes que cuentan con un crédito de tipo consumo. Por efectos de confidencialidad de la institución no se revelará el nombre de la misma. Esta muestra cuenta con 352,777 diferentes clientes, y 32 variables demográficas y financieras, 1 variable objetivo y 1 identificador por cliente. En la Cuadro 4.1 se encuentran desglosados los nombres de cada una de las variables que se utilizaron.

Nombres de Variables	Descripción
NCta	Número identificador de la cuenta
MAX_BK_12M	Máxima Mora que ha tenido el cliente en los últimos 12 meses
*BK_T0_*	Bucket que tenía la cuenta en el periodo de evaluación
V_1PV_0	Veces que el cliente ha estado en bucket 1
V_2PV_0	Veces que el cliente ha estado en bucket 2
V_3PV_0	Veces que el cliente ha estado en bucket 3
V_4PV_0	Veces que el cliente ha estado en bucket 4
V_5PV_0	Veces que el cliente ha estado en bucket 5
V_6PV_0	Veces que el cliente ha estado en bucket 6
SAL_0	Saldo que el cliente al momento de la evaluación
SAL_1	Saldo que el cliente un mes antes de la evaluación
SAL_2	Saldo que el cliente dos meses antes de la evaluación
SAL_3	Saldo que el cliente tres meses antes de la evaluación
SAL_4	Saldo que el cliente cuatro meses antes de la evaluación
SAL_5	Saldo que el cliente cinco meses antes de la evaluación
MAX SALDO	Máximo saldo que el cliente a tenido hasta el momento de la fecha de evaluación
LC_T0	Línea de crédito que tenía en el momento de la evaluación
REV_0	Revolvencia que el cliente tiene al momento de la evaluación
REV_1	Revolvencia que el cliente tiene un mes antes del momento de la evaluación
REV_2	Revolvencia que el cliente tiene dos meses antes del momento de la evaluación
REV_3	Revolvencia que el cliente tiene tres meses antes del momento de la evaluación
REV_4	Revolvencia que el cliente tiene cuatro meses antes del momento de la evaluación
REV_5	Revolvencia que el cliente tiene cinco meses antes del momento de la evaluación
UTIL_0	Utilización de la línea de crédito al momento de la evaluación
UTIL_1	Utilización de la línea de crédito un mes antes del momento de la evaluación
UTIL_2	Utilización de la línea de crédito dos meses antes del momento de la evaluación
UTIL_3	Utilización de la línea de crédito tres meses antes del momento de la evaluación
UTIL_4	Utilización de la línea de crédito cuatro meses antes del momento de la evaluación
UTIL_5	Utilización de la línea de crédito cinco meses antes del momento de la evaluación
FLAG_CASH_T0	Indicador si un cliente ha dispuesto de efectivo valuado en el momento cero
REGION_T0	Entidad federativa donde vive
MOB_T0	Meses de antigüedad que tiene el cliente al momento de la evaluación
EDAD_T0	Edad truncada en años del cliente al momento de la evaluación
Flag_g_b	Indicador de si el cliente pagó o no pagó su crédito 12 meses después de la fecha de evaluación

Cuadro 4.1: Descripción de Variables.

Se puede observar que la base incluye variables como revolvencia hasta 6 meses antes de la valuación, utilización hasta 6 meses antes de la valuación, número de veces que un cliente ha estado en cada uno de los *buckets*, edad, etc. Las variables que se muestran hablan del comportamiento que ha tenido el cliente durante el tiempo que ha permanecido con la institución financiera. Si bien cada una de las variables que se presentan en la base son importantes, una variable por sí misma no aporta suficiente información para clasificar al cliente dentro de un modelo de riesgo. De ahí el interés en continuar con el cálculo del *credit score*.

Conocer los estadísticos descriptivos de las variables forma parte importante del reconocimiento de los datos con los cuales se está trabajando. En el cuadro 4.2 se puede verificar que el máximo bucket que muestra la cartera de clientes durante los últimos 12 meses es de 8. Se puede hablar de que es una cartera joven dentro del banco ya que sus meses máximos como cliente de la institución es de 23, mientras que su promedio es de 10 meses. Ningún cliente tiene un valor nulo dentro de las variables. Esto va a permitir trabajar con herramientas como la regresión logística.

Nombre de la Variable	Media	Std Dev	Varianza	Minimo	Maximo	Moda	Rango	Suma	Número de Registros	Valores Nulos	1st Pctl	5th Pctl	10th Pctl	1er Cuartil	Mediana	3er Cuartil	90th Pctl	95th Pctl	99th Pctl
MAX_BK_12M	0.50	0.72	0.51	0.00	8.00	0.00	8.00	176,351.00	352,777.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	2.00	3.00
BK_T0	0.12	0.33	0.11	0.00	1.00	0.00	1.00	41,904.00	352,777.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00
V_1PV_0	0.91	1.93	3.7	0.00	702	0.00	703.00	319,135.00	352,777.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	3.00	4.00	6.00
V _2PV_0	0.10	0.37	0.14	0.00	6.00	0.00	6.00	32,198.00	352,777.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	2.00
V_3PV_0	0.03	0.17	0.03	0.00	3.00	0.00	3.00	7742.00	352,777.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
V_4PV_0	0.01	0.11	0.02	0.00	4.00	0.00	4.00	3,155.00	352,777.00	0.00	0.00	0.00	0	0.00.	0.00	0.00	0.00	0.00	0.00
V_5PV_0	0.01	0.06	0.01	0.00	3.00	0.00	3.00	1,084.00	352,777.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V_6PV_0	0.01	0.04	0.01	0.00	2.00	0.00	2.00	362.00	352,777.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SAL_0	981.46	774.64	600,066.16	100.01	35,741	199	35,640.99	346,234,815.00	352,777.00	0.00	110.36	158.91	216.15	408.68	809.86	1,321	1,938.03	2,418.66	3,772.69
SAL_1	879.27	780.96	609,886.32	0.00	34,141.00	0.00	34,141.00	310,183,966.00	352,777.00	0.00	0.00	0.00	0.00	305.05	727.49	1,252.93	1,848.58	2,312.29	3,584.81
SAL_2	792.69	766.46	587,452.39	0.00	9,472.14	0.00	9,472.14	279,641,066.00	352,777.00	0.00	0.00	0.00	0.00	185.53	642.46	1,184.39	1,760.29	2,212.38	3,377.02
SAL_3	700.44	870.31	757,432.88	0.00	114,288.03	0.00	114,288.03	247,098,660.00	352,777.00	0.00	0.00	0.00	0.00	0.00	527.00	1,084.11	1,646.77	2,081.48	3,196.09
SAL_4	694.92	764.16	583,929.34	0.00	8,691.7	0.00	8,691.7	245,148,345.00	352,777.00	0.00	0.00	0.00	0.00	0.00	528.00	1,120.06	1,683.6	2,114.97	3,208.43
SAL_5	655.55	765.3	585,674.21	0.00	12,667.9	0.00	12,667.9	23,260,760.00	352,777.00	0.00	0.00	0.00	0.00	0.00	445.8	1,087.67	1,653.01	2,096.44	3,166.98
MAX SALDO	1,289.33	952.86	907,937.01	100.01	114,288.03	199.34	114,188.02	454,844,599.00	352,777.00	0.00	130.93	244.29	365	702.59	1,155.2	1,652.5	2,310.00	2,810.12	4,244.91
LC_T0	2,153.14	1,210.82	1,466,074.68	1.00	14,000.00	1,300.00	13,999.00	759,578,642.00	352,777.00	0.00	700.00	1,000.00	1,100.00	1,300.00	1,800.00	2,500.00	4,000.00	5,000.00	6,300.00
REV_0	33.35	41.35	1709.3	0.00	100.00	0.00	100	11,764,808	352,777	0.00	0.00	0.00	0.00	0.00	0.00	77.00	100.00	100.00	100
REV_1	28.58	39.73	1,578.36	0.00	100.00	0.00	100.00	10,081,493.00	352777.00	0.00	0.00	0.00	0.00	0.00	0.00	68.00	100.00	100.00	100.00
REV_2	28.59	39.64	1,570.58	0.00	100.00	0.00	100.00	10,083,818.00	352777.00	0.00	0.00	0.00	0.00	0.00	0.00	64.00	100.00	100.00	100.00
REV_3	9.70	25.33	641.21	0.00	100.00	0.00	100.00	3,421,189.00	352,777.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	44.00	84.00	100.00
REV_4	39.4	37.00	1368.52	0.00	100.00	0.00	100.00	13,898,186.00	352777.00	0.00	0.00	0.00	0.00	0.00	48.00	76.00	84.00	88.00	100.00
REV_5	36.01	36.95	1,365.15	0.00	100.00	0.00	100.00	12,702,272.00	352,777.00	0.00	0.00	0.00	0.00	0.00	29.00	74.00	84.00	87.00	100.00
UTIL_0	50.62	33.59	1,127.97	1.00	150.00	11.00	149.00	17,854,072.00	352,777.00	0.00	4.00	8.00	11.00	22.00	44.00	75.00	100.00	110.00	138.00
UTIL_1	44.62	34.61	1,197.45	0.00	150.00	0.00	150.00	15,738,432.00	352,777.00	0.00	0.00	0.00	0.00	16.00	38.00	70.00	96.00	108.00	127.00
UTIL_2	39.57	34.81	1,211.23	0.00	150.00	0.00	150.00	13,957,214.00	352,777.00	0.00	0.00	0.00	0.00	9.00	33.00	64.00	91.00	105.00	127.00
UTIL_3	34.41	33.93	1,150.76	0.00	150.00	0.00	150.00	12,137,464.00	352,777.00	0.00	0.00	0.00	0.00	0.00	26.00	57.00	85.00	99.00	127.00
UTIL_4	34.39	35.06	1,229.06	0.00	150.00	0.00	150.00	12,129,623.00	352,777.00	0.00	0.00	0.00	0.00	0.00	26.00	60.00	87.00	100.00	119.00
UTIL_5	32.25	34.99	1,223.81	0.00	150.00	0.00	150.00	11,376,761.00	352,777.00	0.00	0.00	0.00	0.00	0.00	22.00	58.00	86.00	99.00	117.00
FLAG_CASH_T0	0.01	0.09	0.01	0.00	1.00	0.00	1.00	2,777.00	352,777.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
REGION_T0	2.52	1.49	2.20	1.00	5.00	1.00	4.00	888,007.00	352,777.00	0.00	1.00	1.00	1.00	1.00	2.00	4.00	5.00	5.00	5.00
MOB_T0	10.15	7.02	49.21	0.00	23.00	0.00	23.00	3,578,504.00	352,777.00	0.00	0.00	0.00	1.00	4.00	10.00	16.00	21.00	22.00	23.00
EDAD_T0	29.22	10.12	102.23	17.00	104.00	20.00	87.00	10,307,715.00	352,777.00	0.00	18.00	19.00	19.00	21.00	26.00	35.00	45.00	50.00	58.00
Flag_g_b	0.19	0.39	0.16	0.00	1.00	0.00	1.00	65,903.00	352,777.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00

Cuadro 4.2: Estadísticos descriptivos.

4.3. Selección de información

Para efectos de este trabajo se realizará la construcción del score para la muestra total de 352,777 diferentes clientes.

4.4. Variable objetivo

Para generar el score, es indispensable identificar la variable objetivo de acuerdo al propósito del modelo. Dentro de la base se cuenta con una variable binaria, llamada fla_g_b, donde 1 nos indica que el cliente sí tuvo 60 o más días vencidos durante los 12 meses de ventana de desempeño y 0 si no tuvo días vencidos. Con ello se puede determinar como 1="malos clientes" y 0="buenos clientes".

Flag_g_b	Total por clase
0	286,874
1	65,903
Total	352,777

Cuadro 4.3: Buckets.

4.5. Variable: porcentaje de utilización

Gracias a que la información sobre la cartera de clientes contaba con la información necesaria para calcular el porcentaje de utilización, se pudo llevar a cabo el cálculo de la variable. De acuerdo a la siguiente formula.

PORCENTAJE DE UTILIZACIÓN = Saldo / límite de crédito.

4.6. *Credit Score* sin segmentación

Como primer paso, para el caso práctico se desarrolló el *credit score* **sin segmentación** de la cartera de 352,777 diferentes clientes, que sirvió de comparación con el Score de Comportamiento aplicando una segmentación de cartera.

4.6.1. Análisis univariado

En este punto se comenzará a describir el análisis de univariado que se llevó acabo. Para ello, se analizó la relación de cada una de las variables determinadas como independientes de los 352,777 clientes, como saldo, revolvencia, utilización, saldo máximo respecto con la variable objetivo y se encontraron los siguientes casos.

3 tipos de casos:

- Cuando la variable se comporta de forma decreciente, y es considerada para el modelo.
- Cuando la variable se comporta de forma creciente, también es importante para el score.
- Cuando la variable presenta una trayectoria tipo plana, este comportamiento de la variable se descarta como parte del modelo, ya que no aporta significancia.

Ya en el análisis de las variables se observó, por ejemplo: las variables utilización en el mes 0 (fig.4.2), máximo bucket en 12 meses (fig.4.3), veces 1 pago vencido (fig.4.4) y revolvencia en el 1er mes (fig.4.5) no tiene un crecimiento totalmente creciente por lo que fue necesario ajustarlas, de tal forma que en su comportamiento no se presentaran picos.

Para realizar el ajuste en las variables fue necesario establecer intervalos, esto se realizó generando veintiles para las variables continuas. Posteriormente obteniendo su porcentaje de clientes malos en cada uno de los intervalos. El porcentaje de malos se obtuvo con la finalidad de saber en que punto del comportamiento de la variable era necesario un agrupamiento de intervalos.

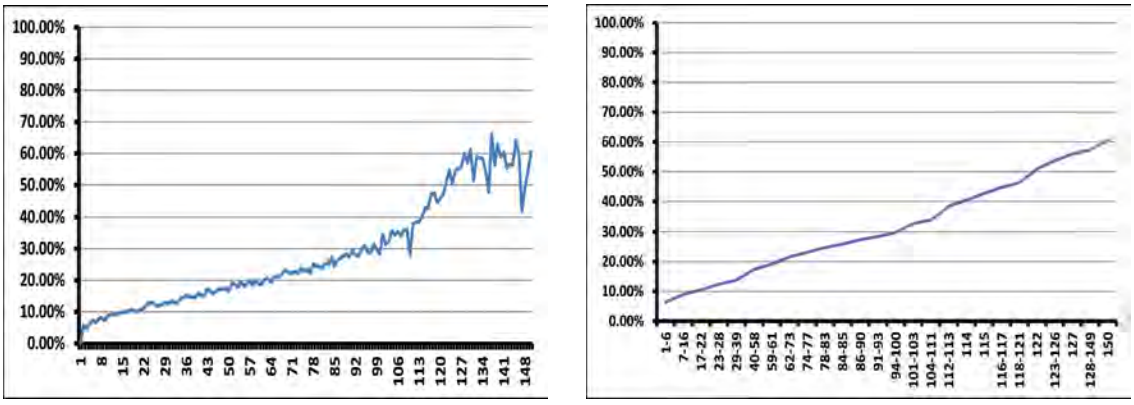
Para el caso de las variables discretas, no fue necesario separar la variable en veintiles. Sólo se calculó su porcentaje de malos, para saber donde era necesario agrupar sus puntos.

En la tabla 4.4 se muestran los veintiles obtenidos para la variable Variable saldo en el tiempo 0 antes de ser ajustada, agrupando veintiles de acuerdo a su porcentaje de clientes malos.

Intervalos	Clientes buenos	Clientes malos	Total general	% Malos
0-158.91	15,947	1,693	17,640	9.6
158.92-216.15	15,853	1,786	17,639	10.13
216.16-275.99	15,659	1,971	17,630	11.18
276-338.72	15,692	1,955	17,647	11.08
338.73-408.68	15,470	2,169	17,639	12.3
408.69-484	15,448	2,198	17,646	12.46
484.01-560.22	15,206	2,424	17,630	13.75
560.23-641.13	15,114	2,526	17,640	14.32
641.14-724.49	14,907	2,733	17,640	15.5
724.5-809.85	14,830	2,807	17,637	15.92
809.86-901.8	14,511	3,129	17,640	17.74
901.81-995.99	14,322	3,316	17,638	18.81
996-1093.95	14,061	3,579	17,640	20.29
1,093.96-1,200.64	13,713	3926	17,639	22.26
1,200.65-1,321	13,463	4,175	17,638	23.68
1,321.02-1,463.22	13,146	4,493	17,639	25.48
1,463.23-1,655.43	12,782	4,856	17,638	27.54
1,655.45-1,938.03	12,657	4,983	17,640	28.25
1,938.04-2,418.59	12,342	5,296	17,638	30.03
2,418.66-35,741	11,751	5,888	17,639	33.39

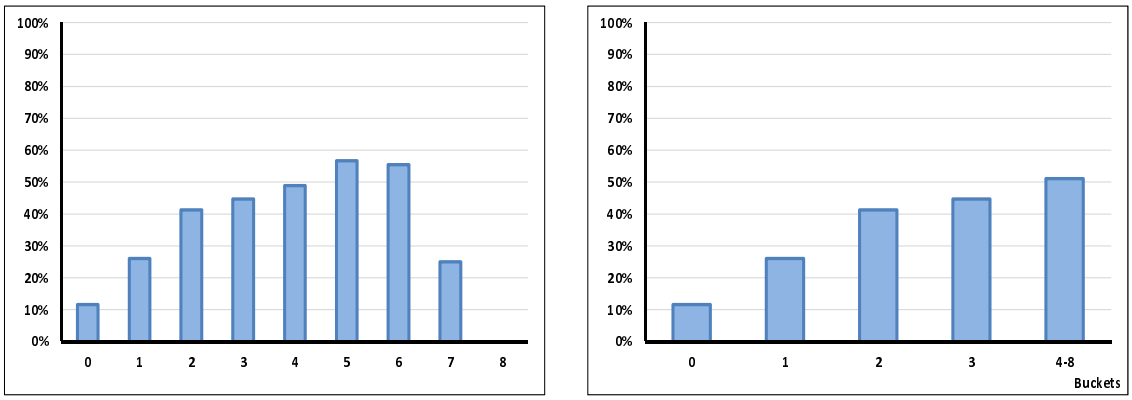
Cuadro 4.4: Variable saldo en tiempo 0 en veintiles.

En la fig.(a) para la variable de Utilización en el Mes 0 la variable muestra picos en su comportamiento, en la fig.(b) se muestra su ajuste, el cual sólo fue en algunos puntos como en los rango 2-3 y 4-5 de su comportamiento figura.4.2.



(a) Antes de ajuste. (b) Después de ajuste.

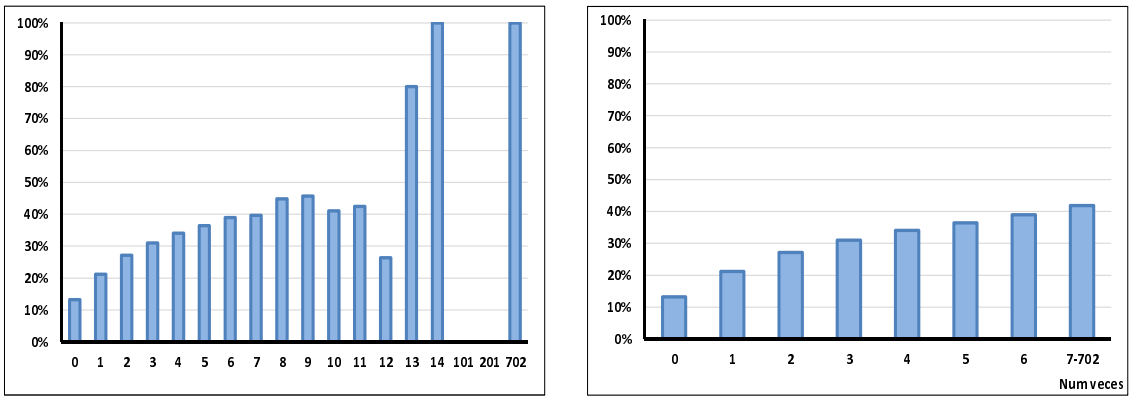
Figura 4.2: Utilización en mes 0.



(a) Antes de ajuste. (b) Después de ajuste.

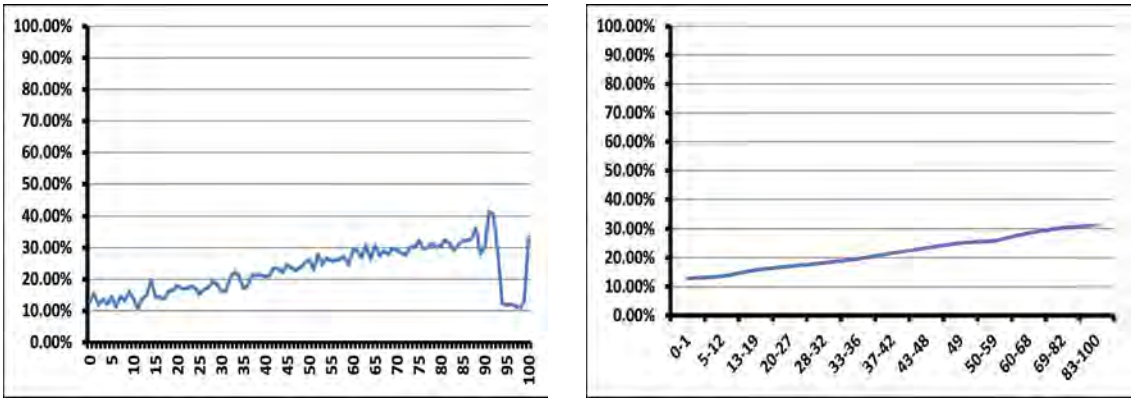
Figura 4.3: Máximo bucket en 12 meses.

Para la variable de número de veces que el cliente cayó en un pago vencido fig. 4.4, se realizó un ajuste en los puntos 101 y 202 de su rango. El ajuste se ver reflejado en la fig. 4.4, dejando la variable con una forma creciente.



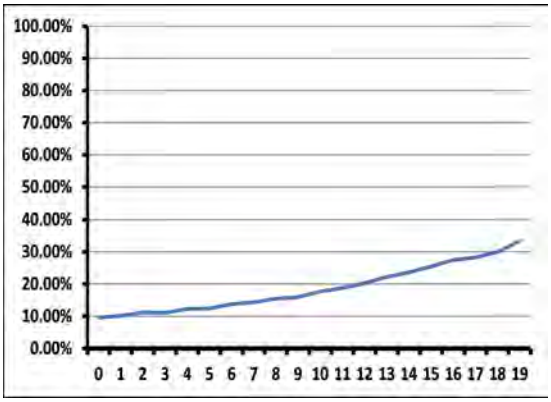
(a) Antes de ajuste. (b) Después de ajuste.

Figura 4.4: Veces 1 pago vencido.

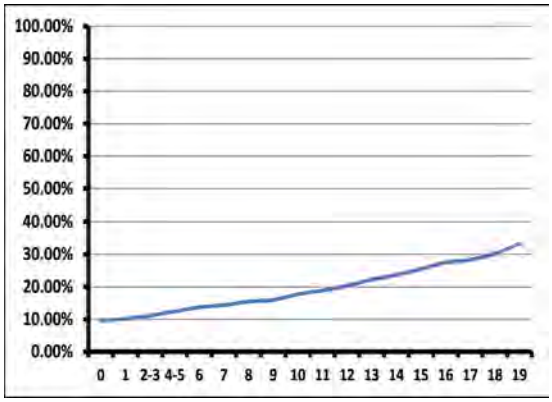


(a) Antes de ajuste. (b) Después de ajuste.

Figura 4.5: Revolvencia 1 mes antes a la valuación.



(a) Antes de ajuste.



(b) Después de ajuste.

Figura 4.6: Saldo en el mes 0.

Como se mencionó se realizó un análisis univariado de 32 variables por cliente que contenía la base de información. En las gráficas 4.7-4.35 se presentan las 29 variables que se consideraron para el score de comportamiento sin segmentación, una vez que se revisó la relación con la variable objetivo.

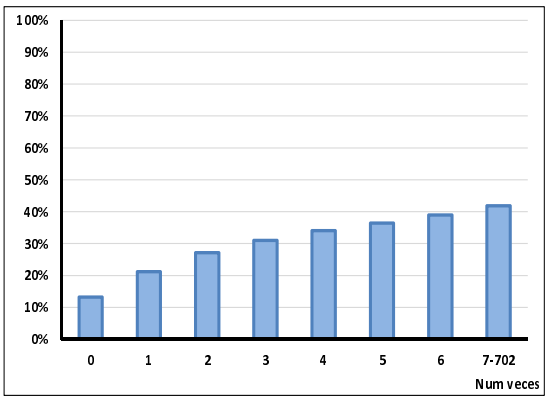


Figura 4.7: Veces 1 pago vencido.

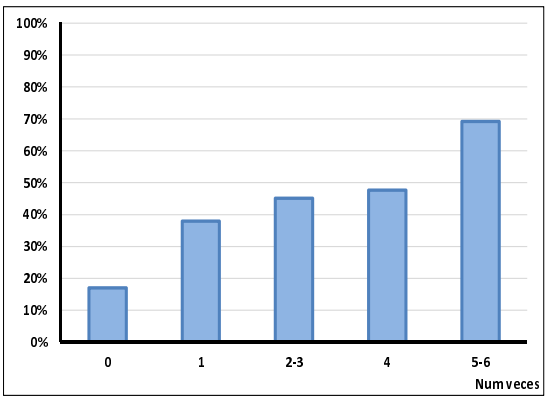


Figura 4.8: Veces 2 pagos vencidos.

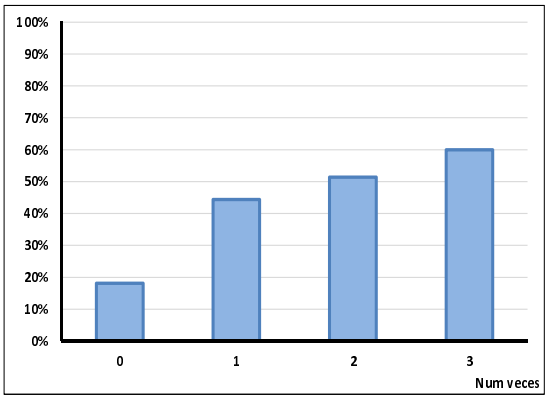


Figura 4.9: Veces 3 pagos vencidos.

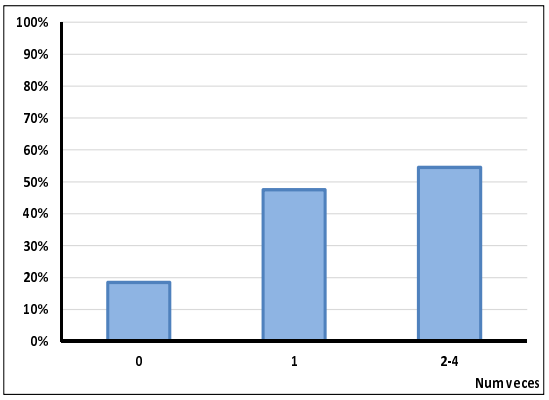


Figura 4.10: Veces 4 pagos vencidos.

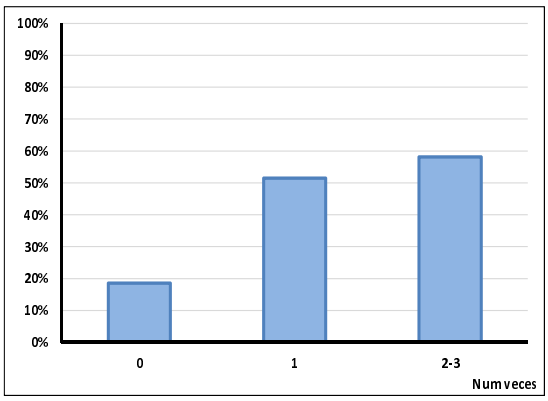


Figura 4.11: Veces 5 pagos vencidos.

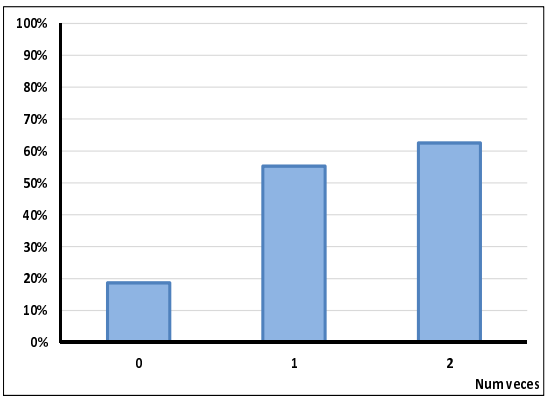


Figura 4.12: Veces 6 pagos vencidos.

Los pagos vencidos muestran un tendencia creciente, el número de veces con 2 pagos vencidos tiene 70 % de clientes malo, es el más alto de entre los 6 tipos de pagos vencidos que se consideraron.

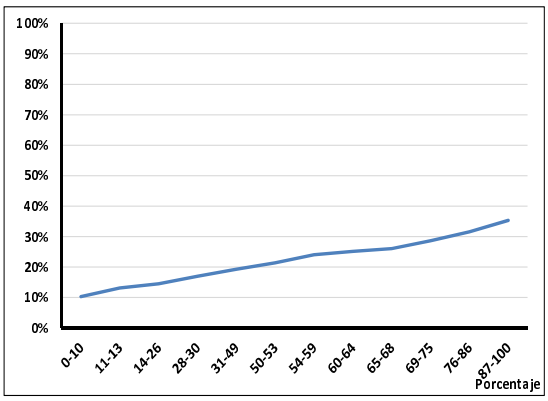


Figura 4.13: Revolvencia mes 0.

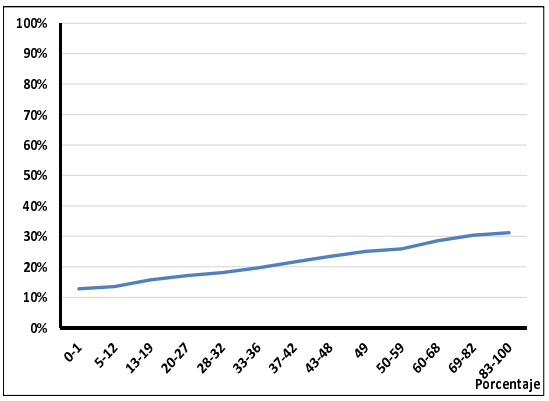


Figura 4.14: Revolvencia 1 mes antes.

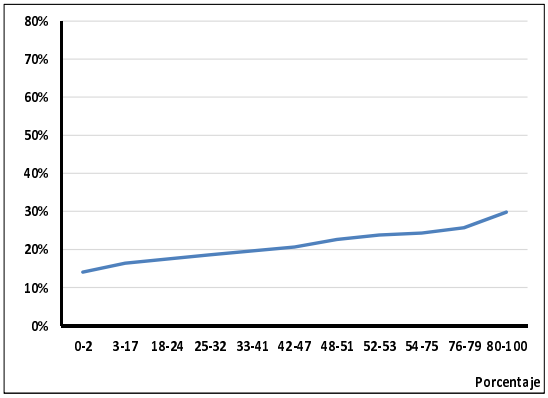


Figura 4.15: Revolvencia 2 meses antes.

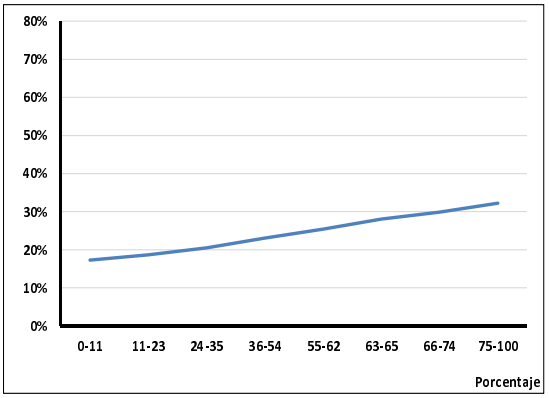


Figura 4.16: Revolvencia 3 meses antes.

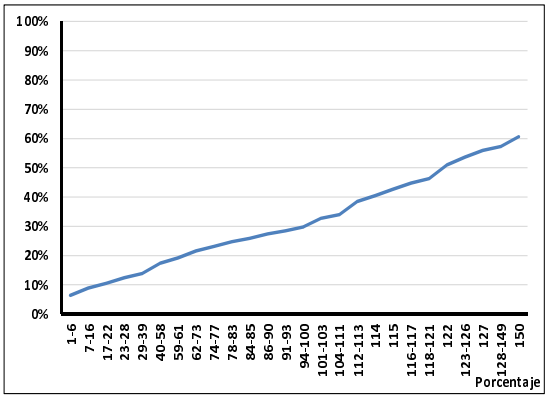


Figura 4.17: Utilización en el mes 0.

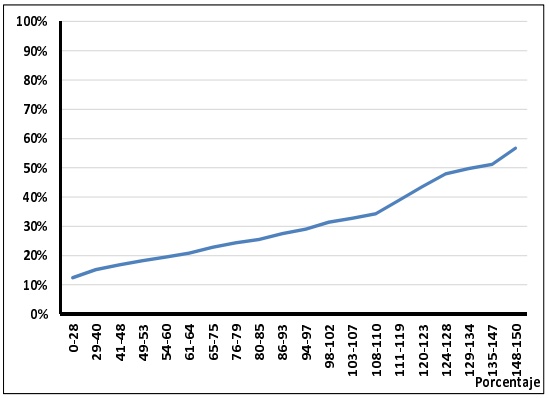


Figura 4.18: Utilización 1 mes anterior.

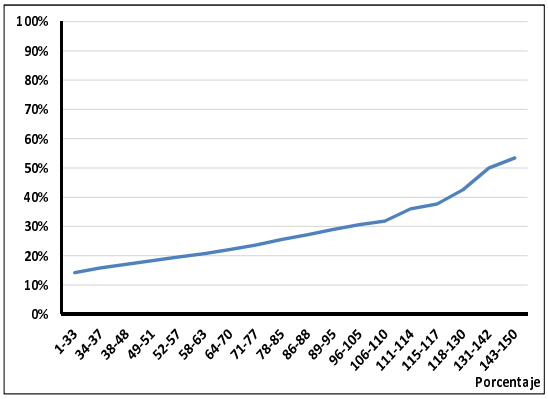


Figura 4.19: Utilización 2 meses anteriores.

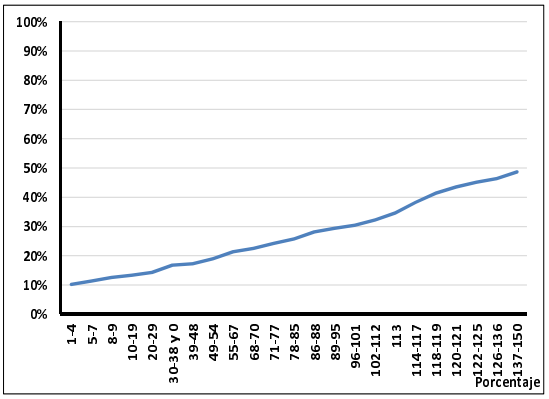


Figura 4.20: Utilización 3 meses anteriores.

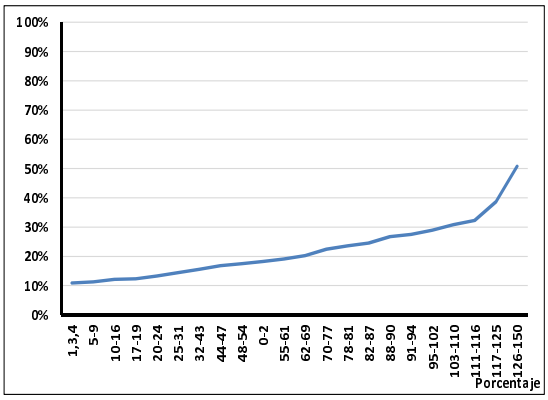


Figura 4.21: Utilización 4 meses anteriores.

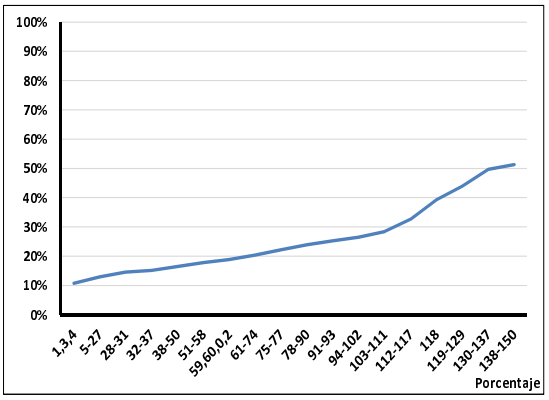


Figura 4.22: Utilización 5 meses anteriores.

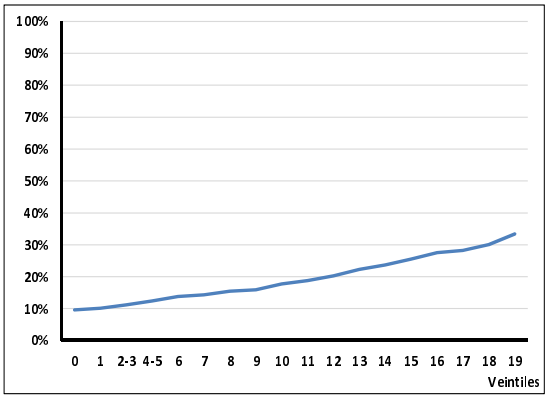


Figura 4.23: Saldo del mes 0.

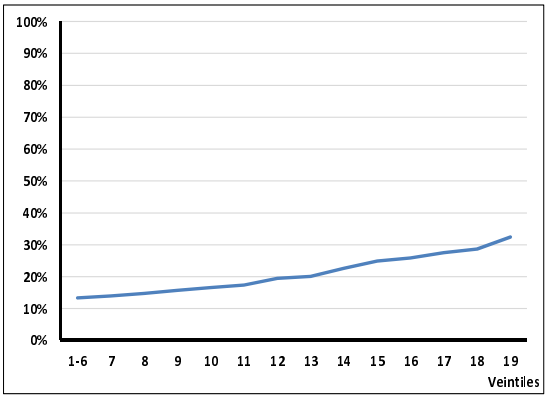


Figura 4.24: Saldo 1 meses anterior.

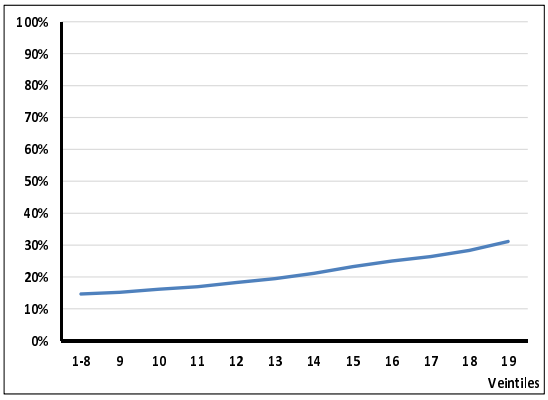


Figura 4.25: Saldo 2 meses anteriores.

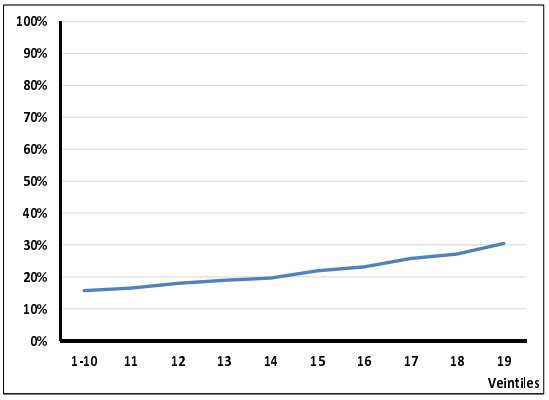


Figura 4.26: Saldo 3 meses anteriores.

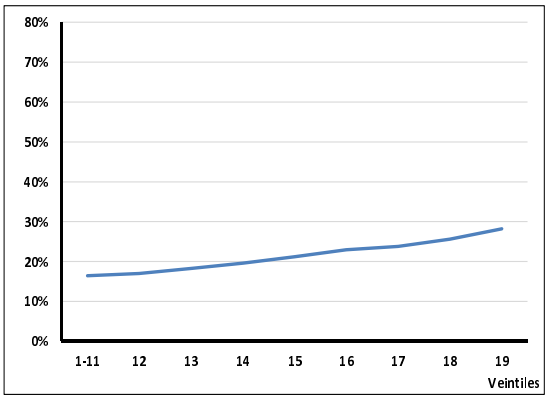


Figura 4.27: Saldo 4 meses anteriores.

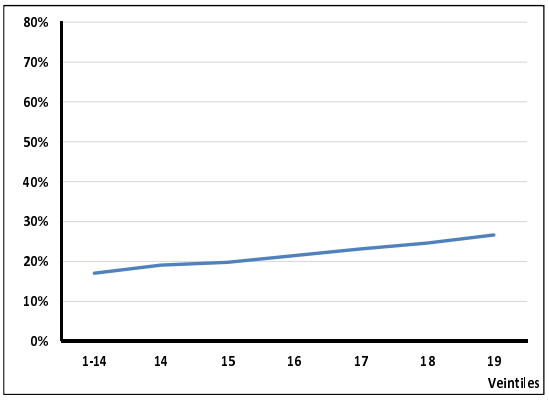


Figura 4.28: Saldo 5 meses anteriores.

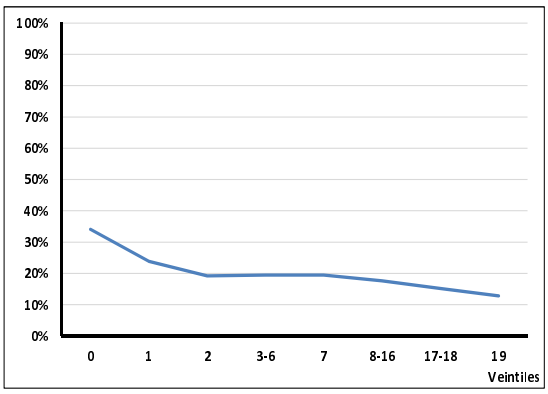


Figura 4.29: Límite de crédito en el mes 0.

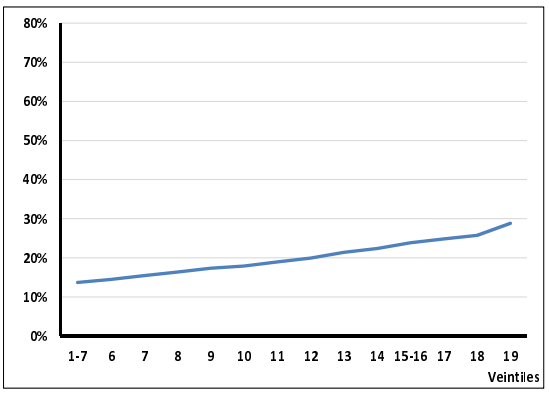


Figura 4.30: Máximo Saldo.

El límite de crédito en el mes 0 tuvo un comportamiento decreciente, es decir que conforme el veintil iba incrementando el porcentaje de clientes malos también.

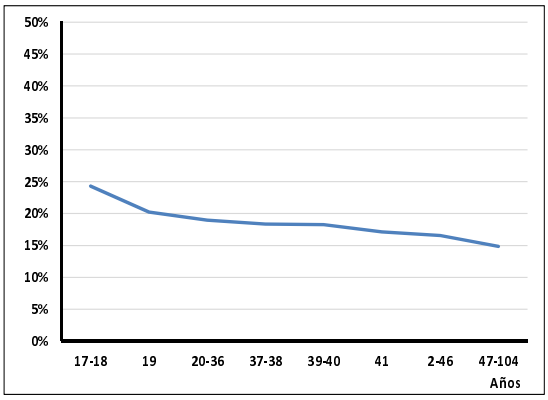


Figura 4.31: Edad.

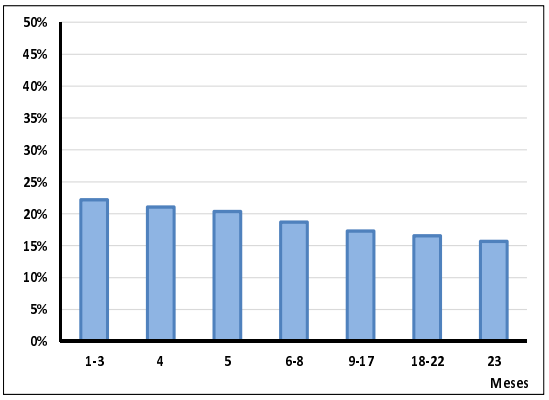


Figura 4.32: Months on Books.

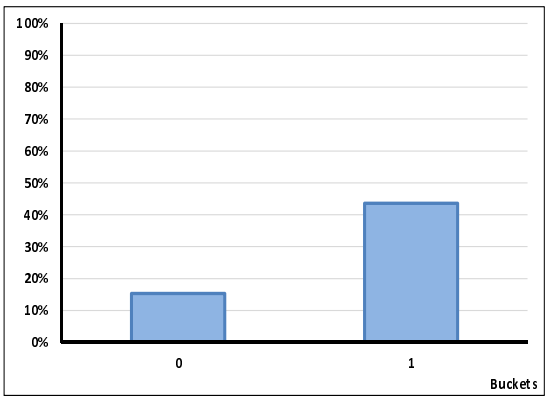


Figura 4.33: Bucket 0.

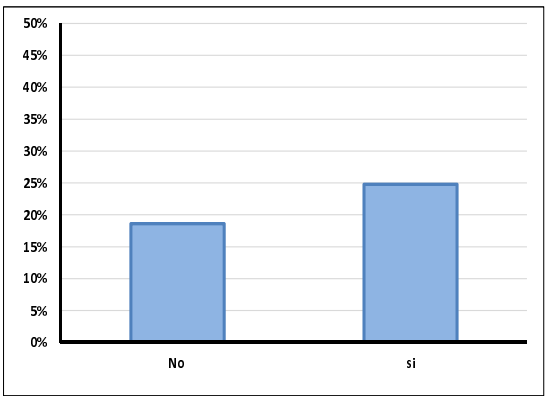


Figura 4.34: Disposición en efectivo.

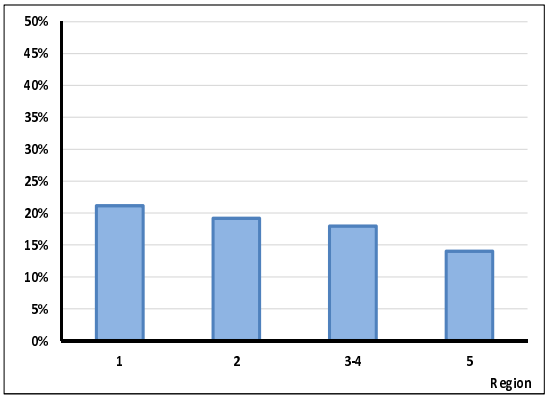


Figura 4.35: Entidad federativa.

4.6.2. Regresión logística aplicada

Como se mencionó en el capítulo 2 el paso que preside al análisis univariado es utilizar la herramienta de regresión.

La regresión se calculó con las 29 variables de la base de datos que mostraron un comportamiento creciente o decreciente una vez ajustada, es decir, las variables que se mostraron en la parte superior. Para la construcción del Score fue imprescindible conocer las variables que mejor describen a la variable dependiente $flag\_g\_b$, para ello se requirió calcular una regresión.

El tipo de regresión que se utilizó para este caso es una regresión logística, con un método *"hacia adelante"* de selección de variables. El método comienza por un modelo que no contiene ninguna variable explicativa y se añade como primera de ellas a la que presente un mayor coeficiente de correlación (en valor absoluto) con la variable dependiente. En los pasos sucesivos se va incorporando al modelo aquella variable que presenta un mayor coeficiente de correlación parcial con la variable dependiente dadas las independientes ya incluidas en el modelo. El procedimiento se detiene cuando el incremento en el coeficiente de determinación debido a la inclusión de una nueva variable explicativa en el modelo ya no es importante.

Al realizar el cálculo de la regresión logística con la cartera sin segmentar, se obtuvieron 18 variables con significancia estadística y el intercepto. Los resultados se muestran en la tabla 4.5.

PARÁMETRO	DF	ESTIMADOR	Pr>JiSq
INTERCEPTO	1	11.791	<.0001
MAX_BK_12M	1	-3.1561	<.0001
BK_T0	1	-1.758	<.0001
V_1PV_0	1	-1.2966	<.0001
V_2PV_0	1	-0.829	<.0001
V_3PV_0	1	0.4072	0.0052
V_4PV_0	1	1.1891	<.0001
REV_0	1	-3.6242	<.0001
REV_1	1	-2.0554	<.0001
REV_2	1	-0.4229	<.0001
REV_3	1	-0.3251	0.0055
UTIL_0	1	-1.9316	<.0001
UTIL_2	1	2.1301	<.0001
UTIL_3	1	0.3227	0.0137
UTIL_4EST	1	-1.2269	<.0001
FLAG_CASH_T0	1	-4.4275	<.0001
REGION_T0	1	-5.2433	<.0001
MOB_T0	1	-20.8827	<.0001
EDAD_T0	1	-3.0198	<.0001

Cuadro 4.5: Resultados de la regresión logística.

Donde los coeficientes de la regresión logística (estimador) muestran el cambio en el logaritmo de los momios (*ODDS* se define como la probabilidad de que un evento se presente en una población, frente al riesgo de que ocurra en otra población), cuando se incrementa

en una unidad la variable predictora correspondiente.

En el caso de la Ji-Cuadrada si la significación es menor de 0.05 indica que la variable ayuda a explicar el evento, es decir, las variables independientes explican la variable dependiente.

4.6.3. Calculo del *Credit Score*

Para continuar con el proceso del cálculo de un *credit score*, se utilizará una formula que las instituciones crediticias generalmente consideran para generar su propio score. Una vez que se contó con las variables significativas es necesario aplicar la fórmula para generar la calificación del *Behavior Score*:

$$\text{Puntaje Inicial} + \left(\frac{PDO}{log(\text{Puntaje para duplicar})} \right) * \left(b_o + \sum_{i=1}^{n} b_i * x_i \right)$$

donde :

PDO Point to dubble ODDS

Para este caso en particular se decidió que el PDO es de 45 y el Puntaje inicial es de 450, estos puntajes pueden variar de acuerdo a la necesidades de negocio. En el siguiente ejemplo se muestra el cálculo de score para un sólo cliente.

Considerando los valores de los estimadores de las variables significativas de la regresión logística y de acuerdo a los valores de los variables dentro de la ventana de tiempo para un cliente ya transformadas en porcentaje de clientes malos se realiza el cálculo del score correspondiente:

Variables del Cliente	
NCta	100001
MAX_BK_12M	0.1163
BK_T0	0.15323
V_1PV_0	0.13206
V_2PV_0	0.17065
V_3PV_0	0.18149
V_4PV_0	0.18439
REV_0	0.10324
REV_1	0.12816
REV_2	0.14026
REV_3	0.17272
UTIL_0	0.25895
UTIL_2	0.1417
UTIL_3	0.13313
UTIL_4	0.1212
FLAG_CASH_T0	0.18633
REGION_T0	0.19165
MOB_T0	0.16547
EDAD_T0	0.18984

Resultado de la Regresión Logística	
INTERCEPTO	11.791
MAX_BK_12M_EST	-3.1561
BK_T0_EST	-1.758
V_1PV_0	-1.2966
V_2PV_0	-0.829
V_3PV_0	0.4072
V_4PV_0	1.1891
REV_0	-3.6242
REV_1	-2.0554
REV_2	-0.4229
REV_3	-0.3251
UTIL_0	-1.9316
UTIL_2	2.1301
UTIL_3	0.3227
UTIL_4	-1.2269
FLAG_CASH_T0	-4.4275
REGION_T0	-5.2433
MOB_T0	-20.8827
EDAD_T0	-3.0198

Se realiza la sustitución de la fórmula para calcular el credit score:

$$\mathrm{SC} = 450 + \left(\frac{45}{log(2)}\right)(11.79 + ((0.11630 * -3.15610) + (0.15323 * -1.75800) + ... + (0.16547 * -20.88270)))$$

$$SC = 724$$

Una vez que se obtuvo el *credit score* para cada uno de los clientes se consideró conocer el porcentaje de cliente malos y la distribución de los clientes agrupando por deciles la cartera. En la distribución de la cartera fig. 4.6, se observa que la mayor cantidad de clientes malos recaen sobre el decil 0 y el decil 1. También se puede ver que a menor calificación de score de comportamiento mayor es el porcentaje de malos. Es un score que va de los 425 puntos a 737 puntos.

SCORE	BUENOS	MALOS	Total	% MALOS	%BUENOS ACUM	% MALOS ACUM
425-576	16,966	18,339	35,305	51.9	5.9	27.8
577-607	22,614	12,597	35,211	35.8	13.8	46.9
608-632	26,683	9,575	36,258	26.4	23.1	61.5
633-649	28,080	6,973	35,053	19.9	32.9	72.1
650-661	29,741	5,691	35,432	16.0	43.3	80.7
662-674	30,726	4,352	35,078	12.4	54.0	87.3
675-693	30,750	3,225	33,975	9.5	64.7	92.2
694-716	32,898	2,359	35,257	6.7	76.2	95.8
717-737	34,178	1,785	35,963	5.0	88.1	98.5
738-798	34,238	1,007	35,245	2.9	100.0	100.0
Total	286,874	65,903	352,777			

Cuadro 4.6: Distribución de porcentaje de malos.

Si se tuviera que realizar una campaña de incremento de línea, considerando el *Behavior Score*. Sería riesgoso brindarles un incremento a clientes que tienen un score por debajo de 540 puntos, ya que tienen un alto porcentaje de clientes malos.

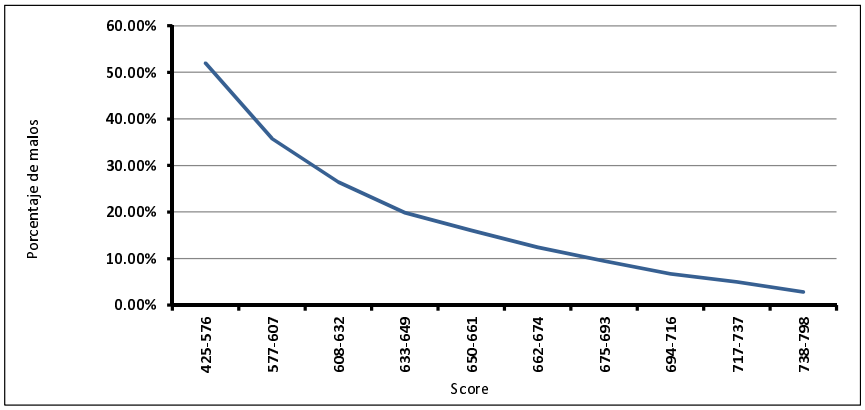


Figura 4.36: Distribución del score sin segmentación.

4.6.4. Medidas de desempeño

En esta sección se mostraran los cálculos del índice de Gini, Kolmogorov Smirnov y Divergencia, para la población no segmentada que se trabajó. Con la finalidad de tener puntos de referencia

para comparar la mejora del modelo realizando una segmentación de cartera.

Índice de Kolmogorov-Smirnov

En el cuadro 4.7 se observa el cálculo del porcentaje de clientes "buenos" y "malos". Para obtener el KS correspondiente a este modelo se consideró el máximo de "Delta", es decir, de la diferencia del acumulado de clientes buenos menos el acumulado de los clientes malos. Y se obtuvo un KS=39.

SCORE	BUENOS	MALOS	% MALOS	%BUENOS ACUM	%MALOS ACUM	DELTA
425-576	16,966	18,339	51.9	5.9	27.8	21.9
577-607	22,614	12,597	35.8	13.8	46.9	33.1
608-632	26,683	9,575	26.4	23.1	61.5	38.4
633-649	28,080	6,973	19.9	32.9	72.1	39.2
650-661	29,741	5,691	16.0	43.3	80.7	37.4
662-674	30,726	4,352	12.4	54.0	87.3	33.3
675-693	30,750	3,225	9.5	64.7	92.2	27.5
694-716	32,898	2,359	6.7	76.2	95.8	19.6
717-737	34,178	1,785	5.0	88.1	98.5	10.4
738-798	34,238	1,007	2.9	100.0	100.0	0.0
Total	286,874	65,903	352,777			**KS=39**

Cuadro 4.7: Índice de Kolmogorov-Smirnov.

En la figura 4.37 se puede observar que el KS=39 se alcanza para clientes que se encuentran en el intervalo de 633-649 puntos de score. También se puede observar la representación gráfica entre las distribuciones acumuladas de clientes buenos y malos en la fig. 4.37.

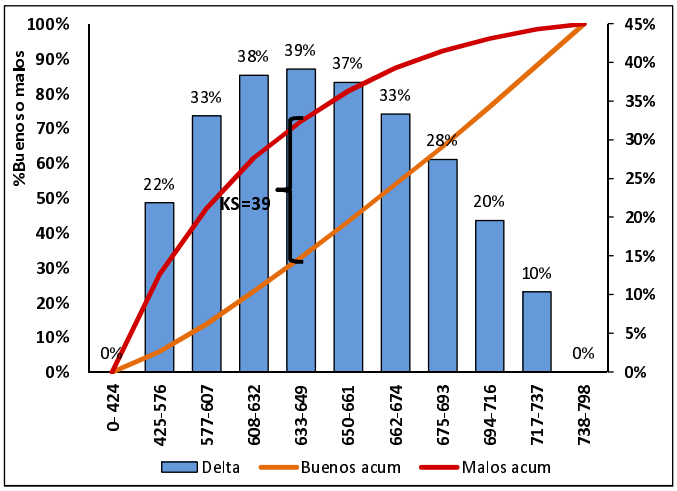


Figura 4.37: Índice KS=39.

Índice de Gini y Divergencia

De acuerdo a las distribuciones de clientes "buenos" y "malos" que mostraron en la tabla anterior, se realizó el cálculo para el índice de Gini y Divergencia, en la tabla 4.8 se puede observar que en términos de porcentaje se obtuvo un Gini de 50.5 y una Divergencia de 9.4:

SCORE	%BUENOS ACUM	%MALOS ACUM	$(x_{k+1}-x_k)(y_{k+1}+y_k)$	Clase	$\bar{X}buenos$	$\bar{X}malos$	σ^2_{goods}	σ^2_{malos}
425-576	5.9	27.8		501	8,491,483.0	9,178,669.5	479,542,453.7	185,913,510.6
577-607	13.8	46.9	5.9	592	13,387,488.0	7,457,424.0	132,764,226.8	1,062,890.5
608-632	23.1	61.5	10.1	620	16,543,460.0	5,936,500.0	63,080,498.6	3,389,350.2
633-649	32.9	72.1	13.1	641	17,999,280.0	4,469,693.0	21,423,885.2	11,053,468.0
650-661	43.3	80.7	15.8	655.5	19,495,225.5	3,730,450.5	5,120,783.9	16,788,716.3
662-674	54.0	87.3	18.0	668	20,524,968.0	2,907,136.00	11,876.3	19,428,003.1
675-693	64.7	92.2	19.2	684	21,033,000.0	2,205,900.00	7,272,122.7	22,117,739.4
694-716	76.2	95.8	21.6	705	23,193,090.0	1,663,095.00	43,536,554.1	2,5423,923.8
717-737	88.1	98.5	23.1	727	24,847,406.0	1,297,695.0	11,6479,468.8	28,255,204.9
738-798	100.0	100.0	23.7	768	26,294,784.0	773,376.0	338,136,010.8	28,021,810.9
			Gini=50.5					**D=9.4**

Cuadro 4.8: Índice de Gini y Divergencia.

En la fig. 4.38 se muestra la representación gráfica del índice de Gini donde se identifica el área bajo la línea de igualdad y el área del score representado como A y de igual forma se representa el área restante correspondiente a B.

La divergencia muestra 9.4 puntos, en la fig. 4.39 se pueden observar ambas distribuciones y sus medias.

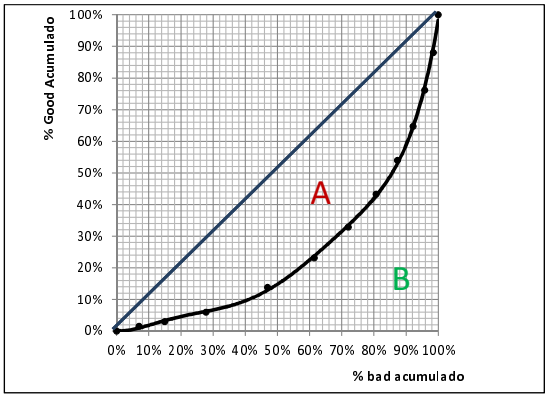


Figura 4.38: Índice de Gini=50.5.

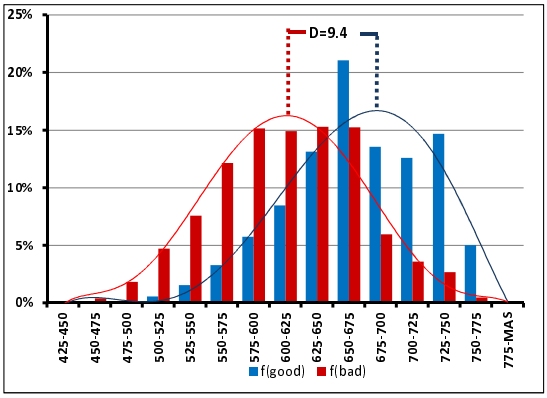


Figura 4.39: Divergencia=9.4.

4.7. *Credit Score* con segmentación

A lo largo de esta sección se describirá el desarrollo de la segmentación de la cartera de clientes que se llevó a cabo, para poder realizar la comparación con el score anterior.

Durante el capítulo 3 de bases estadísticas se habló de la técnica de árboles de decisión para realizar una segmentación de cartera, misma que se utilizará en este segmento.

A partir de la base de datos ya validada, el siguiente paso consistió en un proceso de segmentación por el método de árboles decisión. Para realizar este ejercicio los parámetros del árbol que se consideraron fueron de separación de nodo por Gini (explicado en el capítulo 3), con una profundidad máxima de 3 niveles.

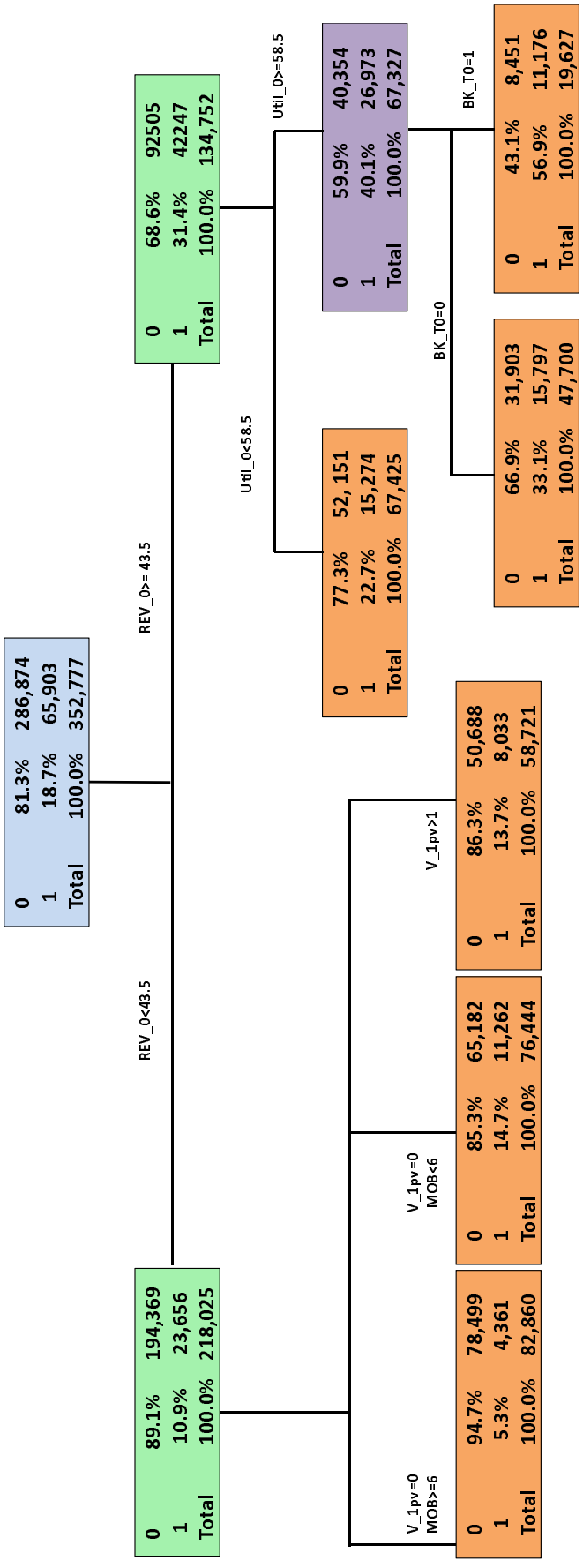


Figura 4.40: Árbol de decisión.

El árbol de decisión propuesto que se obtuvo para efectos de los cálculos de este ejercicio se muestra en la fig. 4.40, el tipo de árbol puede variar dependiendo el objetivo del score y del negocio. En este árbol se pueden observar 3 niveles de profundidad, con 6 ramas.

El árbol considera las variables de revolvencia en el t_0, utilización cero, BK_t0 y número de veces de 1 pago vencido en el mes cero como variables de partición para la segmentación de la cartera. Las variables de partición conforman las reglas para separar la cartera de clientes. Al decantar la base de datos con los filtros del árbol de decisión se obtienen 6 segmentos diferentes.

Para cada una de las 6 diferentes hojas de clientes que se obtuvieron se realizó un análisis univariado, el cálculo de la regresión logística, la construcción del score de comportamiento y sus medidas de desempeño.

Es importante mencionar que se consideraron como puntaje inicial 540 puntos y como *point to double odds* 80 puntos para los 6 diferentes modelos de score.

4.7.1. Hoja 1

En la hoja 1 del árbol se consideraron a los clientes con una revolvencia t_0 <43.5, veces 1 pago vencido=0 y months on books>=6. Considerando por 82,860 clientes, con el 5.3 % de clientes malos y el 94.7 % de clientes buenos.

Análsis univariado

Al igual que se realizó para el score sin segmentación, se revisó que el comportamiento de las variables fuera creciente o decreciente y en la mayoría de los casos se ajustó el comportamiento de las variables para quitar lo picos que tenían. Excluyendo aquellas que tuvieron un comportamiento plano. Las variables que se consideraron se encuentran de la fig.4.41 a la fig.4.64.

A estos clientes se les aplicó una regresión logística, considerando sólo las variables donde su distribución tenía forma creciente o decreciente,

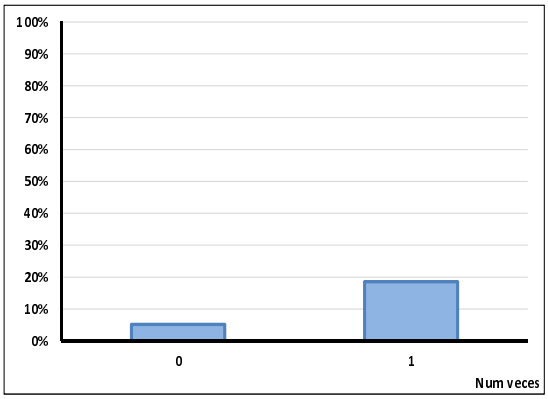


Figura 4.41: Veces 2 pagos vencidos.

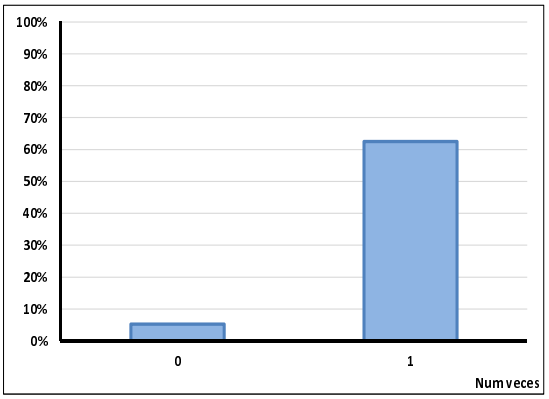


Figura 4.42: Veces 3 pagos vencidos.

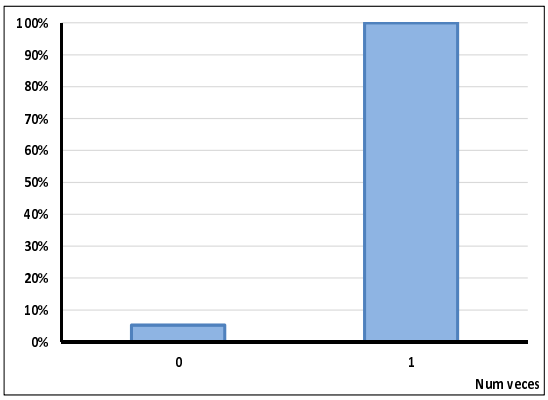


Figura 4.43: Veces 4 pagos vencidos.

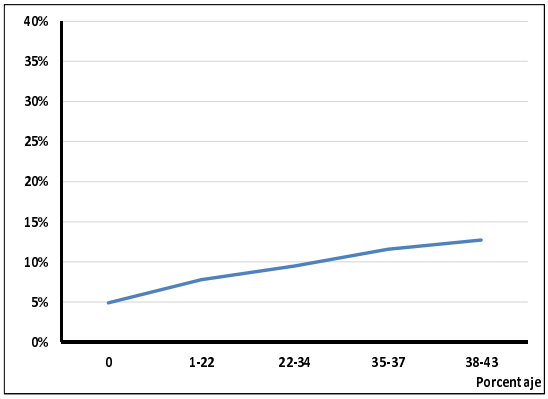


Figura 4.44: Revolvencia mes 0.

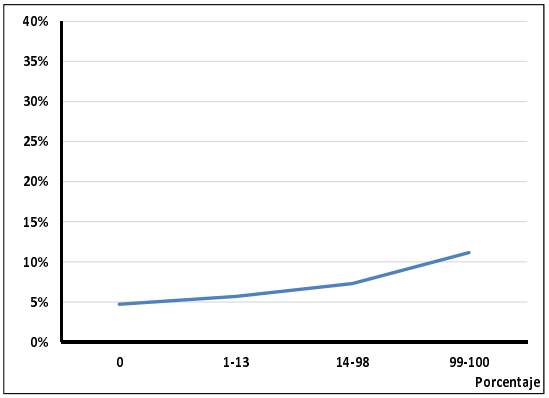


Figura 4.45: Revolvencia 1 mes antes.

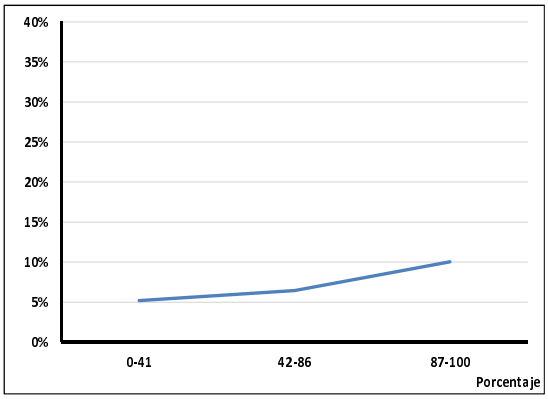


Figura 4.46: Revolvencia 3 mes antes.

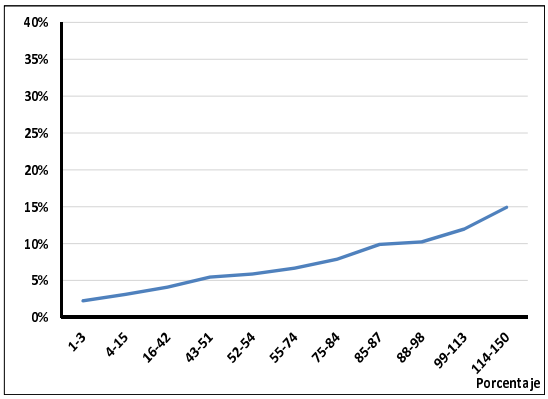


Figura 4.47: Utilización en el mes 0.

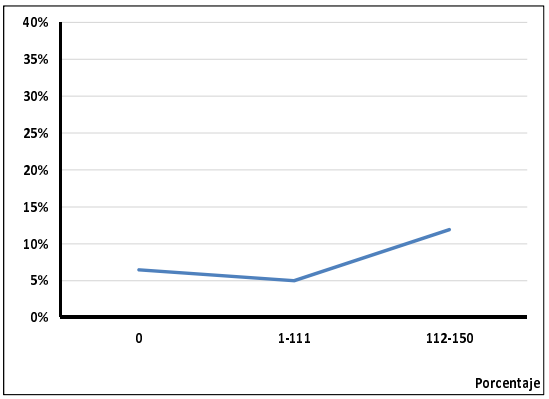


Figura 4.48: Utilización 1 mes anterior.

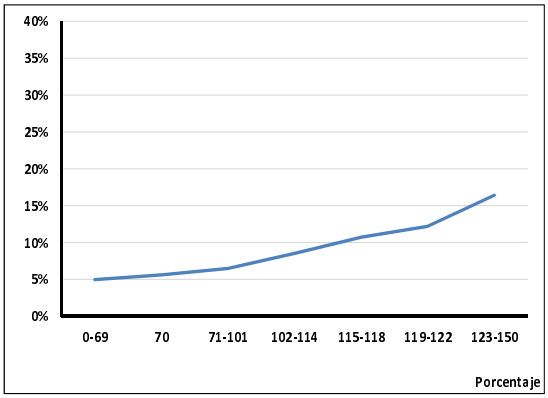


Figura 4.49: Utilización 2 meses anteriores.

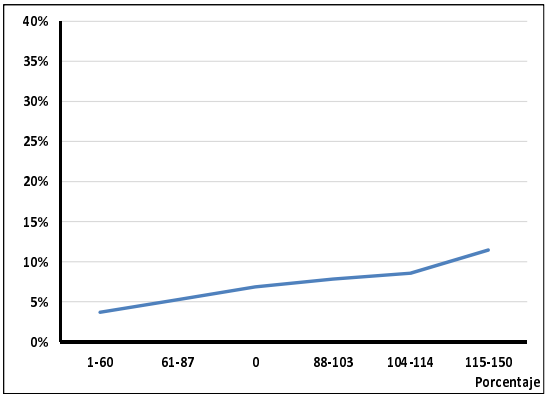


Figura 4.50: Utilización 4 meses anteriores.

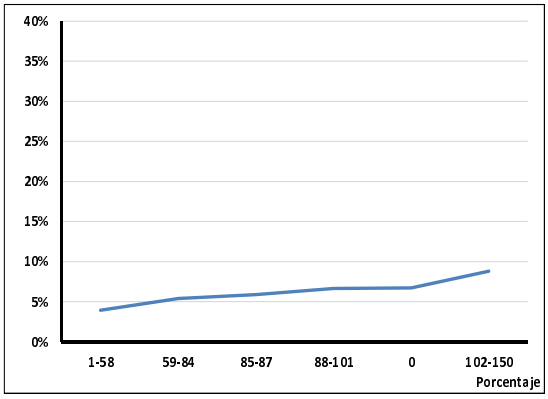


Figura 4.51: Utilización 5 meses anteriores.

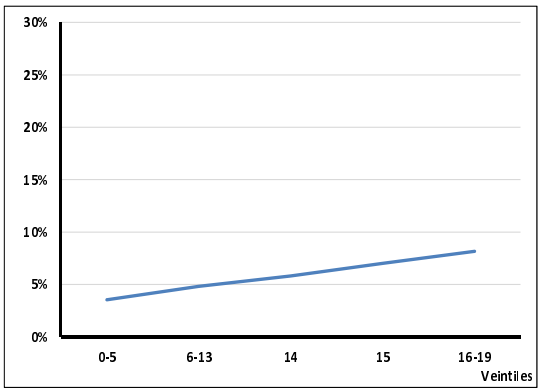


Figura 4.52: Saldo del mes 0.

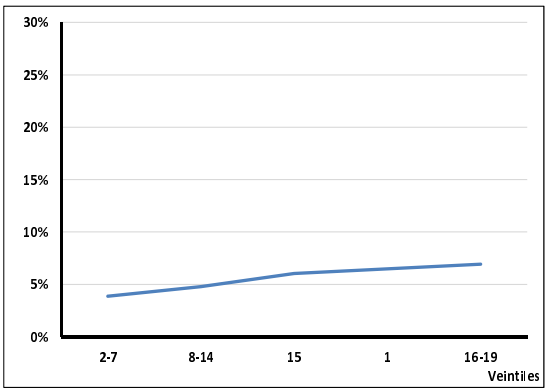


Figura 4.53: Saldo 1 mes anterior.

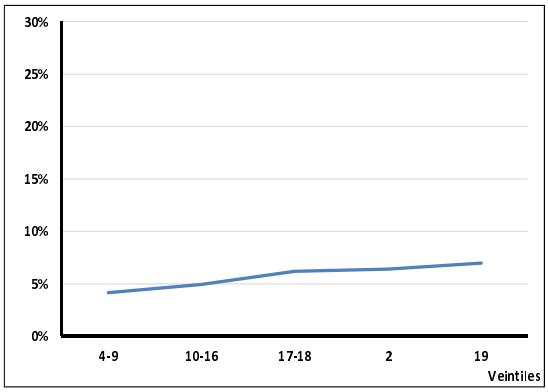


Figura 4.54: Saldo 2 meses anteriores.

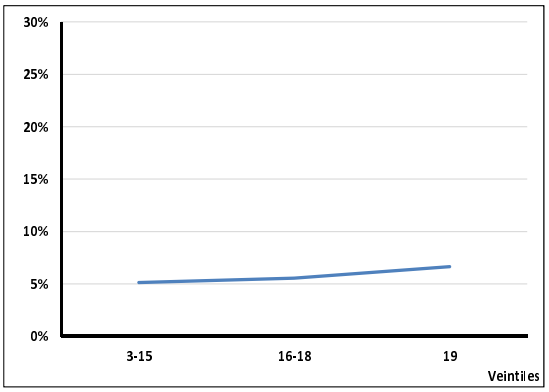


Figura 4.55: Saldo 3 meses anteriores.

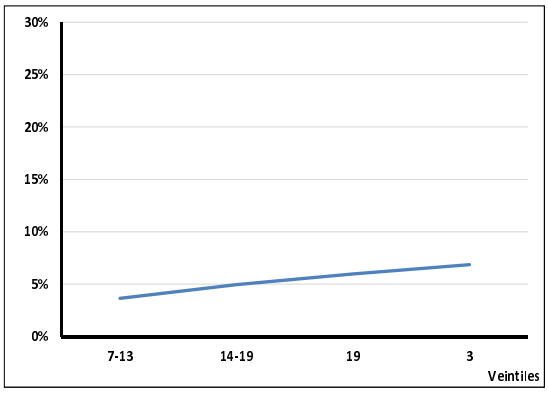


Figura 4.56: Saldo 4 meses anteriores.

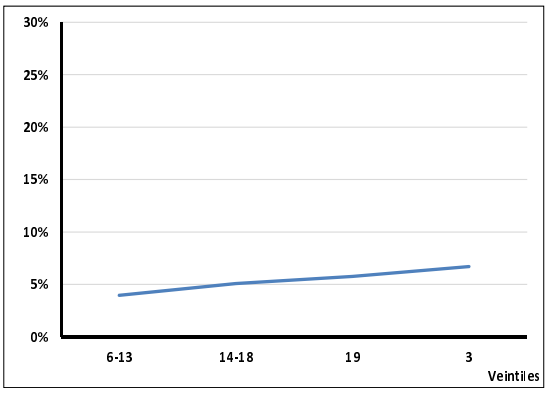


Figura 4.57: Saldo 5 meses anteriores.

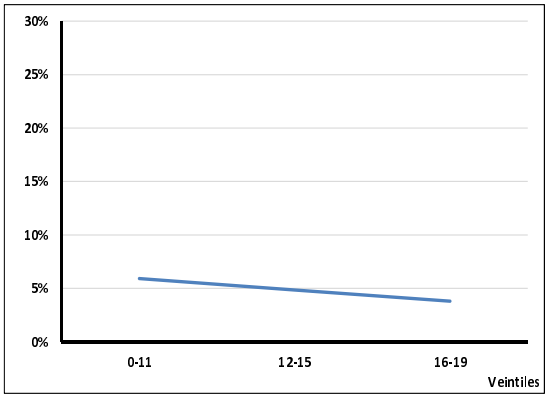


Figura 4.58: Límite de crédito en el mes 0.

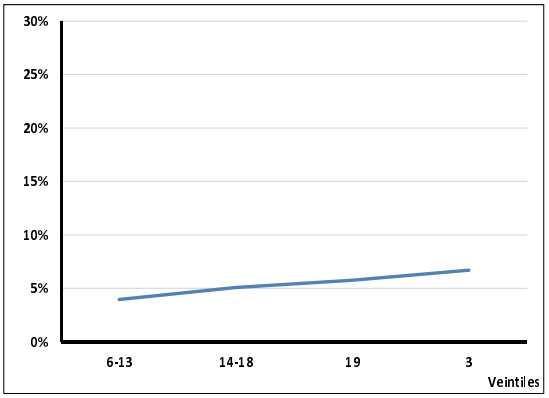


Figura 4.59: Máximo Saldo.

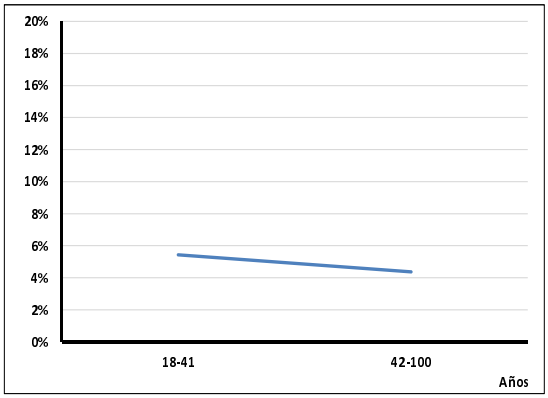


Figura 4.60: Edad.

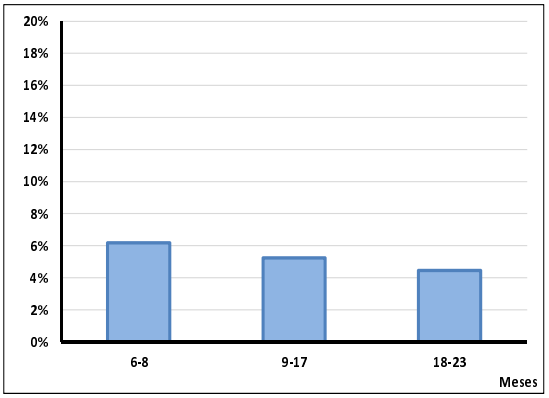


Figura 4.61: Months on Books.

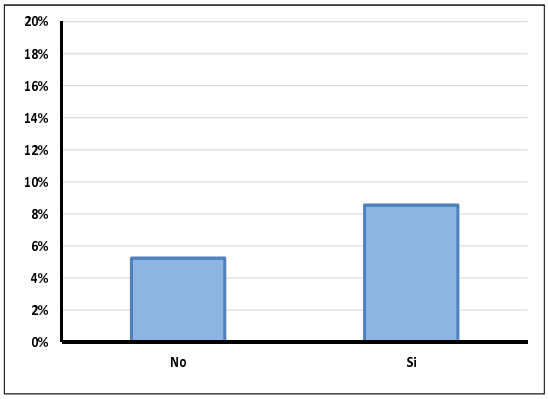


Figura 4.62: Disposición en efectivo.

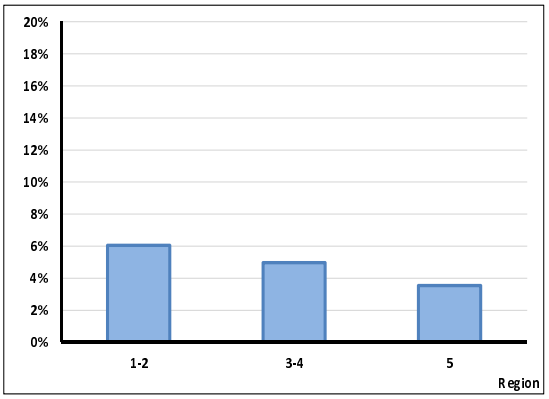


Figura 4.63: Entidad Federativa.

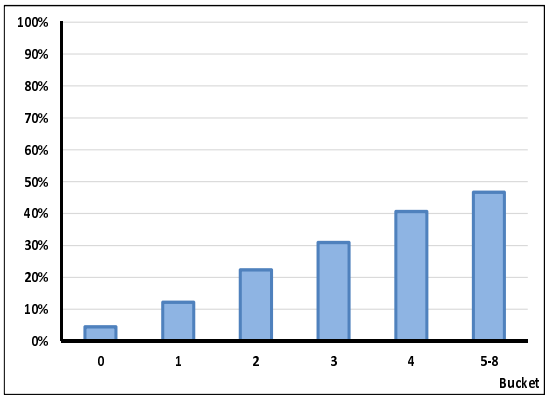


Figura 4.64: Máximo bucket en 12 meses.

Regresión logística

Una vez que se ingresaron al cálculo de la regresión logística las 28 variables que resultaron con un comportamiento creciente o decreciente en el análisis univariado, se obtuvieron los resultados de las variables con significancia estadística para el cálculo de un score de esta hoja 1. Si una variable contaba con un p_value>JiSq menor que 0.05 se consideró de importancia para el modelo, cuadro 4.9.

PARÁMETRO	DF	ESTIMADOR	p_value>JiSq
Intercept	1	10.9741	<.0001
MAX_BK_12M	1	-7.888	<.0001
REV_0	1	-9.743	<.0001
REV_1	1	-8.1103	<.0001
REV_3	1	-9.7861	<.0001
UTIL_0	1	-7.5538	<.0001
FLAG_CASH_T0	1	-16.4636	0.0004
REGION_T0	1	-19.5797	<.0001
MOB_T0	1	-18.4911	<.0001
EDAD_T0	1	-12.5351	0.0061
UTIL_4	1	-6.98	<.0001
SAL_0	1	-10.0599	<.0001
SAL_1	1	-4.0496	0.0211
SAL_2	1	-4.7987	0.0189
SAL_3	1	14.5294	0.0027
SAL_4	1	-5.1549	0.0238
SAL_5	1	-11.8156	<.0001
LC_T0	1	-10.9318	<.0001

Cuadro 4.9: Segmento 1.

Construcción del Score

Con los resultados de la regresión logística y las variables de cada uno de los clientes pertencientes a este segmento. Se realizó el cálculo del *credit score* con la fórmula que se desglosó en el capítulo 2. En la fig.4.65 se puede observar que el rango de 330-761 puntos observa un 15 % de clientes malos, se puede decir que la mayoría de la población de la hoja cuenta con un score bajo.

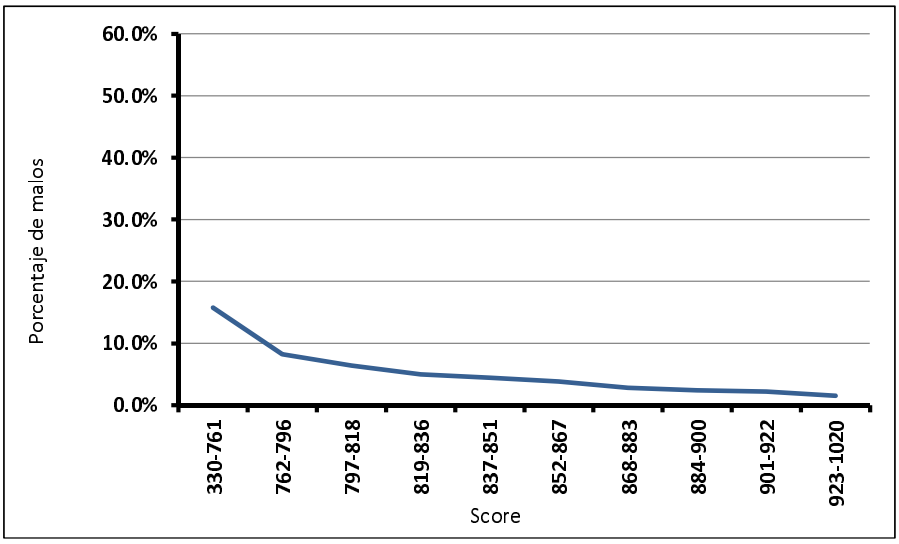


Figura 4.65: Distribución del score de comportamiento.

En el cuadro 4.10 se puede observar la distribución en deciles para este segmento. Se puede ver que para el rango 797-818 puntos de score ya acumula más del 50 % de clientes malos.

SCORE	BUENOS	MALOS	Total	% MALOS	%BUENOS ACUM	% MALOS ACUM
330-761	7,001	1,307	8,308	15.7	8.9	30.0
762-796	7,720	695	8,415	8.3	18.8	45.9
797-818	7,644	523	8,167	6.4	28.5	57.9
819-836	8,068	423	8,491	5.0	38.8	67.6
837-851	7,499	347	7,846	4.4	48.3	75.6
852-867	8,322	332	8,654	3.8	58.9	83.2
868-883	8,069	233	8,302	2.8	69.2	88.5
884-900	7,937	194	8,131	2.4	79.3	93.0
901-922	7,943	178	8,121	2.2	89.4	97.0
923-1,020	8,296	129	8,425	1.5	100	100
Total	78,499	4,361	82,860			

Cuadro 4.10: Distribución de porcentaje de malos.

Medidas de desempeño

Komogorov-Smirnov

En el cuadro 4.11 se muestra su cálculo del Komogorov Sminirnov igual a 29.4:

SCORE	BUENOS	MALOS	% MALOS	%BUENOS ACUM	%MALOS ACUM	DELTA
330-761	7,001	1,307	15.7	8.9	30.0	21.1
762-796	7,720	695	8.3	18.8	45.9	27.2
797-818	7,644	523	6.4	28.5	57.9	29.4
819-836	8,068	423	5.0	38.8	67.6	28.8
837-851	7,499	347	4.4	48.3	75.6	27.2
852-867	8,322	332	3.8	58.9	83.2	24.2
868-883	8,069	233	2.8	69.2	88.5	19.3
884-900	7,937	194	2.4	79.3	93.0	13.6
901-922	7,943	178	2.2	89.4	97.0	7.6
923-1,020	8,296	129	1.5	100	100	0.0
Total	78,499	4,361				**KS=29.4**

Cuadro 4.11: Índice de Kolmogorov-Smirnov.

En la fig. 4.66 se puede ver como el KS más alto se encuentra en intevalo 797-818 puntos de score.

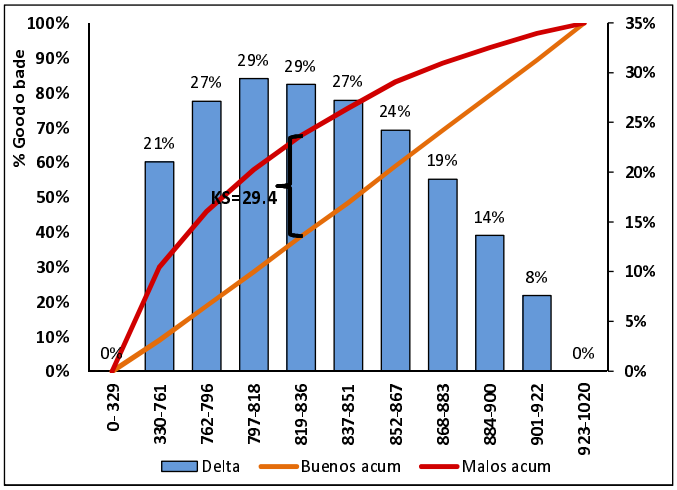


Figura 4.66: Índice KS=29.4.

Gini y Divergencia

En la tabla 4.12 siguiente se muestran las variables necesarias para realizar el cálculo de los indicadores de Gini y Divergencia, así como la media para los clientes buenos y media para los clientes malos.

SCORE	%BUENOS ACUM	%MALOS ACUM	$(x_{k+1}-x_k)(y_{k+1}+y_k)$	Clase	$\bar{X}buenos$	$\bar{X}malos$	σ^2_{goods}	σ^2_{malos}
330-761	8.9	30.0		545.5	3,819,045.5	712,968.5	589,899,258.0	54,513,357.4
762-796	18.8	45.9	7.5	779.0	6,013,880.0	541,405.0	24,884,214.8	595,541.6
797-818	28.5	57.9	10.1	807.5	6,172,530.0	422322.5	6110975.7	1745612.7
819-836	38.8	67.6	12.9	827.5	6,676,270.0	350032.5	552394.6	2558558.6
837-851	48.3	75.6	13.7	844.0	6,329,156.0	292868.0	507373.7	3083911.5
852-867	58.9	83.2	16.8	859.5	7,152,759.0	285354.0	4684448.3	4000619.4
868-883	69.2	88.5	17.6	875.5	7,064,409.5	203991.5	12733812.6	3685777.3
884-900	79.3	93.0	18.3	892.0	7079804.0	173048.0	25091292.2	3926858.3
901-922	89.4	97.0	19.2	911.5	7240044.5	162247.0	45547952.4	4658335.7
923-1,020	100	100	20.8	971.5	8059564.0	125323.5	152824027.9	6344627.2
			Gini=37.0					**D=1.7**

Cuadro 4.12: Índice de Gini y Divergencia.

Como resultado para estos dos indicares se obtuvo un índice de Gini de 37.0, en la fig.4.67 se puede observar la curva de Lorenz y una divergencia de 1.7 en la fig.4.68 se observa las dos distribuciones de clientes buenos y clientes malos.

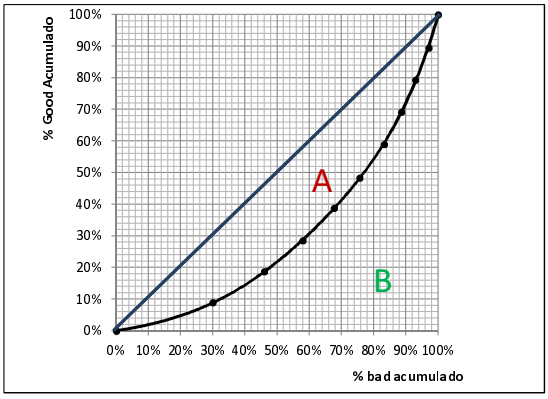


Figura 4.67: Índice de Gini=37.

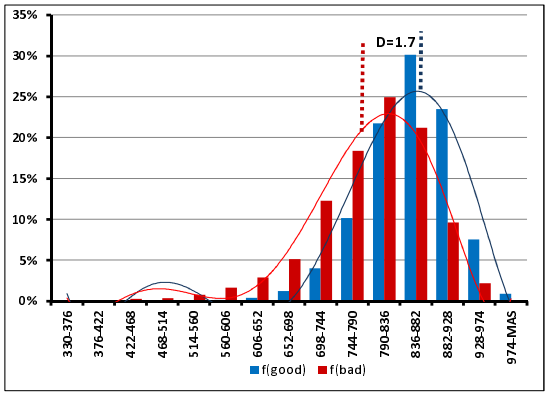


Figura 4.68: Divergencia=1.7.

4.7.2. Hoja 2

La hoja del segmento 2 se compone por los clientes que tienen una revolvencia t_0 <43.5, las veces que tuvo 1 pago vencido = 0 y Mob<6, dando un total de clientes de 76,444.

Análisis univariado

Para este segmento se realizó el análisis univariado, separando las que tienen un comportamiento creciente o decreciente ya con un ajuste en sus intervalos. Obteniendo como resultado las variables de la fig. 4.69 a la fig. 4.94.

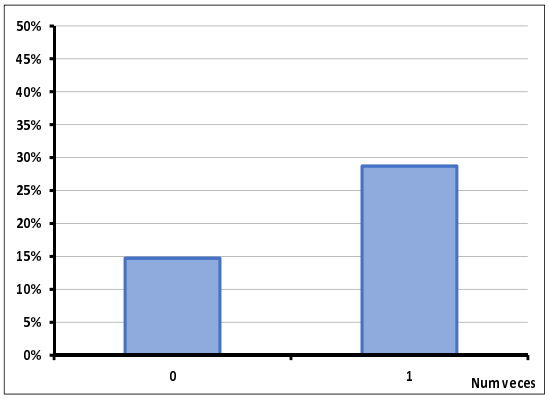


Figura 4.69: Veces 2 pagos vencidos.

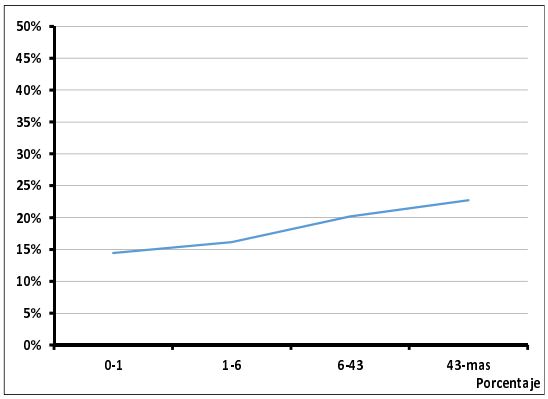


Figura 4.70: Revolvencia mes 0.

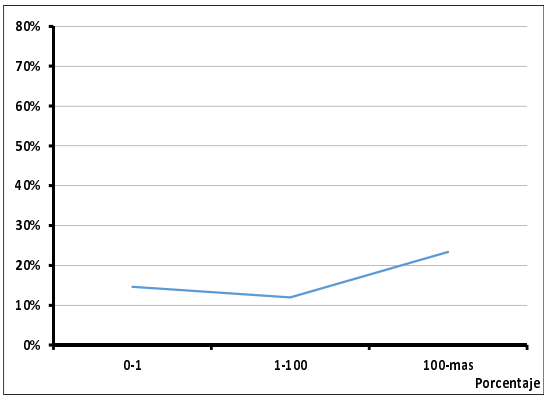


Figura 4.71: Revolvencia 1 mes antes.

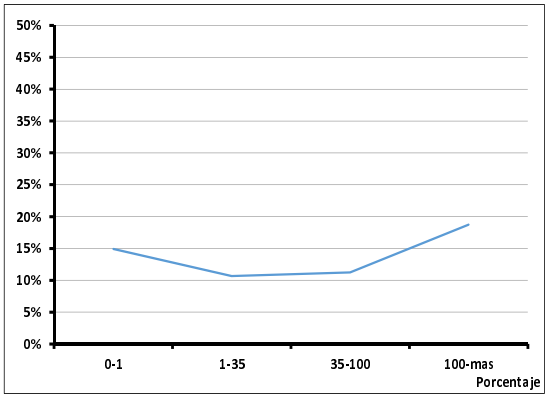


Figura 4.72: Revolvencia 2 meses antes.

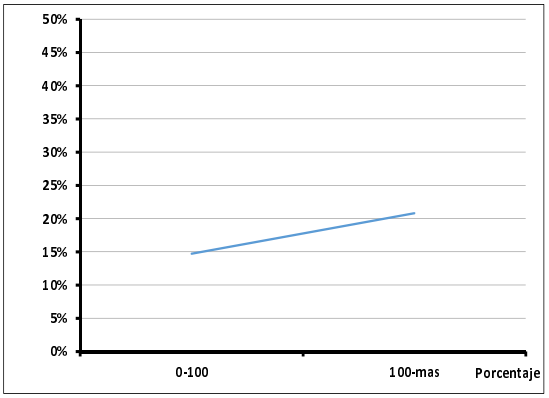


Figura 4.73: Revolvencia 3 meses antes.

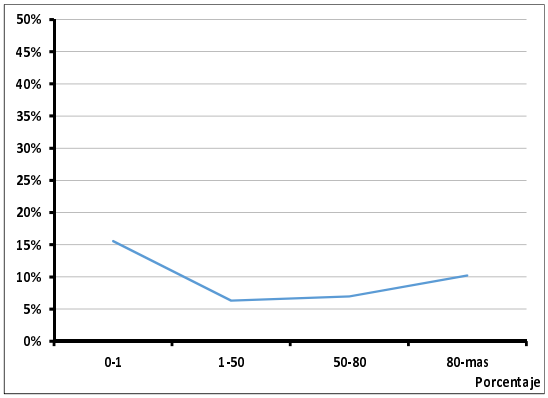


Figura 4.74: Revolvencia 4 meses antes.

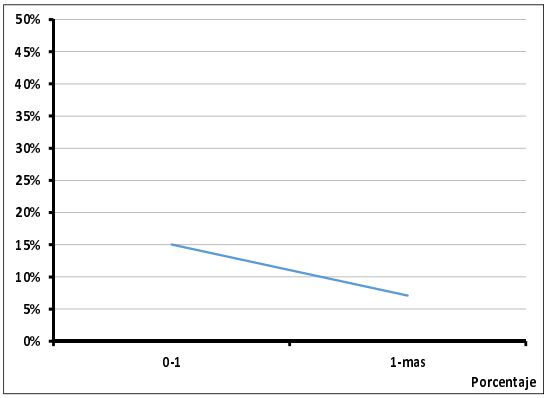


Figura 4.75: Revolvencia 5 meses antes.

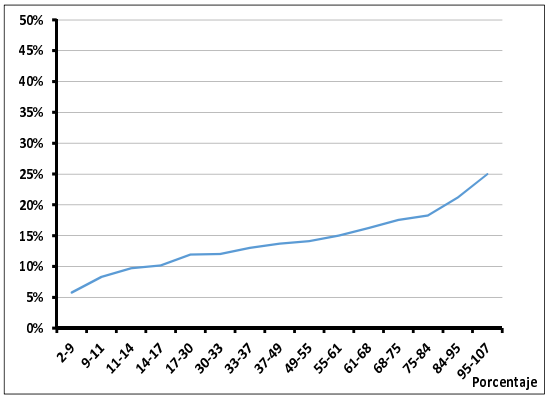


Figura 4.76: Utilización en el mes 0.

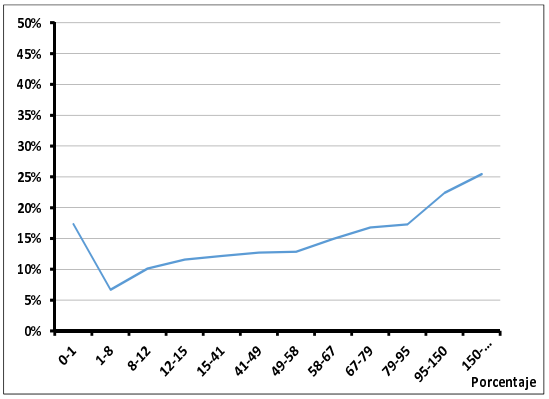


Figura 4.77: Utilización 1 mes anterior.

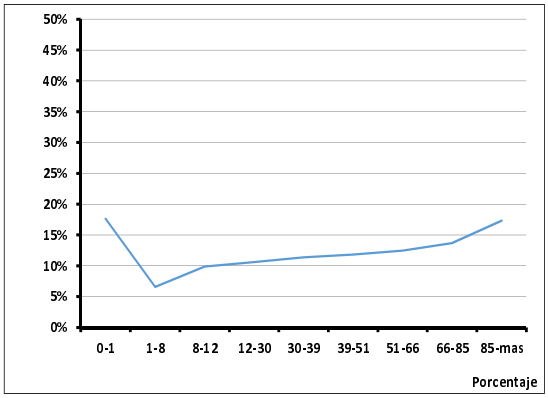


Figura 4.78: Utilización 2 meses anteriores.

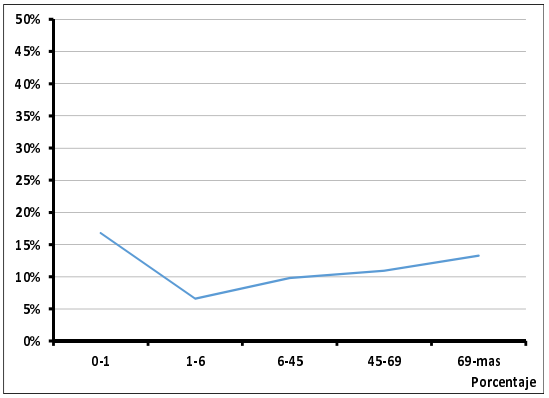


Figura 4.79: Utilización 3 meses anteriores.

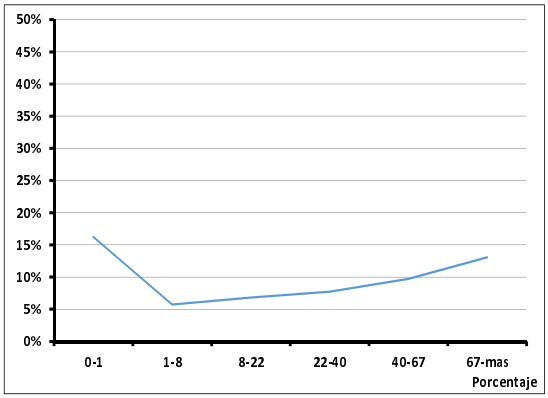


Figura 4.80: Utilización 4 meses anteriores.

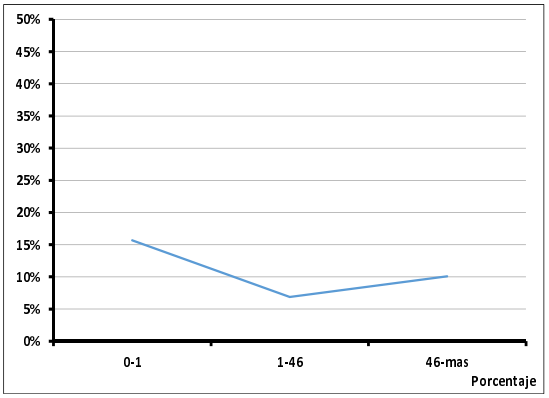


Figura 4.81: Utilización 5 meses anteriores.

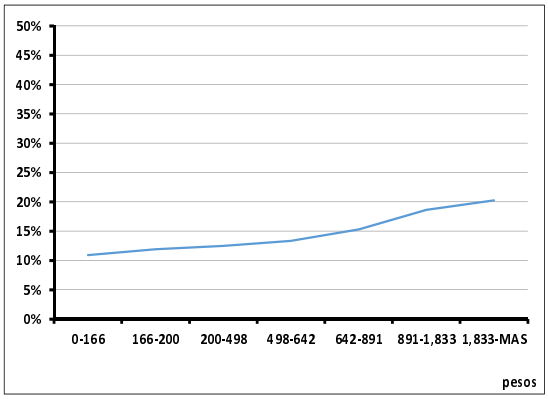


Figura 4.82: Saldo del mes 0.

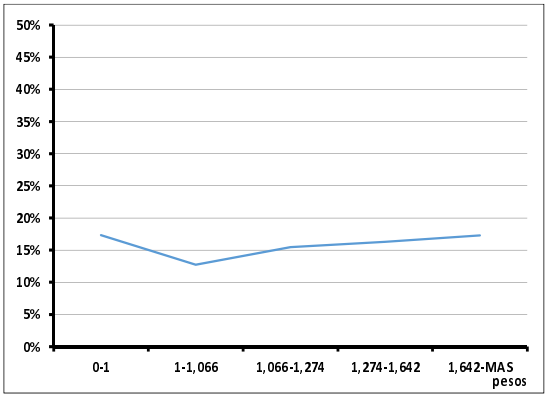


Figura 4.83: Saldo 1 mes anterior.

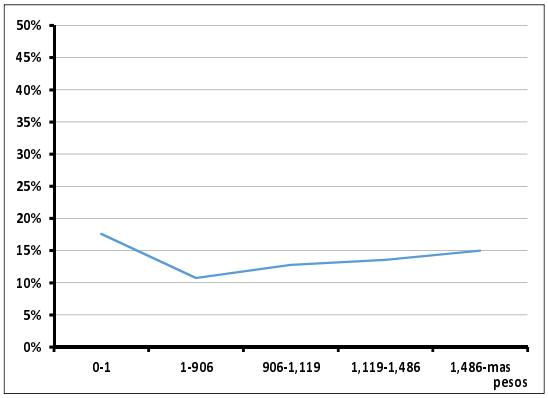


Figura 4.84: Saldo 2 meses anteriores.

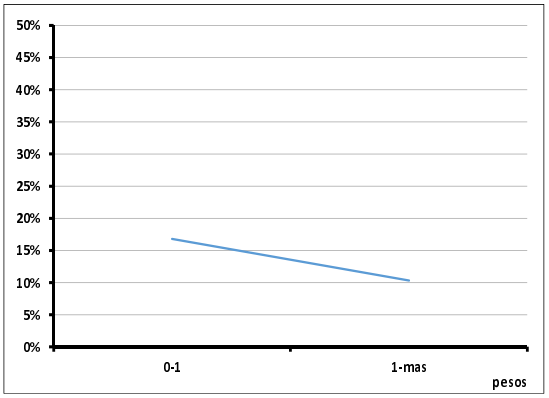


Figura 4.85: Saldo 3 meses anteriores.

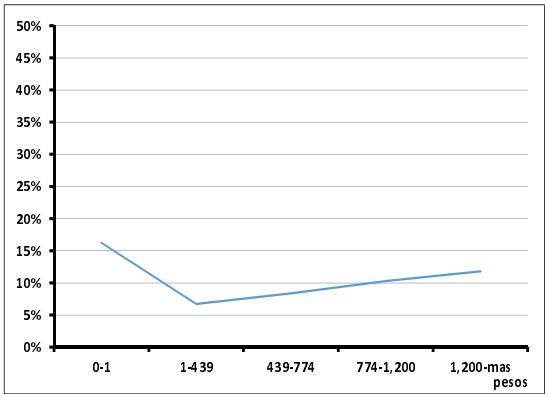


Figura 4.86: Saldo 4 meses anteriores.

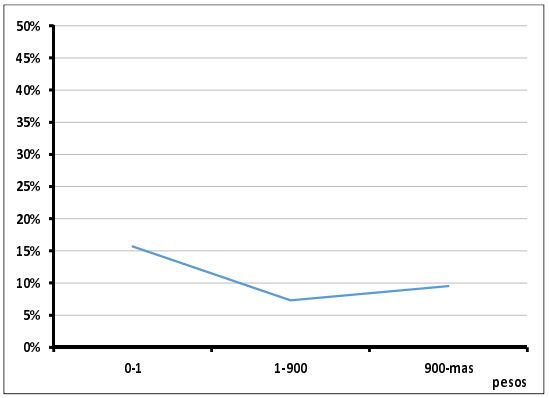


Figura 4.87: Saldo 5 meses anteriores.

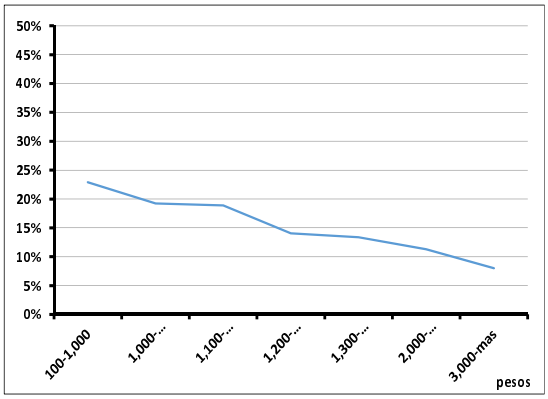


Figura 4.88: Límite de crédito en el mes 0.

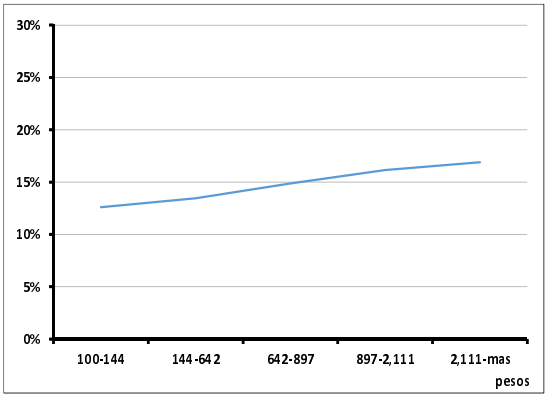


Figura 4.89: Máximo Saldo.

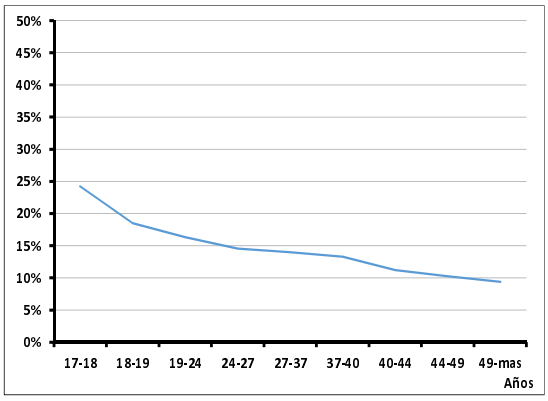


Figura 4.90: Edad.

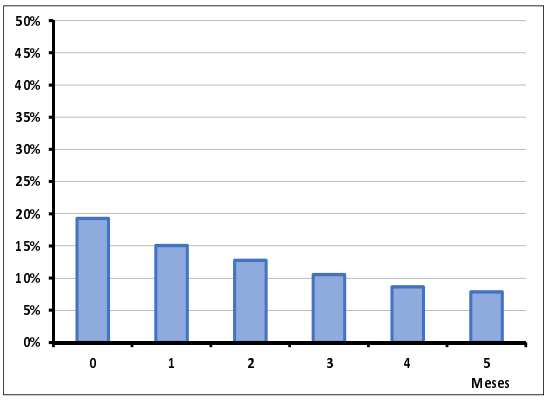


Figura 4.91: Months on books.

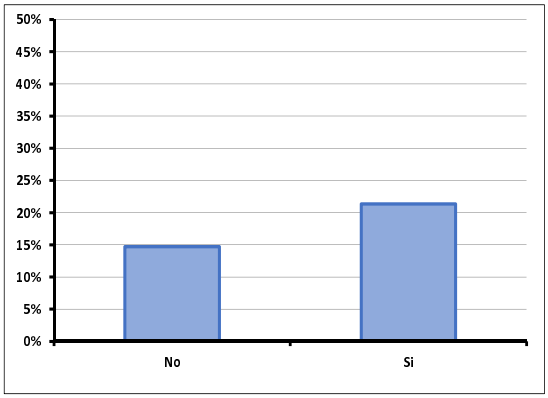


Figura 4.92: Disposición en efectivo.

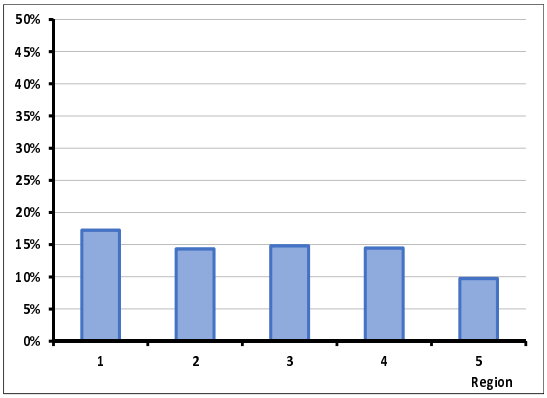


Figura 4.93: Entidad federativa.

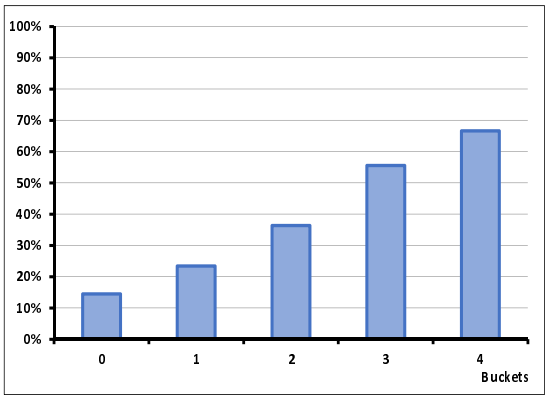


Figura 4.94: Máximo bucket en 12 meses.

Regresión Logística

Con las variables 26 que se consideraron como crecientes o decrecientes se calculó una regresión logística tipo "hacia adelante", obteniendo los resultados que se muestran en la tabla.4.13, donde de acuerdo a lo mencionado se tomaron las variables con con un p_value>JiSq menor que 0.05.

PARÁMETRO	DF	ESTIMADOR	p_value>JiSq
Intercept	1	13.1493	<.0001
MAX_BK_12M	1	-9.0059	<.0001
V_2PV_0	1	4.4251	0.0243
SAL_0	1	-8.4991	<.0001
SAL_1	1	-2.2395	<.0001
REV_0	1	-10.1435	<.0001
REV_1	1	-5.1907	<.0001
REV_3	1	-13.3606	0.0002
REV_5	1	-2.2996	0.0215
UTIL_0	1	-4.2493	<.0001
UTIL_4	1	-2.2004	<.0001
MAX SALDO	1	7.5562	0.0002
FLAG_CASH_T0	1	-9.0662	<.0001
REGION_T0	1	-7.013	<.0001
MOB_T0	1	-7.8203	<.0001
EDAD_T0	1	-4.5243	<.0001
LC_T0	1	-2.6396	<.0001

Cuadro 4.13: Segmento 2.

Construcción del score

Una vez que se obtuvieron las variables representativas para el modelo se procedió a realizar el cálculo del score de comportamiento por cliente para la hoja 2. En el cuadro 4.14 se puede observar la distribución del tipo de clientes para este segmento. Donde en los percentiles 692-707 y 708-724 se encuentra un 17.2 % y 14.7 % de clientes malos de toda la distribución de este segmento.

SCORE	BUENOS	MALOS	Total	% MALOS	%BUENOS ACUM	% MALOS ACUM
321-646	5,250	2,467	7,717	32.0	8.1	21.9
647-674	5,647	1,744	7,391	23.6	16.7	37.4
675-691	6,317	1,521	7,838	19.4	26.4	50.9
692-707	6,212	1,286	7,498	1.2	35.9	62.3
708-724	6,589	1,136	7,725	14.7	46.0	72.4
725-741	6,614	915	7,529	12.2	56.2	80.5
742-761	7,058	779	7,837	9.9	67.0	87.4
762-785	6,969	602	7,571	8.0	77.7	92.8
786-819	7,217	484	7,701	6.3	88.8	97.1
820-981	7,309	328	7,637	4.3	100.0	100.0
Total	65,182	11,262	76,444			

Cuadro 4.14: Distribución de porcentaje de malos.

En la fig. 4.95 se puede ver que en los dos primeros deciles se encuentra un porcentaje de clientes malos entre un 20 % y 30 %.

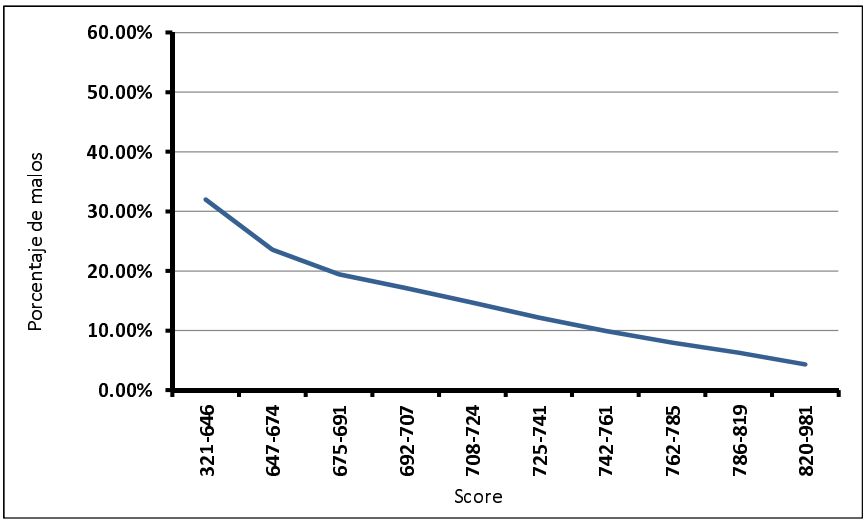


Figura 4.95: Distribución del score de comportamiento.

Medidas de Desempeño

Kolmogorov-Smirnov

El calculo de KS para este segmento fue de 26.4 obtenida la delta del decil 725-741 puntos de score, que se puede observar en el cuadro 4.15.

SCORE	BUENOS	MALOS	% MALOS	%BUENOS ACUM	%MALOS ACUM	DELTA
321-646	5,250	2,467	32.0	8.1	21.9	13.9
647-674	5,647	1,744	23.6	16.7	37.4	20.7
675-691	6,317	1,521	7,838	19.4	26.4	50.9
692-707	6,212	1,286	7,498	17.2	35.9	62.3
708-724	6,589	1,136	14.7	46.0	72.4	24.5
725-741	6,614	915	12.2	56.2	80.5	26.4
742-761	7,058	779	9.9	67.0	87.4	26.4
762-785	6,969	602	8.0	77.7	92.8	24.3
786-819	7,217	484	6.3	88.8	97.1	20.4
820-981	7,309	328	4.3	100.0	100.0	0.0
Total	65,182	11,262				**KS=26.4**

Cuadro 4.15: Índice de Kolmogorov-Smirnov.

En la fig. 4.96 se puede observar que la delta asociada entre la diferencia de los clientes buenos y los clientes malos, de la hoja 2.

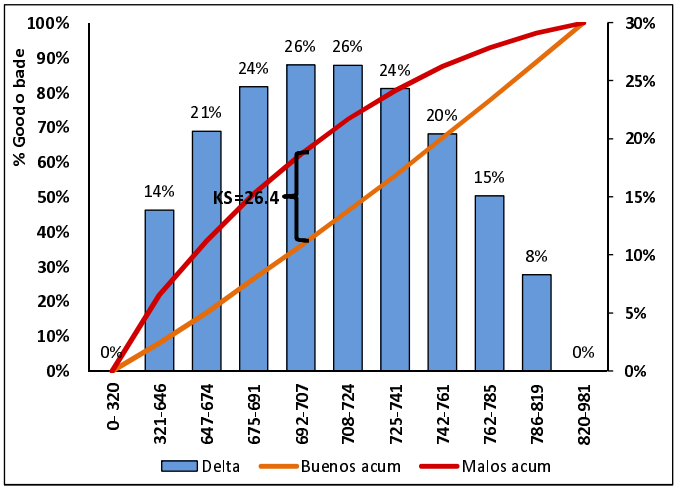


Figura 4.96: Índice KS=26.4.

Gini y Divergencia

Para la hoja 2 se obtuvo un índice de Gini de 34.2 y una divergencia de 4.3, el calcúlo de estas dos medidas de desempeño se puede observar en el cuadro 4.16.

SCORE	%BUENOS ACUM	%MALOS ACUM	$(x_{k+1}-x_k)(y_{k+1}+y_k)$	Clase	$\bar{X}buenos$	$\bar{X}malos$	σ^2_{goods}	σ^2_{malos}
321-646	21.9	13.9	0	483.5	2,538,375.0	1,192,794.5	318,166,703.9	82,288,495.6
647-674	37.4	20.7	5.1	660.5	3,729,843.5	1,151,912.0	27,023,582.8	55,385.7
675-691	26.4	50.9	8.6	683.0	4,314,511.0	1,038,843.0	13,763,194.8	432,594.2
692-707	35.9	62.3	10.8	699.5	4,345,294.0	899,557.0	5,657,016.6	,1431,569.8
708-724	72.4	24.5	13.6	716.0	4,717,724.0	813,376.0	1,232,564.8	2,824,638.1
725-741	80.5	26.4	15.6	733.0	4,848,062.0	670,695.0	73,028.0	4,090,849.0
742-761	87.4	26.4	18.2	751.5	5,304,087.0	585,418.5	3,361,283.9	5,676,661.1
762-785	92.8	24.3	19.3	773.5	5390,521.5	465,647.0	13,383,570.9	6,939,347.3
786-819	97.1	20.4	21.1	802.5	5791,642.5	388,410.0	38,272,975.7	9,000,125.7
820-981	100.0	0.0	22.1	900.5	6581,754.5	295,364.0	213,279,918.4	18,015,977.5
			Gini=34.2					**D=4.3**

Cuadro 4.16: Índice de Gini y divergencia.

Las fig. 4.97 y la fig. 4.98 muestran sus gráficas asociadas. En el índice de Gini se puede observar que por un 50 % de clientes malos se tiene un 25 % de clientes buenos.

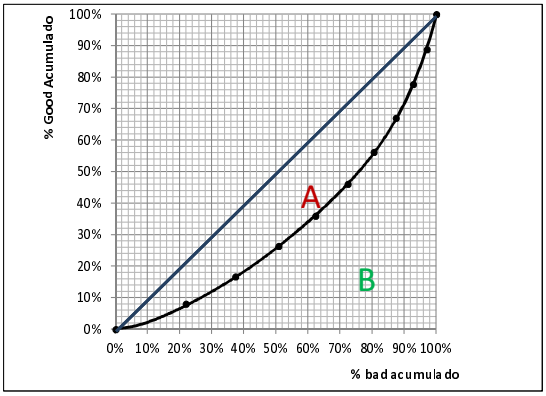


Figura 4.97: Índice de Gini=34.2.

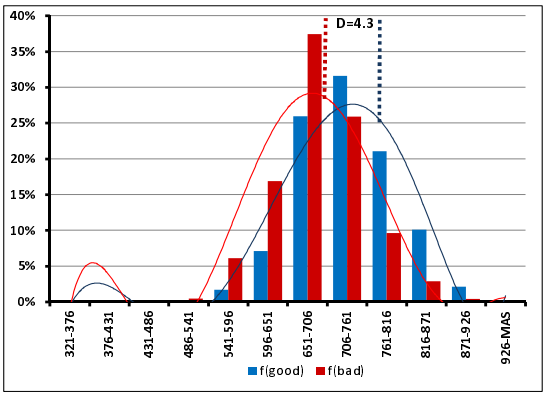


Figura 4.98: Divergencia=4.3.

4.7.3. Hoja 3

En la Hoja 3 se consideraron a los clientes con una revolvencia t_0 <43.5 y con veces 1 pago vencido>1 con un total de 58,721, de los cuales 50,688 clientes cuentan con una variable objetivo=0 y 8,033 clientes con una variable objetivo =1.

Análisis univariado

Se realizó el análisis univariado para la hoja 3, de la fig. 4.99 a la fig.4.128 se pueden observar las variables que se clasificaron con crecientes y decrecientes, necesarias para el siguientes paso del modelo de este segmento.

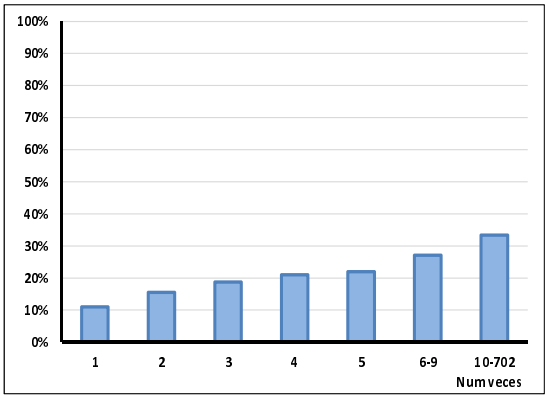


Figura 4.99: Veces con 1 pagos vencidos.

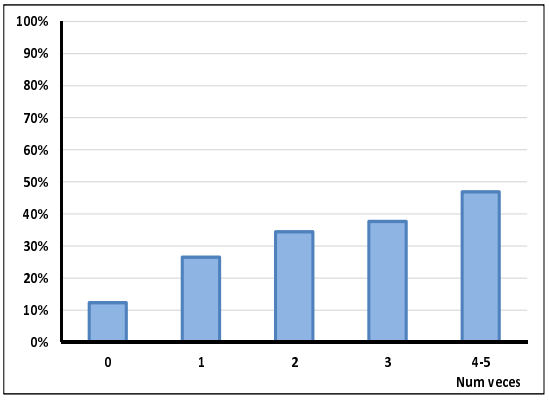


Figura 4.100: Veces con 2 pagos vencidos.

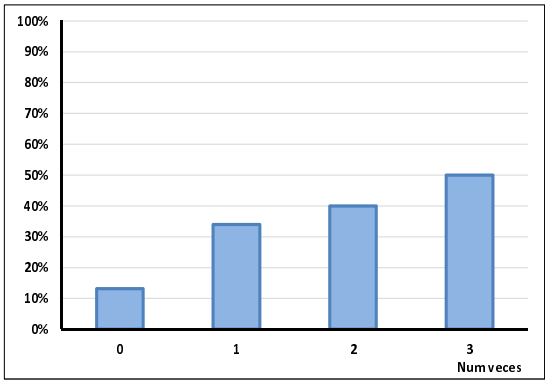


Figura 4.101: Veces con 3 pagos vencidos.

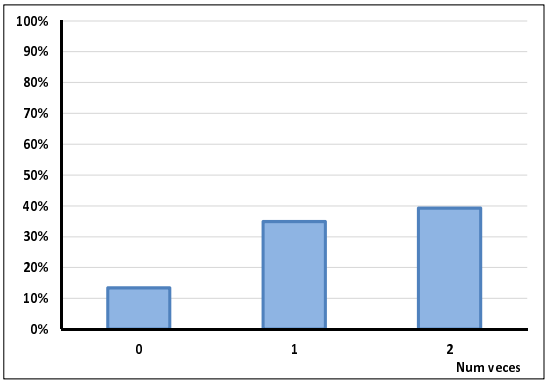


Figura 4.102: Veces con 4 pagos vencidos.

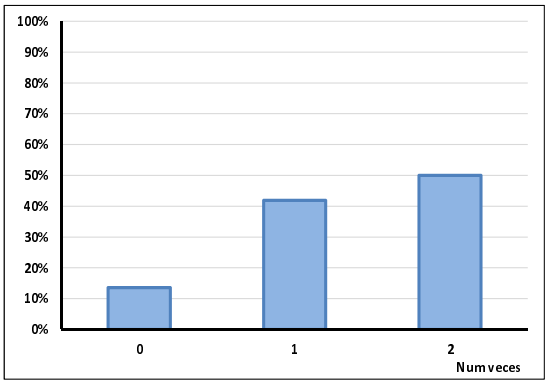


Figura 4.103: Veces con 5 pagos vencidos.

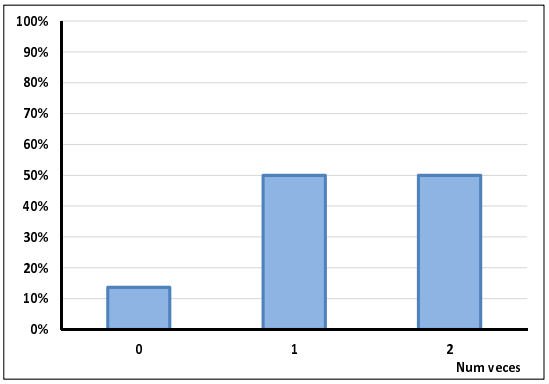


Figura 4.104: Veces con 6 pagos vencidos.

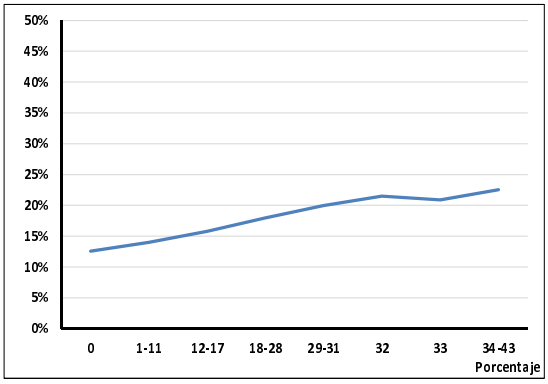


Figura 4.105: Revolvencia mes 0.

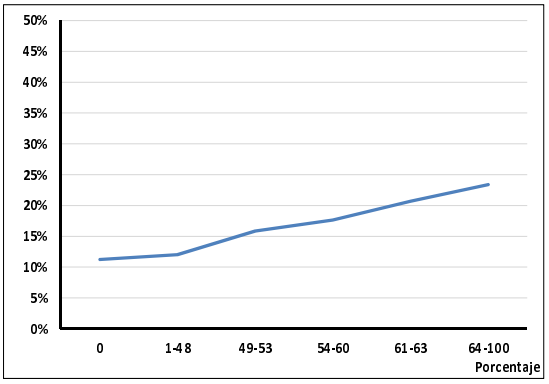


Figura 4.106: Revolvencia 1 mes antes.

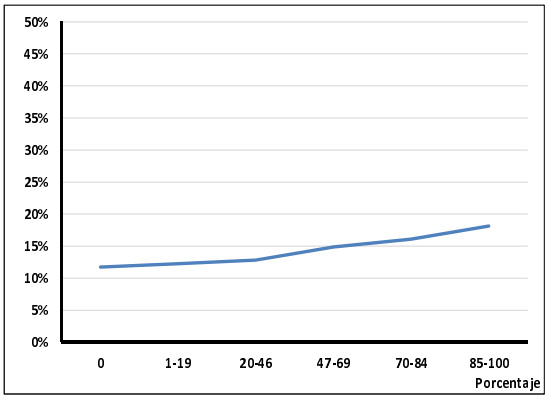


Figura 4.107: Revolvencia 2 meses anteriores.

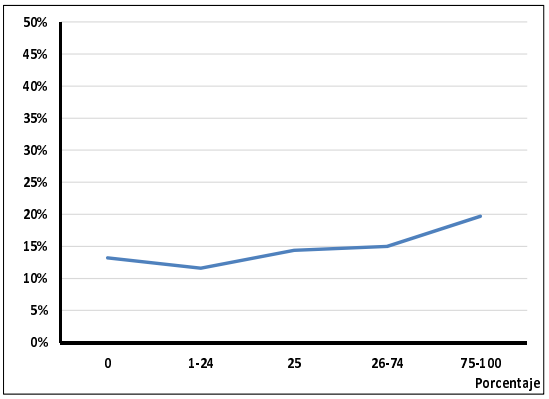


Figura 4.108: Revolvencia 3 meses anteriores.

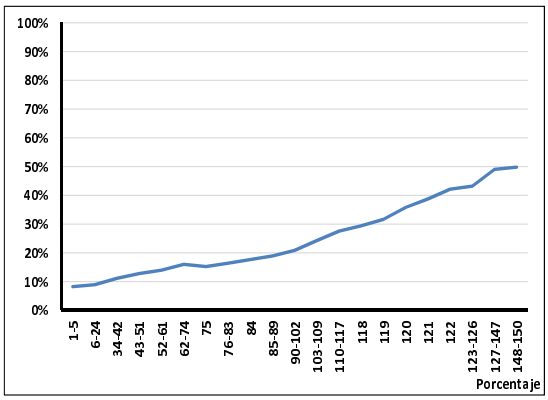


Figura 4.109: Utilización mes 0 de valuación.

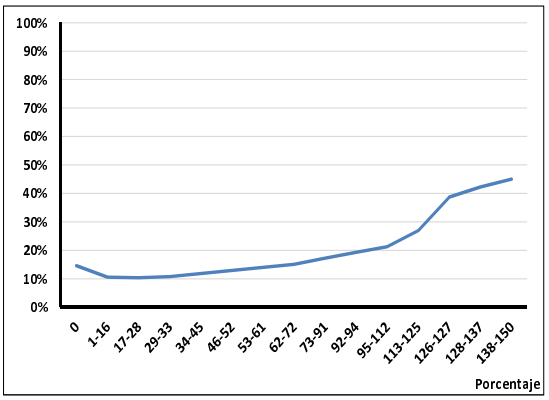


Figura 4.110: Utilización 1 mes antes de la valuación.

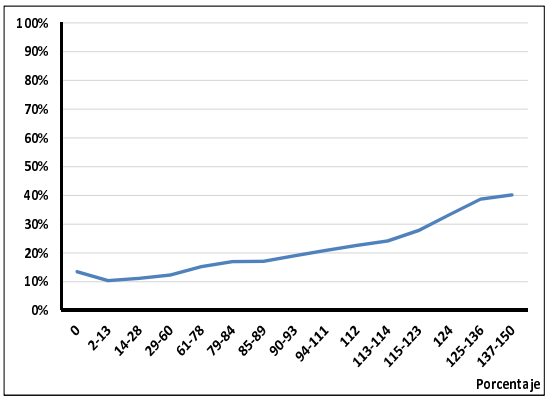


Figura 4.111: Utilización 2 meses anteriores.

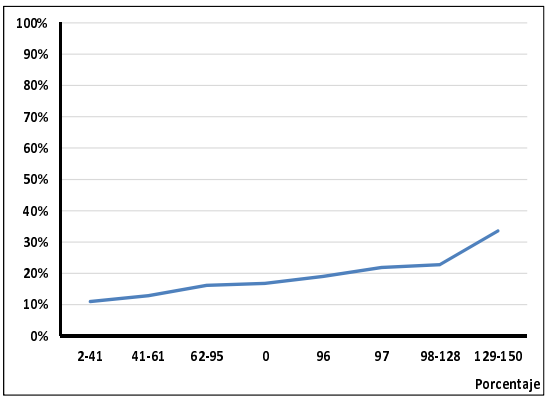


Figura 4.112: Utilización 3 meses anteriores.

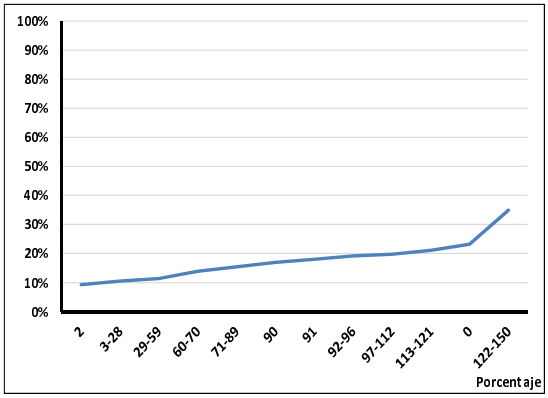


Figura 4.113: Utilización 4 meses anteriores.

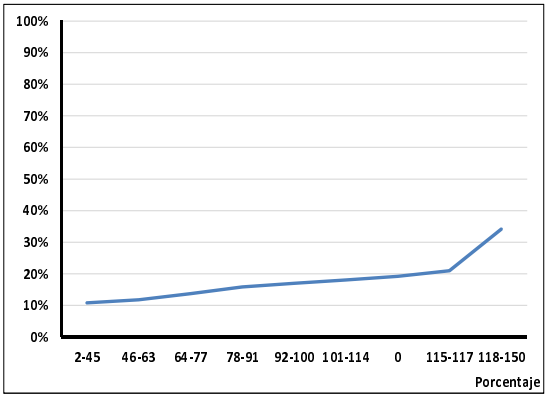


Figura 4.114: Utilización 5 meses anteriores.

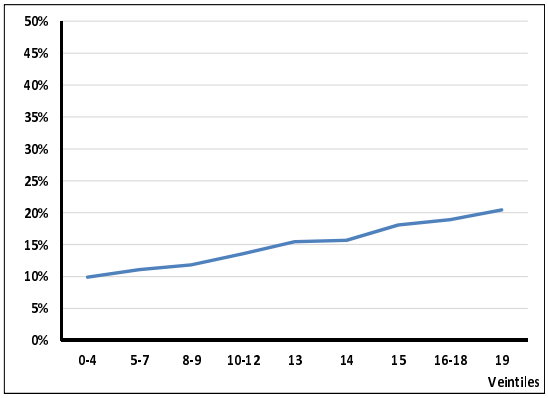


Figura 4.115: Saldo del mes 0.

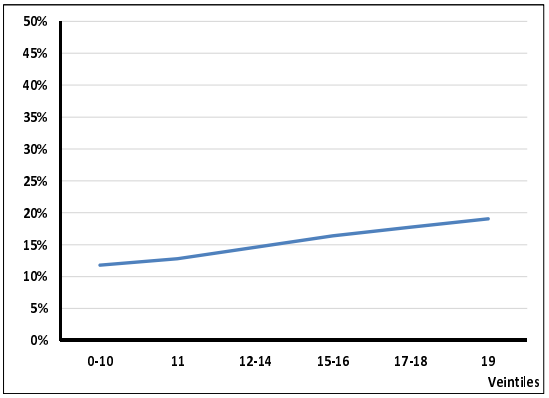


Figura 4.116: Saldo 1 mes anterior.

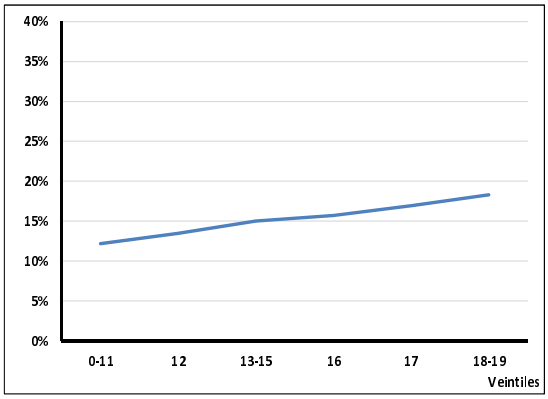


Figura 4.117: Saldo 2 meses anteriores.

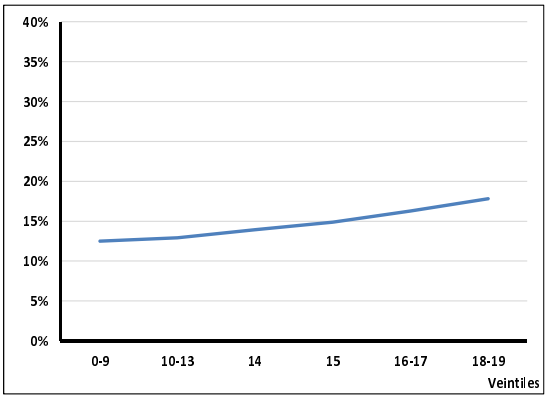


Figura 4.118: Saldo 3 meses anteriores.

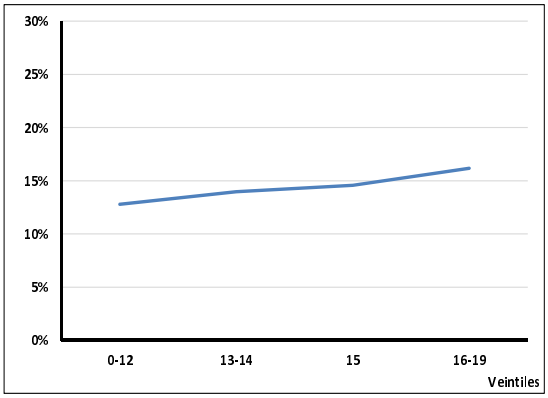


Figura 4.119: Saldo 4 meses anteriores.

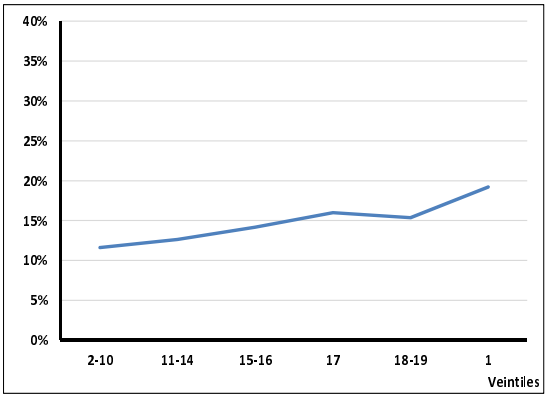


Figura 4.120: Saldo 5 meses anteriores.

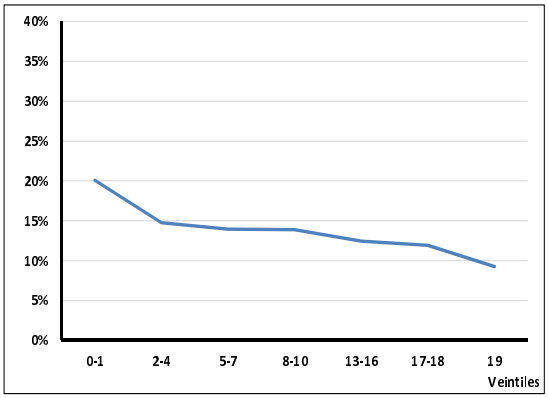


Figura 4.121: Límite de crédito en el mes 0.

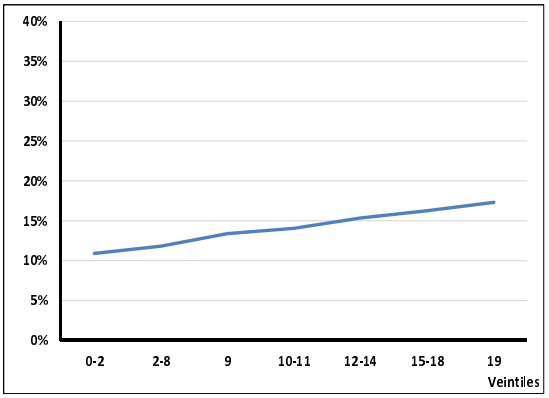


Figura 4.122: Máximo Saldo hasta el mes 0.

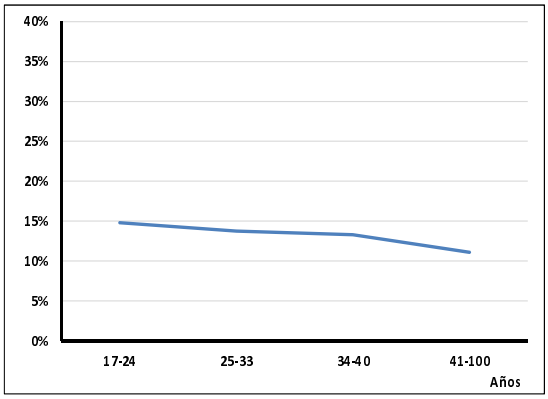


Figura 4.123: Edad en el mes 0.

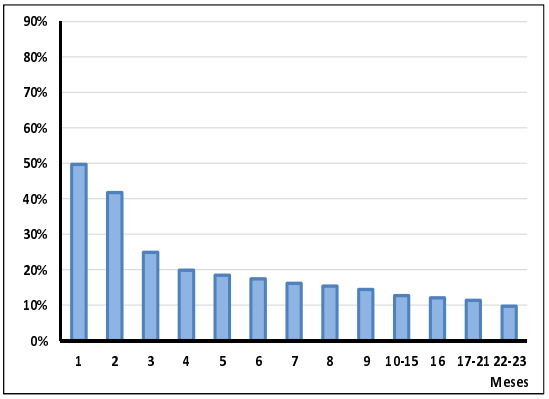


Figura 4.124: Months on Books en el mes 0.

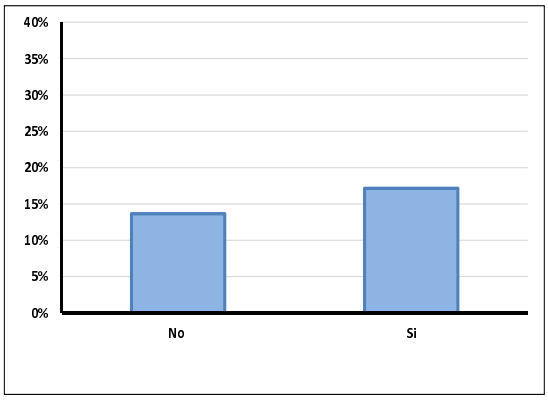


Figura 4.125: Disposición en efectivo.

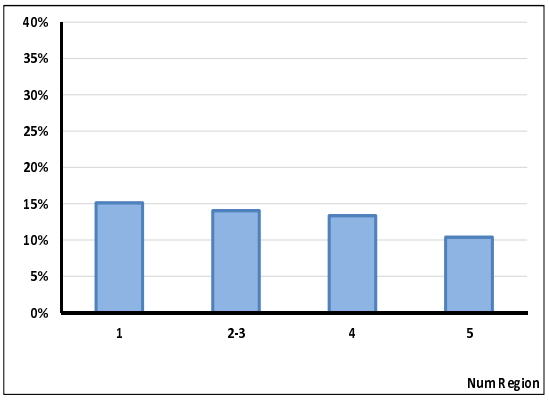


Figura 4.126: Entidad Federativa.

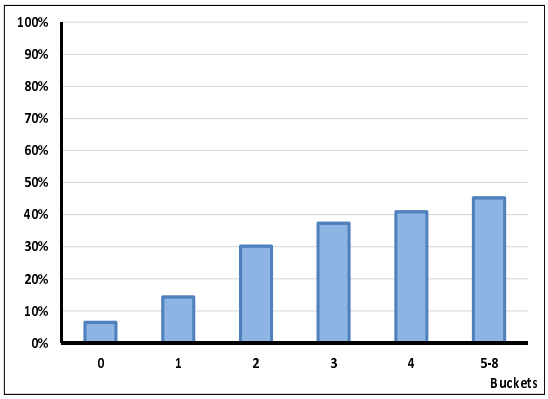


Figura 4.127: Máximo bucket en 12 meses.

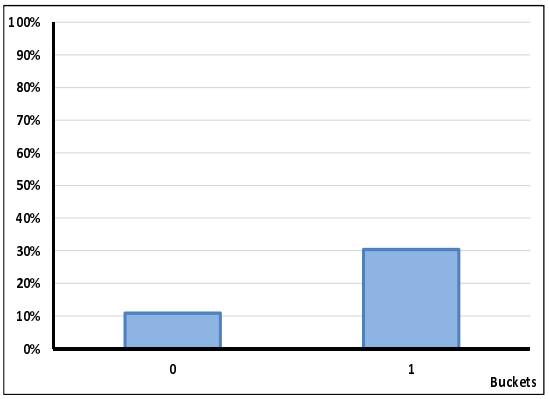


Figura 4.128: Bk en el mes 0.

Regresión logística

Con las 30 variables que se mostraron en el inciso anterior, se calculó la regresión logística "hacia adelante" para la esta hoja. Obteniendo los resultados que se muestran en el cuadro 4.17.

PARÁMETRO	DF	ESTIMADOR	Pr>JiSq
Intercept	1	10.9725	<.0001
MAX_BK_12M	1	-3.281	<.0001
BK_T0	1	-4.1104	<.0001
V_1PV_0	1	-3.1385	<.0001
V_2PV_0	1	-2.6147	<.0001
V_3PV_0	1	-0.8983	0.0219
V_4PV_0	1	1.6736	0.0026
SAL_0	1	-5.8231	<.0001
SAL_4	1	-5.1143	<.0001
REV_0	1	-2.8384	<.0001
REV_1	1	-4.2689	<.0001
REV_2	1	-3.4444	<.0001
REV_3	1	-2.9091	<.0001
UTIL_0	1	-2.4278	<.0001
UTIL_5	1	-1.372	<.0001
MAX SALDO	1	2.5293	0.0224
FLAG_CASH_T0	1	-9.3746	0.015
REGION_T0	1	-6.7211	<.0001
MOB_T0	1	-4.7094	<.0001
EDAD_T0	1	-6.1889	<.0001

Cuadro 4.17: Segmento 3.

Construcción del score

Ya con los resultados de la regresión logística, se procedió a realizar el cálculo de *credit score* por cada uno de los clientes del segmento, en el cuadro 4.18 se puede apreciar la distribución del porcentaje de clientes malos que se obtuvo. Donde se puede ver que en primer decil cuenta con un 40 % de clientes malos.

SCORE	BUENOS	MALOS	Total	% MALOS	%BUENOS ACUM	% MALOS ACUM
295-633	3,529	2,369	5,898	40.2	7.0	29.5
634-683	4,350	1,455	5,805	25.1	15.5	47.6
684-715	4,792	1,029	5,821	17.7	25.0	60.4
716-739	5,087	766	5,853	13.1	35.0	6.9
740-759	5,218	639	5,857	10.9	45.3	77.9
760-777	5,531	511	6,042	8.5	56.2	84.3
778-793	5,319	404	5,723	7.1	66.7	89.3
794-810	5,546	369	5,915	6.2	77.7	93.9
786-819	5,662	286	5,948	4.8	88.8	97.4
811-832	5,654	205	5,859	3.5	100.0	100.0
Total	50,688	8,033	5,8721			

Cuadro 4.18: Distribución de porcentaje de malos.

En la gráfica **??** se puede observar la forma decreciente de la distribución del score, donde conforme incrementa el score, va disminuyendo el porcentaje de clientes malos.

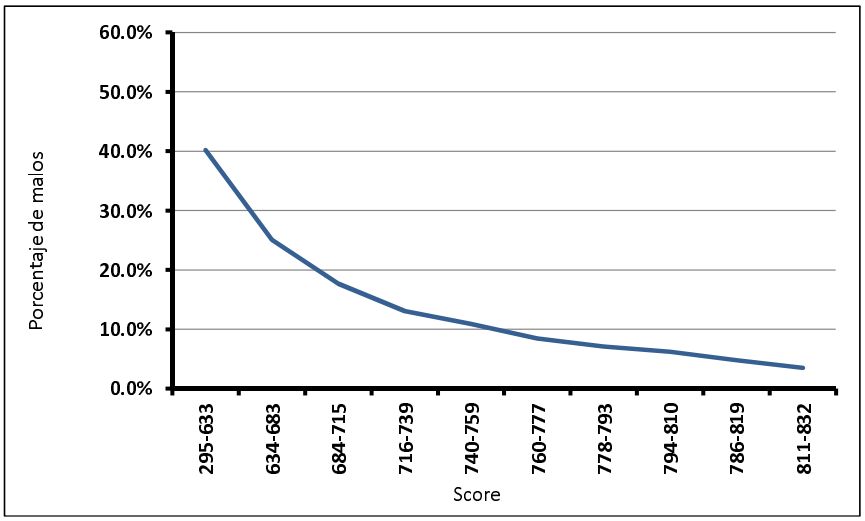


Figura 4.129: Distribución del score de comportamiento.

Medidas de Desempeño

Kolmogorov-Smirnov

Esta hoja cuenta con un KS=35.4 que representa el más alto de entre las 6 hojas de la segmentación. en la tabla 4.19 se puede observar los calculos necesarios para su obtención.

SCORE	BUENOS	MALOS	% MALOS	%BUENOS ACUM	%MALOS ACUM	DELTA
295-633	3,529	2,369	40.2	7.0	29.5	22.5
634-683	4,350	1,455	25.1	15.5	47.6	32.1
684-715	4,792	1,029	17.7	25.0	60.4	35.4
716-739	5,087	766	13.1	35.0	6.9	34.9
740-759	5,218	639	10.9	45.3	77.9	32.6
760-777	5,531	511	8.5	56.2	84.3	28.0
778-793	5,319	404	7.1	66.7	89.3	22.6
794-810	5,546	369	6.2	77.7	93.9	16.2
786-819	5,662	286	4.8	88.8	97.4	8.6
811-832	5,654	205	3.5	100.0	100.0	0.0
Total	50,688	8,033				**KS=35.4**

Cuadro 4.19: Índice de Kolmogorov-Smirnov.

En la gráfica de KS se puede ver el comportamiento de delta generado de la diferencia entre el porcentaje de clientes buenos y el porcentaje de clientes malos. Obteniendo su punto más alto en el decil 684-715.

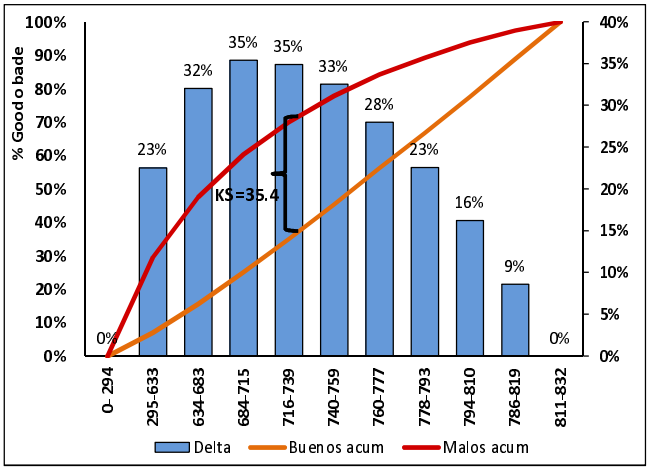


Figura 4.130: Índice KS=35.4.

Indice de Gini y divergencia

Este segmento obtuvo un índice de Gini=44.5, el máximo índice de Gini de entre los 6 segmetos diferentes. Una divergencia=2.0, en la tabla **??** se puede observar el cálculo de estas dos medidas de desempeño.

SCORE	%BUENOS ACUM	%MALOS ACUM	$(x_{k+1}-x_k)(y_{k+1}+y_k)$	Clase	$\bar{X}buenos$	$\bar{X}malos$	σ^2_{goods}	σ^2_{malos}
295-633	7.0	29.5	0	464.0	1,637,456.0	1,099,216.0	269,654,110.6	81,494,579.4
634-683	15.5	47.6	6.6	658.5	2,864,475.0	958,117.5	2,9196,089.6	118,550.5
684-715	25.0	60.4	10.3	699.5	3,352,004.0	719,785.5	8,026,001.8	2,575,229.2
716-739	35.0	6.9	13.1	727.5	3,700,792.5	557,265.0	849,843.4	4,663,513.0
740-759	45.3	77.9	15.2	749.5	3,910,891.0	478,930.5	429,709.0	6,393,389.1
760-777	56.2	84.3	17.7	768.5	4,250,573.5	392,703.5	4,359,491.2	7,239,496.1
778-793	66.7	89.3	18.2	785.5	4,178,074.5	317,342.0	10,806,793.7	7,475,298.0
794-810	77.7	93.9	20.0	802.0	4,447,892	295,938.0	2,1027,388.5	8,584,540.7
786-819	88.8	97.4	21.4	802.5	4,543,755.0	229,515.0	21,817,248.6	6,697,294.7
811-832	100.0	100.0	22.0	821.5	4,644,761.0	168,407.5	37,164,403.1	6,066,590.1
			Gini=44.5					**D=2.0**

Cuadro 4.20: Índice de Gini y divergencia.

En las fig. 4.131 y fig.4.132 se puede apreciar el índice de Gini y Divergencia respectivamente. La divergencia muestra la discrepancia entre las dos distribuciones de clientes buenos y clientes malos.

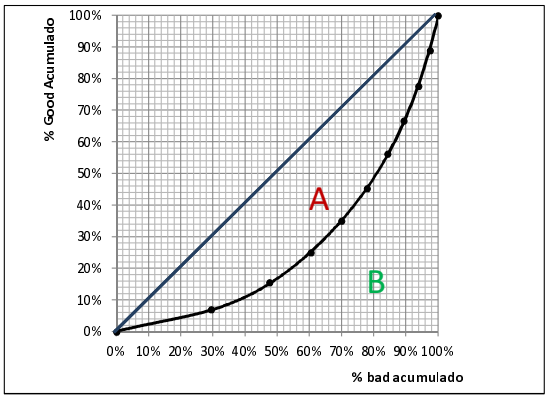


Figura 4.131: Índice de Gini=44.5.

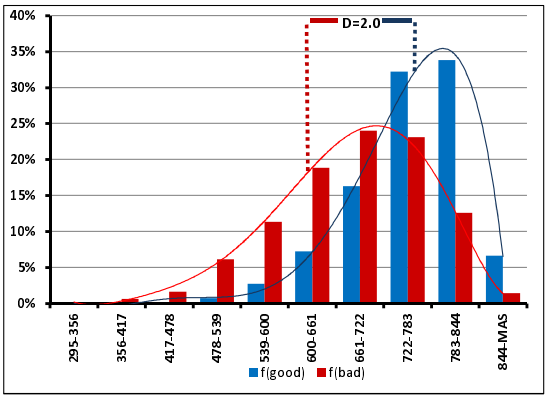


Figura 4.132: Divergencia=2.0.

4.7.4. Hoja 4

La hoja 4 está compuesta por revolvencia t_0 >= 43.5, la utilización en el mes 0< 58.5. Y contiene 67,425 clientes, con 52,151 clientes con variable objetivo=0 y 15,274 con variable objetivo=1.

Análisis univariado

De la fig. 4.133 a la fig. 4.155 se encuentran las variables que se obtuvieron del análisis univariado para el cálculo de la regresión logística.

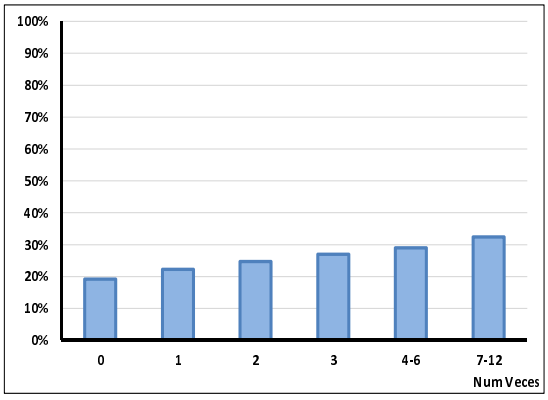


Figura 4.133: Veces con 1 pagos vencidos.

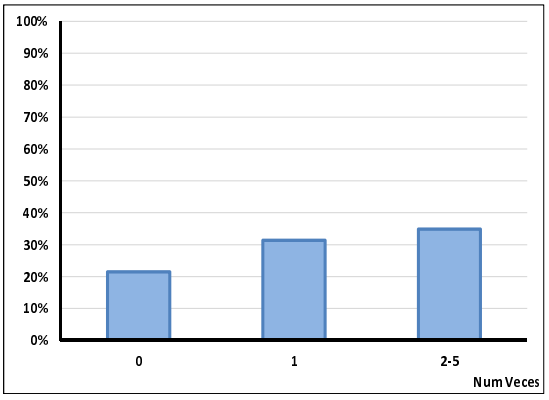


Figura 4.134: Veces con 2 pagos vencidos.

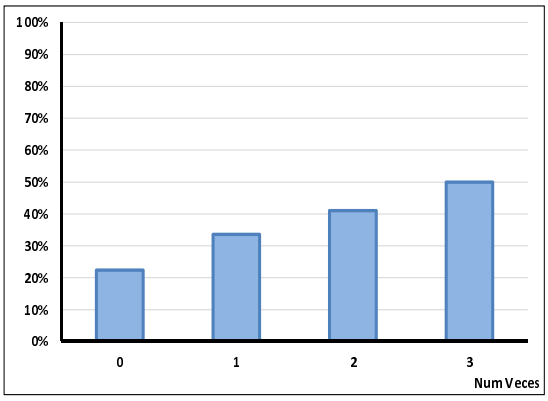


Figura 4.135: Veces con 3 pagos vencidos.

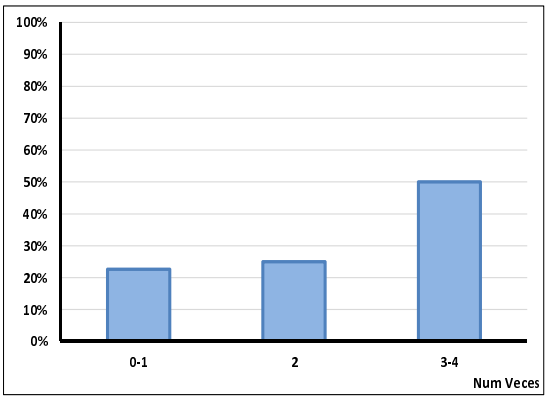


Figura 4.136: Veces con 4 pagos vencidos.

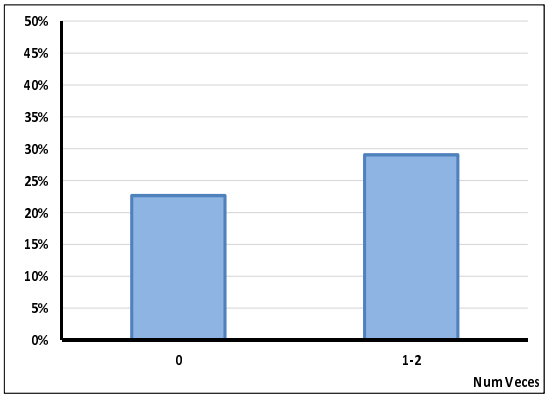


Figura 4.137: Veces con 5 pagos vencidos.

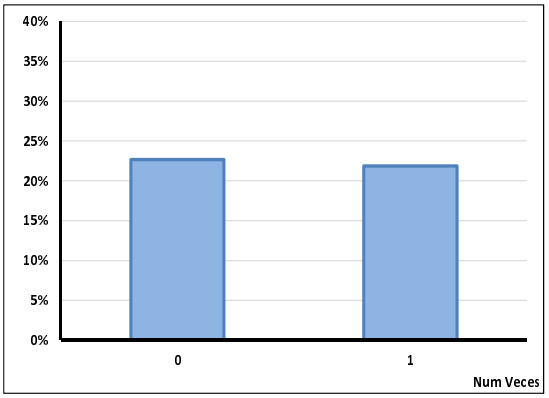


Figura 4.138: Veces con 6 pagos vencidos.

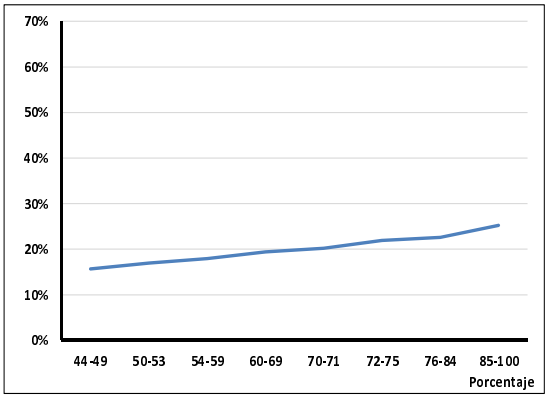


Figura 4.139: Revolvencia mes 0.

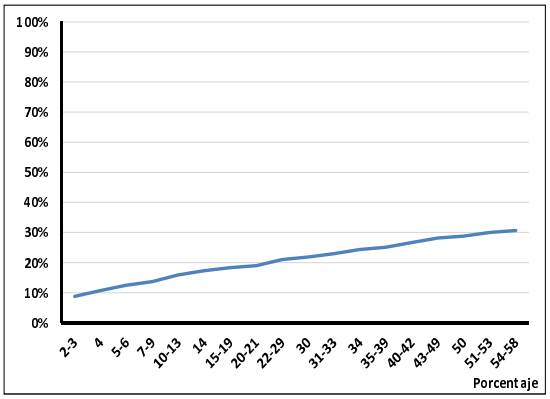


Figura 4.140: Utilización mes 0 de valuación.

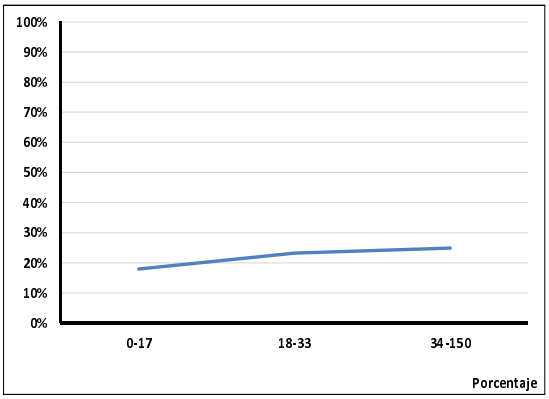


Figura 4.141: Utilización 1 mes antes de la valuación.

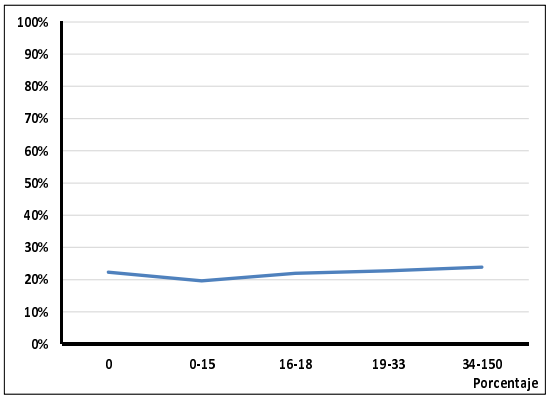


Figura 4.142: Utilización 2 meses anteriores.

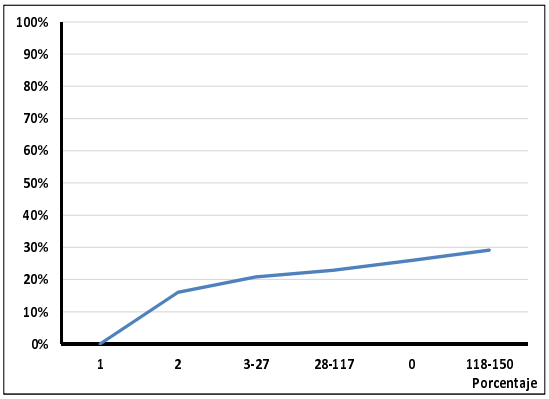


Figura 4.143: Utilización 3 meses anteriores.

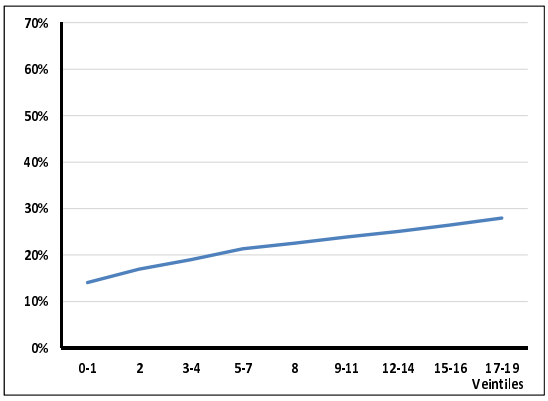


Figura 4.144: Saldo del mes 0.

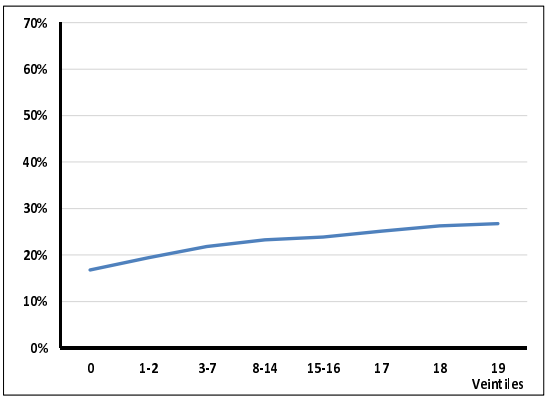


Figura 4.145: Saldo 1 mes anterior.

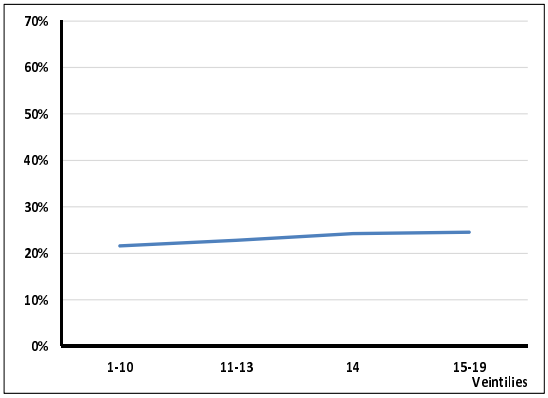


Figura 4.146: Saldo 2 meses anteriores.

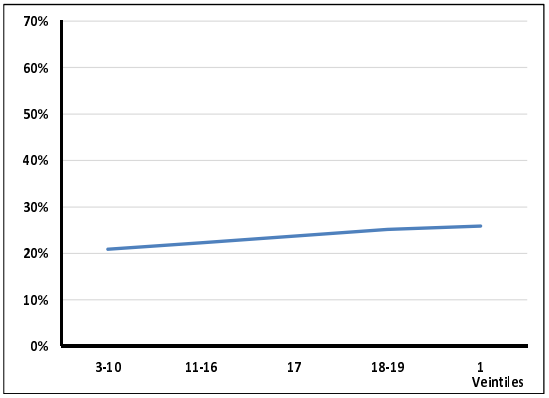


Figura 4.147: Saldo 3 meses anteriores.

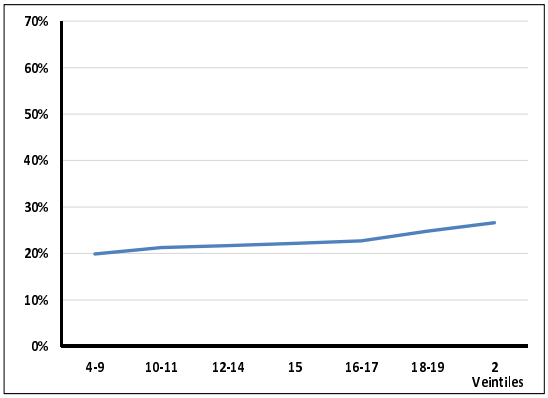


Figura 4.148: Saldo 4 meses anteriores.

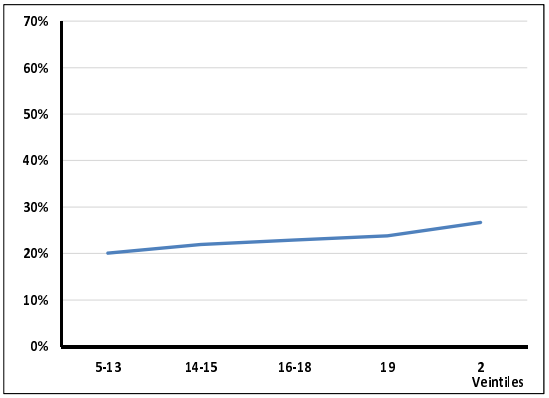


Figura 4.149: Saldo 5 meses anteriores.

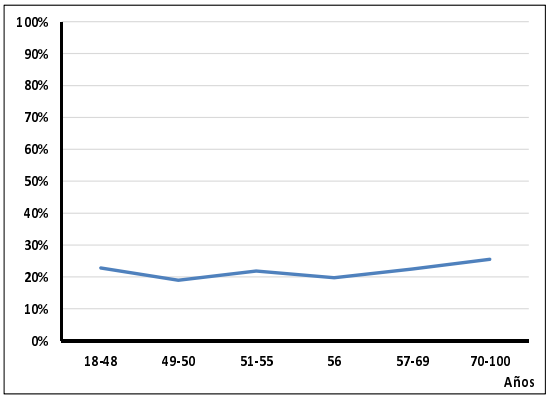


Figura 4.150: Edad en el mes 0.

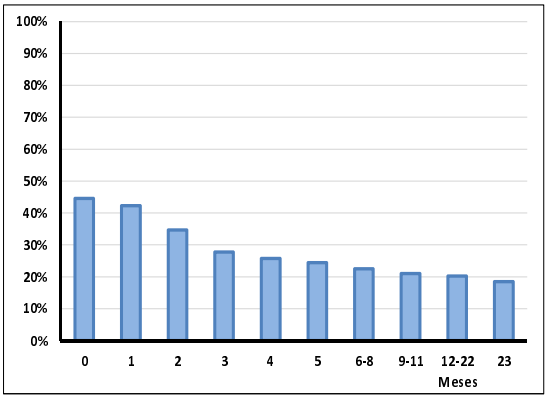


Figura 4.151: Months on Books en el mes 0.

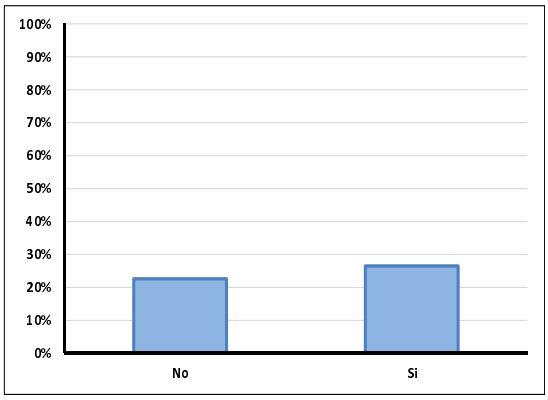


Figura 4.152: Disposición en efectivo.

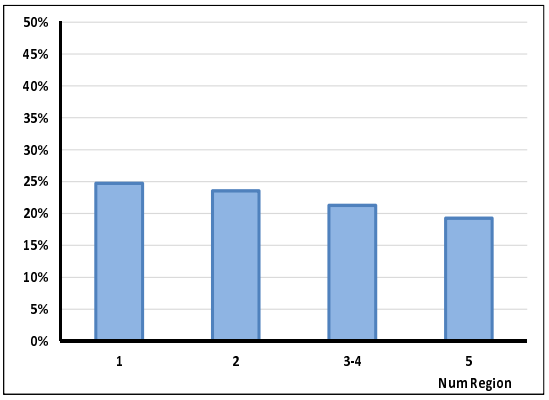


Figura 4.153: Entidad federativa.

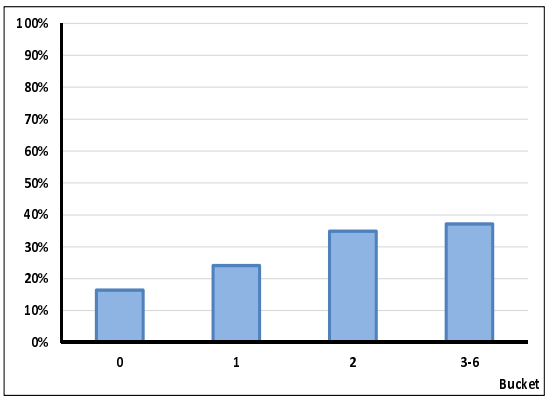


Figura 4.154: Máximo bucket en 12 meses.

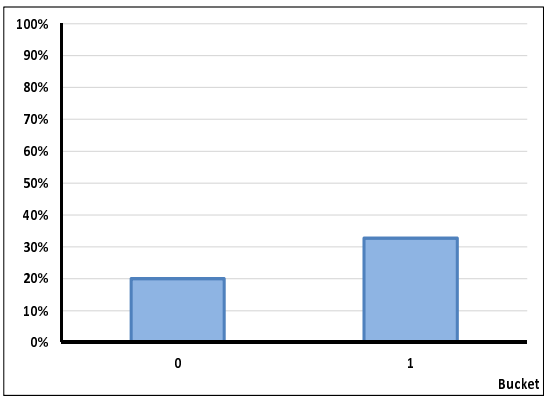


Figura 4.155: Bk en el mes 0.

Regresión logística

Con las 23 variables resultantes del análisis univariado para ésta hoja, se calculó regresión logística obteniendo los resultados del cuadro 4.21.

PARÁMETRO	DF	ESTIMADOR	p_value>JiSq
Intercept	1	11.8258	<.0001
MAX_BK_12M	1	-5.6299	<.0001
BK_T0	1	-3.1716	<.0001
V_1PV_0	1	-2.7166	<.0001
V_2PV_0	1	-0.7593	0.0163
REV_0	1	-3.5508	<.0001
UTIL_0	1	-4.6194	<.0001
UTIL_2	1	3.2405	0.0004
SAL_0	1	-2.3318	<.0001
SAL_1	1	-4.1311	<.0001
SAL_2	1	2.9939	0.0099
SAL_3	1	1.4239	0.0171
SAL_5	1	-2.3571	<.0001
FLAG_CASH_T0	1	-6.4302	0.0113
REGION_T0	1	-5.3763	<.0001
MOB_T0	1	-7.2651	<.0001
EDAD_T0	1	-5.5593	0.0012

Cuadro 4.21: Segmento 4

Construcción del score

Considerando las variables como resultado de la regresión logística y las variables de cada uno de los clientes del segmento. Se calculó el score de comportamiento. Y se realizó una distribución por deciles que muestra el porcentaje de clientes malos que se puede ver en la tabla 4.22. El decil 395-583 tiene un 46.9 % de clientes malos.

SCORE	BUENOS	MALOS	Total	% MALOS	%BUENOS ACUM	% MALOS ACUM
395-583	3,577	3,156	6,733	46.9	6.9	20.7
584-615	4,262	2,406	6,668	36.1	15.0	36.4
616-638	4,746	2,043	6,789	30.1	24.1	49.8
639-658	5,196	1,723	6,919	24.9	34.1	61.1
659-676	5,357	1,431	6,788	21.1	44.4	70.4
677-693	5,285	1,280	6,565	19.5	54.5	78.8
694-710	5,655	1,077	6,732	16.0	65.3	85.9
711-730	5,951	878	6,829	12.9	76.8	91.6
731-756	5,866	726	6,592	11.1	88.0	96.4
757-887	6,256	554	6,810	8.1	100.0	100.0
Total	52,151	15,274	67,425			

Cuadro 4.22: Distribución de porcentaje de malos.

En la fig.4.156 se puede observar el comportamiento decreciente de la distribución del porcentaje de malos. Teniendo su máximo en el decil 395-583.

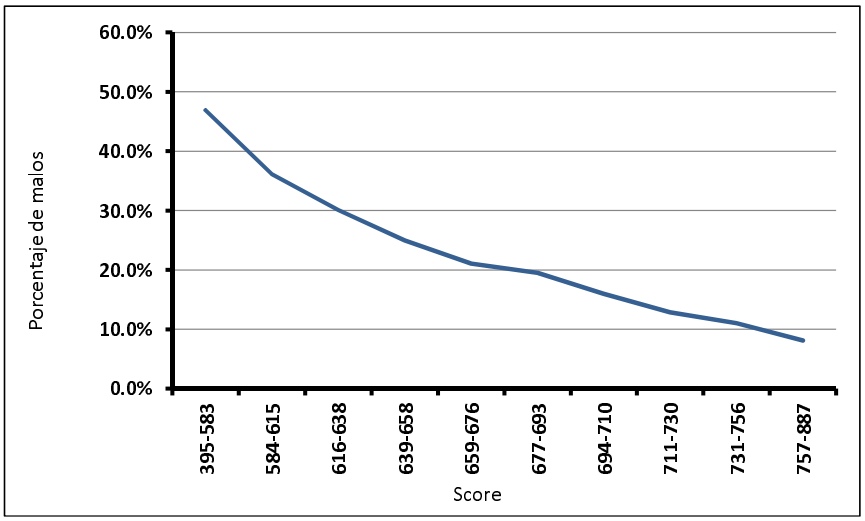


Figura 4.156: Distribución del score de comportamiento

Medidas de Desempeño

Kolmogorov Smirnov

Como medida de desempeño se obtuvo un KS=27.0, considerando el máximo entre un delta de 13.8 y 0.0 de los 9 deciles restantes. Se puede observar en el cuadro 4.23.

SCORE	BUENOS	MALOS	% MALOS	%BUENOS ACUM	%MALOS ACUM	DELTA
395-583	3,577	3,156	46.9	6.9	20.7	13.8
584-615	4,262	2,406	36.1	15.0	36.4	21.4
616-638	4,746	2,043	30.1	24.1	49.8	25.7
639-658	5,196	1,723	24.9	34.1	61.1	27.0
659-676	5,357	1,431	21.1	44.4	70.4	26.1
677-693	5,285	1,280	19.5	54.5	78.8	24.3
694-710	5,655	1,077	16.0	65.3	85.9	20.5
711-730	5,951	878	12.9	76.8	91.6	14.9
731-756	5,866	726	11.1	88.0	96.4	8.4
757-887	6,256	554	8.1	100.0	100.0	0.0
Total	52,151	15,274				**KS=27.0**

Cuadro 4.23: Índice de Kolmogorov-Smirnov.

En la fig. 4.157 de Kolmogorov Smirnov se puede observar la distribución del porcentaje de clientes malos versus la distribución de clientes buenos. En el decil 639-658 se encuentra el delta de 27.0.

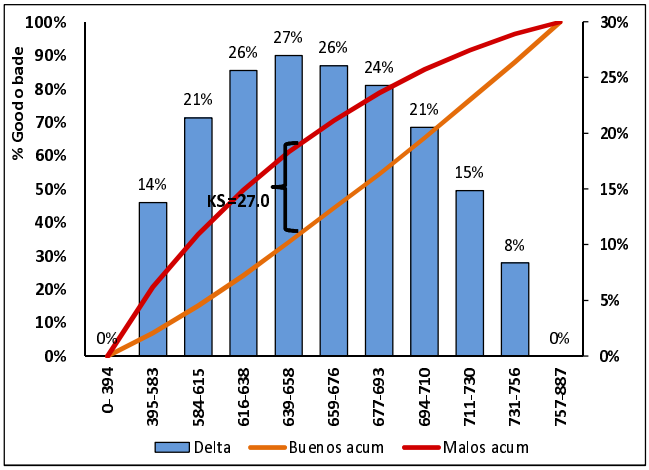


Figura 4.157: Índice KS=27.

Gini y Divergencia

En tabla 4.24 se muestran las variables correspondiente para el cálculo del índice de Gini=35.0 y Divergencia= 6.2.

SCORE	%BUENOS ACUM	%MALOS ACUM	$(x_{k+1}-x_k)(y_{k+1}+y_k)$	Clase	$\bar{X}buenos$	$\bar{X}malos$	σ^2_{goods}	σ^2_{malos}
395-583	6.9	20.7	0	489	1,749,153.0	1,543,284.0	134,298,642.9	6,1423,063.5
584-615	15.0	36.4	4.7	599.5	2,555,069.0	1,442,397.0	29,549,022.3	2,024,478.5
616-638	24.1	49.8	7.8	627.0	2,975,742.0	1,280,961.0	14,759,041.7	4642.2
639-658	34.1	61.1	11.0	648.5	3,369,606.0	1,117,365.5	6,100,733.2	688690.2
659-676	44.4	70.4	13.5	667.5	3,575,797.5	955,192.5	1,248,363.0	2,175,725.3
677-693	54.5	78.8	15.1	685.0	,3620,225.0	876,800.0	26389.1	4,085,010.0
694-710	65.3	85.9	17.9	702.0	,3969,810.0	756,054.0	2092168.5	5,817,052.0
711-730	76.8	91.6	20.3	720.5	4,287,695.5	632,599.0	8,473,606.5	7,430,196.9
731-756	88.0	96.4	21.1	743.5	4,361,371.0	539,781.0	21,637,829.3	9,600,114.7
757-887	100.0	100.0	23.6	822.0	5,142,432.0	455,388.0	121,280,441.5	20,741,420.4
			Gini=35.0					**D=6.2**

Cuadro 4.24: Índice de Gini y divergencia

En las fig. 4.158 se muestra el área de la curva de Lorenz donde se muestra que para el 90 % de clientes malos se tiene entre el 60 % y 70 % de clientes buenos.

La fig. 4.159 muestra la distribución tanto de clietes buenos y malos, siendo ésta la alta de entre las 6 hojas del árbol. Identificando su medias de 682.8 y 628.5 respectivamente.

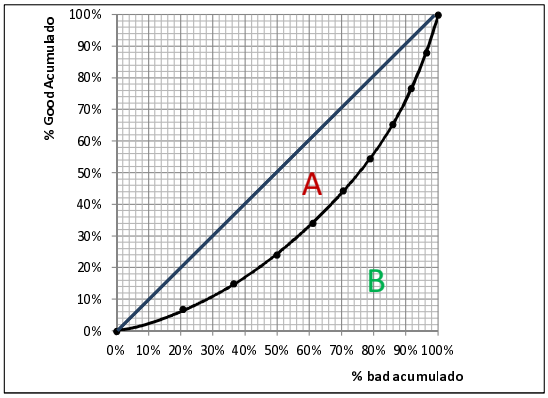


Figura 4.158: Índice de Gini=35.0.

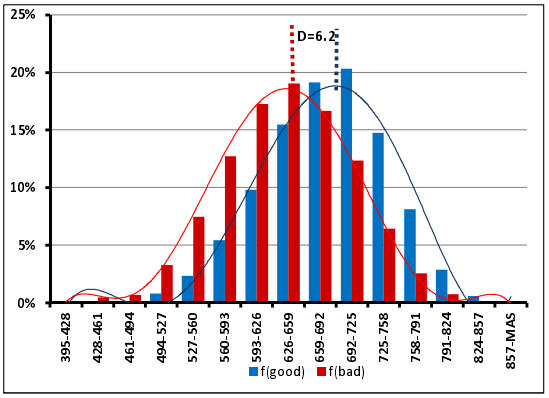


Figura 4.159: Divergencia=6.2.

4.7.5. Hoja 5

En la hoja 5 se consideraron los clientes con revolvencia t_0 >=43.5, utilización en el mes 0 >= 58.5 y Bucket en el mes 0=0, dando como resultado un total de 47,700 clientes, con 31,903 clientes con target=0 y 15,797 clientes con target=1.

Análisis univariado

A estos clientes se les aplicó un análisis univariado a cada una de las variables, de la fig. 4.160 a la fig. 4.183 se muestran las variables que se consideraron con un comportamiento creciente o decreciente para este segmento.

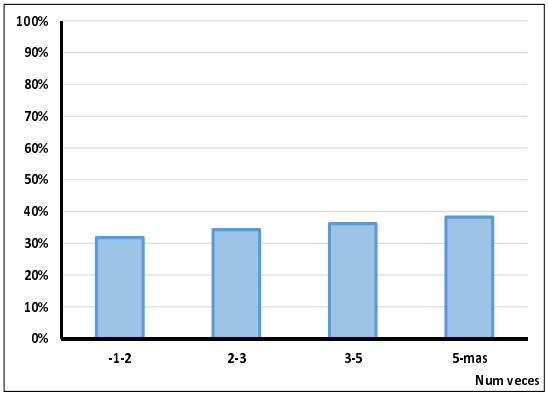


Figura 4.160: Veces con 1 pagos vencidos.

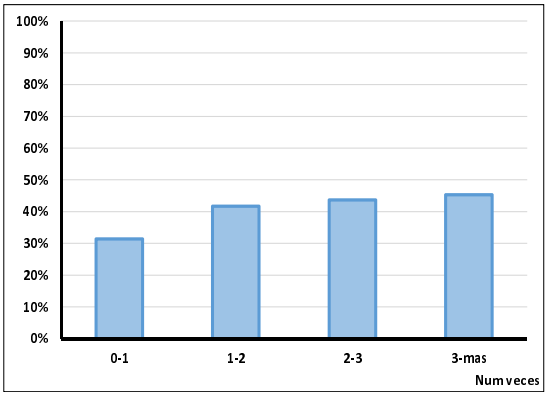


Figura 4.161: Veces con 2 pagos vencidos.

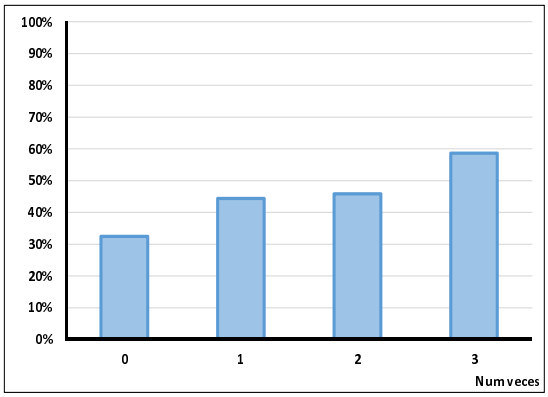


Figura 4.162: Veces con 3 pagos vencidos.

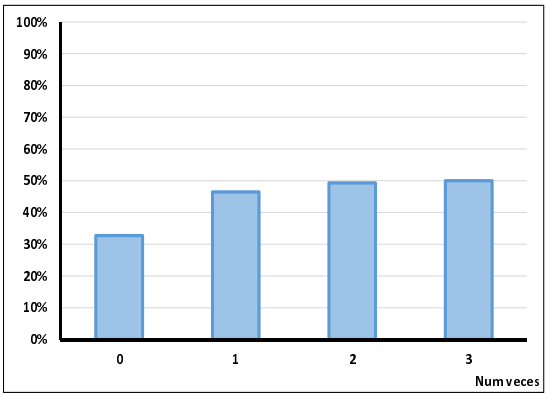


Figura 4.163: Veces con 4 pagos vencidos.

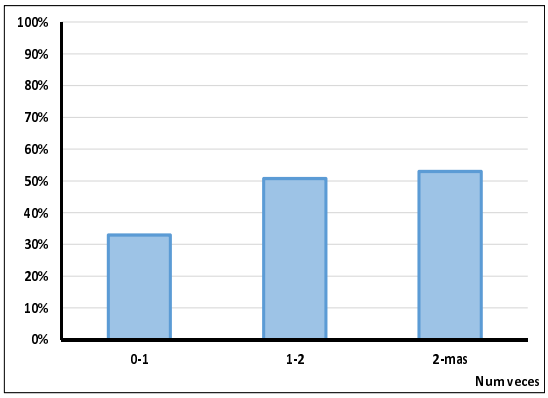


Figura 4.164: Veces con 5 pagos vencidos.

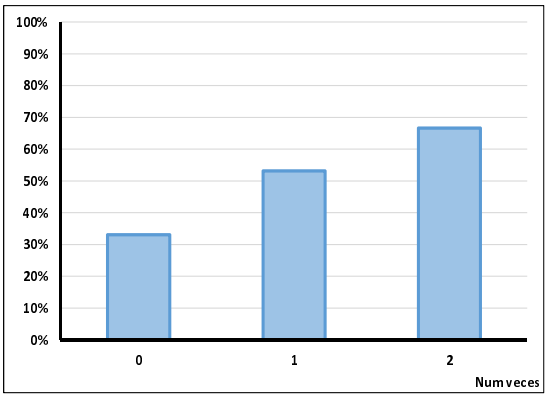


Figura 4.165: Veces con 6 pagos vencidos.

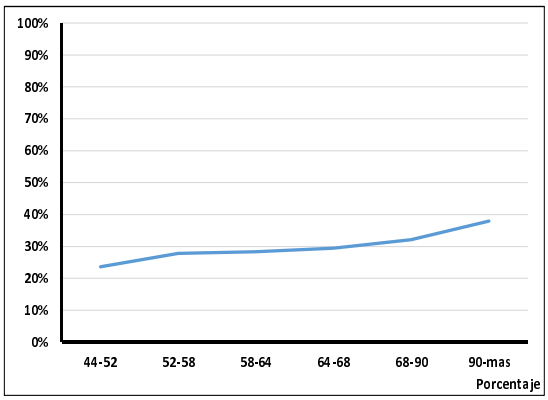


Figura 4.166: Revolvencia mes 0.

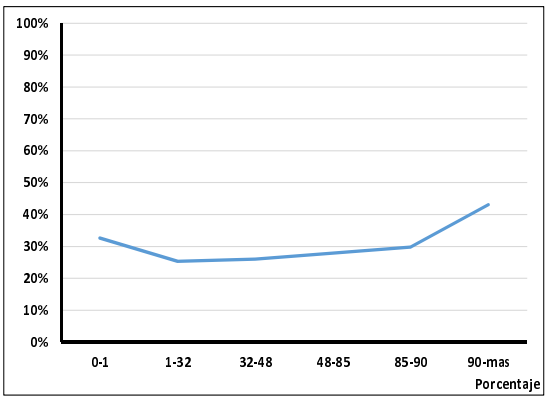


Figura 4.167: Revolvencia 1 mes antes.

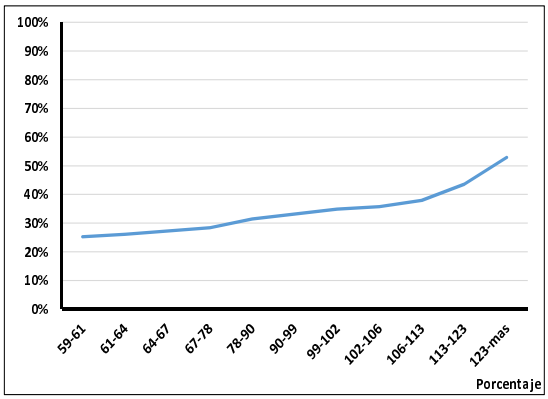


Figura 4.168: Utilización mes 0 de valuación.

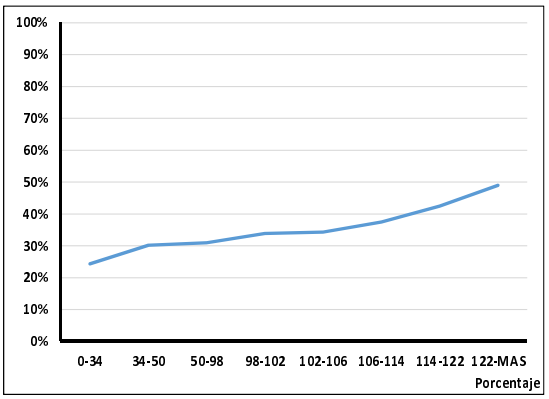


Figura 4.169: Utilización 1 mes antes de la valuación.

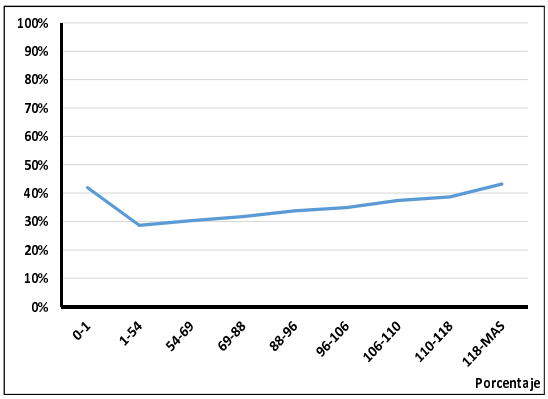


Figura 4.170: Utilización 2 meses anteriores.

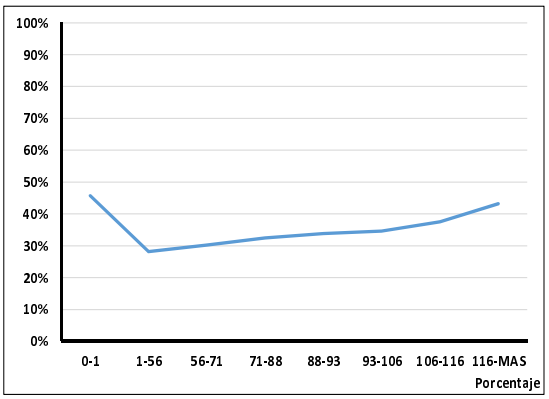


Figura 4.171: Utilización 3 meses anteriores.

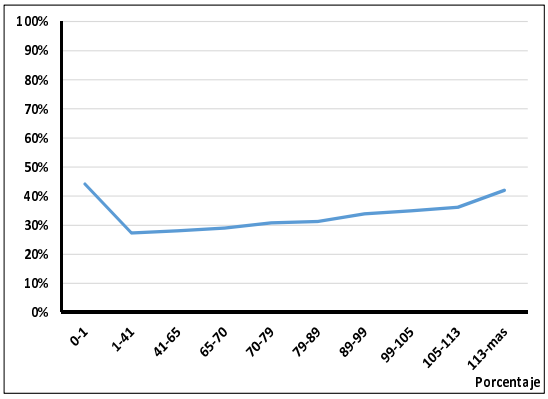


Figura 4.172: Utilización 4 meses anteriores.

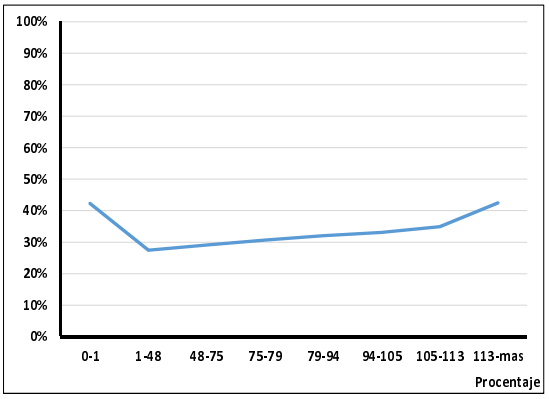


Figura 4.173: Utilización 5 meses anteriores.

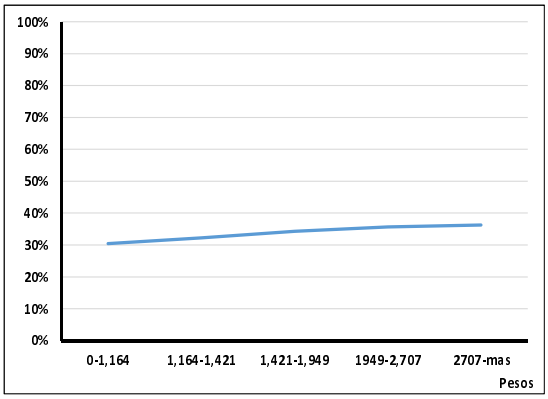


Figura 4.174: Saldo del mes 0.

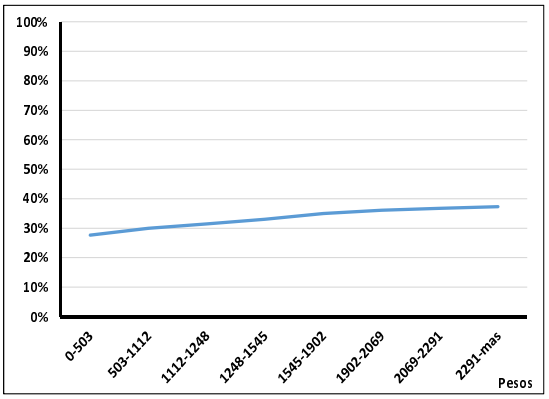


Figura 4.175: Saldo 1 mes anterior.

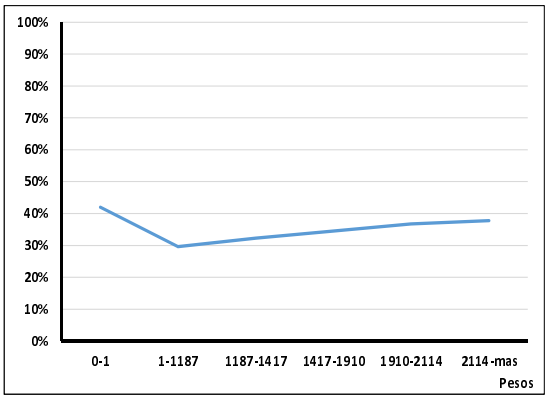


Figura 4.176: Saldo 2 meses anteriores.

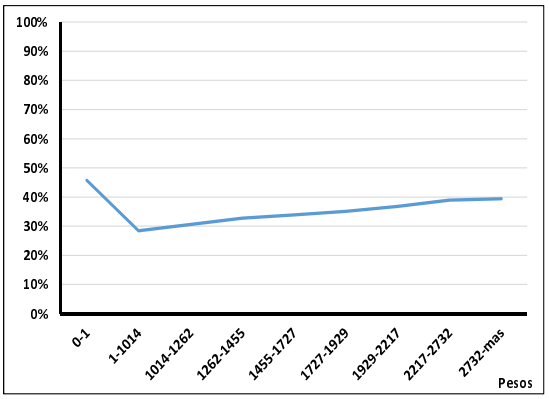


Figura 4.177: Saldo 3 meses anteriores.

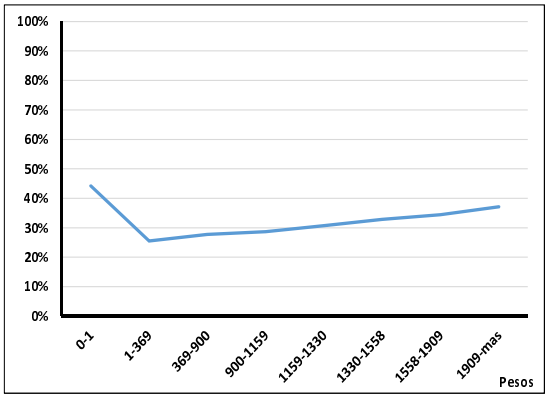


Figura 4.178: Saldo 4 meses anteriores.

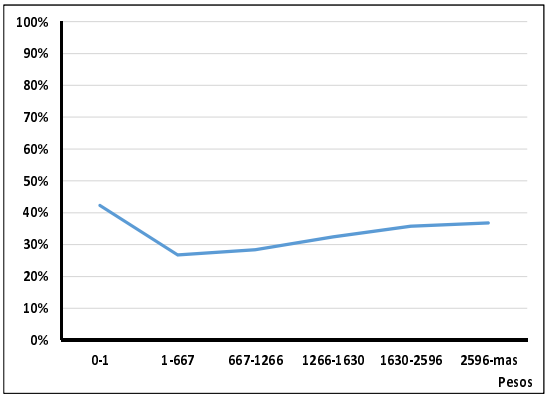


Figura 4.179: Saldo 5 meses anteriores.

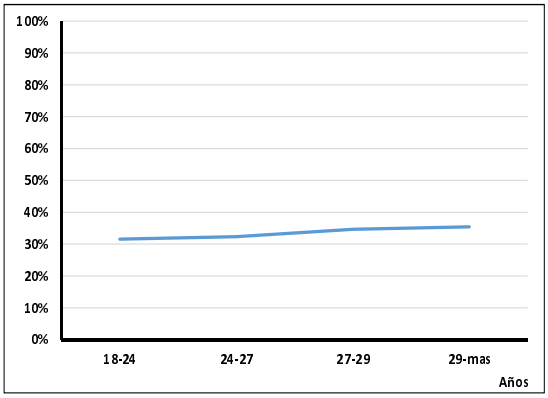


Figura 4.180: Edad en el mes 0.

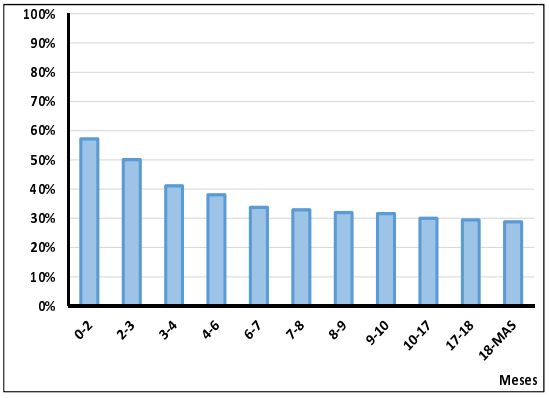


Figura 4.181: Months on Books en el mes 0.

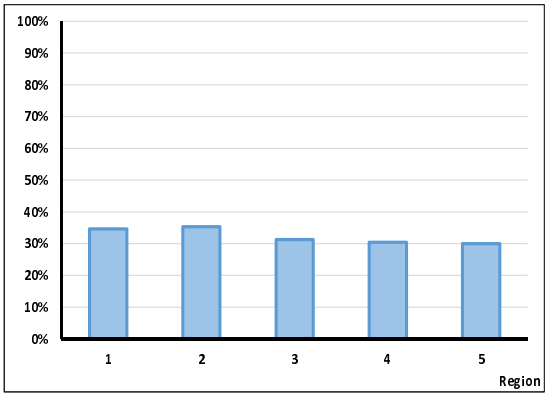


Figura 4.182: Entidad federativa.

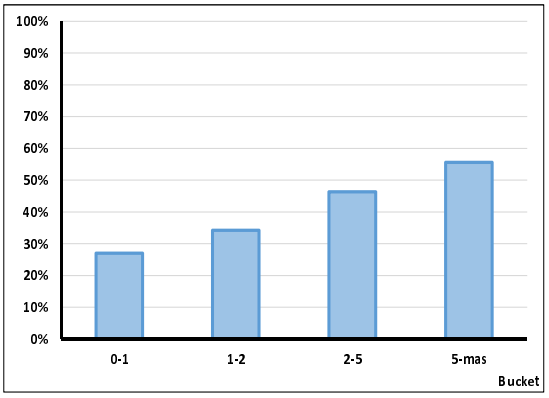


Figura 4.183: Máximo bucket en 12 meses.

Regresión logística

Con las 23 variables crecientes y decrecientes se calculó la regresión logística "hacia adelante", los resultados de las 12 variables significativas se muestran en l cuadro 4.25.

PARÁMETRO	DF	ESTIMADOR	p_value>JiSq
Intercept	1	12.2374	<.0001
MAX_BK_12M	1	-4.916	<.0001
V_1PV_0	1	-3.6559	<.0001
V_3PV_0	1	1.2918	0.0014
SAL_1	1	-5.7746	<.0001
SAL_4	1	-1.0048	<.0001
REV_0	1	-3.4566	<.0001
REV_1	1	-3.204	<.0001
UTIL_0	1	-2.1908	<.0001
REGION_T0	1	-4.4061	<.0001
MOB_T0	1	-5.6883	<.0001
EDAD_T0	1	-1.6483	0.0114

Cuadro 4.25: Segmento 5.

Construcción del score

Con las 12 variables resultado de la regresión logística se realizó el calculo de *credit score*, en el cuadro 4.26 se muestra la distribución de los percentiles y la distribución del porcentaje de clientes malos.

SCORE	BUENOS	MALOS	Total	% MALOS	%BUENOS ACUM	% MALOS ACUM
340-525	1,917	2,794	4,711	59.3	6.0	17.7
526-562	2,507	2,340	4,847	48.3	13.9	32.5
563-586	2,703	2,031	4,734	42.9	22.3	45.4
587-605	2,960	1,757	4,717	37.2	31.6	56.5
606-622	3,250	1,603	4,853	33.0	41.8	66.6
623-637	3,250	1,353	4,603	29.4	52.0	75.2
638-653	3,669	1,358	5,027	27.0	63.5	83.8
654-669	3,498	1,056	4,554	23.2	74.5	90.5
670-690	3,936	9,19	4,855	18.9	86.8	96.3
691-783	4,213	586	4,799	12.2	100.0	100.0
Total	31,903	15,797	47,700			

Cuadro 4.26: Distribución de porcentaje de malos.

En la fig. 4.184 se puede apreciar la distribución decreciente de score de Comportamiento en deciles. Teniendo como máximo decil 340-525 puntos y como mínimo decil 691-783 puntos.

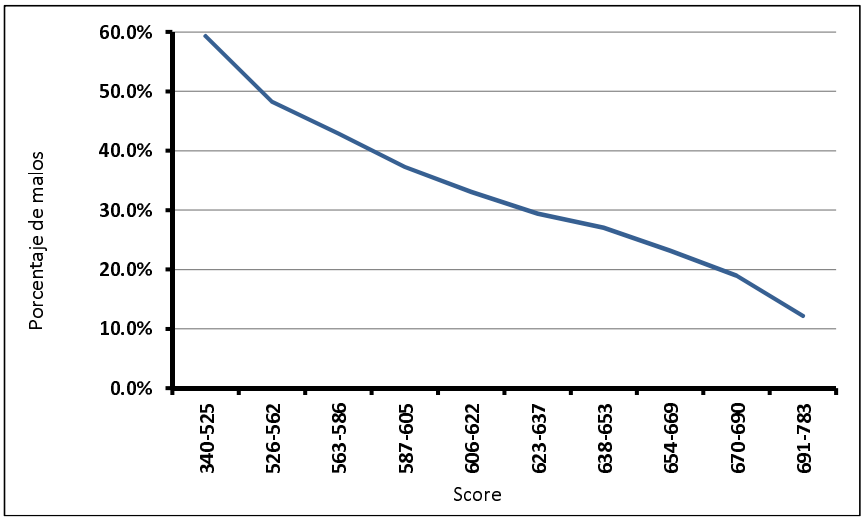


Figura 4.184: Distribución del score de comportamiento.

Medidas de desempeño

Kolmogorov Smirnov

Para la hoja 5 se obtuvo un KS=24.9, mismo que se puede ver el cuadro 4.27, así como la forma en que se calculó.

SCORE	BUENOS	MALOS	% MALOS	%BUENOS ACUM	%MALOS ACUM	DELTA
340-525	1,917	2,794	59.3	6.0	17.7	11.7
526-562	2,507	2,340	48.3	13.9	32.5	18.6
563-586	2,703	2,031	42.9	22.3	45.4	23.0
587-605	2,960	1,757	37.2	31.6	56.5	24.9
606-622	3,250	1,603	33.0	41.8	66.6	24.8
623-637	3,250	1,353	29.4	52.0	75.2	23.2
638-653	3,669	1,358	27.0	63.5	83.8	20.3
654-669	3,498	1,056	23.2	74.5	90.5	16.0
670-690	3,936	9,19	18.9	86.8	96.3	9.5
691-783	4,213	586	12.2	100.0	100.0	0.0
Total	31,903					**KS=24.9**

Cuadro 4.27: Índice de Kolmogorov-Smirnov.

En la fig.4.185 se puede observar la diferencia de porcentajes entre la distribución de clientes buenos y la distribución de clientes malos, obteniendo una delta máxima en el decil 587-605 de 24.9.

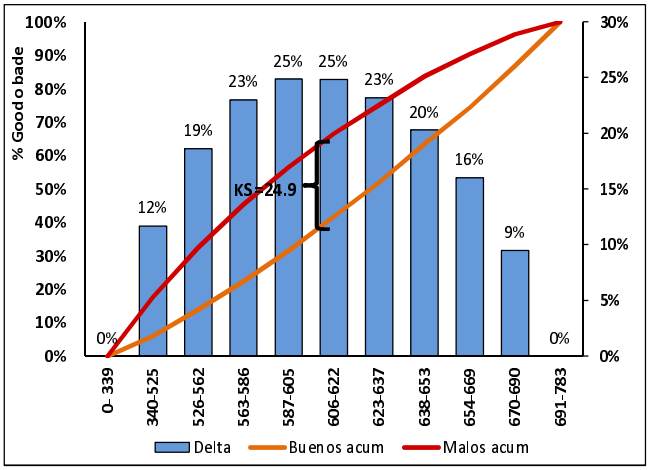


Figura 4.185: Índice KS=24.9.

Gini y divergencia

El cuadro 4.28 muestra el cálculo del índice de Gini=33.1 y con una divergencia=3.1, con una media de 627.4 para clientes buenos y una media de 580.1 para clientes malos.

SCORE	%BUENOS ACUM	%MALOS ACUM	$(x_{k+1}-x_k)(y_{k+1}+y_k)$	Clase	$\bar{X}buenos$	$\bar{X}malos$	σ^2_{goods}	σ^2_{malos}
340-525	6.0	17.7	0	432.5	829102.5	1208405.0	72836264.2	60873266.7
526-562	13.9	32.5	3.9	544.0	1363808.0	1272960.0	17447149.5	3050300.7
563-586	22.3	45.4	6.6	574.5	1552873.5	1166809.5	7570641.1	63798.5
587-605	31.6	56.5	9.4	596.0	1764160.0	1047172.0	2922692.5	443926.2
606-622	41.8	66.6	12.5	614.0	1995500.0	984242.0	585562.8	1841675.9
623-637	52.0	75.2	14.4	630.0	2047500.0	852390.0	21585.4	3368352.8
638-653	63.5	83.8	18.3	645.5	2368339.5	876589.0	1198966.8	5807553.5
654-669	74.5	90.5	19.2	661.5	2313927.0	698544.0	4062057.7	6996210.6
670-690	86.8	96.3	23.0	680.0	2676480.0	624920.0	10880503.9	9170771.4
691-783	100.0	100.0	25.9	737.0	3104981.0	431882.0	50586121.4	14425060.4
			Gini=33.3					**D=3.1**

Cuadro 4.28: Índice de Gini y divergencia.

La fig.4.186 muestra la gráfica del índice de Gini de 33.3 puntos. Si bien se puede observar que la distribución del score es buena ya que se encuentra más cerca de la línea de igualdad.

En la fig. 4.187 se puede observar la divergencia de éste score, así como la media de 627.4 para clientes buenos y una media de 580.1 para clientes malos, así como la representación gráfica de ambos tipos de clientes.

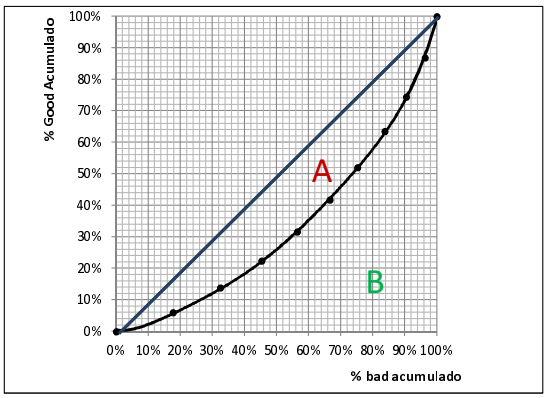


Figura 4.186: Índice de Gini=33.3.

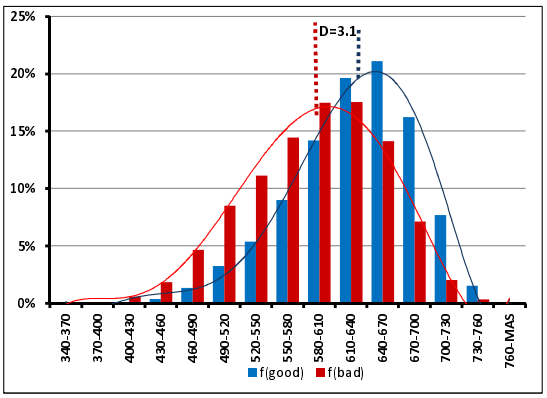


Figura 4.187: Divergencia=3.1.

4.7.6. Hoja 6

La hoja 6 compuesta por los clientes que su revolvencia t_0 >= 43.5, utilización en el mes 0 >= 58.5 y que el bucket en el mes 0 >= 1. Obteniendo 19,627 de clientes para este segmento.

Análisis univariado

Las varibles dela fig. 4.188 a la fig. 4.215 se consideraron con un comportamiento creciente o decreciente ajustadas después de aplicarles el análisis univariado. Se suavizó el comportamiento de las variables agrupando los puntos donde mostraban picos en su recorrido. Estas variables son las utilizadas para el cálculo de la regresión logística.

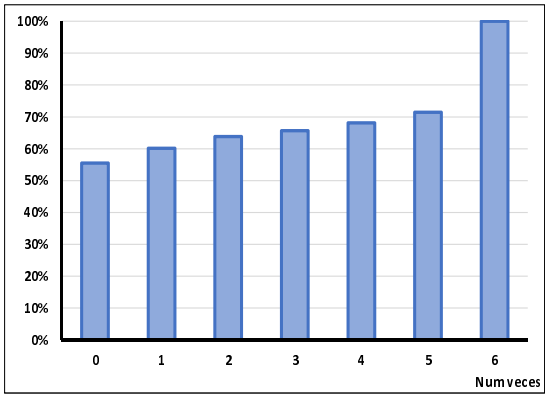


Figura 4.188: Veces con 2 pagos vencidos.

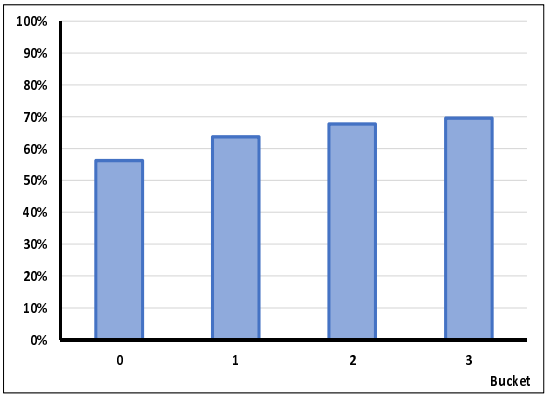


Figura 4.189: Veces con 3 pagos vencidos.

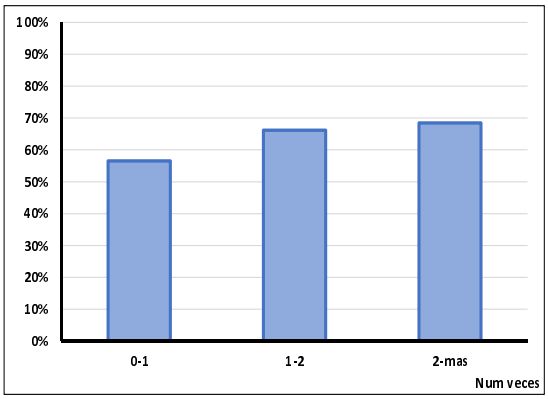


Figura 4.190: Veces con 4 pagos vencidos.

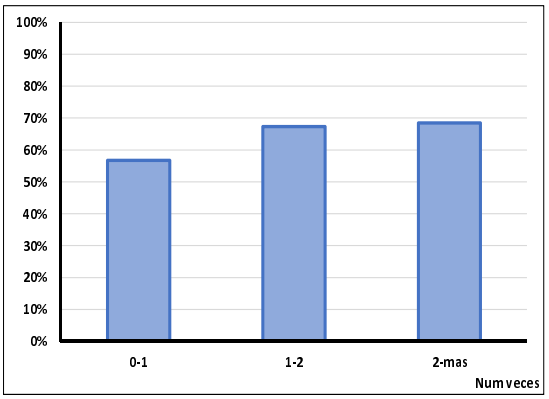


Figura 4.191: Veces con 5 pagos vencidos.

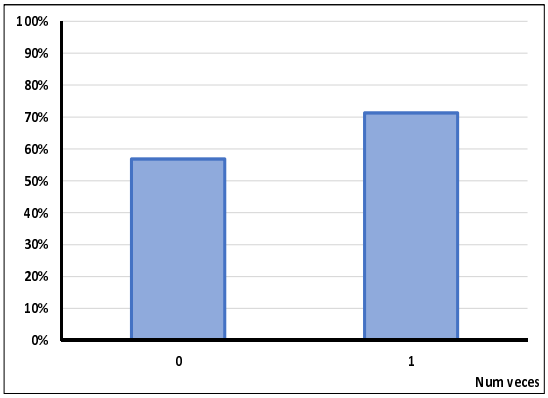


Figura 4.192: Veces con 6 pagos vencidos.

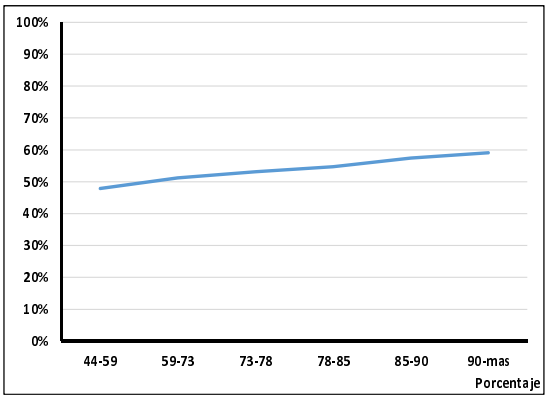


Figura 4.193: Revolvencia mes 0.

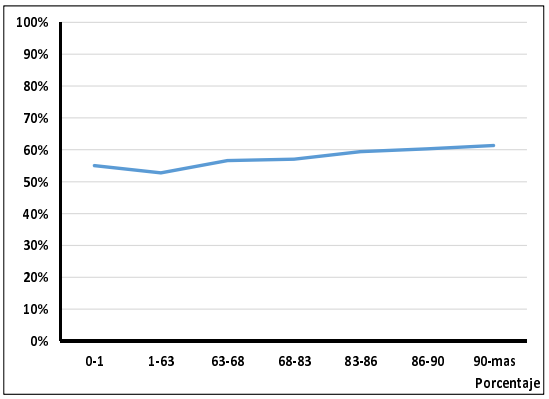


Figura 4.194: Revolvencia 1 mes antes.

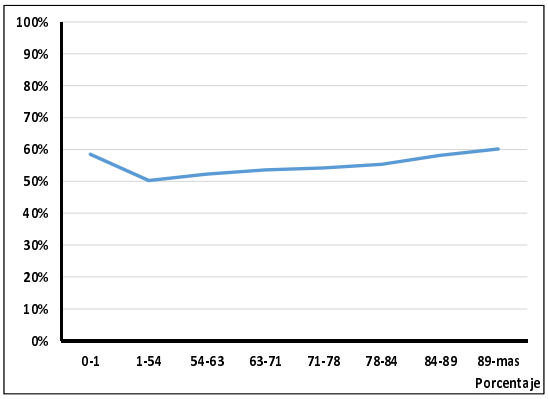


Figura 4.195: Revolvencia 2 meses anteriores.

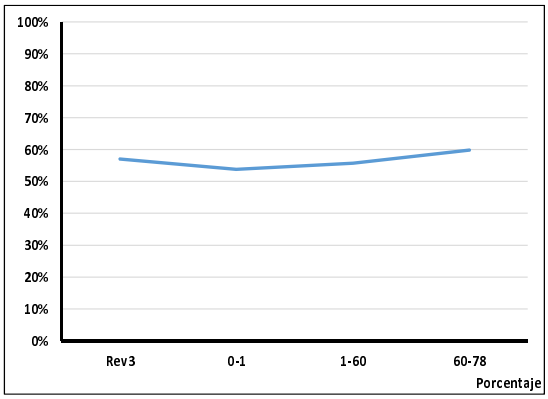


Figura 4.196: Revolvencia 3 meses anteriores.

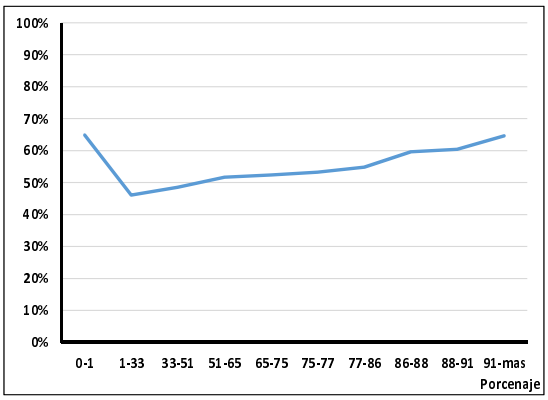


Figura 4.197: Revolvencia 4 meses anteriores.

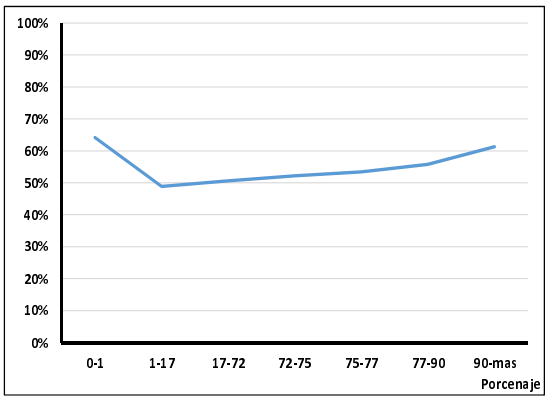


Figura 4.198: Revolvencia 5 meses anteriores.

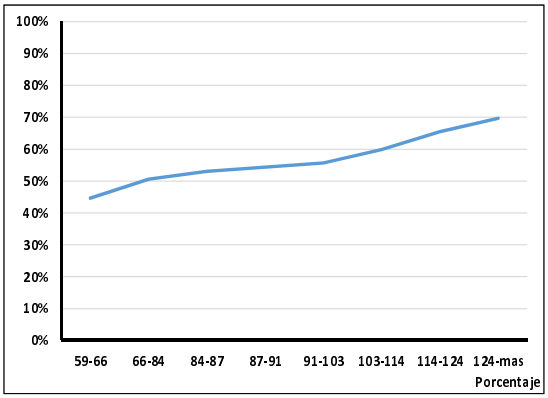


Figura 4.199: Utilización mes 0 de valuación.

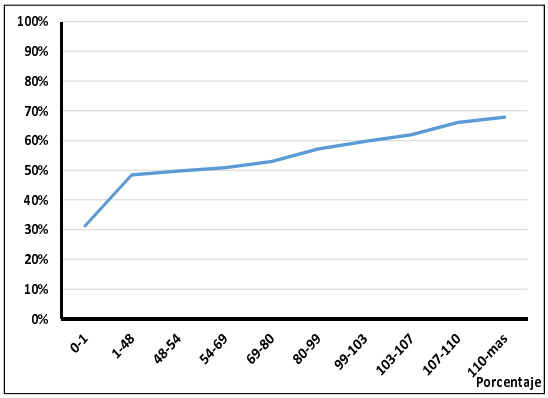


Figura 4.200: Utilización 1 mes antes de la valuación.

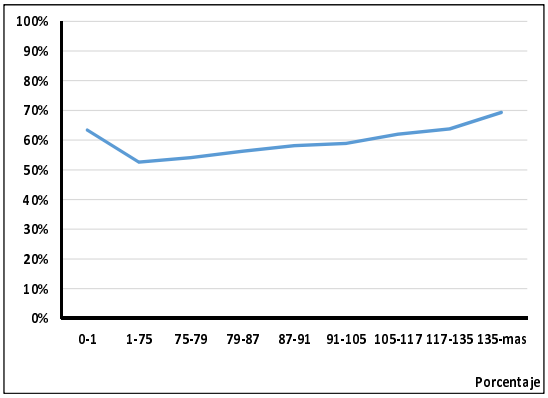


Figura 4.201: Utilización 2 meses anteriores.

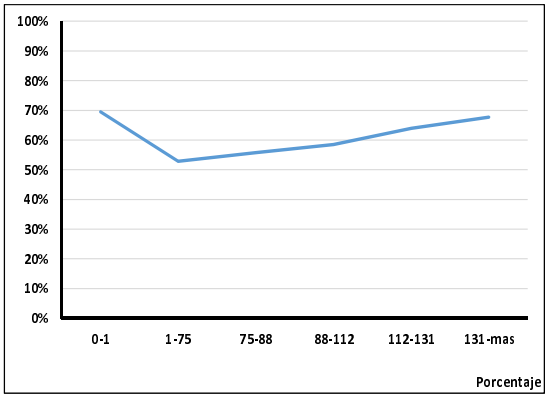


Figura 4.202: Utilización 3 meses anteriores.

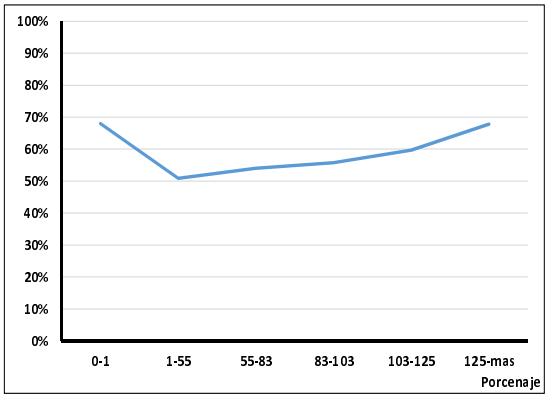


Figura 4.203: Utilización 4 meses anteriores.

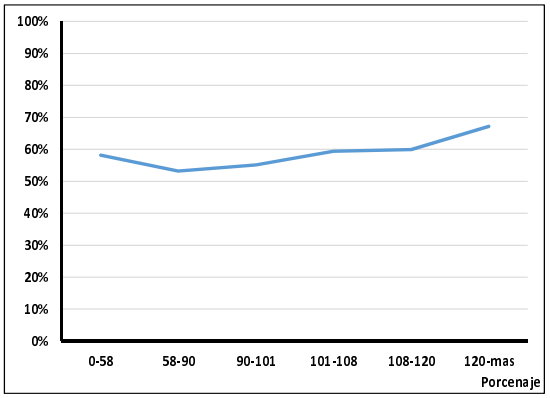


Figura 4.204: Utilización 5 meses anteriores.

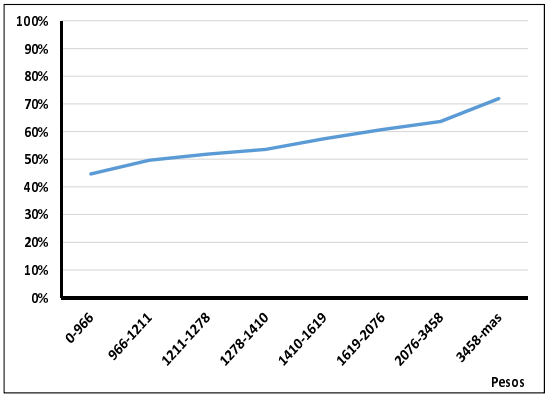


Figura 4.205: Saldo del mes 0.

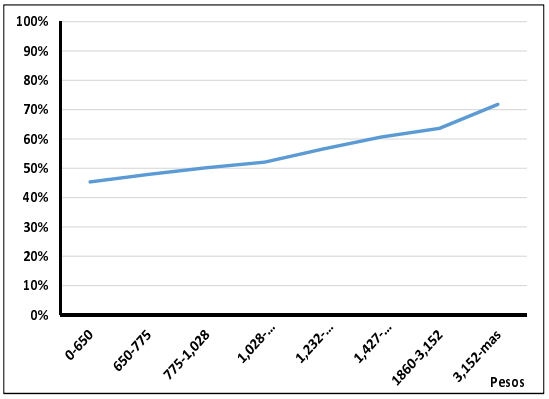


Figura 4.206: Saldo 1 mes anterior.

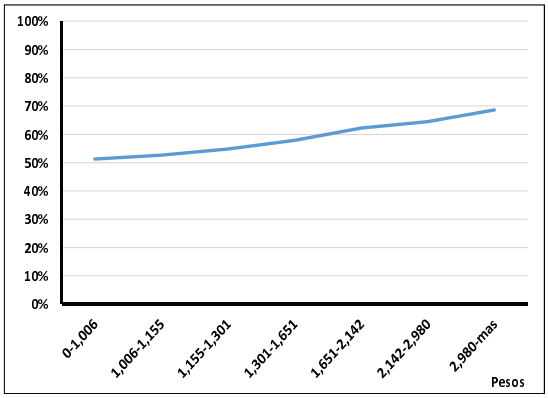


Figura 4.207: Saldo 2 meses anteriores.

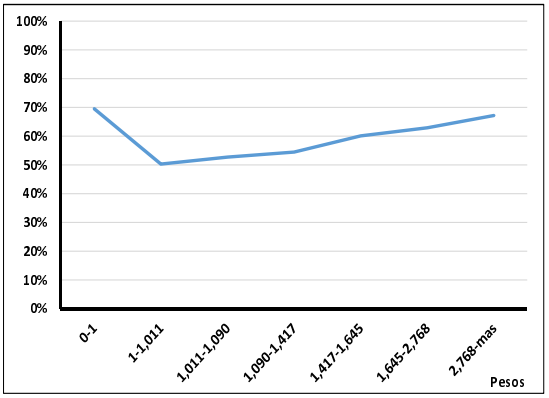


Figura 4.208: Saldo 3 meses anteriores.

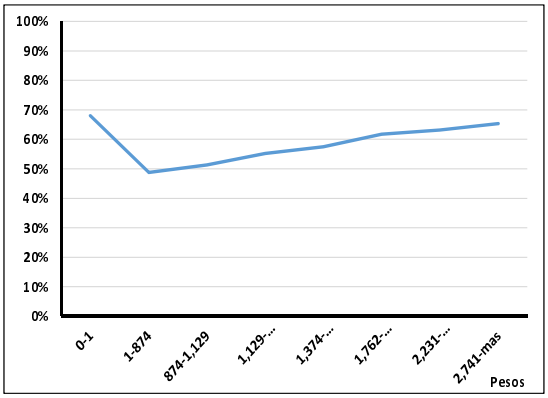


Figura 4.209: Saldo 4 meses anteriores.

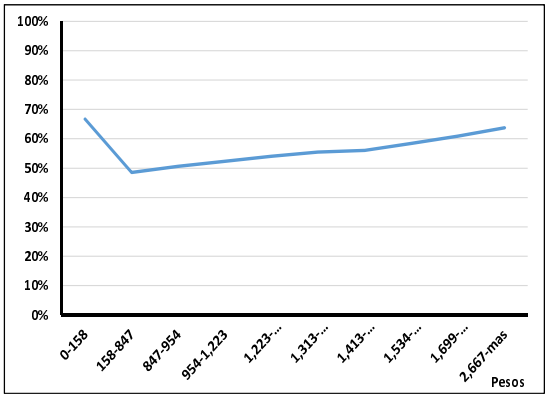


Figura 4.210: Saldo 5 meses anteriores.

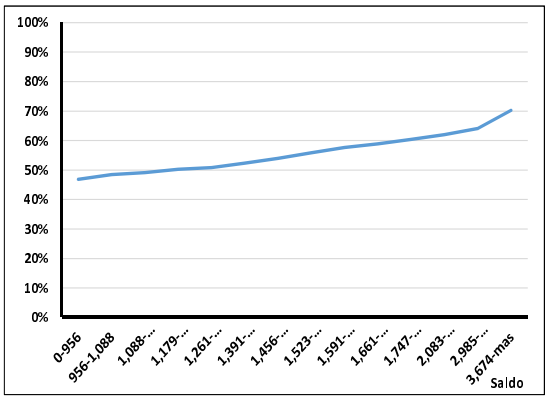


Figura 4.211: Máximo Saldo.

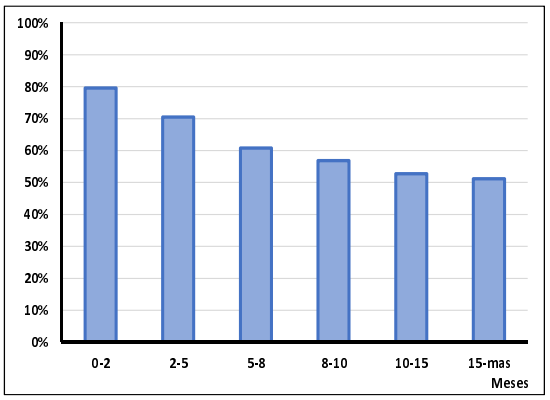


Figura 4.212: Months on Books en el mes 0.

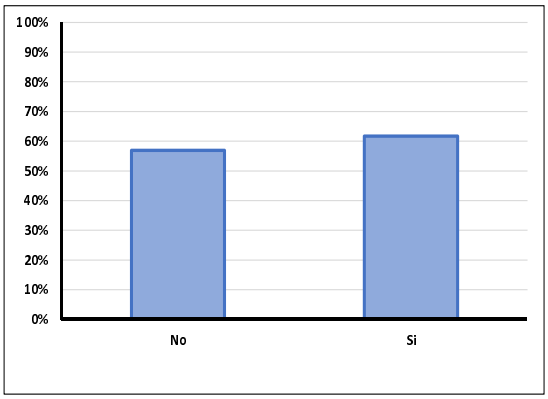


Figura 4.213: Disposición en efectivo.

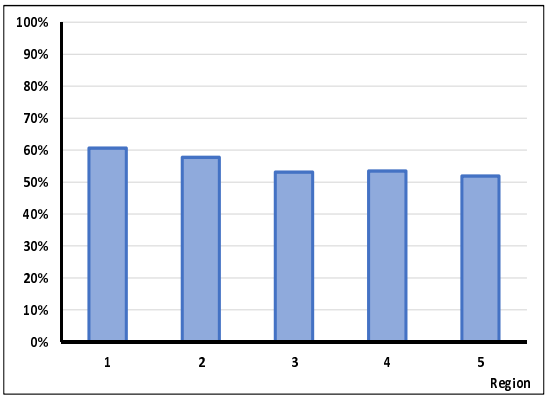


Figura 4.214: Entidad federativa.

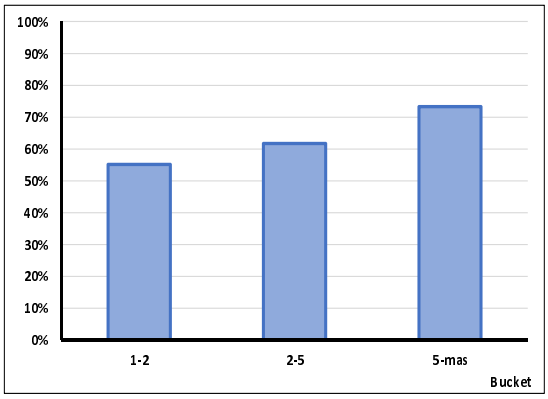


Figura 4.215: Máximo bucket en 12 meses.

Regresión logística

En el cuadro 4.29 se puede ver el resultado de aplicar una regresión logística de tipo "hacia adelante", considerando las 28 variables de análisis univariado, obteniendo 13 variables y un intercepto, para construir el score de comportamiento para el segmento.

PARÁMETRO	DF	ESTIMADOR	p_value>JiSq
Intercept	1	21.8641	<.0001
MAX_BK_12M	1	-3.2238	<.0001
V_2PV_0	1	-4.1535	<.0001
SAL_0	1	-2.4278	<.0001
SAL_2	1	-2.5219	<.0001
REV_0	1	-3.0476	<.0001
REV_1	1	-6.0734	<.0001
REV_2	1	-2.8354	<.0001
REV_3	1	-3.1209	0.0007
UTIL_0	1	-2.4468	<.0001
UTIL_2	1	1.7253	0.0008
UTIL_4	1	-0.7226	0.0196
REGION_T0	1	-4.5334	<.0001
MOB_T0	1	-5.5631	<.0001

Cuadro 4.29: Segmento 6.

Construcción del score

Del cálculo de score de comportamiento se generaron los deciles que se pueden ver en el cuadro 4.30, donde se puede observar que ningún decil tiene un porcentaje de malos menor al 50 %.

SCORE	BUENOS	MALOS	Total	% MALOS	%BUENOS ACUM	% MALOS ACUM
241-424	363	1,605	1,968	81.6	4.3	14.4
425-457	541	1,422	1,963	72.4	10.7	27.1
458-480	646	1,347	1,993	67.6	18.4	39.1
481-498	702	,1223	1,925	63.5	26.7	50.1
499-514	799	1,161	1,960	59.2	36.1	60.5
515-530	929	1,088	2,017	53.9	47.1	70.2
531-546	973	996	1,969	50.6	58.6	79.1
547-564	1,037	866	1,903	45.5	70.9	86.9
565-587	1,148	841	1,989	42.3	84.5	94.4
588-697	1,313	627	1,940	32.3	100.0	100.0
Total	8,451	11,176	19,627			

Cuadro 4.30: Distribución de porcentaje de malos.

En la gráfica 4.216 se puede apreciar que el score va de 242 a 697 puntos, es decir que no tiene clientes con una buena calificación.

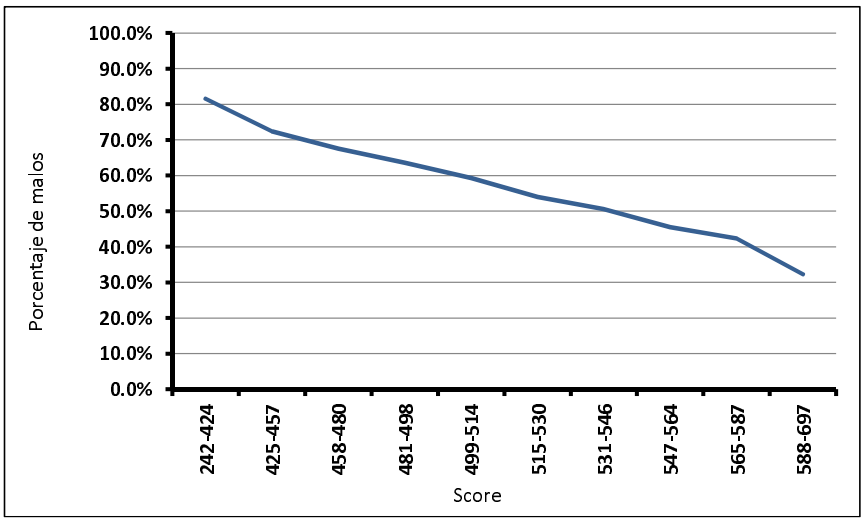


Figura 4.216: Distribución del score de comportamiento.

Medidas de Desempeño

Kolmogorov-Smirnov

Se obtuvo un KS=24.4 como medida de desempeño, el cuadro 4.31 muestra el cálculo de índice.

SCORE	BUENOS	MALOS	% MALOS	%BUENOS ACUM	%MALOS ACUM	DELTA
241-424	363	1,605	81.6	4.3	14.4	10.1
425-457	541	1,422	72.4	10.7	27.1	1.4
458-480	646	1,347	67.6	18.4	39.1	20.8
481-498	702	,1223	63.5	26.7	50.1	23.4
499-514	799	1,161	59.2	36.1	60.5	24.4
515-530	929	1,088	53.9	47.1	70.2	23.1
531-546	973	996	50.6	58.6	79.1	20.5
547-564	1,037	866	45.5	70.9	86.9	16.0
565-587	1,148	841	42.3	84.5	94.4	9.9
588-697	1,313	627	32.3	100.0	100.0	0.0
Total	8,451	11,176				**KS=24.4**

Cuadro 4.31: Índice de Kolmogorov-Smirnov.

En la fig. 4.217 se muestra la distribución del delta de cada uno se los deciles de la hoja, así como la distribución del porcentaje de clientes malos y clientes buenos.

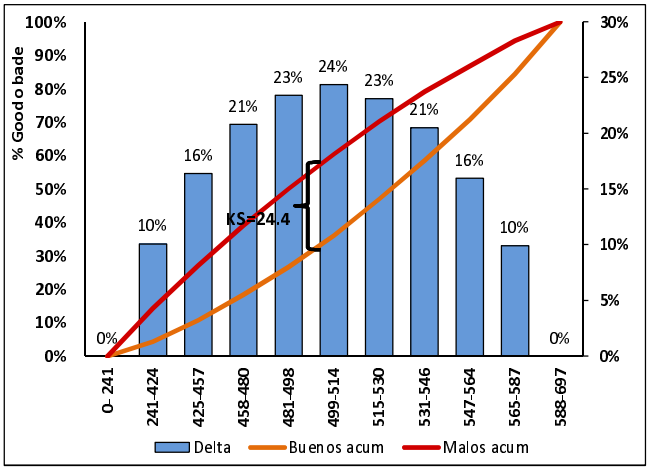


Figura 4.217: Índice KS=24.4.

Gini y divergencia

En el cuadro 4.32 se muestran los cálculos que apoyaron para obtener un índice de Gini de 32.3 y una divergencia de 2.8.

SCORE	%BUENOS ACUM	%MALOS ACUM	$(x_{k+1}-x_k)(y_{k+1}+y_k)$	Clase	$\bar{X}buenos$	$\bar{X}malos$	σ^2_{goods}	σ^2_{malos}
241-424	4.3	14.4	0.0	332.5	120,697.5	533,662.5	14,531,763.0	38,741,309.1
425-457	10.7	27.1	2.7	441.0	238,581.0	627,102.0	4,537,409.7	3,123,009.9
458-480	18.4	39.1	5.1	469.0	302,974.0	631,743.0	2,611,482.8	479,316.7
481-498	26.7	50.1	7.4	489.5	343,629.0	598,658.5	1,302,892.6	3,274.5
499-514	36.1	60.5	10.5	506.5	404693.5	588046.5	543,494.5	403,227.7
515-530	47.1	70.2	14.4	522.5	485,402.5	568,480.0	94,411.0	1,305,242.6
531-546	58.6	79.1	17.2	538.5	523,960.5	536,346.0	34088.7	2,553,776.1
547-564	70.9	86.9	20.4	555.5	576,053.5	481,063.0	5,44,716.1	3,961,660.4
565-587	84.5	94.4	24.6	576.0	661,248.0	484,416.0	2,164,220.9	6,532,890.4
588-697	100.0	100.0	30.2	642.5	843,602.5	402,847.5	15,863,912.0	14,993,060.6
			Gini=32.3					**D=2.8**

Cuadro 4.32: Índice de Gini y divergencia.

En las fig. 4.218 se muestran la curva de Lorenz, cuando se tiene el 90% de malos clientes, ya se acumuló el 70% de buenos clientes.

En fig. 4.219 se muestra la divergencia para el score de esta hoja. Las distribuciones de clientes buenos y malos, sobre los deciles del score de comportamiento, con medias de 532.6 para clientes buenos y 487.9 para clientes malos.

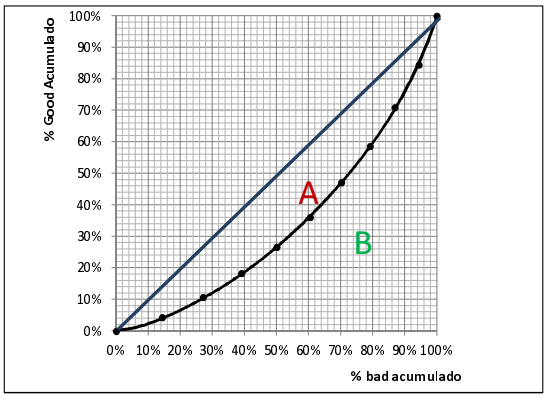


Figura 4.218: Índice de Gini=32.3.

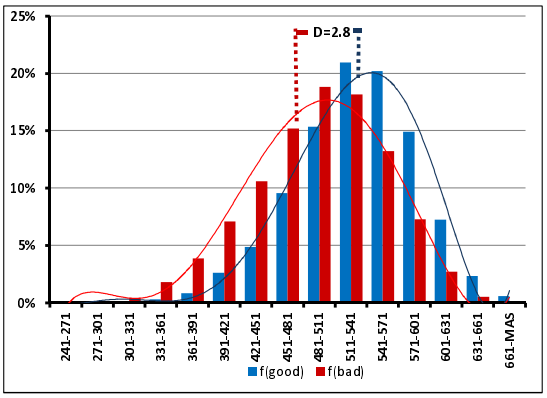


Figura 4.219: Divergencia=2.8.

4.7.7. Unificación de las 6 hojas del árbol de decisión

La fig.4.33 muestra los deciles y la distribución de clientes malos de los 352,777 clientes.

SCORE	BUENOS	MALOS	Total	% MALOS	%BUENOS ACUM	% MALOS ACUM
241-570	15,505	19,612	35,117	55.8	5.4	29.8
571-625	22,597	12,644	35,241	35.9	13.3	48.9
626-662	25,910	9,269	35,179	26.3	22.3	63.0
663-692	28,386	7,236	35,622	20.3	32.2	74.0
693-721	30,109	5,482	35,591	15.4	42.7	82.3
722-753	31,260	4,196	35,456	11.8	53.6	88.7
754-787	31,810	2,957	34,767	8.5	64.7	93.2
788-823	32,674	2,181	34,855	6.3	76.1	96.5
824-867	34,010	1,521	35,531	4.3	87.9	98.8
968-1,020	34,613	805	35,418	2.3	100.0	100.0
Total	286,874	65,903	352,777			

Cuadro 4.33: Distribución de porcentaje de malos.

En la fig. 4.220 se puede observar el comportamiento decreciente del porcentaje de malos. El score máximo que se obtuvo en el modelo fue de 1,020 puntos.

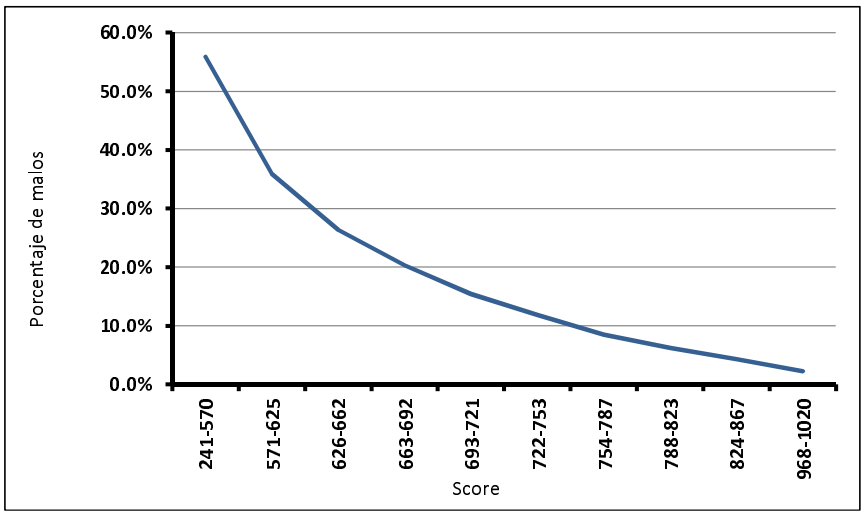


Figura 4.220: Distribución del score de comportamiento.

Medidas de desempeño

Índice de Kolmogorov-Smirnov

Para poder obtener objetivamente una conclusión se calcula el índice de Kolmogorov-Smirnov total. Y para conocer más sobre su comportamiento se muestra la distribución de su porcentaje de clientes malos en el cuadro 4.34.

SCORE	BUENOS	MALOS	% MALOS	%BUENOS ACUM	%MALOS ACUM	DELTA
241-570	15,505	19,612	55.8	5.4	29.8	24.4
571-625	22,597	12,644	35.9	13.3	48.9	35.7
626-662	25,910	9,269	26.3	22.3	63.0	40.7
663-692	28,386	7,236	20.3	32.2	74.0	41.8
693-721	30,109	5,482	15.4	42.7	82.3	39.6
722-753	31,260	4,196	11.8	53.6	88.7	35.1
754-787	31,810	2,957	8.5	64.7	93.2	28.5
788-823	32,674	2,181	6.3	76.1	96.5	20.4
824-867	34,010	1,521	4.3	87.9	98.8	10.8
968-1,020	34,613	805	2.3	100.0	100.0	0.0
Total	286,874	65,903	352,777			**KS=42**

Cuadro 4.34: Índice de Kolmogorov-Smirnov.

En la fig. 4.221 se puede observar la distribución de clientes malos y de buenos, así como el comportamiento de sus diferencias por decil. El KS=42 se encuentra en el decil 663-692.

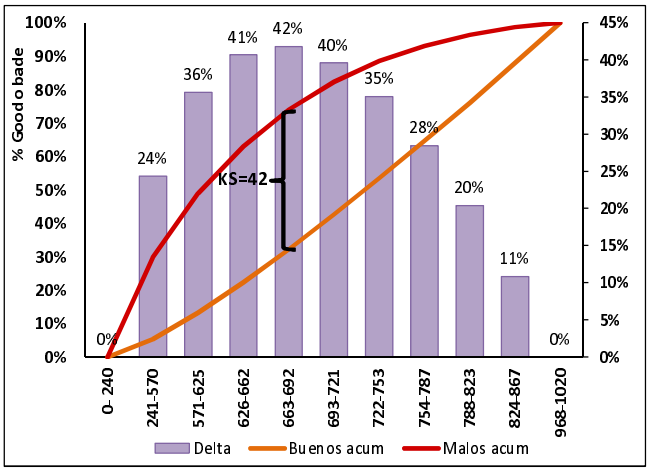


Figura 4.221: Índice KS=42.

Índice de Gini y Divergencia

De igual forma se consideró el total de 352,777 clientes para el cálculo de índice de Gini, se obtuvo un 53.8.Y una divergencia= 25.4 considerando unas medias para los buenos clientes de 746.1 y 599.0 para clientes malos.

SCORE	%BUENOS ACUM	%MALOS ACUM	$(x_{k+1}-x_k)(y_{k+1}+y_k)$	Clase	$\bar{X}buenos$	$\bar{X}malos$	σ^2_{goods}	σ^2_{malos}
241-570	5.4	29.8	0.0	405.5	6,287,277.5	7,952,666.0	1,799,142,221.2	734,452,216.0
571-625	13.3	48.9	6.2	598.0	13,513,006.0	7561,112.0	495,907,948.5	13,097.1
626-662	22.3	63.0	10.1	644.0	16,686,040.0	5,969,236.0	270,313,195.4	18,754,911.6
663-692	32.2	74.0	13.6	677.5	19,231,515.0	4,902,390.0	13,3742,932.2	44,569,865.9
693-721	42.7	82.3	16.4	707.0	21,287,063.0	3,875,774.0	46,127,421.8	63,921,018.8
722-753	53.6	88.7	18.6	737.5	23,054,250.0	3,094,550.0	2,334,062.7	80,468,078.9
754-787	64.7	93.2	20.2	770.5	24,509,605.0	2,278,368.5	18,874,878.4	86,954,010.1
788-823	76.1	96.5	21.6	805.5	26,318,907.0	1,756,795.5	115,126,713.2	92,986,749.7
824-867	87.9	98.8	23.1	845.5	28,755,455.0	1,286,005.5	335,754,192.1	92,406,064.6
968-1,020	100.0	100.0	24.0	994.0	3,4405,322.0	800,170.0	2,126,418,707.0	125,588,830.2
			Gini=53.8					**D=25.4**

Cuadro 4.35: Índice de Gini y divergencia.

Para el índice de Gini, la curva de Lorenz fig.4.223 muestra que para un 70 % de clientes malos se tiene un 20 % de clientes buenos.

La fig.4.223 muestra la distribución tanto de clientes buenos y malos, la ubicación de sus medias. Y los puntos en los que coinciden ambas distribuciones.

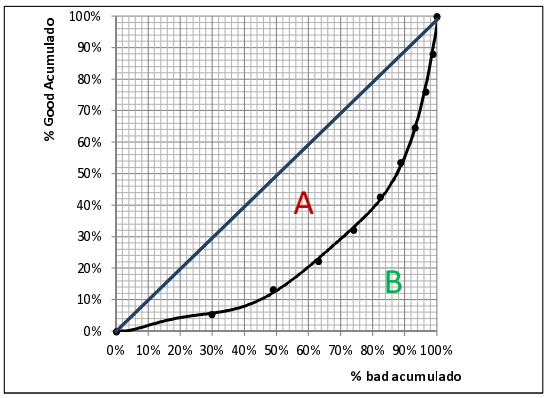


Figura 4.222: Índice de Gini=53.8.

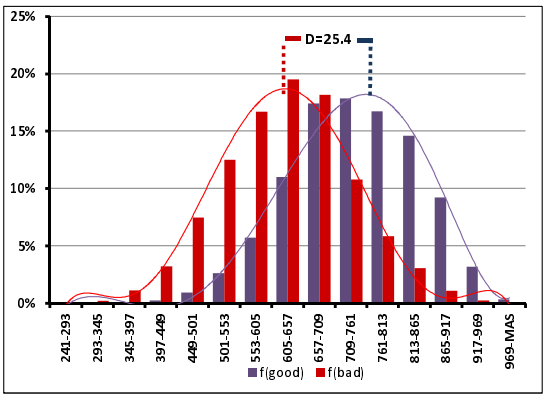


Figura 4.223: Divergencia=25.4.

Capítulo 5

Conclusiones

En la fig. 5.1 se muestra un resumen de las variables resultado del análisis univariado aplicado a cada uno de las hojas. Se puede observar que las mismas variables no tuvieron un comportamiento creciente o decreciente para todos los casos, incluyendo cuando la cartera no fue segmentada.

		1	2	3	4	5	6
Variables	Cartera Sin Segmentar	REV_0<43.5 V_1PV_0=0 MOB_T0>=6	REV_0<43.5 V_1PV_0=0 MOB_T0<6	REV_0<43.5 V_1PV_0>1	REV_0>=43.5 Util_0<58.5	REV_0>=43.5 Util_0>=58.5 BK_T0=0	REV_0>=43.5 Util_0>=58.5 BK_T0=1
MAX_BK_12M	x	x	x	x	x	x	x
BK_T0	x	x		x	x		
V_1PV_0	x	x		x	x	x	
V_2PV_0	x	x	x	x	x	x	x
V_3PV_0	x	x		x	x	x	x
V_4PV_0	x	x		x	x	x	x
V_5PV_0	x	x		x	x	x	x
V_6PV_0	x	x		x	x	x	x
SAL_0	x	x	x	x	x	x	x
SAL_1	x	x	x	x	x	x	x
SAL_2	x	x	x	x	x	x	x
SAL_3	x	x	x	x	x	x	x
SAL_4	x	x	x	x	x	x	x
SAL_5	x	x	x	x	x	x	x
MAX SALDO	x	x	x	x			x
LC_T0	x	x	x	x			
REV_0	x	x	x	x	x	x	x
REV_1	x	x	x	x		x	x
REV_2	x		x	x			x
REV_3	x	x	x	x			x
REV_4			x				x
REV_5			x				x
UTIL_0	x	x	x	x	x	x	x
UTIL_1	x	x	x	x	x	x	x
UTIL_2	x	x	x	x	x	x	x
UTIL_3	x		x	x	x	x	x
UTIL_4	x	x	x	x		x	x
UTIL_5	x	x	x	x		x	x
FLAG_CASH_T0	x	x	x	x	x		x
REGION_T0	x	x	x	x	x	x	x
MOB_T0	x	x	x	x	x	x	x
EDAD_T0	x	x	x	x	x	x	

Figura 5.1: Resumen de análisis univariado.

En la fig.5.2 se muestran las variables resultado de las regresiones logísticas de tipo “hacia adelante” que se aplicaron a cada una de las 6 hojas. Se puede notar que las variables que resultaron significativas para un segmento, no necesariamente aplicaron para otro. Como es el caso de la variable “Revolvencia tres meses antes de la valuación”, esta variable fue significativa para el score

que se desarrolló con la cartera sin segmentar, hoja1, hoja2, hoja3 y hoja6, y no fue significativa para las hojas 4 y hoja 5.

		1	2	3	4	5	6
Variables	Cartera Sin Segmentar	REV_0<43.5 V_1PV_0=0 MOB_T0>=6	REV_0<43.5 V_1PV_0=0 MOB_T0<6	REV_0<43.5 V_1PV_0>1	REV_0>=43.5 Util_0<58.5	REV_0>=43.5 Util_0>=58.5 BK_T0=0	REV_0>=43.5 Util_0>=58.5 BK_T0=1
MAX_BK_12M	x	x	x	x	x	x	x
BK_T0	x			x	x		
V_1PV_0	x			x	x	x	
V_2PV_0	x		x	x	x		x
V_3PV_0	x			x		x	
V_4PV_0	x			x			
V_5PV_0							
V_6PV_0							
SAL_0		x	x	x	x		x
SAL_1		x	x		x	x	
SAL_2		x			x		x
SAL_3		x			x		
SAL_4		x		x		x	
SAL_5		x			x		
MAX SALDO			x	x			
LC_T0		x	x				
REV_0	x	x	x	x	x	x	x
REV_1	x	x	x	x		x	x
REV_2	x			x			x
REV_3	x	x	x	x			x
REV_4							
REV_5			x				
UTIL_0	x	x	x	x	x	x	x
UTIL_1							
UTIL_2	x				x		x
UTIL_3	x						
UTIL_4	x	x					x
UTIL_5			x	x			
FLAG_CASH_T0	x	x	x	x	x		
REGION_T0	x	x	x	x	x	x	x
MOB_T0	x	x	x	x	x	x	x
EDAD_T0	x	x	x	x	x	x	

Figura 5.2: Resumen variables resultado de las regresiones logísticas.

Ya que se realizó por un lado un *credit score* sin segmentación donde obtuvimos un índice de Kolmogorov-Smirnov de 39 y que posteriormente se aplicó una segmentación de la población de la muestra con la cual contamos. Obteniendo 6 diferentes bases de clientes, que se analizaron por separado, también se generó una regresión logística para cada una de ellas. Finalmente se concentraron en una sola base, calculando un índice de Kolmogorov-Smirnov de 42. Podemos observar una mejora de 3 puntos, es decir que mejoró la diferencia entre la distribución de malos clientes que el modelo rechaza y el porcentaje de buenos clientes que el modelo rechaza.

Y de acuerdo a los estándares internacionales por el incremento de 3 puntos en el KS el modelo pasó de “satisfactorio” a “bueno”. Este resultado nos indica que en un futuro o con otras técnicas de segmentación es posible mejorar el modelo de score, para clasificarse como “muy bueno” o “extraordinario”[8].

Los índices de Gini y divergencia también muestran una diferencia:

	Score Sin Segmentación	Score Con Segmentación
Indice de Gini	50.5	53.8
Divergencia	9.4	25.4

Cuadro 5.1: Tabla comparativa Gini y divergencia.

Como el índice de Gini del score Con segmentacion 53.8>50.5 del score sin segmentación se puede decir que la segmentación generó 3 % más de iniquidad en el score de comportamiento. Sin embargo la divergencia incrementó de 9.4 a un 25.4 lo que nos indica que generó una mayor diferencia entre la distribución de clientes buenos y de malos clientes.

Para poder tener una mejor apreciación de la mejora del modelo, se construyó una matriz de confusión para el modelo de la cartera Sin Segmentación, fig. 5.2. Se puede observar que 230,247 clientes que son buenos clientes, el modelo de score Sin Segmentación los consideró como buenos clientes también. Y que 37,410 clientes que de acuerdo a la variable objetivo están clasificados como malos clientes, el modelo sin segmentación los consideró como malos clientes.

		Real		
		Clientes Buenos	**Clientes Malos**	**Total**
Modelo	**Clientes Buenos**	230,247	28,493	258,740
	Clientes Malos	56,627	37,410	94,037
	Total general	286,874	65,903	352,777

Cuadro 5.2: Matriz de confusión cartera sin segmentación.

También se calculó el porcentaje de error del modelo y la precisión del modelo:

$$\text{Error rate} = (28,493 + 56,627)/352,777 = 24\,\%$$
$$\text{Accurancy rate} = (230,247 + 37,410)/352,777 = 76\,\%$$

De igual forma para el score con segmentación de la cartera, se desarrolló la matriz de confusión y se calculó el porcentaje de error del modelo y la precisión del modelo:

		Real		
		Clientes Buenos	**Clientes Malos**	**Total**
Modelo	**Clientes Buenos**	249,297	33,878	283,175
	Clientes Malos	37,577	32,025	69,602
	Total general	286,874	65,903	352,777

Cuadro 5.3: Matriz de confusión cartera segmentada.

$$\text{Error rate} = (33,878 + 37,577)/352,777 = 20\,\%$$
$$\text{Accurancy rate} = (249,297 + 32,025)/352,777 = 80\,\%$$

A partir de estos resultados se muestra que el error de clasificar incorrectamente a un cliente bueno como malo o malo como bueno, con el modelo de la cartera sin segmentación es mayor que el porcentaje de error del modelo segmentando la cartera. Así como la tasa de precisión del modelo muestra que es mejor segmentar la cartera para generar un score de comportamiento apropiado para cada cliente.

El impacto de esta diferencia se puede observar, cuando se busca aplicar una campaña comercial dirigida, es decir a una población de la cartera en particular. Para ejemplificarlo se pretende enviar una campaña comercial, con incrementos de línea. Para llevar a cabo esta campaña el área de control de riesgo ha decidido dar un 15 % de su incremento de línea a los clientes que cuenten con un score mayor o igual a 630 puntos.

Si se considera que la media de líneas de crédito de los clientes es de 30,000 pesos. Tenemos un incremento de 4,500 pesos por cliente. Entonces:

BC	Num. de Clientes	Media de LC	Total
KS 39 %	249,402	30,000	1,122,309,000
KS 42 %	278,281	30,000	1,252,264,500

Cuadro 5.4: Incrementos de línea de crédito.

En el cuadro 5.4 se muestra que el número de clientes que captura el modelo con un KS 42 % es mayor que el de KS 39 %, esto se genera a razón de que existe una mejora del modelo. Si se consideran los 278,281 clientes del KS 42 %, se estaría hablando que el riesgo expuesto de la institución financiera sería de 129,955,500 pesos al incrementar las líneas de crédito, considerando una mejor clasificación de clientes como buenos. Ya que si, en caso contrario, se considerara el modelo KS 39 % entonces la institución estaría dejando de percibir 129,955,500 pesos. Un impacto generoso en las ganancias de la empresa.

Dado el caso práctico basado en la teoría de los capítulos 1 y 2, se puede concluir que es mejor realizar una segmentación de clientes para establecer los diferentes perfiles con los que cuenta una cartera de clientes. Para así realizar mantenimientos de cuenta dirigidos. Que como consecuencia pueden hacer que una institución de crédito gane, ahorre e invierta en los canales y clientes correctos, para controlar mejor las pérdidas derivadas del incumplimiento de pago de los clientes hacia el prestamista.

Bibliografía

[1] Gabriel Núñez Antonio. Crédito al consumo: La estadística aplicada a un problema de riesgo crediticio. In *Revista Mexicana de Investigación Actuarial Aplicada*. 2011.

[2] Viterbo Berberena. *Árboles y Redes*. Universidad Anáhuac, 2012.

[3] CreditoRealMX. ¿Qué es un crédito de consumo? http://www.creditoreal.com.mx/contenidos/edufin/que-es-un-cr2014. Accedido Jun-2015.

[4] Tipos de Org Portal Educativo. Tipos de crédito. http://www.tiposde.org/economia-y-finanzas/479-tipos-de-creditos, 2015. Accedido Julio-2015.

[5] Importancia. Importancia del crédito. http://www.importancia.org/credito.php, 2015. Accedido Julio-2015.

[6] Compara On line. Algunas características de un crédito de consumo. https://www.comparaonline.cl/blog/finanzas/credito-consumo/2013/06/algunas-caracteristicas-de-un-credito-de-consumo/, 2015. Accedido Julio-2015.

[7] SEQC. *Regresión Logística*. 2012.

[8] Slide Share. Medidas de forma: Grado de concentración. http://es.slideshare.net/tonipita/leccion-7-grado-de-concentracion-indice-de-gini, 2012. Accedido Feb-2016.

[9] Nieto Murillo Soraida. *Crédito al Consumo: La Estadística aplicada a un problema de Riesgo Crediticio*. Universidad Autónoma Metropolitana, 2010.

[10] Aluja Tomás. *La minería de Datos, entre la estadística y la inteligencia artificial*. Universidad Politécnica de Cataluña, 2001.

[11] TransUnion. Preguntas frecuentes sobre los modelos de score de transunion. https://www.transunion.com/docs/DR-Scoring-FAQ, 2012. Accedido Mar-2016.

[12] Espín Garcia Osvaldo; Rodríguez Caballero Carlos Vladimir. *Metodología para un Scoring de Clientes Sin Referencias Crediticias*. Cuadernos de Economía, 2013.

[13] Coloma Pablo; Weber Richard; Guajardo José y Miranda Jaime. *Modelos Analíticos para el manejo del riesgo de crédito*. Trend Management, 2006.

[14] Bonilla María; Olmeda Ignacio y Puertas Rosa. *Modelos Paramétricos y No Paramétricos en Problemas de Credit Scoring*. Revista Española de financiación y contabilidad, 2003.